INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAG013

Thèse de doctorat

ENSAE

IP PARIS

Contributions to high-dimensional, infinite-dimensional and nonlinear statistics

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 15 décembre 2023, par

Gabriel Romon

Composition du Jury :

Matthieu Lerasle Professeur, ENSAE (CREST)	Président
Stanislav Minsker Professeur, University of Southern California	Rapporteur
Antoine Godichon-Baggioni Maître de conférences, Sorbonne Université (LPSM)	Rapporteur
Quentin Paris Professeur, HSE University	Examinateur
Victor-Emmanuel Brunel Professeur, ENSAE (CREST)	Directeur de thèse
Guillaume Lecué Professeur, ESSEC Business School	Invité

Remerciements

En premier lieu, je tiens à remercier Victor-Emmanuel pour sa pédagogie en tant que professeur et pour avoir accepté d'encadrer cette thèse. Je me souviendrai longtemps de nos discussions et je te suis reconnaissant pour les opportunités que tu as pu m'offrir.

Je tiens ensuite à remercier Pierre Bellec. Tu as fait figure pour moi de mentor, aussi bien en mathématiques qu'en code. Au cours du stage à Rutgers tu m'as appris à écrire un article; les savoir-faire que tu m'as inculqués ont été très utiles par la suite. Ta générosité, ta patience et tes nombreux conseils, aussi bien mathématiques qu'académiques, m'ont beaucoup aidé.

J'aimerais exprimer ma reconnaissance envers les rapporteurs, Antoine Godichon-Baggioni et Stanislav Minsker, qui ont accepté d'examiner ce manuscrit dans un délai court. Je souhaite également remercier Quentin Paris, Matthieu Lerasle et Guillaume Lecué de leur présence en tant que membres du jury.

Pour leur aide à un moment ou à un autre, je tiens à remercier en particulier ces anciens ou actuels doctorants du CREST: Gautier, Badr, Avetik, Nicolas, Jules, Suzanne, Théo, Nayel.

J'adresse mes remerciements à la communauté open-source qui est la pierre angulaire de nombreux outils de travail que j'utilise au quotidien. Je pense en particulier à LaTeX-Workshop, Skim et DjVuLibre.

Finalement je souhaite remercier ma famille et mes proches pour leur soutien pendant ces trois années.

Contents

List of Figures

Chapter 1

Introduction

Three topics are explored in this thesis: inference in high-dimensional multi-task regression, geometric quantiles in infinite-dimensional Banach spaces, and convex Fréchet ℓ-means in metric trees. These topics are not unrelated with one another; we will see later in the introduction that they are bound by the thread of inference in M-estimation. Each topic has a dedicated chapter in the manuscript; the following introduction provides a broad overview for each theme, its goal is to provide important context and background information. We warn the reader that notation may change between chapters.

1.1 High-dimensional sparse multi-task regression

1.1.1 The Lasso and debiasing

In the linear Gaussian regression model with n observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, each response $y_i \in \mathbb{R}$ is a linear function of the feature vector $\boldsymbol{x}_i \in \mathbb{R}^p$, contaminated by a Gaussian noise $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i,$$

with $\boldsymbol{\beta}^* \in \mathbb{R}^p$ being the unknown coefficient vector. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ and $\boldsymbol{X}$ denote the design matrix with rows $\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top$ (which may be fixed or random). The model rewrites in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}.$$

In the low-dimensional setting where $p \leq n$ and $\boldsymbol{X}$ has full rank, a classical estimator of $\boldsymbol{\beta}^*$ is the ordinary least-squares estimator $\widehat{\boldsymbol{\beta}}^{(\text{ols})} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, which solves the least-squares optimization problem

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

$\widehat{\boldsymbol{\beta}}^{(\text{ols})}$ is an unbiased estimator of $\boldsymbol{\beta}^*$ and its distribution is $\mathcal{N}_p(\boldsymbol{0}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$. This enables statistical inference on $\boldsymbol{\beta}^*$, i.e., hypothesis testing or the construction of confidence intervals and confidence regions for coefficients of $\boldsymbol{\beta}^*$ [8].

The high-dimensional setting where the number of covariates p can be much larger than n has attracted considerable attention in the last two decades. In this setting, the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible and this calls for the introduction of other estimators. When it is believed that only a subset of the covariates contribute to the response, i.e., when the coefficient vector $\boldsymbol{\beta}^*$ is s-sparse, an appropriate estimator is the Lasso [251] $\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}$ which solves the penalized minimization problem

$$\operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where $\lambda > 0$ is a regularization parameter chosen by the user.

The performance of the Lasso for prediction, estimation and support recovery has been extensively studied in the 2000s; see, e.g., [106, 56, 35, 188, 282, 268]. These results are foundational, and yet they are insufficient for statistical inference on low-dimensional functions of $\boldsymbol{\beta}^*$, e.g., for inference on a single coefficient. For instance, oracle inequalities [269, Equation (7.26)] yield confidence intervals for $\boldsymbol{\beta}_1^*$ of size $\asymp \sqrt{(s\log p)/n}$, which is far from optimal. Besides, $\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}$ does not have a tractable limit distribution, even in the low dimensional setting [153].

In contrast with the least-squares estimator, the Lasso is provably biased [139, Corollary 11] and the bias is greater for coefficients of $\boldsymbol{\beta}^*$ with large magnitude. This has motivated the construction of other estimators based on $\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}$ which may have nicer inferential properties.

In the 2010s, under the assumption of random design with i.i.d. rows having covariance $\boldsymbol{\Sigma}$, a first line of research [280, 51, 256, 139] introduced debiased estimators $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})}$ of the form $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})} = \widehat{\boldsymbol{\beta}}^{(\mathrm{L})} + \frac{1}{n}\boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(\mathrm{L})})$ where $\boldsymbol{M}\in\mathbb{R}^{p\times p}$ is chosen as an estimate of the precision matrix $\boldsymbol{\Sigma}^{-1}$. By leveraging $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})}$, these works develop confidence intervals for single coefficients of $\boldsymbol{\beta}^*$ and in the regime $s \lesssim \sqrt{n}/\log p$. [141] extended the sparsity requirement to $s \lesssim n/(\log p)^2$, [284, 42, 54, 55, 285, 26] construct confidence intervals for general linear functionals $\boldsymbol{a}^\top\boldsymbol{\beta}^*$ where $\boldsymbol{a}\in\mathbb{R}^p$. Most recently, [193, 26, 27] establish the need for a degrees-of-freedom adjustment to deal with larger sparsity: $\boldsymbol{M}$ should ideally be chosen as $\boldsymbol{\Sigma}^{-1}/(1-\|\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}\|_0/n)$.

1.1.2 Group Lasso and multi-task regression

A more general structure assumption on the parameter vector $\boldsymbol{\beta}^*$ is group-sparsity: the set of feature indices $[1,p]$ is partitioned into m groups $G_1,\dots,G_m \subset [1,p]$ known a priori and there are only few indices $k\in[1,m]$ such that $\{\boldsymbol{\beta}_j^* : j\in G_k\} \neq \{0\}$. Inside a group, it is therefore understood that either all the covariates are relevant, or they are all simultaneously excluded.

An appropriate estimator in this case is the Group Lasso [275] $\widehat{\boldsymbol{\beta}}^{(\mathrm{g})}$ which solves the minimization problem penalized by the $\ell^{2,1}$ norm $\|\boldsymbol{\beta}\|_{2,1} = \sum_{k=1}^m\big(\sum_{j\in G_k}\boldsymbol{\beta}_j^2\big)^{1/2}$:

$$\operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{2,1}.$$

The prediction and estimation performance of the estimator were analyzed in [201, 130, 174]. The aforementioned debiasing methodology was extended to the group setting in [187, 194, 255, 242, 27].

A problem related to group-sparse estimation is that of sparse multi-task regression. We consider the multivariate Gaussian linear model with T responses or tasks

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}$$

where $\boldsymbol{B}^* \in \mathbb{R}^{p\times T}$, $\boldsymbol{Y} \in \mathbb{R}^{n\times T}$ and $\boldsymbol{E}$ has i.i.d. $\mathcal{N}_T(\mathbf{0}, \boldsymbol{S})$ rows. The structural assumption is that $\boldsymbol{B}^*$ is row-sparse, i.e., there are many features that are irrelevant across all tasks. Multi-task regression can be recast as group-sparse regression

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}}^* + \bar{\boldsymbol{\varepsilon}}$$

where $\bar{\boldsymbol{y}} = \mathbf{vec}(\boldsymbol{Y}) \in \mathbb{R}^{\bar{n}}$, $\bar{\boldsymbol{\varepsilon}} = \mathbf{vec}(\boldsymbol{E}) \in \mathbb{R}^{\bar{n}}$, $\bar{\boldsymbol{X}} \in \mathbb{R}^{\bar{n}\times\bar{p}}$ is block-diagonal with blocks of $\boldsymbol{X}$, $\bar{p} = pT$, $\bar{n} = nT$ and the features $\{1, ..., \bar{p}\}$ are partitioned into p groups with equal sizes.

The block-diagonal design $\bar{\boldsymbol{X}}$ is an obstacle that precludes straightforward application of inference results for the Group Lasso. Inference in the multi-task regression model has been taken on in [67], which extends the debiasing methodology of [280, 256].

1.1.3 Summary of our results

Chapter 3 is based on Bellec and Romon [23], which is currently under review.

The inferential goals of the chapter are twofold. First, we construct valid confidence intervals for a linear functional $\boldsymbol{a}^\top \boldsymbol{B}^* \boldsymbol{e}_1$ of the unknown coefficient on the first task, by leveraging responses on all tasks simultaneously. Second, we construct valid confidence ellipsoids for rows $\boldsymbol{e}_j^\top \boldsymbol{B}^* \in \mathbb{R}^{1\times T}$ of the unknown coefficient matrix $\boldsymbol{B}^*$, for instance to provide hypothesis tests on the nullity of the j-th row of $\boldsymbol{B}^*$, or equivalently testing that the signal does not depend on the j-th covariate.

In order to achieve these statistical goals, we introduce a new object, the data-driven symmetric *interaction matrix* $\widehat{\mathbf{A}} \in \mathbb{R}^{T\times T}$. Introduction of the matrix $\widehat{\mathbf{A}}$ is key to equip the estimator $\widehat{\boldsymbol{B}}$ with the aforementioned inference capabilities. This data-driven matrix $\widehat{\mathbf{A}}$ generalizes, to the multi-task setting, the effective degrees-of-freedom and other scalar adjustments in single-task linear models.

While previous proposals in grouped-variables regression require row-sparsity $s \lesssim \sqrt{n}$ up to constants depending on T and logarithmic factors in (n, p) for unknown $\boldsymbol{\Sigma}$, the debiasing scheme using the interaction matrix provides confidence intervals and χ^2_T confidence ellipsoids under the conditions $\min(T^2, \log^8 p)/n \to 0$ and

$$\frac{sT + s\log(p/s) + \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0 \log p}{n} \to 0, \qquad \frac{\min(s, \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0)}{\sqrt{n}}\sqrt{[T + \log(p/s)]\log p} \to 0,$$

allowing for row-sparsity $s \ggg \sqrt{n}$ when $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0 \sqrt{T} \lll \sqrt{n}$ up to logarithmic factors.

1.2 M-estimation, infinite dimension and geometric quantiles

The ordinary least-squares, Lasso and Group Lasso estimators introduced in the previous section are defined as solutions of a minimization problem. This is a particular instance of M-estimation, on which we focus next.

1.2.1 M-estimation: classical results

A population parameter $\theta_\star$ is often defined in an implicit fashion as a minimizer of an objective function ϕ of the following type:

$$\begin{aligned} \phi \colon \Theta &\to \mathbb{R} \\ \theta &\mapsto \int_{\mathcal{X}} \varphi(x,\theta) d\mu(x), \end{aligned} \tag{1.1}$$

where Θ is the parameter space, $(\mathcal{X}, \mathcal{A}, \mu)$ is a probability space and $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$ is a contrast function that is integrable in the first argument. Given an i.i.d. sample $X_1, \ldots, X_n \sim \mu$, an M-estimator $\widehat{\theta}_n$ of $\theta_\star$ is defined as a minimizer of the empirical objective function $\widehat{\phi}_n$

$$\widehat{\phi}_n : \theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i, \theta), \tag{1.2}$$

which is obtained by replacing the unknown population measure μ in (1.1) with the empirical measure $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

The random elements $X_1, X_2, \ldots$ are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. To simplify the exposition we will not dwell on measurability issues in this introduction. We assume therefore from the outset that $\widehat{\theta}_n$ is measurable between the σ-algebras $\mathcal{F}$ and $\mathcal{A}$.

This general framework of estimation was first formulated by Huber [131], who used the letter "M" as a shorthand for "minimize". Classical examples of M-estimation in the Euclidean setting (i.e., Θ is a subset of the Euclidean space $\mathbb{R}^d$, which has the standard Hilbert structure given by the dot product) include the case where:

1. $\Theta = \mathcal{X} = \mathbb{R}^d$ with $d \geq 1$, μ is a Borel probability measure with finite first moment and $\varphi : (x, \theta) \mapsto \|x - \theta\|_2^2 - \|x\|_2^2$. Here $\theta_\star$ is the mean of μ, i.e., $\theta_\star = \int_{\mathbb{R}^d} x d\mu(x)$ and $\widehat{\theta}_n$ is the sample mean: $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

2. $\Theta = \mathcal{X} = \mathbb{R}$, μ is a Borel probability measure and $\varphi : (x, \theta) \mapsto |x - \theta| - |x|$. The (possibly infinitely many) minimizers of ϕ are the medians of μ and the minimizers of $\widehat{\phi}_n$ are the usual sample medians, which are expressed in terms of the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$: when n is odd $X_{(\lfloor \frac{n}{2} \rfloor + 1)}$ is the unique sample median, and when n is even the argmin set is the interval $[X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$.

3. $\Theta = \mathcal{X} = \mathbb{R}$, μ is a Borel probability measure with finite first moment and $\varphi : (x, \theta) \mapsto (x - \theta)^2 \mathbb{1}_{|x-\theta| \leq c} + (2c|x - \theta| - c^2)\mathbb{1}_{|x-\theta|>c}$ where $c \geq 0$. This contrast function was introduced by Huber [131] for robustness purposes.

4. $\Theta \subset \mathbb{R}^d$ with $d \geq 1$, μ is contained in a parametric family $(\mathbb{P}_\theta)_{\theta\in\Theta}$ (i.e., $\mu = \mathbb{P}_{\theta_0}$ for some $\theta_0 \in \Theta$), the family is dominated by a sigma-finite measure ν, the corresponding densities $(f_\theta)_{\theta\in\Theta}$ are positive ν-a.e., for each $\theta \in \Theta$,
$$\int_{\mathcal{X}} |\ln f_\theta(x)| f_{\theta_0}(x) d\nu(x) < \infty$$
and $\varphi : (x, \theta) \mapsto -\ln f_\theta(x)$. Any minimizer $\theta_\star$ verifies $\mathbb{P}_{\theta_\star} = \mu$ and it is unique if and only if the family $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is identifiable. Minimization of $\widehat{\phi}_n$ coincides with the classical maximum likelihood estimation.

5. $\Theta = \mathbb{R}^d$, $\mathcal{X} = \mathbb{R}^{d+1}$ with $d \geq 1$, μ is a Borel probability measure with finite second moment, the random vector $(X, Y) \sim \mu$ satisfies $\mathbb{E}[Y|X] = \theta_0^\top X$ for some $\theta_0 \in \mathbb{R}^d$ and $\varphi : \big((x, y), \theta\big) \mapsto (y - \theta^\top x)^2$. In that case, M-estimation is the same as ordinary least-squares.

Consistency, rate of convergence and limit distribution

Assume that ϕ has a unique minimizer $\theta_\star$ which is the parameter of interest. A first step towards successful estimation of $\theta_\star$ is the consistency of the sequence $(\widehat{\theta}_n)_{n\geq1}$, i.e., some form of stochastic convergence to $\theta_\star$. To quantify this behavior we require from now on that Θ be a metric space with metric d, and we say that $(\widehat{\theta}_n)_{n\geq1}$ is strongly consistent (resp., weakly consistent) if $d(\widehat{\theta}_n, \theta_\star)$ converges almost surely (resp., in probability) to 0.

Given the generality of the estimation framework (1.1), statisticians have strived to establish general consistency results that encompass a wide range of contrast functions φ. The following textbook consistency conditions are given in [261, 259].

Proposition 1.1 ([261, Corollary 3.2.3], [259, Theorem 5.7]). *1. If*
$$\sup_{\theta\in\Theta} |\widehat{\phi}_n(\theta) - \phi(\theta)| \xrightarrow[n\to\infty]{\mathbb{P}} 0 \tag{1.3}$$
and
$$\forall \varepsilon > 0, \inf_{\substack{\theta\in\Theta \\ \|\theta-\theta_\star\|\geq\varepsilon}} \phi(\theta) > \phi(\theta_\star), \tag{1.4}$$
then $(\widehat{\theta}_n)_{n\geq1}$ is weakly consistent.

2. If (1.3) is replaced with uniform convergence over compact sets, if (1.4) holds and assuming
$$\forall \varepsilon > 0, \exists K \, compact \,, \forall n \geq 1, \; \mathbb{P}(\widehat{\theta}_n \in K) \geq 1 - \varepsilon,$$
then $(\widehat{\theta}_n)_{n\geq1}$ is weakly consistent.

Proposition 1.2 ([259, Theorem 5.14]). *If $\theta \mapsto \varphi(x, \theta)$ is lower semicontinuous for μ-almost every x and $\forall\theta \in \Theta, \exists r > 0, \mathbb{E}\Big[\inf_{\substack{\alpha\in\Theta \\ \|\alpha-\theta\|\geq r}} \varphi(X_1, \alpha)\Big] < \infty$, then for every $\varepsilon > 0$ and every compact $K \subset \Theta$,*
$$\mathbb{P}\big(\{d(\widehat{\theta}_n, \theta_\star) \geq \varepsilon\} \cap \{\widehat{\theta}_n \in K\}\big) \xrightarrow[n\to\infty]{} 0. \tag{1.5}$$

Let us comment briefly on these results. The stochastic uniform convergence condition (1.3) is equivalent to the class of functions $\big(\varphi(\cdot,\theta)\big)_{\theta\in\Theta}$ being μ-Glivenko–Cantelli. This can be ascertained using tools from the theory of empirical processes, such as bracketing and random L^1-entropy numbers [261, Part 2]. For example, the class is Glivenko–Cantelli if it is pointwise compact, meaning that it is dominated by an integrable function, that Θ is a compact metric space and $\theta \mapsto \varphi(x,\theta)$ is continuous for every $x \in \mathcal{X}$ [259, Example 19.8].

Under condition (1.4), if θ is separated away from $\theta_\star$ then $\phi(\theta)$ cannot get arbitrarily close to the minimum value of ϕ. In that case the minimizer $\theta_\star$ is said to be "well-separated". Assuming that ϕ is lower semicontinuous and Θ is a compact metric space, $\theta_\star$ is automatically well-separated. The convergence statement (1.5) yields weak consistency if one can exhibit a compact subset K such that $\lim_{n\to\infty} \mathbb{P}(\widehat{\theta}_n \in K) = 1$.

Upon this discussion, it appears that the compact subsets of Θ as well as the compactness of Θ itself play a role in establishing consistency. Further general conditions can be found in [132, 212, 120, 83]. They are similar in spirit to the ones displayed previously and rely also on compactness.

After consistency is obtained, it is interesting to quantify the rate of convergence, i.e., finding a sequence of positive reals $(r_n)_{n\geq 1}$ such that $\lim_n r_n = \infty$ and $r_n d(\widehat{\theta}_n, \theta_\star) = O_{\mathbb{P}}(1)$. A general result for this purpose is [261, Corollary 3.2.6], which requires that ϕ grows at least locally quadratically, i.e., $\phi(\theta) \geq \phi(\theta_\star) + c d(\theta,\theta_\star)^2$ for some constant $c > 0$ and every θ in some neighborhood of $\theta_\star$, and that there is some control of the empirical process indexed by the class $\mathcal{M}_\delta = \{\varphi(\cdot,\theta) - \varphi(\cdot,\theta_\star) : d(\theta,\theta_\star) < \delta\}$ where δ ranges in a neighborhood of 0. The growth condition easily follows when Θ is a Euclidean space and ϕ is twice-differentiable at $\theta_\star$ with nonsingular Hessian matrix $\nabla^2\phi(\theta_\star)$. The second condition can be verified by bounding a uniform-entropy integral or a bracketing integral of $\mathcal{M}_\delta$. In the special case where Θ is Euclidean and the contrast is Hölder continuous in the second variable, i.e., for every θ_1, θ_2 in a neighborhood of $\theta_\star$ and μ-almost every $x \in \mathcal{X}$,

$$|\varphi(x,\theta_1) - \varphi(x,\theta_2)| \leq C(x)\|\theta_1 - \theta_2\|^\alpha, \tag{1.6}$$

the bracketing integral of $\mathcal{M}_\delta$ is easily bounded using the covering numbers of balls in $\mathbb{R}^d$. In practice, to obtain a rate of convergence it is therefore convenient that Θ be a Euclidean space. In non-Euclidean spaces such as infinite-dimensional normed spaces, it is sometimes appropriate to consider a sieved M-estimator instead of $\widehat{\theta}_n$: given an increasing sequence of subsets $\Theta_n \subset \Theta$, minimization is carried over Θ_n instead of the whole parameter space. Sieved M-estimation can be seen as a form of regularization, which may help avoid overfitting. Rates for sieved M-estimators are found in [261, Chapter 3.4].

The rate r_n has the correct order if additionally $\big(r_n d(\widehat{\theta}_n, \theta_\star)\big)_{n\geq 1}$ converges in distribution. Generic results of this kind are usually stated in the setting where Θ is a Euclidean space. Such a result is [261, Theorem 3.2.10], which leverages empirical process theory. Another one based on linearization is [261, Theorem 3.2.16], which has the more precise conclusion $r_n(\widehat{\theta}_n - \theta_\star) = -[\nabla^2\phi(\theta_\star)]^{-1} Z_n + o_{\mathbb{P}}(1)$ where $(Z_n)_{n\geq 1}$ is a tight sequence of random vectors. It is often (but not always, see [151]) the case that

$r_n = \sqrt{n}$ and the stochastic expansion is explicit:

$$\sqrt{n}(\widehat{\theta}_n - \theta_\star) = -[\nabla^2\phi(\theta_\star)]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\nabla_\theta\varphi(X_i,\theta_\star) + o_{\mathbb{P}}(1). \tag{1.7}$$

The central limit theorem implies that $\sqrt{n}(\widehat{\theta}_n - \theta_\star)$ is asymptotically normal: it converges in distribution to a centered multivariate normal with covariance matrix $[\nabla^2\phi(\theta_\star)]^{-1}\mathbb{E}\big[\nabla_\theta\varphi(X_1,\theta_\star)\nabla_\theta\varphi(X_1,\theta_\star)^\top\big][\nabla^2\phi(\theta_\star)]^{-1}$. This holds for example when the contrast satisfies (1.6) with $\alpha = 1$ [261, Example 3.2.22]. Further results along the same lines can be found in [132, 214, 215].

The convex case

The following setting is common: Θ is an open convex subset of $\mathbb{R}^d$ and the contrast is convex in the second argument, i.e., for μ-almost every $x \in \mathcal{X}$, the function $\theta \mapsto \varphi(x,\theta)$ is convex. Convexity greatly simplifies the statements of consistency and asymptotic normality, as observed in [108, 199, 119].

Proposition 1.3 ([108, Theorem 5.1],[199, Theorem 1])**.** *Under the convexity assumption, $(\widehat{\theta}_n)_{n\geq 1}$ is strongly consistent.*

Proposition 1.4 ([199, Theorem 4])**.** *For each $x \in \mathcal{X}$ and $\theta \in \Theta$, let $g(x,\theta)$ denote a subgradient at θ of $\varphi(x,\cdot)$. Assume that $\mathbb{E}[\|g(X,\theta)\|_2^2] < \infty$ for each θ in a neighborhood of $\theta_\star$, and that ϕ is twice-differentiable at $\theta_\star$ with nonsingular Hessian. Then the stochastic expansion* (1.7) *holds.*

These results are generic, and yet the assumptions are considerably simpler than those needed for the statements mentioned above that exploit empirical process theory.

1.2.2 The challenge of infinite dimension

In this subsection the parameter space Θ has a linear structure and it is infinite-dimensional. More precisely, Θ is an infinite-dimensional Banach space, or less generally an infinite-dimensional Hilbert space. The theory of probability in Banach spaces flourished in the 1970s; see, e.g., the monographs [161, 9, 252, 168].

An infinite-dimensional parameter space arises naturally when the data lives in a function space, for instance when modeling curves (e.g., radar waveforms, spectrometric data, electricity consumption, ECGs, EEGs, stock prices). Functional data analysis is the corresponding area of research and it has gained considerable traction since the 1990s (see, e.g., [217, 93, 124, 127, 270]). In the literature, modeling is usually performed in the Hilbert space L^2, however there has recently been interest in non-Hilbertian spaces as well [74]. Another case for infinite-dimensional statistics is kernel methods [69, 121, 281, 198, 173].

A salient feature of infinite-dimensional normed spaces is that closed balls and spheres are not compact in the norm topology [7, Theorem 5.26]. As a consequence, compact subsets of Θ have empty interior. A straightforward way of generating compact subsets is to take a finite-dimensional subspace $V \subset \Theta$ and consider its closed and bounded subsets $K \subset V$. Conversely, the following result shows that all compact subsets of Θ are approximately finite-dimensional.

Proposition 1.5 ([9, Lemma 4.3], [168, Lemma 2.2])**.** *Let Θ be a Banach space and $K \subset \Theta$. K is a compact subset of Θ with respect to the norm topology if and only if the following conditions hold:*

1. *K is closed and bounded.*

2. *For every $\varepsilon > 0$, there exists a finite-dimensional vector subspace V such that for every $x \in K$, $d(x, V) < \varepsilon$.*

Compact subsets in infinite dimension are therefore somewhat pathological, and for the statistician this may be understood as some *curse of infinite dimensionality* (not to be confused with the synonymous curse from functional data analysis [93, 99]). Indeed, it was seen earlier that classical consistency results in M-estimation are most easily obtained by leveraging compactness. The control of covering or bracketing numbers underpins many of the aforementioned theorems on consistency, convergence rate and limit distribution. It is sometimes achieved by controlling the covering or bracketing numbers of the index set. For instance, when the Hölder assumption (1.6) holds and $\Theta = \mathbb{R}^d$, it is possible to exploit bounds on the covering number of balls in $\mathbb{R}^d$. In contrast, when Θ is infinite-dimensional, balls are not totally bounded and such technique fails. It is also worth mentioning that the limit distribution result [261, Theorem 3.2.10] hinges on total boundedness of balls, and is therefore not applicable in infinite dimension.

Unfortunately, the convex case is not spared by the curse either. The remarkably transparent Propositions 1.3 and 1.4 rely crucially on the following result from convex analysis.

Proposition 1.6 ([287, Corollary 2.2.23])**.** *Let E be a Banach space and Θ be an open convex subset of E. Let $(f_n)_{n\geq 1}$ be a sequence of convex functions on Θ that converges pointwise to some f. Then $(f_n)_{n\geq 1}$ converges uniformly on compact subsets of Θ to f.*

In the proofs of Propositions 1.3 and 1.4, uniform convergence is applied naturally on closed balls, which are compact in finite dimension.

We were not successful in developing a general theory of M-estimation in infinite dimension, even in the convex case. Infinite-dimensional M-estimation was investigated by van der Vaart [257, 258, 260] who formulates results with assumptions similar to those in [261]. For the study of regression M-estimators in infinite dimension, see the successive works [79, 80, 78, 160]. More recently, Sinova et al. [238] aim to develop a general theory of M-estimation in Hilbert spaces, with an emphasis on the infinite-dimensional function space L^2. As noted by the authors, their consistency result in the norm topology [238, Theorem 3.4] covers only finite-dimensional spaces.

Our contribution to infinite-dimensional M-estimation is the study of a specific M-estimator, which is introduced in the next subsection.

1.2.3 Quantiles: from $\mathbb{R}$ to infinite dimension

Univariate quantiles

Given a probability measure μ on $\mathbb{R}$ and $p \in (0,1)$ an elementary parameter of location is the p-th quantile of μ, which is usually defined as any $\alpha \in \mathbb{R}$ that satisfies both

$$\mu((-\infty, \alpha]) \geq p \quad \text{and} \quad \mu([\alpha, \infty)) \geq 1 - p. \tag{1.8}$$

A prominent special case is that of the median, i.e., when $p = \frac{1}{2}$. Quantiles are important since they provide a measure of the degree of centrality and the median can be interpreted as a central tendency of the distribution. Quantiles have applications in hypothesis testing [233], in regression [154] and in robust statistics [133, 192, 177].

It is well-known (see, e.g., [259, p.44]) that the p-th quantile fits the setting of M-estimation: with the preceding notation, let $\Theta = \mathcal{X} = \mathbb{R}$, assume that μ has a finite first moment and consider the contrast function $\varphi : (x, \alpha) \mapsto (1-p)(x-\alpha)^+ + p(x-\alpha)^-$. Alternatively, it is possible to drop the moment assumption by defining the contrast

$$\varphi : (x, \alpha) \mapsto |x - \alpha| - |x| - (2p - 1)\alpha. \tag{1.9}$$

Geometric quantiles

The definition (1.8) relies on the intervals $(-\infty, \alpha]$ and $[\alpha, \infty)$, thus one may argue that it hinges on $\mathbb{R}$ being a totally ordered set. Measuring centrality is an important topic in multivariate statistics, hence the need for a generalization of quantiles to higher dimension. Since there is no natural order on $\mathbb{R}^d$, extending the formulation (1.8) is impractical. One might be tempted to define a coordinatewise p-th quantile, i.e., to consider the vector of univariate p-th quantiles on each coordinate. The resulting parameter of location is translation equivariant: if T is a translation mapping, X is a random vector with distribution μ and α is a coordinatewise quantile of X, then $T(\alpha)$ is a coordinatewise quantile of $T(X)$. However, a major downside of this parameter is its dependence on the coordinate system, meaning that it is not rotation equivariant. There is a rich literature on the subject of generalizing measures of centrality and outlyingness to higher dimension; see, e.g., [239, 288, 232, 197] and the references therein.

The univariate contrast (1.9) is defined in terms of the absolute value which is a special case of a norm, and the linear function $\alpha \mapsto (2p-1)\alpha$, which is a special case of a linear functional with operator norm less than 1. Consider a normed vector space $(E, \|\cdot\|)$ and let $\Theta = \mathcal{X} = E$. A natural generalization of the contrast (1.9) is therefore

$$\varphi : (x, \alpha) \mapsto \|x - \alpha\| - \|x\| - \ell(\alpha), \tag{1.10}$$

where ℓ is an element of the continuous dual space with dual norm less than 1. The corresponding M-estimator is called geometric quantile or spatial quantile. The geometric median (i.e., when $\ell = 0$) was introduced in the two-dimensional Euclidean setting by Weber [272] in 1909 and was later reintroduced in the same setting by Gini and Galvani [103, 225] as well as Haldane [110]. Valadier [253, 254] extended the concept to any reflexive Banach space and Kemperman [148] performed a systematic study of existence and uniqueness in general Banach spaces. Chaudhuri [65] and Koltchinskii [156, 157] defined geometric quantiles in Banach spaces by adding the linear functional ℓ in the objective function.

1.2.4 Summary of our results

Chapter 4 is based on Romon [223], which is currently under review.

We study large-sample properties of geometric quantiles in infinite-dimensional Banach spaces.

We begin with new descriptive results for population medians: we prove existence of a geometric median in a wide variety of L^1 spaces, thus improving on Kemperman [148, Corollary 3.2], and we characterize the set of medians in the degenerate case where the measure μ is supported on an affine line.

Estimation is performed using an approximate M-estimator: instead of exactly minimizing the empirical objective (1.2), we require only minimization up to an additive error ϵ_n, which may be random.

When the population quantile is not uniquely defined we leverage the theory of variational convergence to obtain asymptotic statements on subsequences in the weak topology.

When the population quantile is unique, we show strong consistency of the estimator in the norm topology. Our result holds under minimal assumptions on μ and in any separable, uniformly convex space (e.g., separable Hilbert spaces, L^p, $W^{k,p}$ with $p \in (1, \infty)$). It is a significant improvement on the result by Chakraborty and Chaudhuri [64, Theorem 4.2.2], which is only valid in separable Hilbert spaces and requires extra distributional assumptions.

In a separable Hilbert space, we obtain novel expansions of the kind (1.7) (which are known as Bahadur–Kiefer representations). An immediate consequence is the asymptotic normality of the geometric quantile. Our central limit theorem is formulated under assumptions that exactly match those of the finite-dimensional case, and it is therefore a major improvement on Gervini's normality statement [100, Theorem 6].

1.3 M-estimation in metric trees

1.3.1 M-estimation in metric spaces and Fréchet means

In the previous section, much emphasis was put on the case where the data lives in a normed vector space E. Such modeling is sometimes not realistic, as the data may reside in a nonlinear subset $S \subset E$, and the metric induced from the norm may not be meaningful. Examples range from data on the sphere $\mathbb{S}^d$, which is the object of directional statistics [181, 171], to data in the form of symmetric positive definite matrices (which are used to model covariance matrices, e.g., in diffusion tensor imaging [165] and have also found some use for image segmentation in computer vision [221, 60]), to data in measure spaces (e.g., Wasserstein spaces, which are the cornerstone of optimal transport [211]) or data in quotient spaces (e.g., when a practioner is interested in shapes of objects, the data can be analyzed modulo translations, rotations and scalings, hence it belongs to a quotient space, also called a shape space [149, 31]).

The M-estimation framework introduced in the previous section is general and it applies also in the nonlinear setting [48]. As mentioned in [138], examples include :

1. Extending principal component analysis to manifold-valued data [94, 134, 135, 137]: for example in [137], the first geodesic principal component in the planar shape space Σ_2^k is defined via a minimization problem where $\Theta = \Gamma(\Sigma_2^k)$ is the space of geodesics on Σ_2^k and $\mathcal{X} = \mathbb{S}_2^k$ is the pre-shape sphere.

2. Extending Euclidean measures of central tendency (e.g., the mean and median) to the setting of metric spaces: $\Theta = \mathcal{X} = E$, where (E, d) is a metric space. This is our subject of exposition for the rest of this section.

The mean of a measure μ on $\mathbb{R}^d$ is defined most elementarily as the vector of the means for each univariate marginal. In infinite-dimensional Banach spaces, the mean is classically defined as a Bochner integral [75]. In both cases the definition relies crucially on the linear structure of the ambient space. As stated earlier in Section 1.2.1, when $\Theta = \mathcal{X} = \mathbb{R}^d$ the contrast function $\varphi : (x, \theta) \mapsto \|x - \theta\|_2^2 - \|x\|_2^2$ yields the mean of the measure μ. This contrast function provides a natural extension of the concept of mean to metric spaces: fix some arbitrary $o \in E$, assume that $\int_E d(x, o) d\mu(x) < \infty$ and define

$$\varphi : (x, \theta) \mapsto d(x, \theta)^2 - d(x, o)^2. \tag{1.11}$$

Assuming a finite second moment, i.e., if $\int_E d(x, o)^2 d\mu(x) < \infty$ for some (and thus for every) $o \in E$, the contrast can be replaced by the simpler $(x, \theta) \mapsto d(x, \theta)^2$ and one can define a Fréchet variance. For convenience we let $M(\mu)$ denote the set of minimizers of the corresponding objective function. Elements of $M(\mu)$ are known as Fréchet means [95], barycenters or centers of mass. The existence and uniqueness of Fréchet means hinges on the geometry of the space E and it is a longstanding topic of research; see, e.g., [147, 164, 243, 3, 202, 273, 166, 5, 167].

Regarding estimation, we focus solely on the M-estimator obtained by minimization of the empirical objective (1.2), but another popular estimator is the inductive mean introduced by Sturm [243]. Since the Fréchet mean may not be uniquely defined, consistency results quantify some form of closeness between the stochastic set $M(\widehat{\mu}_n)$ and the true set $M(\mu)$ as n goes to infinity. Ziezold [286] proved consistency when E is a separable metric space, while Bhattacharya and Patrangenaru [33] do so when E has the Heine–Borel property, i.e., when every closed and bounded set is compact. Central limit theorems have been developed in the setting where E is a Riemannian manifold [34, 31, 32, 87]. Non-Euclideanity of the space allows for new asymptotic phenomena such as stickiness [126, 136] and smeariness [125, 86].

1.3.2 Fréchet means in Hadamard spaces

Next, we briefly introduce nonpositive curvature in the sense of Alexandrov, a geometric feature of the space E that plays a key role in the analysis of Fréchet means. Extensive expositions can be found in the monographs [145, 45, 52, 15].

Let $x, y \in E$. A constant speed geodesic from x to y is a map γ from some interval $[a, b] \subset \mathbb{R}$ to E such that $\gamma(a) = x$, $\gamma(b) = y$ and $d(\gamma(t_1), \gamma(t_2)) = v|t_1 - t_2|$ for some $v \in [0, \infty)$ and every $t_1, t_2 \in [a, b]$. The real number v is called the speed of the geodesic γ. The image of γ is denoted by $[x, y]$ and it is referred to as a geodesic segment joining x and y. For the sake of legibility, we will often write γ_t in lieu of $\gamma(t)$.

If for every $x, y \in E$ there exists a geodesic segment joining x and y, then (E, d) is said to be a geodesic metric space. If in addition such a geodesic segment is unique, then (E, d) is called uniquely geodesic. Let us illustrate with some examples. When E is a normed vector space with norm $\|\cdot\|$, the line segment $\{(1-t)x + ty : t \in [0, 1]\}$ is a geodesic segment corresponding to the geodesic $t \mapsto x + t(y - x)$ defined on $[0, 1]$ with speed $\|y - x\|$. Thus a normed vector space is a geodesic space, but it need not be uniquely geodesic (consider for instance $\mathbb{R}^2$ with the ℓ^1 norm). The sphere $\mathbb{S}^2$ is classically equipped with the angular metric, i.e., the standard Riemannian metric. The sphere is geodesic; more precisely, a geodesic segment is a minor arc of a great circle. $\mathbb{S}^2$ is not uniquely geodesic however: if x and y are antipodal, there are infinitely many geodesic segments between x and y.

Before we can define nonpositive curvature, we need the concept of comparison triangle. Assume from now on that E is uniquely geodesic and fix $x, y, z \in E$. The geodesic triangle with vertices x, y, z is the union of geodesic segments $[x, y] \cup [y, z] \cup [z, x]$. By the triangle inequality, it is possible to construct a triangle in $\mathbb{R}^2$ with vertices $\bar{x}, \bar{y}, \bar{z} \in \mathbb{R}^2$ such that $d(x, y) = \|\bar{x} - \bar{y}\|$, $d(y, z) = \|\bar{y} - \bar{z}\|$ and $d(z, x) = \|\bar{z} - \bar{x}\|$. The triangle $\triangle\bar{x}\bar{y}\bar{z}$ is called a comparison triangle for $\triangle xyz$ and it is unique up to isometries. If $\gamma : [0, 1] \to E$ is the geodesic from x to y, note that $d(\gamma_t, x) = \|(1-t)\bar{x} + t\bar{y} - x\|$ hence the convex combination $(1-t)\bar{x} + t\bar{y}$ can be interpreted as a comparison point for γ_t in $\triangle\bar{x}\bar{y}\bar{z}$.

The space (E, d) is said to have nonpositive curvature in the sense of Alexandrov if for every $x, y, z \in E$ and every geodesic $\gamma : [0, 1] \to E$ from x to y we have the inequality

$$d(\gamma_t, z) \leq \|(1-t)\bar{x} + t\bar{y} - \bar{z}\|. \tag{1.12}$$

Alternatively, (E, d) is said to be CAT(0), which stands for Cartan–Alexandrov–Toponogov. Intuitively, (1.12) states that every geodesic triangle is thinner than its Euclidean comparison triangle, as seen in Figure 1.1.

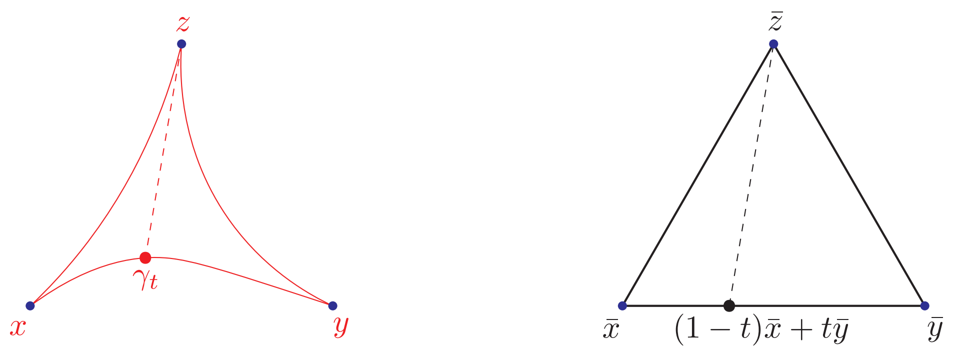

Figure 1.1: Geodesic triangle (left) and a corresponding comparison triangle (right) for a CAT(0) space.

If (E, d) is complete and CAT(0), then it is called a Hadamard space. Examples of Hadamard spaces are Hilbert spaces, convex subsets thereof, and complete simply connected Riemannian manifolds with nonpositive sectional curvature. Note that the aforementioned sphere $\mathbb{S}^2$ is not CAT(0). In Hadamard spaces it is possible to develop a theory of convex analysis, convex optimization and probability that generalizes to nonlinear settings the classical results known in Hilbert spaces.

From now on we assume that (E, d) is Hadamard. A subset $C \subset E$ is said to be geodesically convex if for every $x, y \in C$, the geodesic segment $[x, y]$ is a subset of C. Given such a subset, a function $f : C \to \mathbb{R}$ is said to be geodesically convex if for $x, y \in C$ and every geodesic γ from x to y defined on $[0, 1]$, we have $f(\gamma_t) \le (1-t)f(x)+tf(y)$. After squaring (1.12) and some algebra, the following equivalent inequality is reached:

$$d(\gamma_t, z)^2 \le (1-t)d(x, z)^2 + td(y, z)^2 - t(1-t)d(x, y)^2, \tag{1.13}$$

thus for each $z \in E$, the function $d(\cdot, z)^2$ is geodesically (strongly) convex.

Sturm [243] made apparent the allure of Hadamard spaces and the importance of (1.13) for the study of the Fréchet mean. First, in this case the Fréchet mean exists and it is unique: $M(\mu) = \{\theta_\star\}$. Second, because of (1.13) the minimizer $\theta_\star$ is well-separated (as defined in Section 1.2.1) and this is quantified by the so-called variance inequality [243, Proposition 4.4]

$$\phi(\theta) \ge \phi(\theta_\star) + d(\theta, \theta_\star)^2 \tag{1.14}$$

which holds for every $\theta \in E$.

Estimating the rate of convergence of empirical Fréchet means has attracted much attention recently, especially in Hadamard spaces without Riemannian structure where central limit theorems are not available. The inequality (1.14) is a key technical ingredient in all the following works, mentioned in chronological order. Ahidar-Coutrix et al. [5] consider the case where E is bounded and under a strong metric entropy assumption they establish that for some constants C_1, C_2 and every $n \ge 1$, $t > 0$,

$$\mathbb{P}\Big(\sqrt{n}d(\widehat{\theta}_n, \theta_\star) \ge C_1 \max(C_2, \sqrt{t})\Big) \le 2e^{-t}.$$

Under a weaker entropy condition they obtain non-parametric rates. Schötz [230] deals with unbounded E and obtains under a weak entropy condition that for some constant C_3 and every $n \ge 1$, $t > 0$,

$$\mathbb{P}\big(\sqrt{n}d(\widehat{\theta}_n, \theta_\star) \ge t\big) \le \frac{C_3}{t^2}.$$

Under a strong entropy condition he also shows the asymptotic statement that for some constant $\beta > 0$,

$$\mathbb{E}[d(\widehat{\theta}_n, \theta_\star)^2] = O\Big(\frac{\log(n)^\beta}{n}\Big).$$

Le Gouic et al. [167] impose that the Hadamard space have curvature bounded from below by some $\kappa \le 0$ (see, e.g., [52] for the definition) and with σ^2 denoting the Fréchet variance they establish that for every $n \ge 1$,

$$\mathbb{E}[d(\widehat{\theta}_n, \theta_\star)^2] \le \frac{\sigma^2}{n}. \tag{1.15}$$

Yun and Park [276] obtain results similar to those of Schötz. Brunel et al. [49] define a notion of sub-Gaussian distribution in metric spaces. Assuming that μ is sub-Gaussian (e.g., if μ has bounded support) and E has curvature bounded below (to exploit (1.15)), they show that for some constant C_4 and every $n \geq 1$, $t > 0$,

$$\mathbb{P}\big(\sqrt{n}d(\widehat{\theta}_n, \theta_\star) \geq \sigma + t\big) \leq e^{-C_4 t^2}. \tag{1.16}$$

Most recently, Escande [88] proved by a clever stability argument that (1.15) holds, up to a universal constant, without the lower bound assumption on curvature. He obtains a bound similar to (1.16) under a sub-exponential tail assumption.

When $E = \mathbb{R}^d$ it is well-known that the sample mean suffers from a major defect: it is easily influenced by outlying observations. This motivates the need for other parameters of central tendency, such as the median. In the nonlinear setting, for $p \in [1, \infty)$ and assuming that $\int_E d(x, o)^{p-1} d\mu(x) < \infty$ for some $o \in E$, the location parameter corresponding to the contrast function

$$\varphi : (x, \theta) \mapsto d(x, \theta)^p - d(x, o)^p$$

is referred to as the Fréchet p-mean. When $p = 2$ this is the classical Fréchet mean, and when $p = 1$ it is called Fréchet median. For $p \neq 2$, statistical results are limited to consistency [137, 231], and asymptotic normality when E is a Riemannian manifold [48]. It is difficult to obtain an analog of the variance inequality (1.14), hence the lack of results on convergence rates.

1.3.3 Metric trees

An important incarnation of Hadamard spaces is the metric tree. Consider a tree T in the graph-theoretic sense, i.e., an undirected connected acyclic graph with weighted edges. The weights are interpreted as lengths of the edges, so that the tree is equipped with the shortest path metric d, thus giving rise to the metric tree (T, d). A rigorous construction and topological properties of T can be found in [45, p.7]. Metric trees are important in applications since they can be used to model networks such as road, river, communication or distribution networks.

There is little statistical literature on Fréchet p-means in the specific setting of metric trees. Basrak [18] focuses on the Fréchet mean in a binary metric tree, and he establishes a central limit theorem for the inductive mean. Risser et al. [96, 98] seek to compute Fréchet means on metric graphs, while Hotz et al. [126] develop laws of large numbers and central limit theorems when the ambient space is an open book. A special case of an open book is the m-spider, which can be viewed as a peculiar kind of metric tree.

An adjacent topic that has attracted greater attention is that of stratified spaces [138], i.e., spaces that are finite unions of disjoint subspaces. Examples of CAT(0) stratified spaces include open books [126] and the Billera–Holmes–Vogtmann tree space [37]. The elements of this space are trees and it was introduced in order to evaluate the proximity between multiple phylogenetic trees.

1.3.4 Summary of our results

Chapter 5 is based on Romon and Brunel [224], which is currently under review.

The ambient space E that we consider is a metric tree. We add mild assumptions on E so that it is a compact Hadamard space. We consider location parameters defined via the generic contrast $\varphi : (x, \theta) \mapsto \ell\big(d(x, \theta)\big)$, where $\ell : [0, \infty) \to [0, \infty)$ is a convex nondecreasing function. We call them Fréchet ℓ-means.

We leverage the geodesic convexity of the objective function ϕ and the geometry of the tree to define a notion of directional derivative for ϕ. This helps us locate and characterize Fréchet ℓ-means.

Estimation is performed using the standard M-estimator. We extend to metric trees the notion of stickiness defined by Hotz et al. [126]: a Fréchet ℓ-mean is either sticky or partly sticky. We show that empirical stickiness is a non-asymptotic phenomenon that we quantify with exponential bounds. As an immediate consequence we obtain a sticky law of large numbers.

Then, we focus on Fréchet medians. We begin by providing more precise results on their location and uniqueness. In the partly sticky case, we develop non-asymptotic concentration bounds and central limit theorems.

Chapter 2

Introduction en français

Trois problèmes sont abordés dans ce manuscrit: l'inférence pour la régression multi-tâche en grande dimension, les quantiles géométriques dans les espaces de Banach de dimension infinie, et les ℓ-moyennes de Fréchet dans les arbres métriques. Ces problèmes ne sont pas sans rapport: nous verrons plus tard dans l'introduction qu'ils sont connectés par le fil de l'inférence pour la M-estimation. Chaque sujet a un chapitre dédié dans le manuscrit. L'introduction qui suit fournit une vue d'ensemble pour chacun des thèmes, le but étant de fournir des éléments de contexte importants. Le lecteur est averti que les notations employées peuvent changer d'un chapitre à l'autre.

2.1 Régression multi-tâche sparse en grande dimension

2.1.1 Le Lasso et son débiaisage

Dans le modèle de régression linéaire Gaussien avec n observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, chaque réponse $y_i \in \mathbb{R}$ est une fonction linéaire du vecteur $\boldsymbol{x}_i \in \mathbb{R}^p$, contaminée par un bruit Gaussien $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i,$$

avec $\boldsymbol{\beta}^* \in \mathbb{R}^p$ le vecteur du paramètre inconnu. En définissant $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\varepsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ et $\boldsymbol{X}$ la matrice de design dont les lignes sont $\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top$ (qui peuvent être déterministes ou aléatoires), le modèle se réécrit sous forme matricielle comme

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}.$$

Dans le cadre de faible dimension où $p \leq n$ et $\boldsymbol{X}$ est de rang plein, un estimateur classique de $\boldsymbol{\beta}^*$ est celui des moindres carrés ordinaires $\widehat{\boldsymbol{\beta}}^{(\text{ols})} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, qui est solution du problème des moindres carrés

$$\operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

$\widehat{\boldsymbol{\beta}}^{(\mathrm{ols})}$ est un estimateur sans biais de $\boldsymbol{\beta}^*$ et sa loi est $\mathcal{N}_p(\mathbf{0}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$. Ceci rend possible l'inférence statistique sur $\boldsymbol{\beta}^*$, i.e., les tests d'hypothèse et la construction d'intervalles de confiance et de régions de confiance pour les coefficients de $\boldsymbol{\beta}^*$ [8].

Le cadre de la grande dimension où p peut être beaucoup plus grand que n a attiré une grande attention au cours des deux dernières décennies. Dans ce cadre, la matrice $\boldsymbol{X}^\top \boldsymbol{X}$ n'est pas inversible, ce qui nécessite l'introduction d'autres estimateurs. Quand il est suspecté que seul un petit nombre de variables explicatives contribuent à la réponse, i.e., quand le vecteur $\boldsymbol{\beta}^*$ est s-sparse, un estimateur approprié est le Lasso [251] $\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}$ qui est solution du problème de minimisation

$$\operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

où $\lambda > 0$ est un paramètre de régularisation choisi par l'utilisateur.

La performance du Lasso pour la prédiction, l'estimation et la récupération du support a été étudiée en détail dans les années 2000; voir, e.g., [106, 56, 35, 188, 282, 268]. Ces résultats sont fondamentaux, et pourtant ils sont insuffisants pour l'inférence statistique sur des quantités $f(\boldsymbol{\beta}^*)$ où $f : \mathbb{R}^p \to \mathbb{R}^d$ avec d petit devant p, e.g., pour l'inférence sur un seul coefficient du vecteur $\boldsymbol{\beta}^*$. Par exemple, les inégalités oracle [269, Equation (7.26)] donnent des intervalles de confiance pour $\boldsymbol{\beta}_1^*$ qui sont de taille $\asymp \sqrt{(s \log p)/n}$, ce qui est loin d'être optimal. Par ailleurs, $\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}$ n'a pas de loi limite raisonnable, même dans le cadre de faible dimension [153].

Contrairement aux moindres carrés, on peut prouver que le Lasso est biaisé [139, Corollary 11] et le biais est plus grand pour les gros coefficients de $\boldsymbol{\beta}^*$. Ceci a motivé la construction d'autres estimateurs en partant de $\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}$, qui aient de meilleures propriétés inférentielles.

Dans les années 2010, sous l'hypothèse d'un design aléatoire avec des lignes i.i.d. de covariance $\boldsymbol{\Sigma}$, les travaux pionniers [280, 51, 256, 139] ont introduit des estimateurs débiaisés $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})}$ de la forme $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})} = \widehat{\boldsymbol{\beta}}^{(\mathrm{L})} + \frac{1}{n}\boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(\mathrm{L})})$ où $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ est choisi de sorte à approcher la matrice de précision $\boldsymbol{\Sigma}^{-1}$. En exploitant $\widehat{\boldsymbol{\beta}}^{(\mathrm{d})}$, ces travaux construisent des intervalles de confiance pour un seul coefficient de $\boldsymbol{\beta}^*$ dans le régime $s \lesssim \sqrt{n}/\log p$. [141] relâche l'hypothèse de sparsité à $s \lesssim n/(\log p)^2$, [284, 42, 54, 55, 285, 26] construisent des intervalles de confiance pour des fonctionnelles générales $\boldsymbol{a}^\top \boldsymbol{\beta}^*$ où $\boldsymbol{a} \in \mathbb{R}^p$. Plus récemment, [193, 26, 27] établissent la nécessité d'un ajustement par degrés de liberté pour traiter des sparsités plus grandes: $\boldsymbol{M}$ devrait idéalement être égal à $\boldsymbol{\Sigma}^{-1}/(1 - \|\widehat{\boldsymbol{\beta}}^{(\mathrm{L})}\|_0/n)$.

2.1.2 Group Lasso et régression multi-tâche

Une hypothèse structurelle plus générale sur le paramètre $\boldsymbol{\beta}^*$ est la groupe-sparsité: l'ensemble d'indices $[1, p]$ est partitionné en m groupes $G_1, \ldots, G_m \subset [1, p]$ connus a priori et il n'y a qu'un petit nombre d'indices $k \in [1, m]$ tels que $\{\boldsymbol{\beta}_j^* : j \in G_k\} \neq \{0\}$. À l'intérieur d'un groupe, toutes les variables sont soit pertinentes soit toutes exclues.

Un estimateur approprié dans ce cas est le Group Lasso [275] $\widehat{\boldsymbol{\beta}}^{(\mathrm{g})}$ qui est solution du problème de minimisation pénalisé par la norme $\ell^{2,1}$ $\|\boldsymbol{\beta}\|_{2,1} = \sum_{k=1}^{m} \big(\sum_{j \in G_k} \boldsymbol{\beta}_j^2 \big)^{1/2}$:

$$\operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_{2,1}.$$

La performance en prédiction et estimation de cet estimateur est analysée dans [201, 130, 174]. La méthodologie susmentionnée pour le débiaisage a été étendue au cadre groupé dans [187, 194, 255, 242, 27].

Un problème connexe est celui de la régression multi-tâche. Nous considérons le modèle linéaire multivarié Gaussien avec T réponses ou tâches

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}$$

où $\boldsymbol{B}^* \in \mathbb{R}^{p \times T}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times T}$ et $\boldsymbol{E}$ a des lignes i.i.d. $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{S})$. L'hypothèse structurelle est que $\boldsymbol{B}^*$ est row-sparse, i.e., de nombreuses variables sont sans intérêt, et ceci pour toutes les tâches. La régression multi-tâche peut être considérée comme un cas de régression groupe-sparse

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{X}} \bar{\boldsymbol{\beta}}^* + \bar{\boldsymbol{\varepsilon}}$$

où $\bar{\boldsymbol{y}} = \mathbf{vec}(\boldsymbol{Y}) \in \mathbb{R}^{\bar{n}}$, $\bar{\boldsymbol{\varepsilon}} = \mathbf{vec}(\boldsymbol{E}) \in \mathbb{R}^{\bar{n}}$, $\bar{\boldsymbol{X}} \in \mathbb{R}^{\bar{n} \times \bar{p}}$ est diagonale par blocs avec des blocs $\boldsymbol{X}$, $\bar{p} = pT$, $\bar{n} = nT$ et les indices $\{1, ..., \bar{p}\}$ sont partitionnés en p groupes de tailles égales.

Le fait que le design $\bar{\boldsymbol{X}}$ soit diagonal par blocs est un obstacle qui empêche l'application directe des résultats inférentiels sur le Group Lasso. L'inférence pour le modèle de régression multi-tâche a été abordée dans [67], qui étend la méthodologie de débiaisage introduite dans [280, 256].

2.1.3 Résumé de nos résultats

Le Chapitre 3 est tiré de Bellec et Romon [23], qui est actuellement en cours d'examen dans une revue.

Les objectifs inférentiels de ce chapitre sont doubles. Premièrement nous construisons des intervalles pour une fonctionnelle du vecteur correspondant à la première tâche $\boldsymbol{a}^\top \boldsymbol{B}^* \boldsymbol{e}_1$, en exploitant les réponses pour chaque tâche simultanément. Deuxièmement, nous construisons des ellipsoïdes de confiance pour les lignes $\boldsymbol{e}_j^\top \boldsymbol{B}^* \in \mathbb{R}^{1 \times T}$ de $\boldsymbol{B}^*$, ce qui nous permet de formuler des tests d'hypothèse pour la nullité de la j-ème ligne de $\boldsymbol{B}^*$, ou de façon équivalente de tester que le signal ne dépend pas de la j-ème variable.

Afin de réaliser ces objectifs, nous introduisons un nouvel objet qui ne dépend que des données: la *matrice d'interaction* $\widehat{\mathbf{A}} \in \mathbb{R}^{T \times T}$. L'introduction de cette matrice est cruciale pour l'obtention de résultats inférentiels. $\widehat{\mathbf{A}} \in \mathbb{R}^{T \times T}$ généralise au cadre multi-tâche les méthodes de débiaisage, en particulier les ajustements par degrés de liberté.

Lorsque $\boldsymbol{\Sigma}$ est supposée inconnue, les résultats issus de la littérature en régression groupée nécessitent une sparsité $s \lesssim \sqrt{n}$ à des constantes près dépendant de T et de facteurs logarithmiques en (n,p). En exploitant la matrice d'interaction nous obtenons des lois limites normales et χ_2 sous les conditions $\min(T^2, \log^8 p)/n \to 0$ et

$$\frac{sT + s\log(p/s) + \|\mathbf{\Sigma}^{-1}\boldsymbol{e}_j\|_0 \log p}{n} \to 0, \quad \frac{\min(s, \|\mathbf{\Sigma}^{-1}\boldsymbol{e}_j\|_0)}{\sqrt{n}} \sqrt{[T + \log(p/s)]\log p} \to 0,$$

ce qui couvre le régime de sparsité $s \ggg \sqrt{n}$ lorsque $\|\mathbf{\Sigma}^{-1}\boldsymbol{e}_j\|_0\sqrt{T} \lll \sqrt{n}$ à des facteurs logarithmiques près.

2.2 M-estimation, dimension infinie et quantiles

Les moindres carrés ordinaires, le Lasso et le Group Lasso introduits dans la section précédente sont définis comme solution d'un problème de minimisation. Il s'agit d'exemples de M-estimation, notion sur laquelle nous nous concentrons ci-après.

2.2.1 M-estimation: résultats classiques

Un paramètre de population $\theta_\star$ est souvent défini de manière implicite comme un minimiseur d'une fonction objectif ϕ du type suivant:

$$\begin{aligned}\phi \colon \Theta &\to \mathbb{R} \\ \theta &\mapsto \int_{\mathcal{X}} \varphi(x,\theta) d\mu(x),\end{aligned} \tag{2.1}$$

où Θ est l'espace des paramètres, $(\mathcal{X}, \mathcal{A}, \mu)$ est un espace probabilisé et $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$ est une fonction de contraste intégrable du premier argument. Étant donné un échantillon i.i.d. $X_1, \ldots, X_n \sim \mu$, un M-estimateur $\widehat{\theta}_n$ de $\theta_\star$ est défini comme un minimiseur de la fonction objectif $\widehat{\phi}_n$

$$\widehat{\phi}_n : \theta \mapsto \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i, \theta), \tag{2.2}$$

qui est obtenue en remplaçant la mesure en population μ dans (2.1) par la mesure empirique $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

Les éléments aléatoires $X_1, X_2, \ldots$ sont définis sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. Pour simplifier l'exposition nous ne considérons pas les problèmes de mesurabilité dans cette introduction. Nous supposons par conséquent que $\widehat{\theta}_n$ est mesurable entre les tribus $\mathcal{F}$ et $\mathcal{A}$.

Ce cadre général a été formulé par Huber [131], qui utilise la lettre "M" comme abréviation de "minimize". Des exemples classiques de M-estimation dans le cadre Euclidien (i.e., Θ est une partie de $\mathbb{R}^d$, muni de sa structure hilbertienne classique) incluent le cas où:

1. $\Theta = \mathcal{X} = \mathbb{R}^d$ avec $d \geq 1$, μ est une mesure de probabilité ayant un moment d'ordre 1 et $\varphi : (x,\theta) \mapsto \|x-\theta\|_2^2 - \|x\|_2^2$. Ici $\theta_\star$ est l'espérance de μ, i.e., $\theta_\star = \int_{\mathbb{R}^d} x d\mu(x)$ et $\widehat{\theta}_n$ est la moyenne empirique: $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

2. $\Theta = \mathcal{X} = \mathbb{R}$, μ est une mesure de probabilité et $\varphi : (x, \theta) \mapsto |x - \theta| - |x|$. Les minimiseurs de ϕ (potentiellement en nombre infini) sont les médianes de μ et ceux de $\widehat{\phi}_n$ sont les médianes empiriques usuelles, qui s'écrivent explicitement en termes des statistiques d'ordre $X_{(1)} \leq \ldots \leq X_{(n)}$: lorsque n est impair $X_{(\lfloor \frac{n}{2} \rfloor + 1)}$ est l'unique médiane empirique, et lorsque n est pair l'argmin est l'intervalle $[X_{(\frac{n}{2})}, X_{(\frac{n}{2}+1)}]$.

3. $\Theta = \mathcal{X} = \mathbb{R}$, μ est une mesure de probabilité ayant un moment d'ordre 1 et $\varphi : (x, \theta) \mapsto (x - \theta)^2 \mathbb{1}_{|x-\theta| \leq c} + (2c|x - \theta| - c^2)\mathbb{1}_{|x-\theta|>c}$ où $c \geq 0$. Ce contraste a été introduit par Huber [131] pour des questions de robustesse.

4. $\Theta \subset \mathbb{R}^d$ où $d \geq 1$, μ est contenue dans une famille paramétrique $(\mathbb{P}_\theta)_{\theta \in \Theta}$ (i.e., $\mu = \mathbb{P}_{\theta_0}$ pour un $\theta_0 \in \Theta$), la famille est dominée par une mesure σ-finie ν, les densités correspondantes $(f_\theta)_{\theta \in \Theta}$ sont strictement positives ν-p.p., pour tout $\theta \in \Theta$,
$$\int_{\mathcal{X}} |\ln f_\theta(x)| f_{\theta_0}(x) d\nu(x) < \infty$$
et $\varphi : (x, \theta) \mapsto -\ln f_\theta(x)$. Tout minimiseur de $\theta_\star$ vérifie $\mathbb{P}_{\theta_\star} = \mu$ et il est unique si et seulement si la famille $(\mathbb{P}_\theta)_{\theta \in \Theta}$ est identifiable. Minimiser $\widehat{\phi}_n$ coincide avec l'estimation classique par maximum de vraisemblance.

5. $\Theta = \mathbb{R}^d$, $\mathcal{X} = \mathbb{R}^{d+1}$ with $d \geq 1$, μ est une mesure de probabilité ayant un moment d'ordre 2, le vecteur aléatoire $(X, Y) \sim \mu$ vérifie $\mathbb{E}[Y|X] = \theta_0^\top X$ pour un $\theta_0 \in \mathbb{R}^d$ et $\varphi : \big((x, y), \theta\big) \mapsto (y - \theta^\top x)^2$. Dans ce cas, la M-estimation est un cas particulier des moindres carrés ordinaires.

Consistence, vitesse de convergence et loi limite

Supposons que ϕ a un unique minimiseur $\theta_\star$ qui est le paramètre inconnu d'intérêt. Une première étape vers une estimation satisfaisante de $\theta_\star$ est la consistence de la suite $(\widehat{\theta}_n)_{n \geq 1}$, i.e., une forme de convergence stochastique vers $\theta_\star$. Pour quantifier ce phénomène nous imposons à partir de maintenant que Θ soit un espace métrique avec une distance d, et nous disons que $(\widehat{\theta}_n)_{n \geq 1}$ est fortement consistent (resp., faiblement consistent) si $d(\widehat{\theta}_n, \theta_\star)$ converge presque sûrement (resp., en probabilité) vers 0.

Eu égard à la généralité du cadre d'estimation (2.1), les statisticiens ont tenté d'établir des conditions générales de consistence qui couvrent une variété de contrastes φ. Les conditions de consistence qui suivent sont classiques et énoncées dans [261, 259].

Proposition 2.1 ([261, Corollary 3.2.3], [259, Theorem 5.7])**.** *1. Si*

$$\sup_{\theta \in \Theta} |\widehat{\phi}_n(\theta) - \phi(\theta)| \xrightarrow[n \to \infty]{\mathbb{P}} 0 \tag{2.3}$$

et

$$\forall \varepsilon > 0, \inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_\star\| \geq \varepsilon}} \phi(\theta) > \phi(\theta_\star), \tag{2.4}$$

alors $(\widehat{\theta}_n)_{n \geq 1}$ est faiblement consistent.

2. *Si* (2.3) *est remplacé par la convergence uniforme sur tout compact, si* (2.4) *est vraie et sous l'hypothèse*

$$\forall \varepsilon > 0, \exists K \, compact \,, \forall n \geq 1, \; \mathbb{P}(\widehat{\theta}_n \in K) \geq 1 - \varepsilon,$$

alors $(\widehat{\theta}_n)_{n\geq 1}$ *est faiblement consistent.*

Proposition 2.2 ([259, Theorem 5.14]). *Si* $\theta \mapsto \varphi(x, \theta)$ *est semicontinue inférieurement pour* μ*-presque tout* x *et* $\forall \theta \in \Theta, \exists r > 0, \mathbb{E}\Big[\inf_{\substack{\alpha \in \Theta \\ \|\alpha - \theta\| \geq r}} \varphi(X_1, \alpha) \Big] < \infty$, *alors pour tout* $\varepsilon > 0$ *et tout compact* $K \subset \Theta$,

$$\mathbb{P}\big(\{ d(\widehat{\theta}_n, \theta_\star) \geq \varepsilon \} \cap \{ \widehat{\theta}_n \in K \} \big) \xrightarrow[n \to \infty]{} 0. \tag{2.5}$$

Commentons brièvement ces résultats. La condition de convergence uniforme stochastique (2.3) est équivalente au fait que la classe de fonctions $\big(\varphi(\cdot, \theta)\big)_{\theta \in \Theta}$ soit μ-Glivenko–Cantelli. Ceci peut être déterminé en utilisant des outils issus de la théorie des processus empiriques, tels que le bracketing et l'entropie [261, Part 2]. Par exemple, la classe est Glivenko–Cantelli si elle est compacte point par point, c'est-à-dire qu'elle est dominée par une fonction intégrable, que Θ est un espace métrique compact et que $\theta \mapsto \varphi(x, \theta)$ est continu pour tout $x \in \mathcal{X}$ [259, Example 19.8].

Sous la condition (2.4), si θ est séparé de $\theta_\star$ alors $\phi(\theta)$ ne peut pas être arbitrairement proche de la valeur minimale de ϕ. Dans ce cas le minimiseur de $\theta_\star$ est dit "bien séparé". Lorsque ϕ is semicontinue inférieurement et Θ est un espace métrique compact, $\theta_\star$ est automatiquement bien séparé. L'énoncé de convergence (2.5) donne la consistence faible si on sait exhiber une partie compacte K telle que $\lim_{n\to\infty} \mathbb{P}(\widehat{\theta}_n \in K) = 1$.

À la lumière de cette discussion, il apparait que les parties compactes de Θ et la compacité de Θ lui-même joue un rôle dans la preuve de la consistence. D'autres conditions générales sont formulées dans [132, 212, 120, 83]. Elles sont similaires en substance aux conditions énoncées précédemment et s'appuient également sur la compacité.

Une fois la consistence obtenue il est intéressant de quantifier la vitesse de convergence, i.e., trouver une suite de réels $(r_n)_{n\geq 1}$ telle que $\lim_n r_n = \infty$ et $r_n d(\widehat{\theta}_n, \theta_\star) = O_{\mathbb{P}}(1)$. Un résultat général dans cette direction est [261, Corollary 3.2.6], qui requiert que ϕ croisse localement à une vitesse au moins quadratique, i.e., $\phi(\theta) \geq \phi(\theta_\star) + c d(\theta, \theta_\star)^2$ pour une constante $c > 0$ et tout θ dans un voisinage de $\theta_\star$, et qu'il y ait un contrôle du processus empirique indexé par la classe $\mathcal{M}_\delta = \{\varphi(\cdot, \theta) - \varphi(\cdot, \theta_\star) : d(\theta, \theta_\star) < \delta\}$ où δ varie dans un voisinage de 0. Cette condition de croissance est immédiate lorsque Θ est un espace Euclidien et ϕ est deux fois différentiable en $\theta_\star$ avec une Hessienne $\nabla^2 \phi(\theta_\star)$ inversible. La seconde condition peut être vérifiée en majorant une intégrale d'entropie uniforme ou une intégrale de bracketing de $\mathcal{M}_\delta$. Dans le cas particulier où Θ est Euclidien et que le contraste est Hölder de la deuxième variable, i.e., pour tout θ_1, θ_2 dans un voisinage de $\theta_\star$ et μ-presque tout $x \in \mathcal{X}$,

$$|\varphi(x, \theta_1) - \varphi(x, \theta_2)| \leq C(x) \|\theta_1 - \theta_2\|^\alpha, \tag{2.6}$$

l'intégrale de bracketing de $\mathcal{M}_\delta$ est facilement majorée en utilisant les covering numbers des boules dans $\mathbb{R}^d$. En pratique, pour obtenir une vitesse de convergence il est donc commode que Θ soit un espace Euclidien. Dans les espaces non-Euclidiens tels que les espaces normés de dimension infinie, il peut être approprié de considérer un M-estimateur tamisé au lieu de $\widehat{\theta}_n$: étant donné une suite croissante de parties $\Theta_n \subset \Theta$, la minimisation est effectué sur Θ_n au lieu de l'espace entier. La M-estimation tamisée peut être vue comme une forme de régularisation, ce qui peut permettre d'éviter le surapprentissage. Des vitesses pour les M-estimateurs tamisés sont énoncées dans [261, Chapter 3.4].

La vitesse r_n a le bon ordre si de surcroît $\left(r_n d(\widehat{\theta}_n, \theta_\star)\right)_{n\geq 1}$ converge en loi. Des résultats génériques de ce type sont ordinairement énoncés dans le cas où Θ est un espace Euclidien. Un tel résultat est [261, Theorem 3.2.10], qui exploite la théorie des processus empiriques. Un autre basé sur la linéarisation est [261, Theorem 3.2.16], qui a la conclusion plus précise $r_n(\widehat{\theta}_n - \theta_\star) = -[\nabla^2\phi(\theta_\star)]^{-1} Z_n + o_{\mathbb{P}}(1)$ où $(Z_n)_{n\geq 1}$ est une suite tendue de vecteurs aléatoires. Il est souvent (mais pas toujours, voir [151]) vrai que $r_n = \sqrt{n}$ et qu'on a le développement stochastique:

$$\sqrt{n}(\widehat{\theta}_n - \theta_\star) = -[\nabla^2\phi(\theta_\star)]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \varphi(X_i, \theta_\star) + o_{\mathbb{P}}(1). \tag{2.7}$$

Le théorème central limite implique alors que $\sqrt{n}(\widehat{\theta}_n - \theta_\star)$ est asymptotiquement normal: il y a convergence en loi vers une normale multivariée centrée de variance-covariance $[\nabla^2\phi(\theta_\star)]^{-1}\mathbb{E}\left[\nabla_\theta\varphi(X_1, \theta_\star)\nabla_\theta\varphi(X_1, \theta_\star)^\top\right][\nabla^2\phi(\theta_\star)]^{-1}$. Ceci est vrai par exemple lorsque le contraste satisfait (2.6) with $\alpha = 1$ [261, Example 3.2.22]. Des résultats du même acabit sont énoncés dans [132, 214, 215].

Le cas convexe

La situation suivante est commune: Θ est un ouvert convexe de $\mathbb{R}^d$ et le contrast est convexe du deuxième argument, i.e., pour μ-presque tout $x \in \mathcal{X}$, la fonction $\theta \mapsto \varphi(x, \theta)$ est convexe. La convexité simplifie grandement les énoncés de consistence et de normalité asymptotique, comme observé dans [108, 199, 119].

Proposition 2.3 ([108, Theorem 5.1],[199, Theorem 1])**.** *Sous l'hypothèse de convexité,* $(\widehat{\theta}_n)_{n\geq 1}$ *est fortement consistent.*

Proposition 2.4 ([199, Theorem 4])**.** *Pour tout $x \in \mathcal{X}$ et $\theta \in \Theta$, on note $g(x, \theta)$ un sous-gradient en θ de $\varphi(x, \cdot)$. Supposons que $\mathbb{E}[\|g(X, \theta)\|_2^2] < \infty$ pour tout θ dans un voisinage de $\theta_\star$, et que ϕ est deux fois différentiable en $\theta_\star$ avec une Hessienne invertible. Alors on a le développement stochastique* (2.7)*.*

Ces énoncés sont génériques, et pourtant les hypothèses sont considérablement plus simples que celles requises pour les résultats évoqués précédemment qui exploitent la théorie des processus empiriques.

2.2.2 Le défi de la dimension infinie

Dans cette sous-section l'espace des paramètres Θ est doté d'une structure linéaire et il est supposé de dimension infinie. Plus précisément, Θ est un espace de Banach de dimension infinie, ou plus spécifiquement un espace de Hilbert de dimension infinie. La théorie des probabilités dans les espaces de Banach s'est développée dans les années 1970; voir, e.g., les monographies [161, 9, 252, 168].

La dimension infinie est un cadre naturel lorsque les données vivent dans un espace fonctionnel, par exemple dans la modélisation de courbes (e.g., les données spectrométriques, la consommation d'électricité, les électrocardiogrammes, le cours de la bourse). L'analyse des données fonctionnelles est le domaine de recherche correspondant et elle s'est considérablement développée depuis les années 1990 (voir, e.g., [217, 93, 124, 127, 270]). Dans la littérature, la modélisation est faite habituellement dans l'espace de Hilbert L^2, toutefois il y a un intérêt récent pour les espaces non-Hilbertiens [74]. Un autre contexte pour la statistique de dimension infinie est celui des méthodes à noyaux [69, 121, 281, 198, 173].

Une caractéristique essentielle des espaces normés de dimension infinie est que les boules fermées et les sphères ne sont pas compactes pour la topologie normique [7, Theorem 5.26]. Par conséquent, les parties compactes de Θ sont d'intérieur vide. Un moyen pour créer des compacts est de fixer un sous-espace vectoriel $V \subset \Theta$ de dimension finie et de considérer ses parties $K \subset V$ qui sont à la fois fermées et bornées. Réciproquement, le résultat suivant montre que tous les compacts de Θ sont approximativement de dimension finie.

Proposition 2.5 ([9, Lemma 4.3], [168, Lemma 2.2])**.** *Soit Θ un Banach et $K \subset \Theta$. K est une partie compacte de Θ pour la topologie normique si et seulement si les deux conditions suivantes sont vérifiées:*

1. *K est fermé et borné.*
2. *Pour tout $\varepsilon > 0$, il existe V un sous-espace vectoriel de dimension finie tel que pour tout $x \in K$, $d(x, V) < \varepsilon$.*

Les compacts en dimension infinie sont donc plutôt pathologiques, et le statisticien peut interpréter ce phénomène comme un *fléau de la dimension infinie* (à ne pas confondre avec le fléau synonyme en analyse des données fonctionnelles [93, 99]). En effet, nous avons vu précédemment que les résultats classiques de consistence en M-estimation sont plus faciles à obtenir en tirant parti de la compacité. Le contrôle des covering numbers ou bracketing numbers sous-tend plusieurs des résultats de consistence, vitesse et loi limite mentionnés au-dessus. Ce contrôle est parfois réalisé en maîtrisant les covering ou bracketing numbers de l'ensemble d'indexation. Par exemple, sous l'hypothèse Hölder (2.6) et quand $\Theta = \mathbb{R}^d$, il est possible d'exploiter les bornes sur le covering number des boules dans $\mathbb{R}^d$. Par opposition, lorsque Θ est de dimension infinie, les boules ne sont pas précompactes et une telle technique ne fonctionne pas. On peut aussi remarquer que l'énoncé de loi limite [261, Theorem 3.2.10] repose sur la précompacité des boules, il n'est donc pas applicable en dimension infinie.

Malheureusement le cas convexe n'est pas épargné par ce fléau. La preuve des Propositions 2.3 and 2.4 s'appuie crucialement sur le résultat suivant d'analyse convexe.

Proposition 2.6 ([287, Corollary 2.2.23])**.** *Soit E un Banach et Θ un ouvert convexe de E. Soit $(f_n)_{n \geq 1}$ une suite de fonctions convexes sur Θ qui converge simplement vers un certain f. Alors $(f_n)_{n \geq 1}$ converge uniformément sur tout compact de Θ vers f.*

Dans les preuves des Propositions 2.3 and 2.4, la convergence uniforme est naturellement appliquée aux boules fermées, qui sont compactes en dimension finie.

Nous n'avons pas réussi à développer une théorie générale de la M-estimation en dimension infinie, même dans le cas convexe. La M-estimation en dimension infinie a été étudiée par van der Vaart [257, 258, 260] qui formule des énoncés avec des hypothèses similaires à celles de [261]. Pour l'étude du M-estimateur en régression, on renvoie aux travaux successifs [79, 80, 78, 160]. Plus récemment, Sinova et al. [238] ont pour objectif de développer une théorie générale de la M-estimation dans les espaces de Hilbert, en mettant l'accent sur l'espace fonctionnel L^2. Comme remarqué par les auteurs, leur théorème de consistence [238, Theorem 3.4] n'est vrai qu'en dimension finie.

Notre contribution à la M-estimation en dimension infinie est l'étude d'un M-estimateur particulier, qui est introduit dans la sous-section qui suit.

2.2.3 Quantiles: de $\mathbb{R}$ à la dimension infinie

Quantiles univariés

Étant donné une mesure de probabilité μ sur $\mathbb{R}$ et $p \in (0,1)$ un paramètre élémentaire de localisation est le p-quantile de μ, qui est habituellement défini comme n'importe quel $\alpha \in \mathbb{R}$ satisfaisant à la fois

$$\mu((-\infty, \alpha]) \geq p \quad \text{et} \quad \mu([\alpha, \infty)) \geq 1 - p. \tag{2.8}$$

Un cas particulier important est la médiane, i.e., lorsque $p = \frac{1}{2}$. Les quantiles sont essentiels car ils fournissent une mesure du degré de centralité, et la médiane peut être interprétée comme une tendance centrale de la loi. Les quantiles ont des applications pour les tests d'hypothèse [233], pour la régression [154] et pour la statistique robuste [133, 192, 177].

Il est connu (voir, e.g., [259, p.44]) que le p-quantile est compatible avec la M-estimation: avec les notations précédentes, soit $\Theta = \mathcal{X} = \mathbb{R}$, supposons que μ a une espérance et considérons le contraste $\varphi : (x, \alpha) \mapsto (1-p)(x-\alpha)^+ + p(x-\alpha)^-$. De manière alternative, on peut omettre l'hypothèse de moment en définissant le contraste

$$\varphi : (x, \alpha) \mapsto |x - \alpha| - |x| - (2p-1)\alpha. \tag{2.9}$$

Quantiles géométriques

La définition (2.8) s'appuie sur les intervalles $(-\infty, \alpha]$ et $[\alpha, \infty)$, et on peut donc affirmer qu'elle repose sur l'ordre de $\mathbb{R}$. Mesurer la centralité est un sujet important en statistique multivariée, d'où la nécessité de généraliser les quantiles aux dimensions supérieures. Comme il n'y a pas d'ordre naturel sur $\mathbb{R}^d$, étendre la formulation (2.8) n'est pas pratique. La littérature afférente aux généralisations des mesures de centralité en dimension supérieure est riche; voir e.g., [239, 288, 232, 197] et les références qui s'y trouvent.

Considérons un espace vectoriel normé $(E, \|\cdot\|)$ et soit $\Theta = \mathcal{X} = E$. Une généralisation naturelle du contraste (2.9) est

$$\varphi : (x, \alpha) \mapsto \|x - \alpha\| - \|x\| - \ell(\alpha), \tag{2.10}$$

où ℓ est une fonctionnelle continue de norme strictement inférieure à 1. Le M-estimateur correspondant est appelé quantile géométrique ou quantile spatial. La médiane géométrique (i.e., lorsque $\ell = 0$) a été introduite dans le cas bidimensionnel Euclidien par Weber [272] en 1909, et a été réintroduit ultérieurement dans le même cadre par Gini et Galvani [103, 225] ainsi que Haldane [110]. Valadier [253, 254] a étendu le concept à n'importe quel Banach réflexif et Kemperman [148] a effectué une étude systématique d'existence et d'unicité dans les espaces de Banach. Chaudhuri [65] et Koltchinskii [156, 157] ont défini les quantiles géométriques dans les Banach en ajoutant la fonctionnelle ℓ à la fonction objectif.

2.2.4 Résumé de nos résultats

Le Chapitre 4 est tiré de Romon [223], qui est actuellement en cours d'examen dans une revue.

Nous étudions les propriétés asymptotiques des quantiles géométriques dans les Banach de dimension infinie.

Nous commençons par des résultats descriptifs sur les médianes en population. L'estimation est effectuée avec un M-estimateur approché. Quand le quantile en population n'est pas unique nous utilisons la théorie de la convergence variationnelle pour obtenir des résultats asymptotiques sur les sous-suites dans la topologie faible. Quand le quantile en population est unique, nous montrons la consistence forte de l'estimateur pour la topologie normique. Notre théorème est valide sous des hypothèses minimales sur μ et dans n'importe quel espace séparable et uniformément convexe. Dans un Hilbert séparable nous obtenons la normalité asymptotique. Notre théorème central limite est formulé sous des hypothèses correspondant exactement à celles de la dimension finie.

2.3 M-estimation dans les arbres métriques

2.3.1 M-estimation dans les espaces métriques et moyennes de Fréchet

Dans la section précédente, l'accent a été mis sur le cas où les données résident dans un espace vectoriel normé E.

Un tel modèle n'est parfois pas réaliste, car les données peuvent se situer dans un sous-ensemble non linéaire $S \subset E$, et la distance induite par la norme peut ne pas être significative.

Les exemples vont des données sur la sphère $\mathbb{S}^d$, qui sont l'objet de la statistique directionnelle [181, 171], aux données sous la forme de matrices symétriques définies positives (utilisées pour modéliser les matrices de covariance, par exemple dans l'imagerie par tenseur de diffusion [165] et qui ont également trouvé une utilisation dans la segmentation d'images en vision par ordinateur [221, 60]), aux données dans les espaces de mesure (comme les espaces de Wasserstein, qui sont au coeur du transport optimal [211]), ou aux données dans des espaces quotient (par exemple, lorsqu'un praticien s'intéresse aux formes d'objets, les données peuvent être analysées modulo les translations, rotations et mises à l'échelle, ce qui les fait appartenir à un espace quotient, également appelé espace de forme [149, 31]).

Le cadre de la M-estimation introduit dans la section précédente est général et s'applique également dans le cadre non linéaire [48]. Comme mentionné dans [138], on peut mentionner par exemple:

1. L'extension de l'analyse en composantes principales aux données à valeurs dans des variétés ([94, 134, 135, 137]): par exemple, dans [137], la première composante principale géodésique dans l'espace de forme planaire Σ_2^k est définie via un problème de minimisation où $\Theta = \Gamma(\Sigma_2^k)$ est l'espace des géodésiques sur Σ_2^k et $\mathcal{X} = \mathbb{S}_2^k$ est la sphère pré-forme.

2. L'extension des mesures Euclidiennes de tendance centrale (comme la moyenne et la médiane) au cadre des espaces métriques : $\Theta = \mathcal{X} = E$, où (E, d) est un espace métrique. C'est le sujet de notre présentation pour le reste de cette section.

L'espérance d'une mesure μ sur $\mathbb{R}^d$ est définie de la manière la plus élémentaire comme le vecteur des moyennes de chaque coordonnée. Dans les espaces de Banach de dimension infinie, la moyenne est classiquement définie comme une intégrale de Bochner [75]. Dans les deux cas, la définition dépend crucialement de la structure linéaire de l'espace ambiant. Comme indiqué précédemment dans la section 2.2.1, lorsque $\Theta = \mathcal{X} = \mathbb{R}^d$, la fonction de contraste $\varphi : (x, \theta) \mapsto \|x - \theta\|_2^2 - \|x\|_2^2$ donne l'espérance de la mesure μ. Cette fonction de contraste offre une extension naturelle du concept de moyenne aux espaces métriques: fixons un point arbitraire $o \in E$, supposons que $\int_E d(x, o) d\mu(x) < \infty$ et définissons

$$\varphi : (x, \theta) \mapsto d(x, \theta)^2 - d(x, o)^2. \tag{2.11}$$

En supposant un moment d'ordre 2, c'est-à-dire si $\int_E d(x, o)^2 d\mu(x) < \infty$ pour un (et donc pour tout) $o \in E$, le contraste peut être remplacé par la fonction plus simple $(x, \theta) \mapsto d(x, \theta)^2$, et on peut définir une variance de Fréchet. Par commodité, nous notons $M(\mu)$ l'ensemble des minimiseurs de la fonction objectif correspondante. Les éléments de $M(\mu)$ sont connus sous le nom de moyennes de Fréchet [95], de barycentres ou de centres de masse. L'existence et l'unicité des moyennes de Fréchet dépendent de la géométrie de l'espace E, et c'est un sujet de recherche de longue date; voir, par exemple, [147, 164, 243, 3, 202, 273, 166, 5, 167].

En ce qui concerne l'estimation, nous nous concentrons uniquement sur le M-estimateur obtenu en minimisant l'objectif empirique (2.2). Cependant, un autre estimateur populaire est la moyenne inductive introduite par Sturm [243]. Étant donné que la moyenne de Fréchet peut ne pas être définie de manière unique, les résultats de consistence quantifient une forme de proximité entre l'ensemble stochastique $(\widehat{\mu}_n)$ et l'ensemble cible $M(\mu)$ lorsque n tend vers l'infini. Ziezold [286] a démontré la consistence lorsque E est un espace métrique séparable, tandis que Bhattacharya et Patrangenaru [33] l'ont fait lorsque E possède la propriété de Heine-Borel, c'est-à-dire lorsque tout ensemble fermé et borné est compact. Des théorèmes central limite ont été développés dans le cas où E est une variété Riemannienne [34, 31, 32, 87]. Le caractère non-Euclidien de l'espace donne lieu à de nouveaux phénomènes asymptotiques tels que la "stickiness" [126, 136] et la "smeariness" [125, 86].

2.3.2 Moyennes de Fréchet dans les espaces de Hadamard

Ensuite, nous introduisons brièvement la notion de courbure négative au sens d'Alexandrov, une caractéristique géométrique de l'espace E qui joue un rôle clé dans l'analyse des moyennes de Fréchet. Des présentations détaillées sont disponibles dans les monographies [145, 45, 52, 15].

Soient $x, y \in E$. Une géodésique à vitesse constante de x à y est une application γ définie sur un intervalle $[a, b] \subset \mathbb{R}$ vers E, telle que $\gamma(a) = x$, $\gamma(b) = y$, et que $d(\gamma(t_1), \gamma(t_2)) = v|t_1 - t_2|$ pour un certain $v \in [0, \infty)$ et pour tout $t_1, t_2 \in [a, b]$. Le nombre réel v est appelé la vitesse de la géodésique γ. L'image de γ est notée $[x, y]$ et est appelée segment géodésique joignant x et y. Pour plus de lisibilité, nous écrirons souvent γ_t au lieu de $\gamma(t)$.

Si pour tout $x, y \in E$, il existe un segment géodésique joignant x et y, alors on dit que (E, d) est un espace métrique géodésique. Si, de plus, un tel segment géodésique est unique, alors on dit que (E, d) est uniquement géodésique. Illustrons cela avec quelques exemples. Lorsque E est un espace vectoriel normé avec la norme $\|\cdot\|$, le segment de droite $(1-t)x + ty : t \in [0, 1]$ est un segment géodésique correspondant à la géodésique $t \mapsto x + t(y - x)$ définie sur $[0, 1]$ ayant pour vitesse $\|y - x\|$. Ainsi, un espace vectoriel normé est un espace métrique géodésique, mais il n'est pas nécessairement uniquement géodésique (cf. $\mathbb{R}^2$ avec la norme ℓ^1). La sphère $\mathbb{S}^2$ est classiquement équipée de la distance angulaire, c'est-à-dire la distance Riemannienne standard. La sphère est géodésique; plus précisément, un segment géodésique est un arc mineur d'un grand cercle. Cependant, $\mathbb{S}^2$ n'est pas uniquement géodésique: si x et y sont antipodaux, il existe une infinité de segments géodésiques entre x et y.

Avant de pouvoir définir la notion de courbure négative, nous avons besoin du concept de triangle de comparaison. Supposons à partir de maintenant que E est uniquement géodésique et fixons $x, y, z \in E$. Le triangle géodésique de sommets x, y, z est l'union de segments géodésiques $[x, y] \cup [y, z] \cup [z, x]$. Par inégalité triangulaire, il est possible de construire un triangle de $\mathbb{R}^2$ de sommets $\bar{x}, \bar{y}, \bar{z} \in \mathbb{R}^2$ de sorte que $d(x, y) = \|\bar{x} - \bar{y}\|$, $d(y, z) = \|\bar{y} - \bar{z}\|$ and $d(z, x) = \|\bar{z} - \bar{x}\|$. Le triangle $\triangle\bar{x}\bar{y}\bar{z}$ est appelé triangle de comparaison pour $\triangle xyz$ et il est unique à isométrie près. Si $\gamma : [0, 1] \to E$ est la géodésique de x à y, nous remarquons que $d(\gamma_t, x) = \|(1-t)\bar{x} + t\bar{y} - x\|$ donc la combinaison convexe $(1-t)\bar{x} + t\bar{y}$ peut être interprétée comme un point de comparaison pour γ_t dans $\triangle\bar{x}\bar{y}\bar{z}$.

On dit que l'espace (E, d) est de courbure négative au sens d'Alexandrov si pour tout $x, y, z \in E$ et toute géodésique $\gamma : [0, 1] \to E$ de x à y nous avons l'inégalité

$$d(\gamma_t, z) \leq \|(1-t)\bar{x} + t\bar{y} - \bar{z}\|. \tag{2.12}$$

De manière alternative, on dit que (E, d) est CAT(0), pour Cartan–Alexandrov–Toponogov. Intuitivement, (1.12) signifie que chaque triangle géodésique est plus mince que son triangle de comparaison Euclidien, comme on peut le voir dans la figure Figure 2.1.

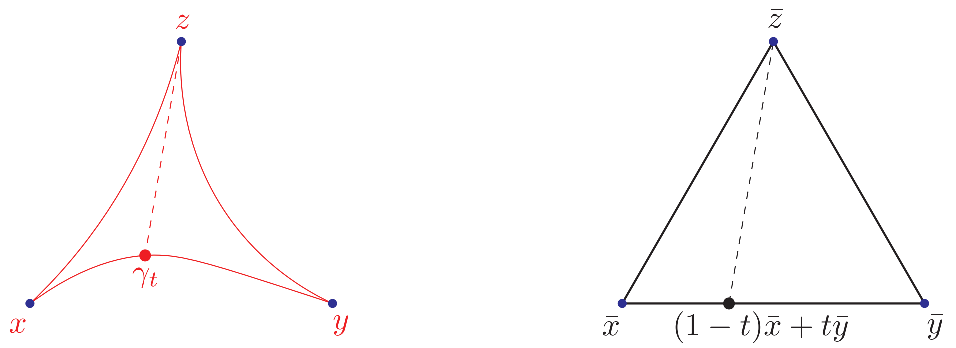

Figure 2.1: Triangle géodésique (gauche) et un triangle de comparaison correspondant (droite) pour un espace CAT(0).

Si (E, d) est complet et CAT(0), alors il est dit espace de Hadamard. Des exemples d'espaces de Hadamard incluent les espaces de Hilbert, les sous-ensembles convexes de ceux-ci, ainsi que les variétés Riemanniennes simplement connexes complètes de courbure sectionnelle négative. Il convient de noter que la sphère $\mathbb{S}^2$ n'est pas CAT(0). Dans les espaces de Hadamard, il est possible de développer une théorie de l'analyse convexe, de l'optimisation convexe et des probabilités qui généralise au cadre non-linéaire les résultats classiques connus dans les espaces de Hilbert.

Nous supposons dans la suite que (E, d) est Hadamard. Une partie $C \subset E$ est dite géodésiquement convexe si pour tout $x, y \in C$, le segment géodésique $[x, y]$ est une partie de C. Étant donné une telle partie, une fonction $f : C \to \mathbb{R}$ est dite géodésiquement convexe si pour tout $x, y \in C$ et tout géodésique γ de x à y définie sur $[0, 1]$, nous avons $f(\gamma_t) \leq (1-t)f(x) + tf(y)$. En développant (2.12) et après quelques calculs, on obtient l'inégalité équivalente:

$$d(\gamma_t, z)^2 \leq (1-t)d(x, z)^2 + td(y, z)^2 - t(1-t)d(x, y)^2, \tag{2.13}$$

donc pour chaque $z \in E$, la fonction $d(\cdot, z)^2$ est géodésiquement (fortement) convexe.

Sturm [243] a mis en évidence l'attrait des espaces de Hadamard et l'importance de l'équation (2.13) pour l'étude de la moyenne de Fréchet. Premièrement, dans ce cas, la moyenne de Fréchet existe et est unique : $M(\mu) = \theta_\star$. Deuxièmement, en raison de l'équation (2.13), le minimiseur $\theta_\star$ est bien séparé (tel que défini dans la section Section 2.2.1), et cela est quantifié par l'inégalité de variance [243, Proposition 4.4]

$$\phi(\theta) \geq \phi(\theta_\star) + d(\theta, \theta_\star)^2 \tag{2.14}$$

valide pour tout $\theta \in E$.

L'évaluation de la vitesse de convergence des moyennes de Fréchet empiriques a récemment attiré beaucoup d'attention, en particulier dans les espaces de Hadamard sans structure Riemannienne, où les théorèmes de la limite centrale ne sont pas disponibles. L'inégalité (2.14) est un ingrédient technique clé dans tous les travaux suivants, mentionnés par ordre chronologique. Ahidar-Coutrix et al. [5] examinent le cas où E est borné et sous une hypothèse forte d'entropie métrique ils montrent qu'il existe des constantes C_1, C_2 telles que pour tous $n \geq 1$, $t > 0$,

$$\mathbb{P}\Big(\sqrt{n}d(\widehat{\theta}_n, \theta_\star) \geq C_1 \max(C_2, \sqrt{t})\Big) \leq 2e^{-t}.$$

Sous une hypothèse plus faible d'entropie ils obtiennent des vitesses non-paramétriques. Schötz [230] considère le cas où E n'est pas forcément borné et il montre sous une hypothèse faible d'entropie que pour une constante C_3 et tous $n \geq 1$, $t > 0$,

$$\mathbb{P}\big(\sqrt{n}d(\widehat{\theta}_n, \theta_\star) \geq t\big) \leq \frac{C_3}{t^2}.$$

Sous une hypothèse forte d'entropie, il montre aussi que pour un certain $\beta > 0$,

$$\mathbb{E}[d(\widehat{\theta}_n, \theta_\star)^2] = O\Big(\frac{\log(n)^\beta}{n}\Big).$$

Le Gouic et al. [167] imposent que l'espace soit de courbure minorée par un certain $\kappa \leq 0$ (voir, e.g., [52] pour une définition) et en notant σ^2 la variance de Fréchet ils obtiennent que pour tout $n \geq 1$,

$$\mathbb{E}[d(\widehat{\theta}_n, \theta_\star)^2] \leq \frac{\sigma^2}{n}. \tag{2.15}$$

Yun and Park [276] montrent des résultats similaires à ceux de Schötz. Brunel et al. [49] définissent une notion de loi sous-Gaussienne dans les espaces métriques. En supposant que μ est sous-Gaussienne (e.g., si μ est à support borné) et que E a une courbure minorée (pour exploiter (1.15)), ils montrent qu'il existe C_4 tel que pour tous $n \geq 1$, $t > 0$,

$$\mathbb{P}\big(\sqrt{n}d(\widehat{\theta}_n, \theta_\star) \geq \sigma + t\big) \leq e^{-C_4 t^2}. \tag{2.16}$$

Escande [88] montre par un argument de stabilité que (2.15) est valide, à une constante universelle près, sans l'hypothèse de courbure minorée. Il obtient une borne similaire à (2.16) sous une hypothèse de queue sous-exponentielle.

Lorsque $E = \mathbb{R}^d$, il est connu que la moyenne empirique souffre d'un défaut majeur : elle est facilement influencée par les observations aberrantes. C'est pourquoi d'autres paramètres de tendance centrale, tels que la médiane, sont fondamentaux. Dans le cadre non linéaire, pour $p \in [1, \infty)$ et en supposant que $\int_E d(x, o)^{p-1} d\mu(x) < \infty$ pour un certain $o \in E$, le paramètre de localisation correspondant à la fonction de contraste

$$\varphi : (x, \theta) \mapsto d(x, \theta)^p - d(x, o)^p$$

est appelé p-moyenne de Fréchet. Lorsque $p = 2$, il s'agit de la moyenne classique de Fréchet, et lorsque $p = 1$, on parle de médiane de Fréchet. Pour $p \neq 2$, les résultats statistiques sont limités à la consistence [137, 231], et à la normalité asymptotique lorsque E est une variété Riemannienne [48]. Il est difficile d'obtenir un analogue de l'inégalité de variance (1.14), d'où le manque de résultats sur les vitesses de convergence.

2.3.3 Les arbres métriques

Une incarnation importante des espaces de Hadamard est l'arbre métrique. Considérons un arbre T au sens de la théorie des graphes, c'est-à-dire un graphe non orienté, connexe et acyclique avec des arêtes pondérées. Les poids sont interprétés comme les longueurs des arêtes, de sorte que l'arbre est équipé de la métrique du plus court chemin

d, donnant ainsi naissance à l'arbre métrique (T, d). Une construction rigoureuse et les propriétés topologiques de T sont énoncées dans [45, p.7]. Les arbres métriques sont importants dans la pratique car ils peuvent être utilisés pour modéliser des réseaux tels que les réseaux routiers, fluviaux, de communication ou de distribution.

Il existe peu de littérature statistique sur les p-moyennes de Fréchet dans le cadre spécifique des arbres métriques. Basrak [18] se concentre sur la moyenne de Fréchet dans un arbre métrique binaire et établit un théorème central limite pour la moyenne inductive. Risser et al. [96, 98] cherchent à calculer les moyennes de Fréchet dans les graphes métriques, tandis que Hotz et al. [126] développent des lois des grands nombres et des théorèmes central limite lorsque l'espace ambiant est un livre ouvert. Un cas particulier d'un livre ouvert est la m-araignée, qui peut être considérée comme un type particulier d'arbre métrique.

Un sujet connexe qui a suscité davantage d'attention est celui des espaces stratifiés [138], c'est-à-dire des espaces qui sont des unions finies de sous-espaces disjoints. Des exemples d'espaces stratifiés CAT(0) comprennent les livres ouverts [126] et l'espace des arbres de Billera–Holmes–Vogtmann [37]. Les éléments de cet espace sont des arbres et il a été introduit pour évaluer la proximité entre plusieurs arbres phylogénétiques.

2.3.4 Résumé de nos résultats

Le Chapitre 5 est tiré de Romon et Brunel [224], qui est actuellement en cours d'examen dans une revue.

L'espace ambiant E que nous considérons est un arbre métrique. Nous faisons des hypothèses raisonnables pour qu'il soit un espace compact et de Hadamard. Nous considérons des paramètres de localisation définis via le contraste générique $\varphi : (x, \theta) \mapsto \ell\big(d(x, \theta)\big)$, où $\ell : [0, \infty) \to [0, \infty)$ est une fonction convexe et croissante. Nous les appelons ℓ-moyennes de Fréchet.

Nous exploitons la convexité géodésique de la fonction objectif ϕ et la géométrie de l'arbre afin de définir une notion de dérivée directionnelle pour ϕ. Ceci nous permet de localiser et caractériser les ℓ-moyennes de Fréchet.

L'estimation est effectuée avec le M-estimateur standard. Nous étendons aux arbres métriques la notion de stickiness définie par Hotz et al. [126]: une ℓ-moyenne de Fréchet est collante ou bien partiellement collante. Nous montrons que la stickiness empirique est un phénomène non-asymptotique que nous quantifions avec des bornes exponentielles. Comme corollaire immédiat nous obtenons une loi forte des grands nombres collante.

Ensuite nous nous concentrons sur les médianes de Fréchet. Nous commençons par fournir des résultats plus précis sur leur localisation et leur unicité. Dans le cas partiellement collant, nous établissons des bornes non-asymptotiques et des théorèmes central limite.

Chapter 3

Chi-square and normal inference in high-dimensional multi-task regression

3.1 Introduction

3.1.1 Model

We consider a multi-task linear regression model with T tasks, with n i.i.d. observations $(\boldsymbol{x}_i, y_i^{(1)}, ..., y_i^{(T)})$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a random feature vector and $y_i^{(1)}, ..., y_i^{(T)}$ are T different scalar responses. We assume that on each task $t = 1, ..., T$, the response $y_i^{(t)}$ satisfies a linear model

$$y_i^{(t)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)} + \epsilon_i^{(t)}, \qquad t = 1, ..., T \tag{3.1}$$

where $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^p$ is the unknown coefficient vector on the task t. Throughout, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the design matrix with n rows $\boldsymbol{x}_1^\top, ..., \boldsymbol{x}_n^\top$. The linear models (3.1) may be rewritten in vector and matrix form

$$\boldsymbol{y}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\varepsilon}^{(t)}, \qquad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E} \tag{3.2}$$

where $\boldsymbol{y}^{(t)} = (y_1^{(t)}, ..., y_n^{(t)})^\top$ and $\boldsymbol{\varepsilon}^{(t)} = (\epsilon_1^{(t)}, ..., \epsilon_n^{(t)})^\top$ are vectors in $\mathbb{R}^n$, $\boldsymbol{Y} \in \mathbb{R}^{n \times T}$ is the response matrix with columns $\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(T)}$, $\boldsymbol{E} \in \mathbb{R}^{n \times T}$ is a noise matrix with columns $\boldsymbol{\varepsilon}^{(1)}, ..., \boldsymbol{\varepsilon}^{(T)}$, and $\boldsymbol{B}^* \in \mathbb{R}^{p \times T}$ is an unknown coefficient matrix with columns $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(T)}$.

Estimation of $\boldsymbol{B}^*$ in the above multi-task model has been well studied during the last decade in the high-dimensional regime where $p \ggg n$, see for instance [174]. This literature on multi-task learning suggests to use a joint convex optimization problem over the tasks in order to estimate $\boldsymbol{B}^*$, namely

$$\widehat{\boldsymbol{B}} = \operatorname*{arg\,min}_{\boldsymbol{B} \in \mathbb{R}^{p \times T}} \Big[\frac{1}{2nT}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + g(\boldsymbol{B})\Big] = \operatorname*{arg\,min}_{\boldsymbol{B} \in \mathbb{R}^{p \times T}} \Big[\frac{1}{2nT}\sum_{t=1}^{T}\sum_{i=1}^{n}(y_i^{(t)} - \boldsymbol{x}_i^\top \boldsymbol{B}\boldsymbol{e}_t)^2 + g(\boldsymbol{B})\Big]$$

where $\boldsymbol{e}_t \in \mathbb{R}^T$ is the t-th canonical basis vector, $\|\cdot\|_F$ is the Frobenius norm of matrices and $g : \mathbb{R}^{p \times T} \to \mathbb{R}$ is a convex penalty function. The role of the convex penalty g is to promote a shared structure on the coefficient vectors $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(T)}$. The most common shared structure is that of row-sparsity where one assumes that only a few features are relevant across all tasks: there is a support set $S \subset \{1, ..., p\}$ of small cardinality (relatively to n, p) such that for every task $t = 1, ..., T$, $\boldsymbol{\beta}_j^{(t)} = 0 \iff j \notin S$. Equivalently, $\boldsymbol{e}_j{}^\top \boldsymbol{B}^* = \boldsymbol{0}_{1\times T}$ if and only if $j \notin S$, i.e., only $|S|$ rows of $\boldsymbol{B}^*$ are nonzero. In this case, the sparsity pattern encoded by $S \subset \{1, ..., p\}$ is shared on all tasks, and previous literature on estimation in this setting uses a penalty proportional to the $\ell_{2,1}$ norm, $g(\boldsymbol{B}) = \lambda \sum_{j=1}^p \|\boldsymbol{B}^\top \boldsymbol{e}_j\|_2$, or alternatively its Elastic-Net version $g(\boldsymbol{B}) = \lambda \sum_{j=1}^p \|\boldsymbol{B}^\top \boldsymbol{e}_j\|_2 + \mu\|\boldsymbol{B}\|_F^2$ for non-negative tuning parameters $\lambda, \mu \geq 0$. If the row-sparsity assumption holds and such $\ell_{2,1}$ penalty is used, estimation of $\boldsymbol{B}^*$ by $\widehat{\boldsymbol{B}}$ is improved compared to estimating $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(T)}$ separately [174].

3.1.2 Noise and residuals: non-trivial correlations for non-separable penalties

Classical multivariate statistics studies the least-squares estimate $\widehat{\boldsymbol{B}}^{(ls)} = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{Y}$, which corresponds to $g(\cdot) = 0$ in the above minimization problem. Here, the estimation on two tasks is independent, as on the t-th task for $t = 1, ..., T$ we have $\widehat{\boldsymbol{B}}^{(ls)}\boldsymbol{e}_t = (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{y}^{(t)}$ for the t-th canonical basis vector $\boldsymbol{e}_t \in \mathbb{R}^T$: the estimator $\widehat{\boldsymbol{B}}^{(ls)}\boldsymbol{e}_t$ of the unknown regression vector $\boldsymbol{\beta}^{(t)}$ on the t-th task only depends on the t-th response $\boldsymbol{y}^{(t)}$, and is independent of the other responses $(\boldsymbol{y}^{(t')})_{t' \in \{1,...,T\} \setminus \{t\}}$. By independence, if the noise $\boldsymbol{E}$ has i.i.d. mean-zero entries, then

$$\mathbb{E}[\boldsymbol{\varepsilon}^{(t')} \boldsymbol{e}_t^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}^{(ls)})^\top] = \boldsymbol{0}_{n\times n} \qquad \forall t \neq t', \tag{3.3}$$

i.e., residual and noise on two different tasks are uncorrelated. A similar story holds for multi-task Ridge regression, which corresponds to $g(\boldsymbol{B}) = \mu\|\boldsymbol{B}\|_F^2$ in the above minimization problem. The optimization problem is separable in the sense that

$$\widehat{\boldsymbol{B}}^{(R)} = \operatorname*{arg\,min}_{\boldsymbol{B}\in\mathbb{R}^{p\times T}} \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2}{2nT} + \mu\|\boldsymbol{B}\|_F^2 \quad \text{and} \quad \widehat{\boldsymbol{B}}^{(R)}\boldsymbol{e}_t = \operatorname*{arg\,min}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{\|\boldsymbol{y}^{(t)} - \boldsymbol{X}\boldsymbol{b}\|_2^2}{2nT} + \mu\|\boldsymbol{b}\|_2^2$$

equivalently define $\widehat{\boldsymbol{B}}^{(R)}$. It follows again that $\widehat{\boldsymbol{B}}^{(R)}\boldsymbol{e}_t$ only depends on the t-th response $\boldsymbol{y}^{(t)}$, and if $\boldsymbol{E}$ has i.i.d. mean-zero entries then (3.3) holds also for $\widehat{\boldsymbol{B}}^{(R)}$ by independence.

The situation is more complex for non-separable penalty functions, for instance if the penalty is proportional to the $\ell_{2,1}$ norm, $g(\boldsymbol{B}) = \lambda \sum_{j=1}^p \|\boldsymbol{B}^\top \boldsymbol{e}_j\|_2$ where $\boldsymbol{e}_j \in \mathbb{R}^p$ is the j-th canonical basis vector. The corresponding estimator studied throughout the chapter is the multi-task Lasso

$$\widehat{\boldsymbol{B}} = \operatorname*{arg\,min}_{\boldsymbol{B}\in\mathbb{R}^{p\times T}} \Big(\frac{1}{2nT}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda\|\boldsymbol{B}\|_{2,1} \Big) \quad \text{where} \quad \|\boldsymbol{B}\|_{2,1} = \sum_{j=1}^p \|\boldsymbol{B}^\top \boldsymbol{e}_j\|_2. \tag{3.4}$$

The estimate $\widehat{\boldsymbol{B}}\boldsymbol{e}_t$ of the unknown vector $\boldsymbol{\beta}^{(t)}$ on the t-th task depends in an intricate way on all the responses including $(\boldsymbol{y}^{(t')})_{t' \in \{1,...,T\} \setminus \{t\}}$. Note that this dependence of

$\widehat{\boldsymbol{B}}\boldsymbol{e}_t$ on all responses is purposeful: we hope to leverage a shared pattern on all tasks (e.g., if $\boldsymbol{B}^*$ is row-sparse and a sparsity pattern is shared by all $\boldsymbol{\beta}^{(t)}, t = 1, ..., T$) in order to improve estimation compared to $\widehat{\boldsymbol{B}}^{(ls)}$ or $\widehat{\boldsymbol{B}}^{(R)}$. In this case, however, (3.3) does not hold and the correlation between the residual on task t and the noise on task t' is non-trivial. Our results below (specifically Lemma 3.24) reveal that for $t, t' \in [T]$,

$$\big((\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_t\big)^\top \boldsymbol{\varepsilon}^{(t')} \approx \begin{cases} \sigma^2(n - \widehat{\mathbf{A}}_{tt'}) & \text{if } t = t' \\ -\sigma^2 \widehat{\mathbf{A}}_{tt'} & \text{if } t \neq t' \end{cases}$$

when the noise $\boldsymbol{E}$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and $\widehat{\mathbf{A}}_{tt'}$ is the (t, t') entry of a symmetric matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{T \times T}$ defined in Section 3.2. This matrix plays a central role in the present chapter to derive asymptotic normality and asymptotic χ^2 results.

3.1.3 Confidence intervals for linear functionals of $\boldsymbol{\beta}^{(1)}$

A first goal of the present work is to provide confidence intervals for linear functionals of the regression vector on the first task. Throughout the chapter, regarding asymptotic normality and confidence intervals, $\boldsymbol{a} \in \mathbb{R}^p$ is a fixed direction of interest and we wish to construct confidence intervals for $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$. For instance, the direction $\boldsymbol{a} \in \mathbb{R}^p$ may be of the following form.

(i) a canonical basis vector $\boldsymbol{e}_j \in \mathbb{R}^p$. For $\boldsymbol{a} = \boldsymbol{e}_j$, the goal is to construct confidence intervals for $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)} = \beta_j^{(1)}$, the coefficient of the j-th feature on the first task. This is the classical goal in statistics where one wishes to provide inference on the effect of the j-th covariate.

(ii) a new feature vector $\boldsymbol{x}_{new} \in \mathbb{R}^p$, that may for instance correspond to the characteristics of a new subject whose responses $y_{new}^{(1)}, ..., y_{new}^{(T)}$ are not known yet. The goal is to provide a confidence interval for $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$ which corresponds to the expected response of Y_{new} conditionally on the feature vector $\boldsymbol{x}_{new}$.

We stress here that the first task ($t = 1$) has a special role: the unknown parameter $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$ only involves the first unknown coefficient vector $\boldsymbol{\beta}^{(1)}$ and not the other coefficient vectors $\boldsymbol{\beta}^{(t)}, t = 2, ..., T$. If a single linear model $\boldsymbol{y}^{(1)} = \boldsymbol{X}\boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon}^{(1)}$ is observed, the construction of confidence intervals for $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$ has been extensively studied. Most related to the present work, [280, 256, 139, 140] initially provided methodologies for de-biasing (or de-sparsifying) the Lasso for construction of confidence intervals in a canonical basis direction $\boldsymbol{a} = \boldsymbol{e}_j$ for sparsity $s \lesssim \sqrt{n}/\log p$, [141] extended the sparsity requirement to $s \lesssim n/(\log p)^2$, [284, 42, 54, 55, 285, 26] studied estimation and construction of confidence intervals in dense direction $\boldsymbol{a} \in \mathbb{R}^p$, and [27] extended the de-biasing methodologies to arbitrary convex penalties.

Of course, one could throw away the responses $\boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(T)}$ and use only the response $\boldsymbol{y}^{(1)}$ with the aforementioned methodologies, since our goal is to construct confidence intervals for $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$. However, throwing away the responses on tasks $2, ..., T$ should intuitively lead to information loss and is not desirable.

3.1.4 Asymptotic χ^2 results and confidence ellipsoids for rows of $\boldsymbol{B}^*$

The second goal of this chapter is to develop confidence ellipsoids for whole rows of the unknown matrix $\boldsymbol{B}^*$. The j-th row of $\boldsymbol{B}^*$ is the vector $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ in $\mathbb{R}^T$ where $\boldsymbol{e}_j \in \mathbb{R}^p$ is the j-th canonical vector. Given a confidence level $\alpha \in (0,1)$, a confidence ellipsoid for $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ is a subset $\hat{\mathcal{E}}_\alpha$ of $\mathbb{R}^T$ constructed from the data such that

$$\mathbb{P}\big((\boldsymbol{B}^*)^\top \boldsymbol{e}_j \in \hat{\mathcal{E}}_\alpha\big) \geq 1 - \alpha - o(1)$$

where $o(1)$ converges to 0 as $n \to +\infty$. Ideally, the confidence ellipsoid enjoys the exact nominal coverage probability $1-\alpha$ asymptotically in the sense that

$$\big|\mathbb{P}\big((\boldsymbol{B}^*)^\top \boldsymbol{e}_j \in \hat{\mathcal{E}}_\alpha\big) - (1-\alpha)\big| \to 0 \tag{3.5}$$

as $n \to +\infty$. Note that one could also consider confidence sets $\hat{\mathcal{E}}_\alpha$ that are not ellipsoids (e.g., hyperrectangles); we focus here on ellipsoids as they are the natural confidence sets stemming from χ^2-distributed pivotal quantities. As in classical multivariate statistics, an advantage of confidence ellipsoids is that they provide simultaneous confidence intervals for every direction $\boldsymbol{b} \in \mathbb{R}^T$, that is, $\mathbb{P}\big(\forall \boldsymbol{b} \in \mathbb{R}^T, \boldsymbol{e}_j^\top \boldsymbol{B}^* \boldsymbol{b} \in \{\boldsymbol{b}^\top \boldsymbol{u}, \boldsymbol{u} \in \hat{\mathcal{E}}_\alpha\}\big) \to 1 - \alpha$ when (3.5) holds and $\hat{\mathcal{E}}$ is closed and convex.

Such a confidence ellipsoid allows to perform hypothesis tests of

$$H_0 : (\boldsymbol{B}^*)^\top \boldsymbol{e}_j = \boldsymbol{0}_{T\times 1} \qquad \text{against} \qquad H_1 : \|(\boldsymbol{B}^*)^\top \boldsymbol{e}_j\|_2 \geq \rho, \tag{3.6}$$

where the null hypothesis corresponds to the signal $\boldsymbol{Y}$ being independent of the j-th feature $\boldsymbol{X}\boldsymbol{e}_j$, and $\rho > 0$ is a separation radius. If a single task is observed ($T = 1$), it is impossible to distinguish between the null $\beta_j = 0$ and the alternative $\beta_j \neq 0$ with constant type I and type II errors unless $|\beta_j| \geq c\sigma n^{-1/2}$ for some constant $c > 0$. This follows by noting that the total variation distance between $\boldsymbol{y}^{H_0} = \boldsymbol{X}\boldsymbol{\beta}^{H_0} + \boldsymbol{\varepsilon}$ and $\boldsymbol{y}^{H_1} = \boldsymbol{X}\boldsymbol{\beta}^{H_1} + \boldsymbol{\varepsilon}$ converges to 0 if $\boldsymbol{\beta}^{H_0}, \boldsymbol{\beta}^{H_1}$ are the same except on coordinate j where $|\beta_j^{H_0} - \beta_j^{H_1}| = a_n$ with $a_n = o(\sigma n^{-1/2})$, $\|\boldsymbol{X}\boldsymbol{e}_j\|^2/n \asymp 1$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$, for instance by Pinsker's inequality and a standard bound on the Kullback Leibler divergence of two multivariate normals. If several tasks are observed as in the setting of interest here, we will see that it is possible to perform the hypothesis test (3.6) in situations where all nonzero coefficients of $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ are of order $o(\sigma n^{-1/2})$, i.e., of indistinguishable order when a single task is observed.

If asymptotic normality results are available for each of the T individual coefficients of $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ (for instance such as those described in the previous subsection), a natural strategy to construct confidence ellipsoids is to sum the square of the T asymptotically normal random variables and hope that the resulting sum has approximately the χ^2 distribution with T degrees-of-freedom. However, throughout the chapter the number of tasks T is allowed to grow to infinity with n which results in some challenges regarding this strategy, as pointed out by [194]. For the sake of illustrating the resulting difficulty, assume that we have established the asymptotic normality of T pivotal random variables $U_1, ..., U_T$ by proving decompositions of the form $U_t = (\hat{\sigma}/\sigma) Z_t + B_t$ where $Z_t \sim \mathcal{N}(0,1)$ and the convergence in probability $\hat{\sigma}/\sigma \xrightarrow{\mathbb{P}} 1$ and $B_t \xrightarrow{\mathbb{P}} 0$ hold, so that Slutsky's theorem ensures that the pivotal quantities are asymptotically normal with $U_t \xrightarrow{d} \mathcal{N}(0,1)$. Denoting by $\chi^2_T = \sum_{t=1}^T Z_t^2$, summing the squares of the pivotal quantities and applying the triangle inequality for the Euclidean norm on $\mathbb{R}^T$ yields

$$\Big|\sqrt{\textstyle\sum_{t=1}^T U_t^2} - \sqrt{\chi^2_T}\Big| \leq |\hat\sigma/\sigma - 1|\sqrt{\chi^2_T} + \sqrt{\textstyle\sum_{t=1}^T B_t^2}. \tag{3.7}$$

While $\mathbb{E}[(\chi^2_T)^{1/2}]$ is of order $\sqrt{T}$, the variance and quantiles of $(\chi^2_T)^{1/2}$ are of constant order (specifically, $\mathbb{P}((\chi^2_T)^{1/2} - \sqrt{T} \leq z_\alpha/\sqrt{2}) \to 1-\alpha$ holds by (3.87) below, and $\mathrm{Var}[(\chi^2_T)^{1/2}] \to 1/2$ by [128]). This implies that a sufficient condition that ensures that $(\sum_{t=1}^T U_t^2)^{1/2}$ and $(\chi^2_T)^{1/2}$ asymptotically share the same quantiles is that $\sum_{t=1}^T B_t^2 \xrightarrow{\mathbb{P}} 0$ and $\sqrt{T}|\hat\sigma/\sigma - 1| \xrightarrow{\mathbb{P}} 0$. While $B_1 \xrightarrow{\mathbb{P}} 0$ and $\hat\sigma/\sigma \xrightarrow{\mathbb{P}} 1$ are sufficient to grant asymptotic normality for U_1 on the first task, the conditions $\sqrt{T}|\hat\sigma/\sigma - 1| \xrightarrow{\mathbb{P}} 0$ and $(\sum_{t=1}^T B_t^2)^{1/2} \xrightarrow{\mathbb{P}} 0$ are much more stringent as they involve the number of tasks T.

3.1.5 Asymptotics and assumptions

We will derive asymptotic normality and asymptotic χ^2_T results for a sequence of multi-task regression problems of increasing dimensions. For each n, we consider the multi-task linear model (3.2) and and the multi-task Lasso estimate $\widehat{\boldsymbol{B}}$ in (3.4) where $\boldsymbol{B}^*$, the number of tasks T, dimension p, tuning parameter λ and row-sparsity s are all functions of n. The dependence in n is implicit and will be omitted to avoid notational burden. We will assume that the sequence of regression problems satisfies the following.

Assumption (A1). *(i)* $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ *is a Gaussian design matrix with i.i.d.* $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ *rows;*

(ii) $\boldsymbol{B}^* \in \mathbb{R}^{p\times T}$ *is a row-sparse unknown matrix with at most* s *nonzero rows;*

(iii) $\boldsymbol{E}$ *is a Gaussian noise matrix with i.i.d.* $\mathcal{N}(0,\sigma^2)$ *entries;*

(iv) $\{s,n,T,p\}$ *are positive and satisfy* $\frac{s}{n}(T + \log\frac{p}{s}) \to 0$ *and* $n \leq p$*, this implies* $\frac{s}{p} \vee \frac{T}{n} \to 0$*;*

(v) The spectrum of $\boldsymbol{\Sigma}$ *is bounded:* $C_{\min} \leq \phi_{\min}(\boldsymbol{\Sigma}) \leq \phi_{\max}(\boldsymbol{\Sigma}) \leq C_{\max}$ *for some constants* $0 < C_{\min} \leq C_{\max}$ *which are independent of* n,p,s,T*;*

(vi) $\boldsymbol{\Sigma}$ *satisfies* $\max_{j=1,\dots,p} \boldsymbol{\Sigma}_{jj} \leq 1$*;*

(vii) For two constants $\eta_1, \eta_2 > 0$*, the tuning parameter* λ *in* (3.4) *is given by*

$$\lambda = (1+\eta_2)\lambda_0, \quad \textit{where} \quad \lambda_0 = \Big(\max_{j=1,\dots,p} \boldsymbol{\Sigma}_{jj}^{1/2}\Big)\frac{\sigma(1+\eta_1)}{\sqrt{nT}}\Big(1 + \sqrt{(2/T)\log(p/s)}\Big). \tag{3.8}$$

3.1.6 Related literature

For integers $\bar{n}, \bar{p} \geq 1$, the multi-task setting above bears resemblance with the single-response linear model of the form

$$\bar{\boldsymbol{y}} = \bar{\boldsymbol{X}}\bar{\boldsymbol{\beta}} + \bar{\boldsymbol{\varepsilon}} \tag{3.9}$$

where $\boldsymbol{y} \in \mathbb{R}^{\bar{n}}$, $\bar{\boldsymbol{\varepsilon}} \in \mathbb{R}^{\bar{n}}$, $\bar{\boldsymbol{X}} \in \mathbb{R}^{\bar{n}\times\bar{p}}$, and the features $\{1, ..., \bar{p}\}$ are partitioned into p groups with equal sizes. Indeed, with $\bar{p} = pT$, $\bar{n} = nT$ and by vectorizing the matrices in (3.1), our multi-task setting is in one-to-one correspondence with the single-response linear model (3.9) with $\bar{\boldsymbol{y}} = \mathbf{vec}(\boldsymbol{Y})$, $\bar{\boldsymbol{\varepsilon}} = \mathbf{vec}(\boldsymbol{E})$, $\bar{\boldsymbol{X}}$ block diagonal with T blocks each equal to $\boldsymbol{X}$, and the partition $(G_1, ..., G_p)$ of $\{1, ..., \bar{p}\}$ into p groups is given by $G_j = \{j + (t-1)p, t = 1, ..., T\}$. With this correspondence, the estimator $\widehat{\boldsymbol{B}}$ is the group Lasso $\hat{\bar{\boldsymbol{\beta}}} = \arg\min_{\bar{\boldsymbol{b}}\in\mathbb{R}^{\bar{p}}} \|\bar{\boldsymbol{y}} - \bar{\boldsymbol{X}}\bar{\boldsymbol{b}}\|^2/(2\bar{n}) + \lambda\|\bar{\boldsymbol{b}}\|_{2,1}$ where $\|\bar{\boldsymbol{b}}\|_{2,1} = \sum_{j=1}^{p} \|\bar{\boldsymbol{b}}_{G_j}\|_2$.

Inference for grouped variables in a single-response linear model (3.9) focuses on estimation, hypothesis tests or confidence sets for the vector $\bar{\boldsymbol{\beta}}_{G_j}$ for a group $G_j \subset \{1, ..., \bar{p}\}$ of interest. In the single task setting (3.9) with grouped variables, [194] extends the de-biasing methodology in [280, 256] to inference for grouped variables and provides χ^2 asymptotic distribution results. The paper [194] already describes some challenges of chi-square inference in high-dimension (cf. the discussion after (3.7)); the multi-task problem of the present work shares some of these challenges, however our approach and proofs have no overlap with that of [194]. The papers [242, 255] give a different extension of the de-biasing methodology of [280, 256] to the group setting, again based on the group Lasso, but here by estimation of the inverse covariance matrix restricted to the group of interest with a multi-task estimator penalized by the nuclear norm. False Discovery Rate control in single-task linear models with grouped variables has been studied in [50] with a group SLOPE estimator. Under weak assumptions (in particular, no assumption on $\boldsymbol{X}$), [187] provides an approach to inference for grouped variables, although the resulting confidence regions are conservative. The papers [179, 180] study group inference in a sequence rejection fashion when the groups are hierarchically ordered. Bootstrap methods based on the group Lasso are studied in [283], without trying to remove the bias. The paper [107] develops conservative inference methods for quantities of the form $(\bar{\boldsymbol{\beta}}_{G_j})^\top \boldsymbol{A}\bar{\boldsymbol{\beta}}_{G_j}$ for a group $G_j \subset [p]$ of interest and a given positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{|G_j|\times|G_j|}$, based on the quadratic program de-biasing methodology given in [280, 139]. Finally, [27] introduces a degrees-of-freedom adjustment for the group Lasso to perform inference on a single coordinate or linear form of the unknown regression vector in (3.9).

Some papers focus on estimation and inference in the multi-task model (3.2). The papers [255, 30] study multi-task models of the form (3.2) where the noise $\boldsymbol{E} \in \mathbb{R}^{n\times T}$ has i.i.d. rows, and the entries within each row are correlated. A multi-task extension of the square-root Lasso is developed to concurrently estimate $\boldsymbol{B}^*$ and the correlations in the noise $\boldsymbol{E}$. Such results on estimating the correlations of the entries in $\boldsymbol{E}$ are useful to de-bias the group Lasso in the single-task model [255]. Support recovery through bounds on the group norm $\|\boldsymbol{B}\|_{2,\infty} = \sup_{j\in[p]} \|\boldsymbol{E}^\top \boldsymbol{e}_j\|_2$ is studied in [183] under a mutual incoherence assumption on $\boldsymbol{X}$. The mutual incoherence assumption requires a row-sparsity level $s \lesssim \sqrt{n}$ if $\boldsymbol{X}$ has i.i.d. entries. Closest to the setup and goals of the present work, [67] extends the de-biasing methodology of [280, 256] to the multi-task setting, using the nodewise Lasso to estimate a column of the precision matrix of the design. This approach requires row-sparsity of $\boldsymbol{B}^*$ of order $s \lesssim \sqrt{n}$ up to logarithmic factors. Although our approach also involves the nodewise Lasso to estimate columns of the precision matrix, the de-biasing methodology significantly differs from [67] and cannot be seen as a straightforward extension of [280, 256]: our approach requires the introduction of a data-driven symmetric matrix $\widehat{\mathbf{A}}$ of size $T \times T$ which captures the interactions between the residuals on different tasks. Introduction of this novel object lets us significantly relax the requirement on the row-sparsity of $\boldsymbol{B}^*$ while obtaining normal and χ^2_T inference results, that are proved to be non-conservative under some assumption on T, s, n, p.

3.1.7 Adjustments in high-dimensional inference

In single-task models, recent literature on high-dimensional inference has highlighted the necessity to adjust classical inference principles with scalar adjustments. To describe such adjustments consider a single-task linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\beta} \in \mathbb{R}^p$, Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n\times n})$ and $\boldsymbol{X}$ with i.i.d. $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ rows, where an initial estimator $\widehat{\boldsymbol{\beta}}^{(init)}$ is available. If one is interested in confidence intervals for the projection $\boldsymbol{a}^\top \boldsymbol{\beta}$ in some direction $\boldsymbol{a}$ normalized with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2 = 1$, a 1-step MLE correction in direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ [278], i.e., maximizing the likelihood over the one-dimensional model $\{\widehat{\boldsymbol{\beta}}^{(init)} + u\boldsymbol{\Sigma}^{-1}\boldsymbol{a}, u \in \mathbb{R}\}$ yields the corrected estimate

$$\boldsymbol{a}^\top \boldsymbol{\beta}^{(init)} + \boldsymbol{z}_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}) \|\boldsymbol{z}_0\|_2^{-2} \tag{3.10}$$

where $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ when $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2 = 1$; and the direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ is the one that maximizes the Fisher information [278]. (Since $\|\boldsymbol{z}_0\|_2^2 \sim \chi^2_n$ concentrates around n, we allow ourselves to replace $\|\boldsymbol{z}_0\|_2^2$ by n in (3.10) in this informal discussion). In high dimensions, this general principle requires a modification that accounts for the degrees-of-freedom of $\widehat{\boldsymbol{\beta}}^{(init)}$: [140, 26] for the Lasso and [27] for general penalty suggest to amplify the correction with the degrees-of-freedom adjustment $(1 - \hat{\mathsf{df}}/n)^{-1}$ and to use the estimate

$$\boldsymbol{a}^\top \boldsymbol{\beta}^{(init)} + (1 - \hat{\mathsf{df}}/n)^{-1} \boldsymbol{z}_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}) n^{-1} \tag{3.11}$$

instead of (3.10). If $\widehat{\boldsymbol{\beta}}^{(init)}$ is the Lasso, the adjustment $(1 - \hat{\mathsf{df}}/n)^{-1}$ is required for efficiency for large sparsity levels [26]. For the Lasso, the data-driven adjustment $(1 - \hat{\mathsf{df}}/n)^{-1}$ may be replaced by a deterministic scalar adjustment, i.e.,

$$\boldsymbol{a}^\top \boldsymbol{\beta}^{(init)} + (1 - \delta^{-1} s_*)^{-1} \boldsymbol{z}_0^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(init)}) n^{-1} \tag{3.12}$$

where $\delta = n/p$ and s_* is the scalar parameter obtained after solving the system of two equations with two unknowns in [193, Proposition 3.1]. The correspondence between $\hat{\mathsf{df}}/n$ and s_* can be seen in [193, Theorem F.1] or [62, Section 3.3]. This system of two nonlinear equations first appeared in [19] for the Lasso and can be extended to permutation invariant penalty functions (see [61] and the references therein) and robust M-estimators [250].

We are not aware of previous proposals to study such high-dimensional adjustments in the multi-task setting, e.g., by extending the data-driven adjustment in (3.11) or the deterministic one in (3.12). One goal of our work is to fill this gap.

3.1.8 Contributions

To summarize Sections 3.1.3 and 3.1.4, the inferential goals of the chapter are twofold:

(i) To construct valid confidence intervals for a linear functional $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$ of the unknown coefficient on the first task, by leveraging responses on all tasks simultaneously.

(ii) To construct valid confidence ellipsoids for rows $\boldsymbol{e}_j^\top \boldsymbol{B}^* \in \mathbb{R}^{1\times T}$ of the unknown coefficient matrix $\boldsymbol{B}^*$, for instance to provide hypothesis tests on the nullity of the j-th row of $\boldsymbol{B}^*$, or equivalently testing that the signal does not depend on the j-th covariate.

In order to achieve these statistical goals, we introduce a new object, the data-driven symmetric matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{T\times T}$. Introduction of the matrix $\widehat{\mathbf{A}}$ is key to equip the estimator $\widehat{\boldsymbol{B}}$ with the inference capabilities (i) and (ii) above, as the theory and simulations of the next sections will show. This data-driven matrix $\widehat{\mathbf{A}}$ generalizes, to the multi-task setting, the effective degrees-of-freedom and other scalar adjustments in single-task linear models discussed in the previous subsection. Since $\widehat{\mathbf{A}}$ is symmetric, $T(T+1)/2$ scalar adjustments are necessary in the multi-task setting and that number of adjustments is unbounded if $T \to +\infty$ as a function of n. The fact that a growing, unbounded number of scalar adjustments would be necessary to achieve the above inference capabilities in the multi-task setting was surprising—at least to us—, since existing works on adjustments in high-dimensional statistics so far only require a bounded number of scalar adjustments.

Our work also includes contributions related to the performance of the multi-task estimator $\widehat{\boldsymbol{B}}$ in (3.4). We improve the logarithmic dependence in tuning parameter λ and the known upper bounds on $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F$ and $\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F$ compared to [174]. We also develop tools to show that the random matrix $\boldsymbol{X}$ enjoys a multi-task Restricted Eigenvalue (RE) condition from [35]. Although the single-task case follows in a straightforward manner from Gordon's escape through a mesh theorem (e.g., [220]), the multi-task version of the RE condition for the random matrix $\boldsymbol{X}$ requires different tools.

3.1.9 Organization

The rest of the chapter is organized as follows. The next section summarizes notation. Section 3.2 describes a new quantity, the interaction matrix $\widehat{\mathbf{A}}$ that plays a major role in our estimates and confidence intervals. Section 3.3.1 constructs confidence intervals for $\boldsymbol{a}^\top \boldsymbol{\beta}^{(1)}$ when the covariance matrix $\boldsymbol{\Sigma}$ of the design is known. Section 3.3.2

extends these results and methodologies when $\boldsymbol{\Sigma}$ is unknown. Section 3.4 develops confidence ellipsoids for rows of $\boldsymbol{B}^*$. Section 3.5 provides an efficient way of computing the interaction matrix. Section 3.6 presents numerical experiments that corroborate our theoretical findings. The proofs are deferred to appendices and some intuition behind the main technical argument is given in Section 3.7.

3.1.10 Notation

Throughout the chapter, the linear model vector and matrix notation (3.2) holds. T, p and s are all non-decreasing functions of n. In all the displays of convergence (e.g., $\to$, lim, $o(\cdot)$, $O(\cdot)$), we implicitly mean that n goes to ∞. Convergence in distribution and in probability are denoted by $\xrightarrow{d}$ and $\xrightarrow{\mathbb{P}}$.

Estimators of the unknown $\boldsymbol{B}^*$ are denoted by $\widehat{\boldsymbol{B}}$. For any real a, $a_+ = \max(0, a)$ and $[k] = \{1, ..., k\}$ for any integer k, e.g., $[n], [p], [T]$. We use indices $i, i', i_1, i_2, ...$ to sum or loop over $[n]$ (i.e., over the n observations), indices $t, t', t_1, t_2, ...$ to sum or loop over $[T]$ (i.e., over the T tasks), indices $j, j', j_1, j_2, ...$ to sum or loop over $[p]$ (i.e., the p covariates). The vectors $\boldsymbol{e}_j \in \mathbb{R}^p$, $\boldsymbol{e}_t \in \mathbb{R}^T$, $\boldsymbol{e}_i \in \mathbb{R}^n$ denote the canonical basis vector of the corresponding index; the size of such canonical vector will be made explicit if it is not clear from context. The identity matrices of sizes $p \times p$, $n \times n$, $T \times T$ are $\boldsymbol{I}_{p\times p}$, $\boldsymbol{I}_{n\times n}$ and $\boldsymbol{I}_{T\times T}$ respectively and $\mathbf{0}_{k\times q}$ is the zero matrix with k rows and q columns.

For any $q \geq 1$, $\|\cdot\|_q$ is the ℓ_q-norm of vector, e.g., $\|\cdot\|_2$ is the Euclidean norm. For any matrix $\boldsymbol{M}$, $\|\boldsymbol{M}\|_F$ is the Frobenius norm and $\|\boldsymbol{M}\|_{op} = \sup_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{M}\boldsymbol{u}\|_2$ the operator norm, also known as the spectral norm. If $\boldsymbol{M}$ is symmetric, $\phi_{\min}(\boldsymbol{M})$ (resp. $\phi_{\max}(\boldsymbol{M})$) denotes the smallest (resp. largest) eigenvalue of $\boldsymbol{M}$. The Moore-Penrose pseudoinverse of matrix $\boldsymbol{M}$ is denoted by $\boldsymbol{M}^\dagger$. The Kronecker product between two matrices $\boldsymbol{U}$, $\boldsymbol{V}$ with $\boldsymbol{U} \in \mathbb{R}^{k\times q}$ is

$$\boldsymbol{U} \otimes \boldsymbol{V} := \begin{pmatrix} u_{11}\boldsymbol{V} & \dots & u_{1q}\boldsymbol{V} \\ \vdots & \vdots & \vdots \\ u_{k1}\boldsymbol{V} & \dots & u_{kq}\boldsymbol{V} \end{pmatrix} \text{ so that } \boldsymbol{I}_{T\times T} \otimes \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X} & \mathbf{0}_{n\times p} & \dots & \mathbf{0}_{n\times p} \\ \mathbf{0}_{n\times p} & \boldsymbol{X} & \dots & \mathbf{0}_{n\times p} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_{n\times p} & \dots & \dots \mathbf{0}_{n\times p} & \boldsymbol{X} \end{pmatrix}$$

for $\boldsymbol{X} \in \mathbb{R}^{n\times p}$. We will use the mixed product property of Kronecker products,

$$(\boldsymbol{U} \otimes \boldsymbol{V})(\boldsymbol{P} \otimes \boldsymbol{Q}) = (\boldsymbol{U}\boldsymbol{P}) \otimes (\boldsymbol{V}\boldsymbol{Q}), \qquad (\boldsymbol{U} \otimes \boldsymbol{V})^\dagger = \boldsymbol{U}^\dagger \otimes \boldsymbol{V}^\dagger \tag{3.13}$$

whenever the dimensions are such that the matrix products $\boldsymbol{U}\boldsymbol{P}$ and $\boldsymbol{V}\boldsymbol{Q}$ make sense. The following trace property also holds

$$\operatorname{Tr}[\boldsymbol{U} \otimes \boldsymbol{V}] = \operatorname{Tr}[\boldsymbol{U}] \operatorname{Tr}[\boldsymbol{V}]. \tag{3.14}$$

If $\|\cdot\|$ denotes a Schatten norm (e.g., Frobenius or spectral norm), then for any $\boldsymbol{U}$, $\boldsymbol{V}$ we have

$$\|\boldsymbol{U} \otimes \boldsymbol{V}\| = \|\boldsymbol{U}\| \|\boldsymbol{V}\|. \tag{3.15}$$

We define the vectorization $\mathbf{vec}(\boldsymbol{U})$ of any matrix $\boldsymbol{U} \in \mathbb{R}^{m\times q}$ by stacking vertically the columns of $\boldsymbol{U}$ into a column vector in $\mathbb{R}^{qm\times 1}$, i.e.,

$$\mathbf{vec}(\boldsymbol{A})^\top = \begin{pmatrix} u_{11} & u_{21} & \dots & u_{m1} & u_{12} & u_{22} & \dots & u_{m2} & \dots & u_{1q} & u_{2q} & \dots & u_{mq} \end{pmatrix}.$$

For any three matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ such that the matrix product $\boldsymbol{ABC}$ makes sense, the above vectorization operator satisfies

$$\mathbf{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A}) \, \mathbf{vec}(\boldsymbol{B}). \tag{3.16}$$

These many properties of Kronecker products are referenced in Section 4.2 of [122].

We consider restrictions of vectors (respectively matrices) by zeroing the corresponding entries (respectively columns). More precisely, if $\boldsymbol{v} \in \mathbb{R}^p$ and $B \subset [p]$ then $\boldsymbol{v}_B \in \mathbb{R}^p$ is the vector with $(\boldsymbol{v}_B)_j = 0$ if $j \notin B$ and $(\boldsymbol{v}_B)_j = v_j$ if $j \in B$. If $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $B \subset [p]$, $\boldsymbol{X}_B \in \mathbb{R}^{n \times p}$ is a matrix of the same dimension as $\boldsymbol{X}$ such that $(\boldsymbol{X}_B)\boldsymbol{e}_j = \mathbf{0}$ if $j \notin B$ and $(\boldsymbol{X}_B)\boldsymbol{e}_j = \boldsymbol{X}\boldsymbol{e}_j$ if $j \in B$, i.e., $\boldsymbol{X}_B$ is a copy of $\boldsymbol{X}$ after having zeroed the columns not indexed in B. Finally, $I\{\Omega\}$ denotes the indicator function of an event Ω, and $I\{i \in B\} = 1$ if $i \in B$ and $I\{i \in B\} = 0$ if $i \notin B$ is the indicator that some index i belongs to B.

3.2 The interaction matrix $\widehat{\mathbf{A}}$ of the Multi-Task Lasso estimator

We consider the multi-task Lasso estimator, with $\ell_{2,1}$ penalty, given (3.4) for some tuning parameter $\lambda > 0$. Let $\hat{S} = \{j \in [p] : \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \neq 0\}$ denote the set of nonzero rows of $\widehat{\boldsymbol{B}}$. We will refer to $\hat{S}$ as the support of $\widehat{\boldsymbol{B}}$ and denote by $|\hat{S}|$ its cardinality. The above estimator is the one commonly used in the multi-task learning literature under a row-sparsity assumption on $\boldsymbol{B}^*$, see, e.g., [174]. Recall that $\boldsymbol{X}_{\hat{S}} \in \mathbb{R}^{n \times p}$ is a copy of $\boldsymbol{X}$ obtained after zeroing the columns not belonging to $\hat{S}$. Define $\tilde{\boldsymbol{X}} := \boldsymbol{I}_{T \times T} \otimes \boldsymbol{X}_{\hat{S}}$ where $\otimes$ denotes the Kronecker product defined in Section 3.1.10, so that $\tilde{\boldsymbol{X}} \in \mathbb{R}^{nT \times pT}$ is block-diagonal with T blocks, each equal to $\boldsymbol{X}_{\hat{S}}$. Consequently $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \boldsymbol{I}_{T \times T} \otimes (\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}}) \in \mathbb{R}^{(pT) \times (pT)}$ is also block-diagonal with T blocks equal to $\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}}$. For any $j \in \hat{S}$, define the matrix

$$\boldsymbol{H}^{(j)} := \lambda \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-1} \left(\boldsymbol{I}_{T \times T} - \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \boldsymbol{e}_j^\top \widehat{\boldsymbol{B}} \, \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-2} \right) \quad \in \mathbb{R}^{T \times T} \tag{3.17}$$

and note that $\boldsymbol{H}^{(j)}$ is proportional to an orthogonal projection of rank $T-1$. The matrix $\boldsymbol{H}^{(j)}$ is the Hessian of $\boldsymbol{u} \mapsto \lambda \|\boldsymbol{u}\|_2$ at $\boldsymbol{u} = \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j$. Finally, let $\tilde{\boldsymbol{H}} \in \mathbb{R}^{(pT) \times (pT)}$ be the matrix defined by $\tilde{\boldsymbol{H}} := \sum_{j \in \hat{S}} \boldsymbol{H}^{(j)} \otimes (\boldsymbol{e}_j \boldsymbol{e}_j^\top)$.

Definition 3.1. The interaction matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{T \times T}$ of the estimator $\widehat{\boldsymbol{B}}$ in (3.4) is defined entrywise by

$$\widehat{\mathbf{A}}_{tt'} := \operatorname{Tr} \left(\left[\, \mathbf{0}_{n \times p(t-1)} \, \middle| \, \boldsymbol{X}_{\hat{S}} \, \middle| \, \mathbf{0}_{n \times p(T-t)} \, \right] \left[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}} \right]^\dagger \begin{bmatrix} \mathbf{0}_{p(t'-1) \times n} \\ \hline (\boldsymbol{X}_{\hat{S}})^\top \\ \hline \mathbf{0}_{p(T-t') \times n} \end{bmatrix} \right) \tag{3.18}$$

for all $t, t' \in [T]$, where $\dagger$ denotes the Moore-Penrose inverse. Equivalently, if $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^T$ then

$$\boldsymbol{u}^\top \widehat{\mathbf{A}} \boldsymbol{v} = \operatorname{Tr} \left(\left[\, u_1 \boldsymbol{X}_{\hat{S}} \, \middle| \, u_2 \boldsymbol{X}_{\hat{S}} \, \middle| \, \dots \, \middle| \, u_T \boldsymbol{X}_{\hat{S}} \, \right] \left[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}} \right]^\dagger \left[\, v_1 \boldsymbol{X}_{\hat{S}} \, \middle| \, v_2 \boldsymbol{X}_{\hat{S}} \, \middle| \, \dots \, \middle| \, v_T \boldsymbol{X}_{\hat{S}} \, \right]^\top \right),$$

or with Kronecker product notation,

$$\boldsymbol{u}^\top \widehat{\mathbf{A}} \boldsymbol{v} = \operatorname{Tr}\left[(\boldsymbol{u}^\top \otimes \boldsymbol{X}_{\hat{S}})[\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger(\boldsymbol{v} \otimes (\boldsymbol{X}_{\hat{S}})^\top)\right]. \tag{3.19}$$

Observe that $\sum_{j\in\hat{S}}(\boldsymbol{e}_j\boldsymbol{e}_j^\top) \otimes \boldsymbol{H}^{(j)}$ is a block-diagonal matrix with p diagonal blocks equal to $I\{j \in \hat{S}\}\boldsymbol{H}^{(j)}$. For $\boldsymbol{A}, \boldsymbol{B}$ any square matrices, $\boldsymbol{A} \otimes \boldsymbol{B} = \boldsymbol{P}(\boldsymbol{B} \otimes \boldsymbol{A})\boldsymbol{P}^\top$ holds for a permutation matrix $\boldsymbol{P}$ that only depends on the dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$. This permutation $\boldsymbol{P}$ is particularly simple and known as a perfect shuffle. It follows that $\boldsymbol{P}\tilde{\boldsymbol{H}}\boldsymbol{P}^\top$ is block diagonal with p diagonal blocks for some permutation matrix $\boldsymbol{P} \in \mathbb{R}^{pT\times pT}$. Thus the matrix

$$\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}} \qquad \in \mathbb{R}^{pT\times pT} \tag{3.20}$$

appearing in (3.18)-(3.19) is the sum of two matrices of size $pT \times pT$, each summand being block diagonal but in a different basis. If $\lambda = 0$ then $\tilde{\boldsymbol{H}} = \mathbf{0}$ and $\widehat{\mathbf{A}}$ is diagonal as $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}$ can be inverted by block. This corresponds to the unregularized least-squares estimate $\widehat{\boldsymbol{B}}^{(ls)}$ discussed in (3.1.2) with $\widehat{\boldsymbol{B}}^{(ls)}\boldsymbol{e}_t$ depending on the t-th response $\boldsymbol{y}^{(t)}$ only. In the case $\lambda > 0$ of interest here, the matrix $\tilde{\boldsymbol{H}}$ induces nonzero entries outside of the T diagonal blocks of $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$, the matrix (3.20) is not diagonal by block and the resulting matrix $\widehat{\mathbf{A}}$ is not diagonal. Additional structure in (3.20) and $\widehat{\mathbf{A}}$ is studied in Section 3.5, which yields an efficient and practical algorithm to compute $\widehat{\mathbf{A}}$.

The interaction matrix plays a major role in the construction of our confidence intervals for $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$ as well as for chi-square inference regions for rows of $\boldsymbol{B}^*$. A high-level interpretation of its role is that $\widehat{\mathbf{A}}$ captures the correlation between the residuals on different tasks. The following proposition summarizes some useful properties of $\widehat{\mathbf{A}}$. Result (iii) is important as our confidence interval for $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$ defined in the next section will involve the inverse of $\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n$. Proposition 3.2 is proved in Section 3.8 of the supplement.

Proposition 3.2. *Let* $\widehat{\mathbf{A}}$ *be defined by* (3.18). *Then*

(i) $\widehat{\mathbf{A}}$ *is symmetric and positive semi-definite.*

(ii) If $\boldsymbol{X}_{\hat{S}}$ *is rank* $|\hat{S}|$ *then the spectral norm of* $\widehat{\mathbf{A}}$ *is bounded from above as* $\|\widehat{\mathbf{A}}\|_{op} \le |\hat{S}|$.

(iii) If $\boldsymbol{X}_{\hat{S}}$ *is rank* $|\hat{S}|$ *and* $|\hat{S}|/n < 1$ *then* $\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n$ *is positive-definite and* $\|\boldsymbol{I}_{T\times T} - (\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op} \le (|\hat{S}|/n)/(1 - |\hat{S}|/n)$.

3.3 Asymptotic normality and confidence intervals in the multi-task setting

3.3.1 Known $\boldsymbol{\Sigma}$: Pivotal random variable, asymptotic normality and confidence intervals

We assume throughout this section that the direction $\boldsymbol{a}$ of interest is normalized with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2 = 1$. This normalization assumption is relaxed in the next Section 3.3.2 where we develop a methodology for unknown $\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}$ is known, our main result is the following where $\widehat{\mathbf{A}}$ denotes the interaction matrix (3.18).

Theorem 3.3. *Let Assumption (A1) be fulfilled. Assume that* $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2^2 = 1$. *If* $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ *then*

$$\frac{n\boldsymbol{a}^T(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b} + \boldsymbol{z}_0^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}}{\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2} \xrightarrow{d} \mathcal{N}(0,1) \tag{3.21}$$

for any $\boldsymbol{b} \in \mathbb{R}^T$. *Hence for* $\boldsymbol{b} = \boldsymbol{e}_1 \in \mathbb{R}^T$, *the parameter* $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$ *of interest satisfies*

$$\frac{n(\boldsymbol{a}^T\widehat{\boldsymbol{B}}\boldsymbol{e}_1 - \boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}) + \boldsymbol{z}_0^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{e}_1}{\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{e}_1\|_2} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.22}$$

Theorem 3.3 is proved in Section 3.11. The left-hand sides of both displays in Theorem 3.3 can be interpreted as Z-scores that have asymptotically standard normal distribution. In the second display, the only unknown quantity on the left hand side is $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$, the parameter of interest (while in the first display, the only unknown quantity is the scalar $\boldsymbol{a}^\top\boldsymbol{B}^*\boldsymbol{b}$). Consequently if $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution such that $\mathbb{P}(|\mathcal{N}(0,1)| \le z_{\alpha/2}) = 1-\alpha$, an asymptotic $1-\alpha$ confidence interval for $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$ is given by $[L_-^\alpha, L_+^\alpha]$ where

$$L_\pm^\alpha = \underbrace{\boldsymbol{a}^T\widehat{\boldsymbol{B}}\boldsymbol{e}_1}_{\substack{\text{initial}\\\text{estimate}}} + \underbrace{\frac{\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \frac{\widehat{\mathbf{A}}}{n})^{-1}\boldsymbol{e}_1}{n}}_{\text{bias correction using the interaction matrix}} \pm \underbrace{\frac{z_{\alpha/2}\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \frac{\widehat{\mathbf{A}}}{n})^{-1}\boldsymbol{e}_1\|_2}{n}}_{\text{confidence interval half-length}}.$$

(3.22) in Theorem 3.3 states that $\mathbb{P}(\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)} \in [L_-^\alpha, L_+^\alpha]) \to (1-\alpha)$ as $n, p \to +\infty$.

The confidence interval is centered at $\boldsymbol{a}^T\widehat{\boldsymbol{B}}\boldsymbol{e}_1$ (which can be interpreted as the initial estimate of $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$ given by the estimator $\widehat{\boldsymbol{B}}$ in (3.4)) plus a de-biasing correction $\boldsymbol{z}_0^T(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}}/n)^{-1}$ that involves the interaction matrix $\widehat{\mathbf{A}}$ through the matrix inverse

$$(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}. \tag{3.23}$$

The fact that penalized estimators such as (3.4) require a de-biasing correction should be expected since it is already the case for $T = 1$ for the Lasso [280, 256, 139, 140, 141, 26] and any regularized least-squares [27]. However, the apparition in the de-biasing correction of the interaction matrix through the matrix inverse (3.23) is surprising at least to us: we did not expect the multi-task de-biasing correction to require a matrix inversion such as (3.23) when initially tackling this problem. The length of the confidence interval above is $2z_{\alpha/2}n^{-1}\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{e}_1\|_2$ when $\boldsymbol{b} = \boldsymbol{e}_1$, and an estimate of this norm is given by the following theorem.

Theorem 3.4. *Let the assumptions and setting of Theorem 3.3 be fulfilled. Then* $\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2^2/n \xrightarrow{\mathbb{P}} \sigma^2$ *when* $\|\boldsymbol{b}\|_2 = 1$.

Consequently the length of the confidence interval is approximately $2z_{\alpha/2}\sigma n^{-1/2}$ which is the typical length for two-sided confidence intervals for an unknown mean μ when observing i.i.d. $Y_1, ..., Y_n$ with $\mathbb{E}[Y_i] = \mu, \mathrm{Var}[Y_i] = \sigma^2$. Theorems 3.3 and 3.4 are proved together in Section 3.11.

Comparison with single-task Lasso on the first task. It is instructive to compare the above confidence interval with the confidence interval induced by a single-task Lasso estimator computed on $(\boldsymbol{X}, \boldsymbol{y}^{(1)})$, i.e., when throwing away the responses $\boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(T)}$ on tasks $2, ..., T$. This is also a good opportunity to analyse the form of $\widehat{\mathbf{A}}$ and the matrix inversion (3.23) in the degenerate case where a single task is observed.

For $T = 1$, a response vector $\boldsymbol{y}^{(1)} = \boldsymbol{X}\boldsymbol{\beta}^{(1)} + \boldsymbol{\varepsilon}^{(1)}$ in $\mathbb{R}^n$ is observed and the estimator (3.4) reduces to the usual Lasso with response vector $\boldsymbol{y}^{(1)}$,

$$\widehat{\boldsymbol{\beta}}^L = \mathop{\arg\min}_{\boldsymbol{b}\in\mathbb{R}^p} \|\boldsymbol{y}^{(1)} - \boldsymbol{X}\boldsymbol{b}\|^2/(2n) + \lambda\|\boldsymbol{b}\|_1.$$

The asymptotic normality result in Theorem 3.3 for $\boldsymbol{b} = 1$ asserts that

$$\frac{n\boldsymbol{a}^\top(\widehat{\boldsymbol{\beta}}^L - \boldsymbol{\beta}^{(1)}) + (1 - \widehat{\mathbf{A}}_{11}/n)^{-1}\boldsymbol{z}_0^\top(\boldsymbol{y}^{(1)} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^L)}{(1 - \widehat{\mathbf{A}}_{11}/n)^{-1}\|\boldsymbol{y}^{(1)} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^L\|_2} \xrightarrow{d} \mathcal{N}(0, 1). \tag{3.24}$$

In the degenerate case $T = 1$, the matrices in (3.17) are all zeros and the matrix $\widehat{\mathbf{A}}$ reduces to a scalar $\widehat{\mathbf{A}}_{11}$ equal to $\mathrm{Tr}[\boldsymbol{X}(\boldsymbol{X}_{\hat{S}^L}^\top\boldsymbol{X}_{\hat{S}^L})^\dagger\boldsymbol{X}^\top] = |\hat{S}^L|$ where $\hat{S}^L$ is the support of the Lasso $\widehat{\boldsymbol{\beta}}^L$. Here $\widehat{\mathbf{A}}_{11}$ is the usual effective degrees-of-freedom for the Lasso. The factor $(1-\widehat{\mathbf{A}}_{11}/n)^{-1} = (1-|\hat{S}^L|/n)^{-1}$ in (3.24) is the degrees-of-freedom adjustment for the Lasso studied in [26], which is required for the asymptotic normality result (3.24) when $s \gtrsim n^{2/3}$ [26]. So Theorem 3.3 reduces to the asymptotic normality result of [26] in the degenerate case $T = 1$, and in this case the matrix inversion (3.23) reduces to a degrees-of-freedom adjustment through the scalar multiplication by $(1 - |\hat{S}^L|/n)^{-1}$. The length of the resulting confidence interval for $\boldsymbol{a}^\top\boldsymbol{\beta}^{(1)}$ when $T = 1$ (or when the tasks $2, ..., T$) are thrown away) is then

$$2z_{\alpha/2}n^{-1}\|\boldsymbol{y}^{(1)} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^L\|_2(1 - |\hat{S}^L|/n)^{-1}. \tag{3.25}$$

We may compare the lengths of the two confidence intervals:

- The confidence interval $[L_-^\alpha, L_+^\alpha]$ based on (3.22) using the responses on all tasks $1, ..., T$ with length $2n^{-1}z_{\alpha/2}\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{e}_1\|_2$, and
- The confidence interval based on (3.24) obtained by throwing away the responses on tasks $2, ..., T$ with length (3.25).

The length of the confidence interval based on $\widehat{\boldsymbol{B}}$ and the responses on all tasks $1, ..., T$ is smaller than the length (3.25) only when

$$\|\boldsymbol{y}^{(1)} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^L\|_2(1 - |\hat{S}^L|/n)^{-1} > \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{e}_1\|_2. \tag{3.26}$$

Our simulations in Section 3.6 (see Figure 3.5) reveal that (3.26) holds, in some situations with significant margins, when s is not too large. Since the comparison (3.26) can be performed by looking at the data, the practitioner should choose the multi-task confidence interval based on (3.22) over the single-task confidence interval based on (3.24) when (3.26) holds. When performing this comparison, two tests are constructed which calls for a Bonferroni correction to avoid invalid coverage due to multiple testing.

3.3.2 Unknown $\boldsymbol{\Sigma}$: Pivotal random variable, asymptotic normality and confidence intervals

The knowledge of $\boldsymbol{\Sigma}$ is not available in most practical situations and the methodology of the previous subsection cannot be applied. Indeed the left hand sides in Theorem 3.3 involve $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ which cannot be directly constructed from the data when $\boldsymbol{\Sigma}$ unknown. Another issue that arises when $\boldsymbol{\Sigma}$ is unknown is that one cannot verify the normalization $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2 = 1$ required in Theorem 3.3. Intuitively, though, if it was possible to estimate both $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ and $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2$ fast enough, replacing these quantities by their estimates in (3.22) should not break asymptotic normality. Following ideas from the early de-biasing literature [280, 140, 256], we consider a direction

$$\boldsymbol{a} = \boldsymbol{e}_j \tag{3.27}$$

for some fixed covariate $j \in \{1, ..., p\}$ and compute the nodewise Lasso

$$\widehat{\boldsymbol{\gamma}}^{(j)} = \operatorname*{arg\,min}_{\boldsymbol{\gamma}\in\mathbb{R}^p} \|\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}_{-j}\boldsymbol{\gamma}\|_2^2/(2n) + \hat{\tau}_j(1+\eta)\sqrt{(2/n)\log p}\|\boldsymbol{\gamma}\|_1 \tag{3.28}$$

for regressing $\boldsymbol{X}\boldsymbol{e}_j$ on $\boldsymbol{X}_{-j}$, where $\boldsymbol{X}_{-j} \in \mathbb{R}^{n\times p}$ is the matrix $\boldsymbol{X}$ with j-th column replaced by a column of zeros, $\hat{\tau}_j$ is a consistent estimate of $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2^{-1}$ and $\eta > 0$ is a small constant. Alternatively, one may use the scale invariant version of (3.28) again for regressing $\boldsymbol{X}\boldsymbol{e}_j$ on $\boldsymbol{X}_{-j}$,

$$\widehat{\boldsymbol{\gamma}}^{(j)} = \operatorname*{arg\,min}_{\boldsymbol{\gamma}\in\mathbb{R}^p:\gamma_j=0} \big(\|\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}_{-j}\boldsymbol{\gamma}\|_2^2/(2n)\big)^{1/2} + (1+\eta)\sqrt{(2/n)\log p}\|\boldsymbol{\gamma}\|_1, \tag{3.29}$$

known as Scaled lasso [244] or square-root Lasso [28], and (3.29) is equal to (3.28) with $\hat{\tau}_j = \|\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}^{(j)}\|_2/\sqrt{n}$. We finally set

$$\widehat{\boldsymbol{z}}_j = \boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}^{(j)}. \tag{3.30}$$

This corresponds to the residuals of the estimator $\widehat{\boldsymbol{\gamma}}^{(j)}$ in the linear model

$$\boldsymbol{X}\boldsymbol{e}_j = \boldsymbol{X}_{-j}\boldsymbol{\gamma}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \tag{3.31}$$

with response vector $\boldsymbol{X}\boldsymbol{e}_j \in \mathbb{R}^n$, design matrix $\boldsymbol{X}_{-j}$, true regression vector $\boldsymbol{\gamma}^{(j)} := -\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2^{-2}(\boldsymbol{I}_p - \boldsymbol{e}_j\boldsymbol{e}_j^\top)\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$ (so that $\boldsymbol{e}_j^\top\boldsymbol{\gamma}^{(j)} = 0$ and $\boldsymbol{e}_k^\top\boldsymbol{\gamma}^{(j)} = -(\boldsymbol{\Sigma}^{-1})_{jj}^{-1}(\boldsymbol{\Sigma}^{-1})_{jk}$ for $k \in [p]\setminus\{j\}$), and Gaussian noise vector $\boldsymbol{\varepsilon}^{(j)} := \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2^{-2}\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$ independent of $\boldsymbol{X}_{-j}$ with distribution $\boldsymbol{\varepsilon}^{(j)} \sim \mathcal{N}_n(\boldsymbol{0}, \tau_j^2\boldsymbol{I}_{n\times n})$ where $\tau_j^2 := \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2^{-2} = (\boldsymbol{\Sigma}^{-1})_{jj}^{-1}$. The relationship between $\boldsymbol{\Sigma}^{-1}$ and $(\boldsymbol{\gamma}^{(j)}, \tau_j)$ is the well known connection between precision matrix and linear regression for multivariate normal random vectors (see, e.g., [188, 245]).

The estimators $\widehat{\boldsymbol{\gamma}}^{(j)}$ in (3.28) and (3.29) both satisfy inequalities

$$\|\boldsymbol{X}_{-j}^\top(\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}^{(j)})\|_\infty = \|\boldsymbol{X}_{-j}^\top\widehat{\boldsymbol{z}}_j\|_\infty \le O_{\mathbb{P}}(1)\tau_j\sqrt{n\log p}, \tag{3.32}$$

$$\|\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma}^{(j)}\|_1 \le O_{\mathbb{P}}(1)\|\boldsymbol{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0\tau_j\sqrt{\log(p)/n} \tag{3.33}$$

provided that $\|\mathbf{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0\log(p)/n \to 0$. Inequality (3.33) is the usual ℓ_1 estimation rate for the Lasso [35] or the Scaled Lasso [245, 28], and $\|\mathbf{\Sigma}^{-1}\|_{op}^{-1}$ represents a high-probability lower bound on the restricted eigenvalue in the linear model (3.31) [220]. Inequality (3.32) follows from the KKT conditions of (3.28) for the Lasso, and from the KKT conditions of (3.29) combined with $\hat\tau_j/\tau_j \xrightarrow{\mathbb{P}} 1$ which holds thanks to properties of the Scaled or square root Lasso [245, 28]. Inequalities (3.32)-(3.33) are the only properties of $\widehat{\boldsymbol{\gamma}}^{(j)}$ that we will use in the proof of the following result. Other estimators $\widehat{\boldsymbol{\gamma}}^{(j)}$ could be used, for instance ones based on the Dantzig selector, as long as (3.32)-(3.33) are satisfied.

Theorem 3.5. *Consider a canonical basis direction $\boldsymbol{e}_j \in \mathbb{R}^p$ for some $j \in [p]$ and let Assumption (A1) be fulfilled. Additionally assume that the sparsity of $\mathbf{\Sigma}^{-1}\boldsymbol{e}_j$ satisfies either*

$$n^{-1/2}\|\mathbf{\Sigma}^{-1}\boldsymbol{e}_j\|_0\sqrt{[T+\log(p/s)]\log p} \to 0. \tag{3.34}$$

or

$$\|\mathbf{\Sigma}^{-1}\boldsymbol{e}_j\|_0\log(p)/n \to 0 \qquad \text{and} \qquad s\sqrt{\log(p)[T+\log(p/s)]/n} \to 0. \tag{3.35}$$

Then for any estimator $\widehat{\boldsymbol{\gamma}}^{(j)}$ satisfying (3.32)-(3.33) and every fixed $\boldsymbol{b} \in \mathbb{R}^T$ we have

$$\frac{n\boldsymbol{e}_j^\top(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\boldsymbol{b} + n(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}\widehat{\boldsymbol{z}}_j^\top(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}}{(\tau_j)^{-1}\ \|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.36}$$

Asymptotic normality (3.36) still holds if τ_j in the denominator is replaced by either $(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j/n)^{1/2}$ or $\hat\tau_j = (\|\widehat{\boldsymbol{z}}_j\|_2/\sqrt{n})$.

Theorem 3.5 is proved in Section 3.13.1.

3.4 Confidence ellipsoids for rows of $\boldsymbol{B}^*$

3.4.1 Known $\mathbf{\Sigma}$

We first construct confidence ellipsoids with the knowledge of $\mathbf{\Sigma}$.

Theorem 3.6. *Define the observable positive semi-definite matrix $\widehat{\mathbf{\Gamma}} = (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}) \in \mathbb{R}^{T\times T}$ as well as*

$$\boldsymbol{\xi} = (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{z}_0 + (n\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}})(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a}. \tag{3.37}$$

Then under Assumption (A1), there exists a random variable χ^2_T with chi-square distribution with T degrees of freedom such that

$$\sqrt{1-\tfrac{T}{n}}\Big\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\Big\|_2 - \sqrt{\chi^2_T} \le o_{\mathbb{P}}(1) + O_{\mathbb{P}}\Big(\min\Big\{\frac{T}{\sqrt{n}}, \frac{s^2\log^2(p/s)}{n\sqrt{T}}\Big\}\Big)$$

as well as

$$-o_{\mathbb{P}}(1) - O_{\mathbb{P}}\Big(\frac{T}{\sqrt{n}} + \frac{sT+s\log(p/s)}{n}\sqrt{T}\Big) \le \sqrt{1-\tfrac{T}{n}}\Big\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\Big\|_2 - \sqrt{\chi^2_T}.$$

Consequently,

(i) $(1-\frac{T}{n})^{\frac{1}{2}}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 - (\chi^2_T)^{1/2} \le o_{\mathbb{P}}(1)$ *holds if additionally* $\min\{\frac{T^2}{n}, \frac{\log^8 p}{n}\} \to 0$, *and*

(ii) $(1-\frac{T}{n})^{\frac{1}{2}}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 - (\chi^2_T)^{1/2} \ge o_{\mathbb{P}}(1)$ *holds if additionally* $\frac{T^2}{n} + \frac{sT+s\log(p/s)}{n}\sqrt{T} \to 0$.

Theorem 3.6 is proved in Section 3.12. The following proposition with $W_n = (1-\frac{T}{n})^{1/2}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2$ relates the $(1-\alpha)$-quantile of $\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2$ to that of $(\chi^2_T)^{1/2}$ when either (i) or (ii) above holds.

Proposition 3.7. *Let $(W_n)_{n\ge1}$ be a sequence of random random variables and χ^2_T a sequence of random variables with chi-square distribution with T degrees-of-freedom, where $T = T_n$ is function of n (in particular, $T \to +\infty$ as $n \to +\infty$ is allowed). If $\alpha \in (0,1)$ is a fixed constant not depending on n, T and $q_{T,\alpha} > 0$ is the quantile defined by $\mathbb{P}((\chi^2_T)^{1/2} \le q_{T,\alpha}) = 1-\alpha$ then*

(i) $W_n - (\chi^2_T)^{1/2} \le o_{\mathbb{P}}(1)$ *implies that* $\mathbb{P}(W_n \le q_{T,\alpha}) \ge 1-\alpha-o(1)$ *and*

(ii) $W_n - (\chi^2_T)^{1/2} \ge -o_{\mathbb{P}}(1)$ *implies that* $\mathbb{P}(W_n \le q_{T,\alpha}) \le 1-\alpha+o(1)$.

Proposition 3.7 is proved in Section 3.12. If $T \to +\infty$, the order of $q_{T,\alpha}$ is given by

$$q_{T,\alpha} - \sqrt{T} \to z_\alpha/\sqrt{2} \tag{3.38}$$

where z_α is the standard normal quantile defined by $\int_{-\infty}^{z_\alpha}(\sqrt{2\pi})^{-1}e^{-u^2/2}du = 1-\alpha$. A short proof of (3.38) is given around (3.87); see [194] for related discussions. However, using $q_{T,\alpha}$ itself to construct confidence sets should be preferred in practice to avoid the approximation error in (3.38).

Combining the above two results provides confidence ellipsoids for the rows of $\boldsymbol{B}^*$, or more generally for the unknown vector $(\boldsymbol{B}^*)^\top\boldsymbol{a} \in \mathbb{R}^T$ for a fixed direction $\boldsymbol{a} \in \mathbb{R}^p$ of interest. Let $\hat{\mathcal{E}}_\alpha$ be the subset of $\mathbb{R}^T$ defined by

$$\hat{\mathcal{E}}_\alpha \coloneqq \Big\{\boldsymbol{\theta} \in \mathbb{R}^T : (1-\tfrac{T}{n})^{1/2}\|\widehat{\mathbf{\Gamma}}^{-1/2}\big[(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{z}_0 + (n\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}})(\widehat{\boldsymbol{B}}^\top\boldsymbol{a}-\boldsymbol{\theta})\big]\|_2 \le q_{T,\alpha}\Big\}.$$

Since $\hat{\mathcal{E}}_\alpha = \{\boldsymbol{\theta} \in \mathbb{R}^T : (\boldsymbol{\theta}-\boldsymbol{u})^\top\boldsymbol{C}(\boldsymbol{\theta}-\boldsymbol{u}) \le 1\}$ where $\boldsymbol{C} = (q_{T,\alpha})^{-2}(1-\frac{T}{n})(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})\widehat{\mathbf{\Gamma}}^{-1}(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})$ and $\boldsymbol{u} = \widehat{\boldsymbol{B}}^\top\boldsymbol{a} + (n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{z}_0$, this set is an ellipsoidal region with center $\boldsymbol{u}$. If

$$\min\Big\{\frac{T^2}{n}, \frac{\log^8 p}{n}\Big\} \to 0 \tag{3.39}$$

additionally to Assumption (A1) as required in case (i) of Theorem 3.6, then $\mathbb{P}[(\boldsymbol{B}^*)^\top\boldsymbol{a} \in \hat{\mathcal{E}}_\alpha] \ge 1-\alpha-o(1)$. If additionally

$$T^2/n + \sqrt{T}\big(sT + s\log(p/s)\big)/n \to 0 \tag{3.40}$$

as required in case (ii) for the lower bound, then $\mathbb{P}[(\boldsymbol{B}^*)^\top\boldsymbol{a} \in \hat{\mathcal{E}}_\alpha] \to 1-\alpha$ and the above confidence ellipsoid provides the exact nominal coverage (i.e., it is provably non-conservative). Note that the upper bound (i) is more important than the lower bound

(ii) since the upper bound (i) guarantees that the type I error in the hypothesis test (3.6) is at most α, i.e., $\mathbb{P}[(\boldsymbol{B}^*)^\top \boldsymbol{a} \in \hat{\mathcal{E}}_\alpha] \geq 1-\alpha-o(1)$. It is thus fortuitous that only the weak additional condition (3.39) is required for the upper bound (i) to guarantee the desired type I error, while the more stringent condition (3.40) is only required to prove non-conservativeness.

The additional assumption (3.39) is satisfied for a large class of growths of (T, n, p). For instance it holds under polynomial growth $p \asymp n^\gamma$ or exponential growth of the form $p \lesssim \exp(n^{1/8-\gamma'})$ for constants $\gamma, \gamma' > 0$, as $\frac{\log^8 p}{n} \to 0$ is then satisfied. Although we believe that the mild condition (3.39) is an artefact of the proof, it is unclear at this point how to relax (3.39) unless a different ellipsoid is considered. In Section 3.4.3, we will construct a different ellipsoid that does not require the extra conditions (3.39) or (3.40) but that has worse performance in simulations.

The radius of $\hat{\mathcal{E}}_\alpha$ i.e., the half-length of its largest axis is given by

$$\phi_{\min}(\boldsymbol{C})^{-1/2} = (1-T/n)^{-1/2} q_{T,\alpha} \|\widehat{\boldsymbol{\Gamma}}^{1/2}(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})^{-1}\|_{op}. \tag{3.41}$$

Since $\|\boldsymbol{I}_{T\times T} - (\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op} = o_{\mathbb{P}}(1)$ by Proposition 3.2 and Lemma 3.14 on the one hand, and all eigenvalues of $\widehat{\boldsymbol{\Gamma}}$ are of order $\sigma^2 n(1+o_{\mathbb{P}}(1))$ by the arguments in the proof of Lemma 3.26 on the other hand, the radius (3.41) is $q_{T,\alpha}\sigma n^{-1/2}(1+o_{\mathbb{P}}(1))$ which is of order $\sigma\sqrt{T/n}$ by (3.38).

The random vector (3.37) involves multiplication by $(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})$ which differs from the pivotal quantity in the asymptotic normality result (3.21). However, Theorem 3.6 still holds with $\boldsymbol{\xi}$ in (3.37) replaced by

$$\check{\boldsymbol{\xi}} = (\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{z}_0 + n(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{a}. \tag{3.42}$$

Indeed, with

$$\big|\|\widehat{\boldsymbol{\Gamma}}^{-1/2}\check{\boldsymbol{\xi}}\|_2 - \|\widehat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2\big| \leq \|\widehat{\boldsymbol{\Gamma}}^{-1/2}(\check{\boldsymbol{\xi}} - \boldsymbol{\xi})\|_2 = \|\widehat{\boldsymbol{\Gamma}}^{-1/2}\big((\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - \boldsymbol{I}_{T\times T}\big)\boldsymbol{\xi}\|_2.$$

Since the eigenvalues of $\widehat{\boldsymbol{\Gamma}}$ are all of order $\sigma^2 n(1+o_{\mathbb{P}}(1))$, since $\|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - \boldsymbol{I}_{T\times T}\|_{op} \leq (1+o_{\mathbb{P}}(1))|\hat{S}|/n$ by Proposition 3.2 and since $\|\boldsymbol{\xi}\|_2 = O_{\mathbb{P}}(\sqrt{\sigma^2 nT})$ by Theorem 3.25, the previous display is $O_{\mathbb{P}}(\sqrt{T}s/n)$ and converges to 0 in probability by Assumption (A1). Under Assumption (A1), Theorem 3.6 thus holds for $\boldsymbol{\xi}$ in (3.37) if and only if it holds for $\check{\boldsymbol{\xi}}$. Furthermore the corresponding ellipsoid,

$$\check{\mathcal{E}}_\alpha = \Big\{\boldsymbol{\theta} \in \mathbb{R}^T : (1-\tfrac{T}{n})^{1/2}\|\widehat{\boldsymbol{\Gamma}}^{-1/2}\big[(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{z}_0 + n(\widehat{\boldsymbol{B}}^\top \boldsymbol{a} - \boldsymbol{\theta})\big]\|_2 \leq q_{T,\alpha}\Big\}$$

enjoys the same properties as $\hat{\mathcal{E}}_\alpha$: Type I error guarantees $\mathbb{P}[(\boldsymbol{B}^*)^\top \boldsymbol{a} \in \check{\mathcal{E}}_\alpha] \geq 1-\alpha-o(1)$ under (3.39), and non-conservativeness $\mathbb{P}[(\boldsymbol{B}^*)^\top \boldsymbol{a} \in \check{\mathcal{E}}_\alpha] \to 1-\alpha$ under (3.40).

3.4.2 Unknown $\boldsymbol{\Sigma}$

A similar result is available if $\boldsymbol{\Sigma}$ is unknown. Consider the notation (3.30) from Section 3.3.2.

Theorem 3.8. *Consider a canonical basis direction $\boldsymbol{e}_j \in \mathbb{R}^p$ for some $j \in [p]$ and let Assumption (A1) be fulfilled. Additionally assume that either (3.34) or (3.35) holds. Then for any estimator $\widehat{\boldsymbol{\gamma}}^{(j)}$ satisfying (3.32)-(3.33),*

$$\frac{\sqrt{n-T}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big\|\widehat{\boldsymbol{\Gamma}}^{-1/2}\Big((\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\widehat{\boldsymbol{z}}_j + \frac{(n\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}})(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{e}_j}{n(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}}\Big)\Big\|_2 \\ \le \sqrt{\chi_T^2} + o_{\mathbb{P}}(1) + O_{\mathbb{P}}(\min\{\tfrac{T}{\sqrt{n}}, \tfrac{\log^2 p}{n^{1/4}}\}) \quad (3.43)$$

where χ_T^2 is a random variable with chi-square distribution with T degrees-of-freedom.

Theorem 3.8 is proved in Section 3.13.2. The corresponding confidence ellipsoid for the j-th row $(\boldsymbol{B}^*)^\top\boldsymbol{e}_j$ of $\boldsymbol{B}^*$ is

$$\hat{\mathcal{E}}_\alpha^j = \Big\{\boldsymbol{\theta}_j \in \mathbb{R}^T : \frac{\sqrt{n-T}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big\|\widehat{\boldsymbol{\Gamma}}^{-1/2}\Big[(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\widehat{\boldsymbol{z}}_j + \frac{(n\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}})(\widehat{\boldsymbol{B}}^\top\boldsymbol{e}_j-\boldsymbol{\theta}_j)}{n(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}}\Big]\Big\|_2 \le q_{T,\alpha}\Big\}.$$

If either one of the condition (3.34) or (3.35) holds on the growth of the sparsity of $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$, this confidence ellipsoid does not require the knowledge of $\boldsymbol{\Sigma}$ and has the same guarantees as those of the previous section.

3.4.3 Relaxing the additional assumptions (3.39) and (3.40)

Instead of normalizing using $\widehat{\boldsymbol{\Gamma}}^{-1/2}$ as in the previous sections, a simple estimate of σ^2 lets us relax the conditions (3.39) and (3.40) that are required in the previous section to ensure $\|\widehat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 = (\chi_T^2)^{1/2} + o_{\mathbb{P}}(1)$.

Theorem 3.9. *Let $\boldsymbol{\xi}, \check{\boldsymbol{\xi}}$ be defined in (3.37) and (3.42) respectively, and let $\hat\sigma^2 = \|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_F^2/(nT)$. Then under Assumption (A1), there exists a random variable χ_T^2 with chi-square distribution with T degrees of freedom such that*

$$(\hat\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}\|_2 = (\chi_T^2)^{1/2} + o_{\mathbb{P}}(1), \qquad (\hat\sigma^2 n)^{-1/2}\|\check{\boldsymbol{\xi}}\|_2 = (\chi_T^2)^{1/2} + o_{\mathbb{P}}(1). \quad (3.44)$$

Theorem 3.10. *Consider a canonical basis direction $\boldsymbol{e}_j \in \mathbb{R}^p$ for some $j \in [p]$ and let Assumption (A1) be fulfilled. Additionally assume that either (3.34) or (3.35) holds. Then for any estimator $\widehat{\boldsymbol{\gamma}}^{(j)}$ satisfying (3.32)-(3.33),*

$$\frac{1}{\|\widehat{\boldsymbol{z}}_j\|_2\hat\sigma}\Big\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\widehat{\boldsymbol{z}}_j + \frac{(n\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}})(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{e}_j}{n(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}}\Big\|_2 = \sqrt{\chi_T^2} + o_{\mathbb{P}}(1) \quad (3.45)$$

where χ_T^2 is a random variable with chi-square distribution with T degrees-of-freedom.

The above asymptotic chi-square results hold under the same assumptions as Theorem 3.3 and Theorem 3.5. The reason for the success of these estimates is that $\hat\sigma$ estimates σ at a rate faster than $T^{-1/2}$: we have $|\hat\sigma/\sigma - 1| = o_{\mathbb{P}}(T^{-1/2})$ by Theorem 3.25. However, simulations in Section 3.6 reveal that the asymptotic $(\chi_T^2)^{1/2}$ estimates of the previous subsections involving the matrix $\widehat{\mathbf{\Gamma}}^{-1/2}$ are more robust to larger sparsity levels, although Assumption (A1) is oblivious to this phenomenon.

The corresponding $1-\alpha$ confidence ellipsoid for $(\boldsymbol{B}^*)^\top \boldsymbol{a}$ based on (3.44) and $\check{\boldsymbol{\xi}}$ is

$$\check{\mathcal{E}}_{\hat\sigma,\alpha} = \Big\{ \boldsymbol{\theta} \in \mathbb{R}^T : \tfrac{1}{\hat\sigma\sqrt{n}} \big\| \big(\boldsymbol{I}_{T\times T} - \tfrac{\widehat{\mathbf{A}}}{n}\big)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{z}_0 + n(\widehat{\boldsymbol{B}}^\top \boldsymbol{a} - \boldsymbol{\theta}) \big\|_2 \le q_{T,\alpha} \Big\} \quad (3.46)$$

and satisfies $\mathbb{P}[(\boldsymbol{B}^*)^\top \boldsymbol{a} \in \check{\mathcal{E}}_{\hat\sigma,\alpha}] \to 1-\alpha$ under Assumption (A1). Similar confidence ellipsoids based on (3.45) can be readily constructed.

3.4.4 Hypothesis testing

We now turn to type II error for the testing problem

$$H_0 : (\boldsymbol{B}^*)^\top \boldsymbol{a} = \mathbf{0}_{T\times 1} \qquad \text{against} \qquad H_1 : \|(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2 \ge \rho_n \quad (3.47)$$

where $\rho_n > 0$ is a separation radius. The hypothesis test (3.47) at level $1-\alpha$ is naturally achieved by rejecting H_0 if and only if $\mathbf{0}_{T\times 1} \notin \check{\mathcal{E}}_{\hat\sigma,\alpha}$ for the ellipsoid in (3.46). Similar rejection procedures can be obtained with $\check{\mathcal{E}}_\alpha$ or $\hat{\mathcal{E}}^j_\alpha$ for the confidence ellipsoids defined in Sections 3.4.1 and 3.4.2.

We can also determine the separation radius ρ_n required so that this testing procedure has nontrivial power (type II error). Focusing here on $\check{\mathcal{E}}_{\hat\sigma,\alpha}$ in (3.46), rejection happens if and only if the following quantity is positive

$$\begin{aligned} &(\hat\sigma^2 n)^{-1} \|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{z}_0 + n\widehat{\boldsymbol{B}}^\top \boldsymbol{a}\|_2^2 - q_{T,\alpha}^2 \\ &= W_n^2 - q_{T,\alpha}^2 \\ &\quad + (\hat\sigma^2 n)^{-1} \|n(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2^2 \\ &\quad + 2(\hat\sigma^2 n)^{-1/2} n \boldsymbol{a}^\top (\boldsymbol{B}^*)^\top \check{\boldsymbol{\xi}}. \end{aligned}$$

where $W_n^2 = (\hat\sigma^2 n)^{-1} \|\check{\boldsymbol{\xi}}\|_2^2$ and $W_n = (\chi^2_T)^{1/2} + o_{\mathbb{P}}(1)$ by (3.44). By Theorem 3.3 applied to $\boldsymbol{b} = (\boldsymbol{B}^*)^\top \boldsymbol{a} \|(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2^{-1}$, the last line is of the form $2\hat\sigma^{-2} \|(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2 \mathcal{N}(0,\sigma^2)$ so that it is of order $\|(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2 O_{\mathbb{P}}(1)$. The second line is positive, of order $\sigma^{-2} n(1 + o_{\mathbb{P}}(1)) \|(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2^2$; this is the quantity that should dominate in order to ensure that the above display is positive. Since the first line $W_n^2 - q_{T,\alpha} = (W_n - q_{T,\alpha})(W_n + q_{T,\alpha})$ is positive with probability at least $\alpha - o(1)$ by Proposition 3.7, we obtain that if $\|(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2 \ge \rho_n$ for $\rho_n/(\sigma n^{-1/2}) \to +\infty$, then the type II error is at most $1 - \alpha + o(1)$. Although this type II error is typically a constant close to 1 (e.g. if $\alpha = 0.05$), this shows that the above test has at most constant type II error as long as the separation radius satisfies $\rho_n \ggg \sigma n^{-1/2}$. We can also find conditions on ρ_n that ensures that the type II error is smaller than any constant. The first line above is of order $(W_n - q_{T,\alpha}) O_{\mathbb{P}}(\sqrt{T}) = O_{\mathbb{P}}(\sqrt{T})$ since $W_n = O_{\mathbb{P}}(\sqrt{T}) + o_{\mathbb{P}}(1)$ and $q_{T,\alpha} = \sqrt{T} + O(1)$ by Proposition 3.7 and (3.38). Thus $\rho_n \ggg T^{1/4}\sigma n^{-1/2}$ is sufficient in order for $(\hat\sigma^2 n)^{-1} \|n(\boldsymbol{B}^*)^\top \boldsymbol{a}\|_2^2$ to dominate both the first and third lines with probability approaching one. In summary, $\rho_n \ggg \sigma n^{-1/2}$ is sufficient to achieve a constant type II error, while $\rho_n \ggg T^{1/4}\sigma n^{-1/2}$ is sufficient to grant a vanishing type II error.

In single task models, coefficients $\boldsymbol{B}^*_{jt}$ of order $o(\sigma n^{-1/2})$ cannot be detected, cf. the discussion after (3.6). Here on the other hand in the multi-task setting with $T \to +\infty$, detection of non-zero vector $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ is possible with constant power even if the individual coefficients in $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ are $u_n \sigma (Tn)^{-1/2}$ for any slowly increasing u_n with $u_n \to +\infty$. If $u_n = o(\sqrt{T})$, the coefficients $(\boldsymbol{B}^*_{jt})_{t=1,...,T}$ are individually impossible to detect, while detection of the row vector $(\boldsymbol{B}^*)^\top \boldsymbol{e}_j$ is possible with constant type I and type II errors. Similarly, if the individual coefficients $(\boldsymbol{B}^*_{jt})_{t=1,...,T}$ are of order $u_n \sigma T^{-1/4} n^{-1/2}$ for any slowly increasing u_n with $u_n \to +\infty$, the above testing procedure for the row vector $(\boldsymbol{B})^\top \boldsymbol{e}_j$ has vanishing type II error.

3.5 Computing the interaction matrix efficiently

Equation (3.18) which defines $\widehat{\mathbf{A}}$ is convenient for theoretical purposes, as the pseudoinverse suppresses invertibility issues and the form (3.18) naturally arises in the proofs, see for instance Lemmas 3.18 to 3.20. However, (3.18) is not computationally tractable as it involves computing a pseudoinverse of size $pT \times pT$. The goal of this section is to provide a computationally tractable representation for $\widehat{\mathbf{A}}$; in particular we will see that one only needs to compute inverses of matrices of size $|\hat{S}| \times |\hat{S}|$. A first step when implementing is to remove all covariates $j \in \{1, ..., p\}$ such that $\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j = \mathbf{0}$, as dropping those indices and the corresponding columns of $\boldsymbol{X}$ does not change the value of $\widehat{\mathbf{A}}$ in (3.18). For the purpose of this section and only in this section, we assume without loss of generality that $\hat{S} = [p]$ and that all variables $j \in [p]$ are such that $\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \neq \mathbf{0}$. However, we will keep the notation $\boldsymbol{X}_{\hat{S}}$ and use summation sign $\sum_{j \in \hat{S}}$ to emphasize that the indices $j \notin \hat{S}$ and corresponding columns of $\boldsymbol{X}$ have been dropped.

Before stating a formal proposition with a computationally friendly representation of the matrix $\widehat{\mathbf{A}}$, we explain the crux of the argument, which relies on the Sherman-Morrison-Woodbury inversion formula. Recall that $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = (\boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}})$ and that for every $j \in \hat{S}$

$$\boldsymbol{H}^{(j)} := \lambda \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-1} \big(\boldsymbol{I}_{T\times T} - \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \boldsymbol{e}_j^\top \widehat{\boldsymbol{B}}\ \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-2}\big) \quad \in \mathbb{R}^{T\times T}, \tag{3.17}$$

as well as $\tilde{\boldsymbol{H}} = \sum_{j\in\hat{S}} \boldsymbol{H}^{(j)} \otimes (\boldsymbol{e}_j \boldsymbol{e}_j^\top)$. By splitting the part of $\boldsymbol{H}^{(j)}$ proportional to the identity and the rank one part, we find

$$\begin{aligned}
\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}} &= (\boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}}) + nT\lambda \sum_{j\in\hat{S}} \frac{\boldsymbol{I}_{T\times T} \otimes \boldsymbol{e}_j \boldsymbol{e}_j^\top}{\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2} - \Big[\frac{\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \boldsymbol{e}_j^\top \widehat{\boldsymbol{B}}}{\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^3} \otimes (\boldsymbol{e}_j \boldsymbol{e}_j^\top)\Big] \\
&= \big(\boldsymbol{I}_{T\times T} \otimes \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)\big) - nT\lambda \sum_{j\in\hat{S}} \Big[\frac{(\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \boldsymbol{e}_j^\top \widehat{\boldsymbol{B}}) \otimes (\boldsymbol{e}_j \boldsymbol{e}_j^\top)}{\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^3}\Big]
\end{aligned}$$

where $\boldsymbol{v} \in \mathbb{R}^{|\hat{S}|}$ is the vector with $v_j = nT\lambda \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-1}$ and $\mathbf{diag}(\boldsymbol{v})$ is the square diagonal matrix with $\boldsymbol{v}$ as its diagonal. By the mixed product property (3.13) we have

$$(\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \boldsymbol{e}_j^\top \widehat{\boldsymbol{B}}) \otimes (\boldsymbol{e}_j \boldsymbol{e}_j^\top) = (\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \otimes \boldsymbol{e}_j)(\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \otimes \boldsymbol{e}_j)^\top$$

so that, with $\boldsymbol{b}^{(j)} = (nT\lambda\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-3})^{1/2}\ \widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \in \mathbb{R}^T$ we obtain

$$\begin{aligned}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}} &= \big(\boldsymbol{I}_{T\times T} \otimes \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)\big) - \sum_{j\in\hat{S}} (\boldsymbol{b}^{(j)} \otimes \boldsymbol{e}_j)(\boldsymbol{b}^{(j)} \otimes \boldsymbol{e}_j)^\top \\ &= \big(\boldsymbol{I}_{T\times T} \otimes \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)\big) - \boldsymbol{U}\boldsymbol{U}^\top \\ &= \boldsymbol{M} - \boldsymbol{U}\boldsymbol{U}^\top,\end{aligned}$$

where $\boldsymbol{U} \in \mathbb{R}^{(|\hat{S}|T)\times|\hat{S}|}$ has columns $(\boldsymbol{b}^{(j)} \otimes \boldsymbol{e}_j)_{j\in\hat{S}}$ and $\boldsymbol{M} = \boldsymbol{I}_{T\times T} \otimes \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)$. If $\boldsymbol{M}$ is invertible and its inverse can be computed efficiently, the inverse of the above display is given by the Sherman-Morrison-Woodbury formula [123]: if the matrix $-\boldsymbol{I}_{|\hat{S}|\times|\hat{S}|} + \boldsymbol{U}^\top \boldsymbol{M}^{-1}\boldsymbol{U}$ is invertible then $\boldsymbol{M} - \boldsymbol{U}\boldsymbol{U}^\top$ is also invertible and

$$(\boldsymbol{M} - \boldsymbol{U}\boldsymbol{U}^\top)^{-1} = \boldsymbol{M}^{-1} - \boldsymbol{M}^{-1}\boldsymbol{U}\left(-\boldsymbol{I}_{|\hat{S}|\times|\hat{S}|} + \boldsymbol{U}^\top \boldsymbol{M}^{-1}\boldsymbol{U}\right)^{-1} \boldsymbol{U}^\top \boldsymbol{M}^{-1}.$$

Since $\boldsymbol{v}$ has positive entries, $\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})$ is always invertible and so is $\boldsymbol{M}$, with

$$\boldsymbol{M}^{-1} = \boldsymbol{I}_{T\times T} \otimes \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)^{-1}. \tag{3.48}$$

Hence we only need to perform two inversions of matrices of size $|\hat{S}|\times|\hat{S}|$: the inversion of $\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})$ and of $-\boldsymbol{I}_{|\hat{S}|\times|\hat{S}|} + \boldsymbol{U}^\top \boldsymbol{M}^{-1}\boldsymbol{U}$.

Proposition 3.11. *With the above notation for $\boldsymbol{v} \in \mathbb{R}^{|\hat{S}|}$ and $\boldsymbol{b}^{(j)} \in \mathbb{R}^T$ for each $j \in \hat{S}$, if the matrix $\boldsymbol{P}$ defined entrywise by*

$$\boldsymbol{P} = (P_{jk})_{(j,k)\in\hat{S}\times\hat{S}}, \quad P_{jk} = -I\{j=k\} + (\boldsymbol{b}^{(j)\top}\boldsymbol{b}^{(k)})\ \big(\boldsymbol{e}_j^\top \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)^{-1}\boldsymbol{e}_k\big)$$

is invertible then

$$\widehat{\mathsf{A}} = \mathrm{Tr}\Big[\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v}))^{-1}\Big]\ \boldsymbol{I}_{T\times T} - \Big[\sum_{j\in\hat{S}} \boldsymbol{b}^{(j)} \sum_{k\in\hat{S}} (\boldsymbol{e}_j^\top \boldsymbol{Q}\boldsymbol{e}_k)(\boldsymbol{e}_j^\top \boldsymbol{P}^{-1}\boldsymbol{e}_k)\boldsymbol{b}^{(k)\top}\Big]$$

where $\boldsymbol{Q} = \big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)^{-1}\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}}\big(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v})\big)^{-1}$.

Proof. By definition of $\widehat{\mathsf{A}}$ and using the above Sherman-Morrison-Woodbury identity

$$\begin{aligned}\widehat{\mathsf{A}}_{t,t'} &= \mathrm{Tr}[(\boldsymbol{e}_t^\top \otimes \boldsymbol{X}_{\hat{S}})(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})^\dagger(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{S}}^\top)] \\ &= \mathrm{Tr}[(\boldsymbol{e}_t^\top \otimes \boldsymbol{X}_{\hat{S}})\boldsymbol{M}^{-1}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{S}}^\top)] \\ &\quad - \mathrm{Tr}\big[(\boldsymbol{e}_t^\top \otimes \boldsymbol{X}_{\hat{S}})\boldsymbol{M}^{-1}\boldsymbol{U}\big(-\boldsymbol{I}_{|\hat{S}|\times|\hat{S}|} + \boldsymbol{U}^\top \boldsymbol{M}^{-1}\boldsymbol{U}\big)^{-1}\boldsymbol{U}^\top \boldsymbol{M}^{-1}(\boldsymbol{e}_{t'} \otimes \boldsymbol{X}_{\hat{S}}^\top)\big] \\ &= \mathrm{Tr}[(\boldsymbol{e}_t^\top \boldsymbol{e}_{t'}) \otimes (\boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v}))^{-1}\boldsymbol{X}_{\hat{S}}^\top)] \\ &\quad - \mathrm{Tr}\big[\boldsymbol{U}^\top \boldsymbol{M}^{-1}((\boldsymbol{e}_{t'}\boldsymbol{e}_t^\top) \otimes \boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}})\boldsymbol{M}^{-1}\boldsymbol{U}\big(-\boldsymbol{I}_{|\hat{S}|\times|\hat{S}|} + \boldsymbol{U}^\top \boldsymbol{M}^{-1}\boldsymbol{U}\big)^{-1}\big].\end{aligned}$$

By (3.14), the first term equals $I\{t=t'\}\,\mathrm{Tr}[\boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v}))^{-1}\boldsymbol{X}_{\hat{S}}^\top]$ which gives the first term in the proposition, proportional to $\boldsymbol{I}_{T\times T}$. Using again the structure of $\boldsymbol{M}^{-1}$ in (3.48), the second summand in the previous display is equal to

$$-\,\mathrm{Tr}\big[\boldsymbol{U}^\top\big((\boldsymbol{e}_{t'}\boldsymbol{e}_t^\top) \otimes \boldsymbol{Q}\big)\boldsymbol{U}\boldsymbol{P}^{-1}\big] \tag{3.49}$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are given in the proposition, after noting that the definition of $\boldsymbol{P}$ is equivalent to $\boldsymbol{P} = -\boldsymbol{I}_{|\hat{S}|\times|\hat{S}|} + \boldsymbol{U}^\top \boldsymbol{M}^{-1}\boldsymbol{U}$. Since $\boldsymbol{U}$ has columns $\boldsymbol{b}^{(j)} \otimes \boldsymbol{e}_j$, the entry $(j,k) \in \hat{S} \times \hat{S}$ of the matrix $\boldsymbol{U}^\top\big((\boldsymbol{e}_{t'}\boldsymbol{e}_t^\top) \otimes \boldsymbol{Q}\big)\boldsymbol{U}$ is equal to

$$(\boldsymbol{b}^{(j)\top}\boldsymbol{e}_{t'}\boldsymbol{e}_t^\top\boldsymbol{b}^{(k)})\ (\boldsymbol{e}_j^\top\boldsymbol{Q}\boldsymbol{e}_k) = (\boldsymbol{e}_{t'}^\top\boldsymbol{b}^{(j)})(\boldsymbol{e}_t^\top\boldsymbol{b}^{(k)})\ (\boldsymbol{e}_j^\top\boldsymbol{Q}\boldsymbol{e}_k).$$

Since $\operatorname{Tr}[\boldsymbol{A}\boldsymbol{B}] = \sum_{j,k} A_{jk}B_{jk}$ for two symmetric matrices of the same size, we obtain

$$\begin{aligned}(3.49) &= -\sum_{j\in\hat{S}}\sum_{k\in\hat{S}}(\boldsymbol{e}_{t'}^\top\boldsymbol{b}^{(j)})(\boldsymbol{e}_t^\top\boldsymbol{b}^{(k)})\ (\boldsymbol{e}_j^\top\boldsymbol{Q}\boldsymbol{e}_k)(\boldsymbol{e}_j^\top\boldsymbol{P}^{-1}\boldsymbol{e}_k)\\ &= \boldsymbol{e}_{t'}^\top\Big[-\sum_{j\in\hat{S}}\boldsymbol{b}^{(j)}\sum_{k\in\hat{S}}\ (\boldsymbol{e}_j^\top\boldsymbol{Q}\boldsymbol{e}_k)(\boldsymbol{e}_j^\top\boldsymbol{P}^{-1}\boldsymbol{e}_k)\boldsymbol{b}^{(k)\top}\Big]\boldsymbol{e}_t.\end{aligned}$$

On the last line, the matrix in bracket is the second matrix in the expression of $\widehat{\mathbf{A}}$. □

We now turn to implementation details. We recommend an approach that makes use of optimized vectorized code as often as possible to compute the quantities in Proposition 3.11, and if available to use a library with Einstein summation routine as this allows the code to mimick the mathematical notation in Proposition 3.11. For concreteness, the following code lets us efficiently compute $\widehat{\mathbf{A}}$ with the Python library Numpy [115], and the Einstein summation function `numpy.einsum` which comes in handy. Assume that $\widehat{\boldsymbol{B}}$ has been computed, the rows in $[p] \setminus \hat{S}$ removed and the result stored in an array `B_S` of size $|\hat{S}| \times T$, that $\boldsymbol{X}$ with the columns in $[p] \setminus \hat{S}$ removed is stored in an array `X_S` of size $n \times |\hat{S}|$, and that the scalar $nT\lambda$ is stored in variable `nTlambda`. Then the vector $\boldsymbol{v}$ and matrix with columns $(\boldsymbol{b}^{(j)})_{j\in\hat{S}}$ in variable `b` can be computed as follows:

```
import numpy as np
norms = np.linalg.norm(B_S, axis=1)                      # shape (|Ŝ|, )
v = nTlambda * norms**(-1)                               # shape (|Ŝ|, )
b = nTlambda**0.5 * np.einsum("j,jt->jt", norms**(-3/2), B_S)# shape (|Ŝ|, T)
```

Finally, matrices $(\boldsymbol{X}_{\hat{S}}^\top\boldsymbol{X}_{\hat{S}} + \mathbf{diag}(\boldsymbol{v}))^{-1}$ and $\boldsymbol{Q}$ are computed using built-in symmetric matrix inversion, while computation of $\boldsymbol{P}$ and $\widehat{\mathbf{A}}$ again resorts to using `np.einsum`:

```
gram = X_S.T @ X_S                                       # shape (|Ŝ|,|Ŝ|)
inverse = np.linalg.invh(gram + np.diag(v))              # shape (|Ŝ|,|Ŝ|)
Q = inverse @ gram @ inverse                             # shape (|Ŝ|,|Ŝ|)
P = - np.eye(p) + np.einsum("jt,kt,jk -> jk", b, b, inverse) # shape (|Ŝ|,|Ŝ|)
A = np.eye(T) * np.einsum("jk,jk->", gram, inverse) \
    - np.einsum("jt,ku,jk,jk->tu", b, b, Q, np.linalg.invh(P))# shape (T, T)
```

In `einsum`, we use indices `t` and `u` to loop over $[T]$, and indices `j` and `k` to loop over $\hat{S}$. All calls to `einsum` can be further optimized by pre-computing the optimal order in which tensor contractions should be performed (see `numpy.einsum_path`).

Empirically, we have observed that this implementation using the Sherman-Morrison-Woodbury identity and the above code is several orders of magnitude faster than a naive one involving sparse matrices and the full inversion of $\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}$.

3.6 Numerical experiments

We run simulations to illustrate the theorems proved in Sections 3.3 and 3.4. The values of the parameters are fixed to $n = 2000$, $p = 6000$, $T = 10$, $\eta_1 = \eta_2 = 0$, $\sigma^2 = 1$. The tuning parameter is $\lambda = \max_j \boldsymbol{\Sigma}_{jj}^{1/2} \frac{1}{\sqrt{nT}}\left(1 + \sqrt{\frac{2}{T}\log\frac{p}{s}}\right)$ (we explain below how $\boldsymbol{\Sigma}$ is constructed). The directions of interest are $\boldsymbol{a} = \boldsymbol{e}_j \in \mathbb{R}^p$ and $\boldsymbol{b} = \boldsymbol{e}_1 \in \mathbb{R}^T$.

3.6.1 Quantile-quantile plots of the pivotal quantities

The goal is to assess how the sparsity of $\boldsymbol{B}^*$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1$ influence the convergence in Theorems 3.3, 3.5, 3.6, 3.8 and 3.10. Denote by s and s_Ω the respective sparsity parameters that will vary in the experiments. Given a target tuple (s, s_Ω) we generate $\boldsymbol{B}^*$ with exactly s non-zero rows and $\boldsymbol{\Sigma}$ with exactly s_Ω non-zero entries on the first column of $\boldsymbol{\Sigma}^{-1}$, so that $s_\Omega = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1\|_0$.

We explain first how $\boldsymbol{\Sigma}$ is constructed so that it satisfies the constraints in Assumption (A1) as well as the sparsity requirement on $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1$. Start by sampling $\boldsymbol{M}$, a $(p-1)\times(p-1)$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Then perform the QR decomposition of $\boldsymbol{M}$ to obtain an orthogonal matrix $\boldsymbol{Q}$, the distribution of which is uniform in the sense of Haar measure on the orthogonal group $O(p-1)$. Next, consider $\boldsymbol{D}$, the diagonal $(p-1)\times(p-1)$ matrix with entries $\{1 + j/(p-2) : j \in \{0,\dots,p-2\}\}$ and set $\boldsymbol{\Lambda} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^\top$. Define the block matrix

$$\widetilde{\boldsymbol{\Lambda}} = \left[\begin{array}{c|c} 3/2 & \boldsymbol{v}^\top \\ \hline \boldsymbol{v} & \boldsymbol{\Lambda} \end{array}\right]$$

where $\boldsymbol{v} \in \mathbb{R}^{p-1}$ is a vector with sparsity $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1\|_0 - 1$ and norm $\|\boldsymbol{v}\|_2 = 1$. This ensures boundedness of the spectrum as the smallest eigenvalue of $\widetilde{\boldsymbol{\Lambda}}$ satisfies the lower bound

$$\lambda_{\min}(\widetilde{\boldsymbol{\Lambda}}) \geq \lambda_{\min}\left(\begin{bmatrix} 3/2 & -\|\boldsymbol{v}\|_2 \\ -\|\boldsymbol{v}\|_2 & \lambda_{\min}(\boldsymbol{\Lambda}) \end{bmatrix}\right) = \frac{5-\sqrt{17}}{4} \gtrsim 0.219,$$

where the last equality follows from $\lambda_{\min}(\boldsymbol{\Lambda}) = 1$ and $\|\boldsymbol{v}\|_2 = 1$. Similarly, the largest eigenvalue of $\widetilde{\boldsymbol{\Lambda}}$ can be bounded above by $\lambda_{\max}(\widetilde{\boldsymbol{\Lambda}}) = \frac{7+\sqrt{17}}{4} \lesssim 2.8$. Finally set $\boldsymbol{\Sigma} = \alpha^{-1}\widetilde{\boldsymbol{\Lambda}}^{-1}$ where α is the greatest diagonal entry of $\widetilde{\boldsymbol{\Lambda}}^{-1}$ so that $\max\{\boldsymbol{\Sigma}_{jj}, j = 1,\dots,p\} = 1$. This construction leads to $\lambda_{\min}(\boldsymbol{\Sigma}) \approx 0.32, \lambda_{\max}(\boldsymbol{\Sigma}) \approx 1.76$ and $(\boldsymbol{\Sigma}^{-1})_{jj} \approx 1.85$.

The row-sparse matrix $\boldsymbol{B}^*$ is constructed as follows. Initialize $\boldsymbol{B}^*$ as a matrix filled with λ's and alter it in two different ways:

(i) *Setting with overlapping supports.* In the first setting, we zero out rows of $\boldsymbol{B}^*$ while forcing an overlap of the supports of $\boldsymbol{B}^*$ and $\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1$ (either $\text{supp}(\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1) \subset \text{supp}(\boldsymbol{B}^*)$ or the reverse inclusion). The intuition is that this makes inequality (3.91) tight. This constraint is therefore expected to slow down convergence.

(ii) *No-overlap setting.* In the second setting this constraint is removed and the support of $\boldsymbol{B}^*$ is picked uniformly at random as a subset of $\{1,\dots,p\} \setminus \text{supp}(\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1)$.

Assume that the tuple (s, s_Ω) is fixed. We sample $N_{sim} = 128$ instances of $(\boldsymbol{X}, \boldsymbol{E})$. For each sample, we compute the estimator $\widehat{\boldsymbol{B}}$ using the function `MultiTaskElasticNet` from the Python library Scikit-learn [207], build the interaction matrix $\widehat{\mathbf{A}}$ using the implementation from Section 3.5 and collect the pivotal quantities appearing in the Theorems. The Q-Q plots and histograms for different pairs (s, s_Ω) are then reported in Figures 3.1 and 3.2 for the overlapping supports setting (i) and Figures 3.3 and 3.4 for the no-overlap setting (ii).

Asymptotic normality is observed empirically on Figure 3.3 in the no-overlap setting, both when $\boldsymbol{\Sigma}$ is known (blue) and unknown (green). The convergence holds up well across a wide range of sparsity levels. In the overlapping supports setting of Figure 3.1, convergence is maintained if $\boldsymbol{\Sigma}$ is known, but in the unknown $\boldsymbol{\Sigma}$ case it deteriorates fast when $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1\|_0$ grows. This suggests that condition (3.34) is not an artefact of the proof.

The picture is different with chi-square results. In the no-overlap setting of Figure 3.4, convergence is observed across all sparsity levels for pivotal quantities in Theorem 3.6 (known $\boldsymbol{\Sigma}$) and Theorem 3.8 (unknown $\boldsymbol{\Sigma}$) whereas an increase in s slows down convergence in Theorem 3.10 (unknown $\boldsymbol{\Sigma}$). In the overlapping supports setting (i) of Figure 3.2, pivotal quantities in Theorems 3.6 and 3.10 exhibit the same behavior as in the previous setting whereas the one from Theorem 3.8 shows increasingly slower convergence as $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1\|_0$ grows. Again, this suggests that condition (3.34) is not an artefact of the proof.

3.6.2 The advantage of multi-task learning for narrower confidence intervals

In Figure 3.5 we illustrate the discussion around (3.26) by comparing the lengths of 95% confidence intervals obtained via multi-task Lasso and single-task Lasso. $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1\|_0$ is set to 5 and the pair (T, s) varies. For a given (T, s) and a sampled $(\boldsymbol{X}, \boldsymbol{E})$ we compute the relative change $(\text{length}_{multi} - \text{length}_{single})/\text{length}_{single}$. We collect these values over $N_{sim} = 128$ samples and obtain the bottom figure. Since the results with or without the overlap constraint in the supports are similar, only the no-overlap setting (ii) is shown. In the upper figure, multi-task confidence interval lengths are pooled together over the samples and we compare them to the aggregate single-task lengths. As a sanity check we observe that multi-task and single-task Lasso coincide when T is equal to 1. For $s = 15$, $\widehat{\mathbf{A}}$-based confidence intervals always have smaller length, which shrinks as T increases. When $T = 20$ we observe a 40% average gain in the width. Exploiting several tasks thus provides better estimates than intervals based on the first task. However, as s grows, this effect fades gradually and when $s = 100$ it is counterbalanced by high variance in the multi-task lengths.

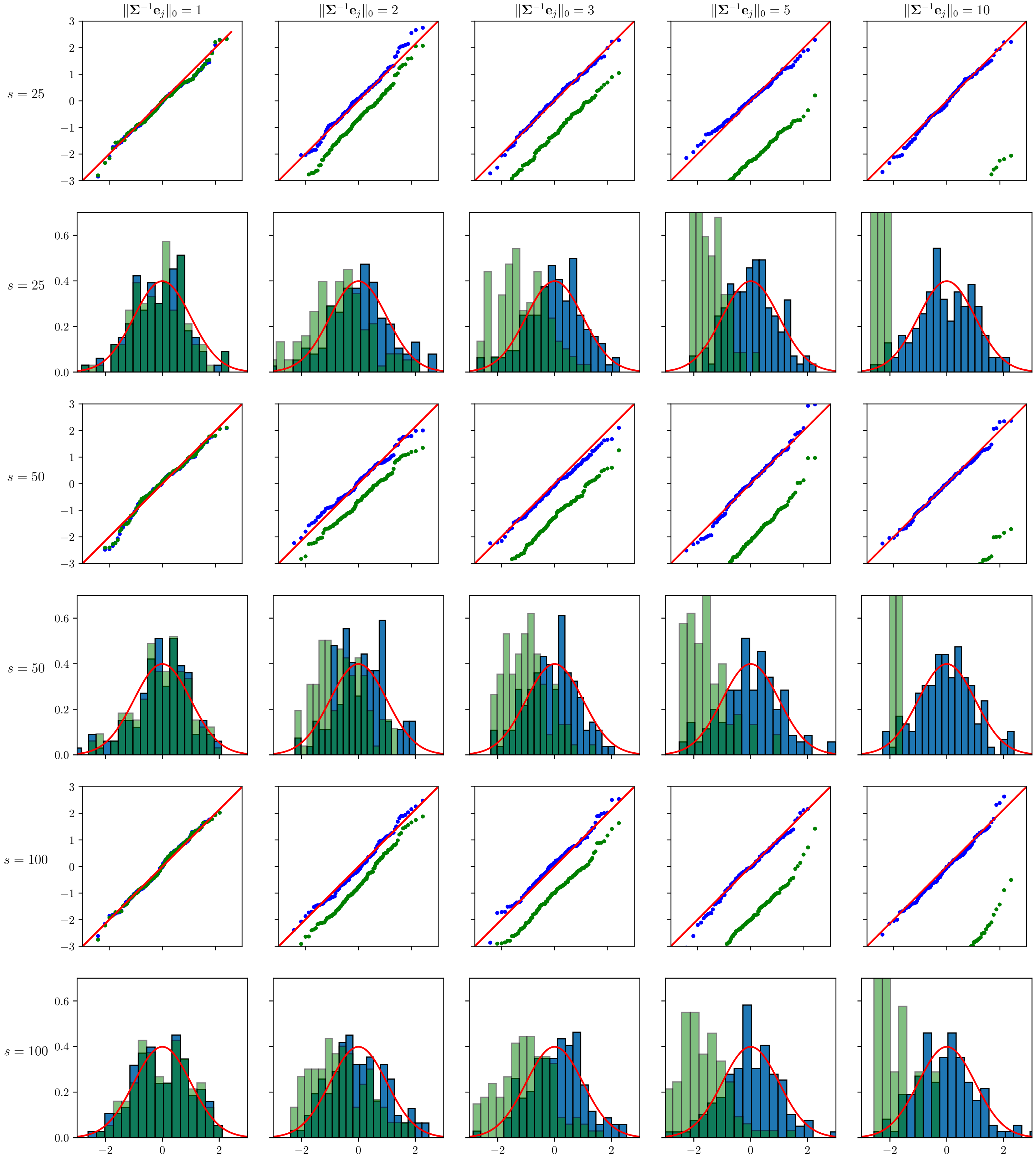

Figure 3.1: QQ-plots and histograms in the unfavorable setting (i) for pivotal quantities in Theorem 3.3 (blue), Theorem 3.5 (green).

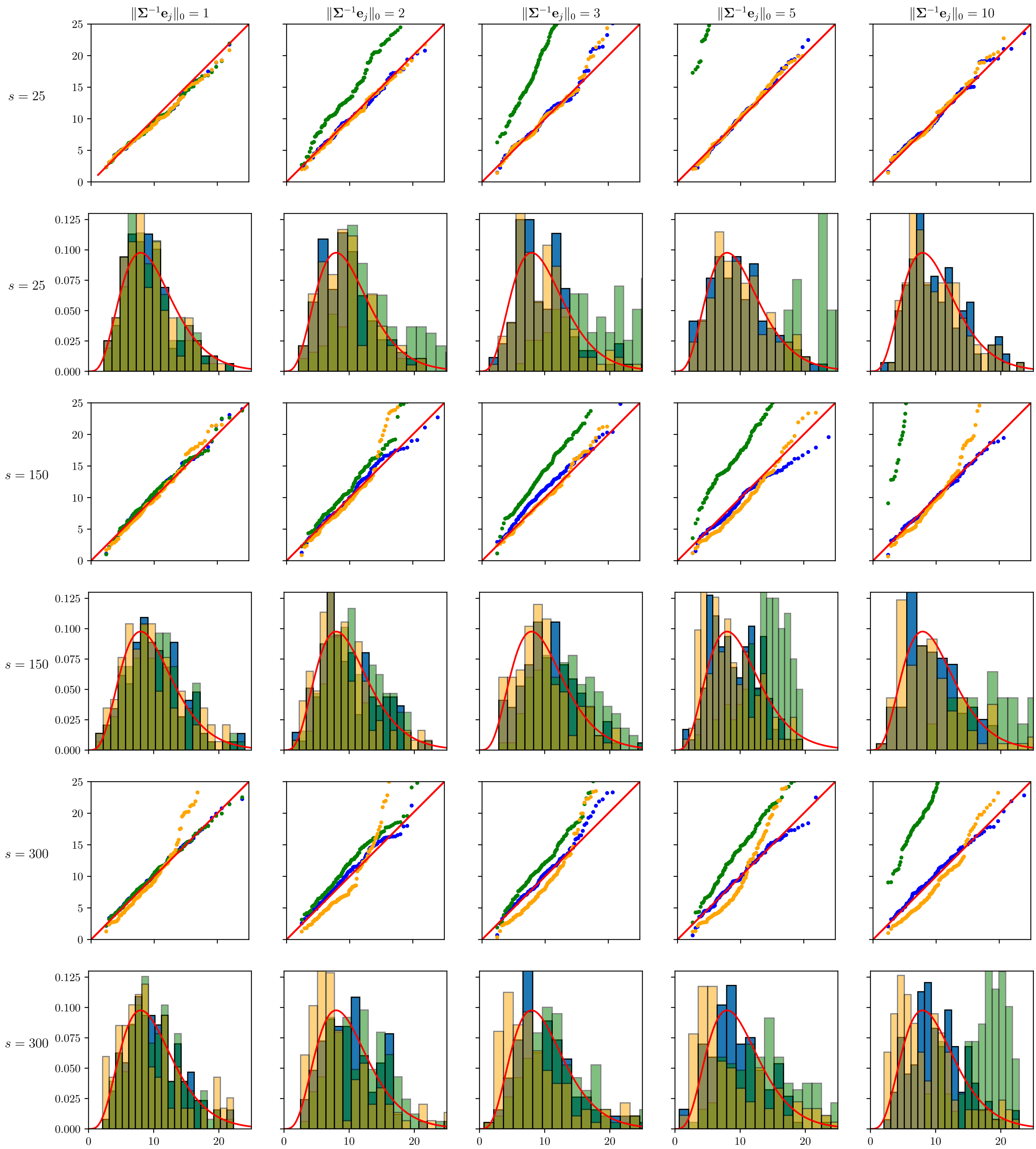

Figure 3.2: QQ-plots and histograms in the unfavorable setting (i) for pivotal quantities in Theorem 3.6 (blue), Theorem 3.8 (green), Theorem 3.10 (orange).

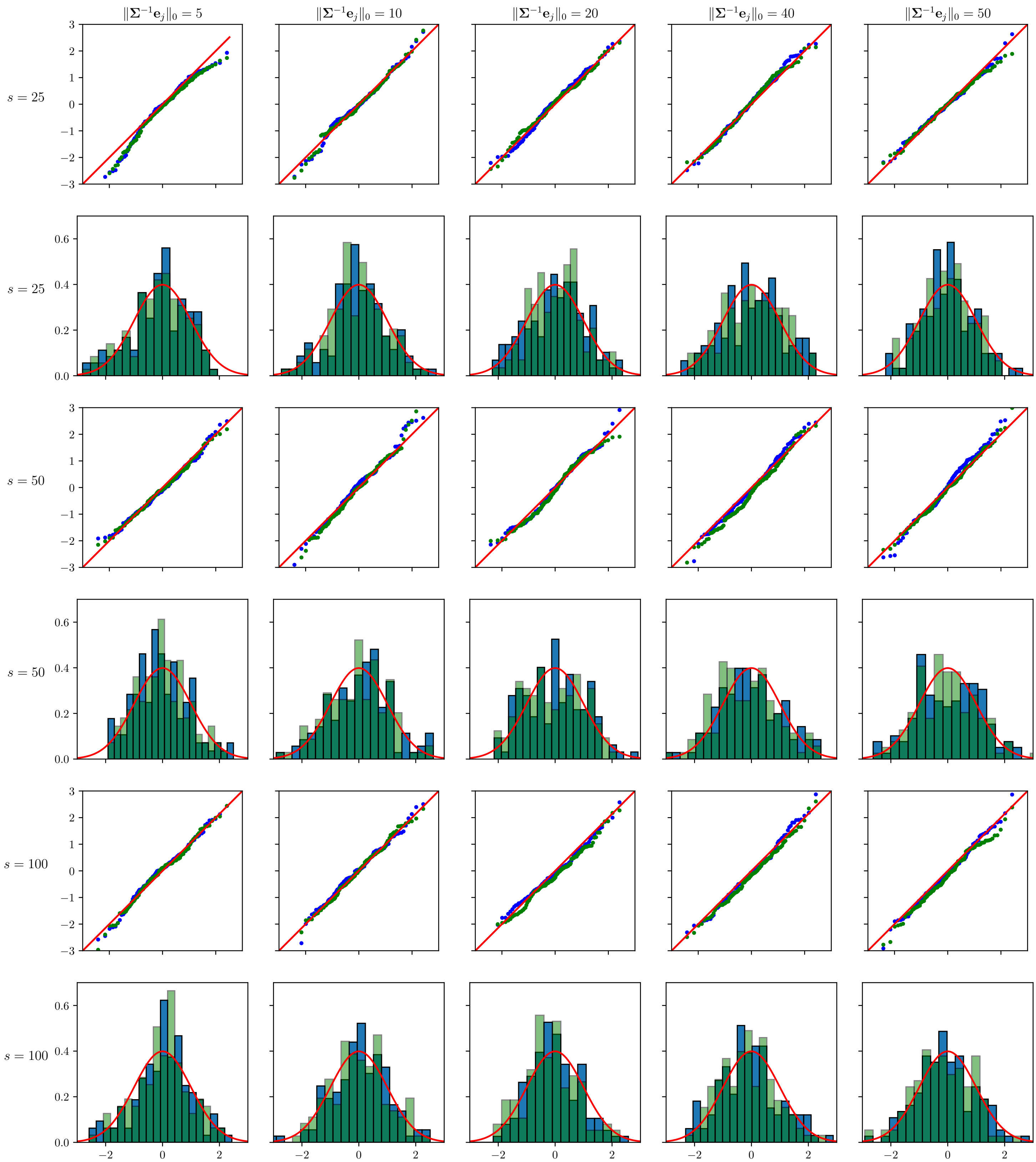

Figure 3.3: QQ-plots and histograms in the favorable setting (ii) for pivotal quantities in Theorem 3.3 (blue), Theorem 3.5 (green).

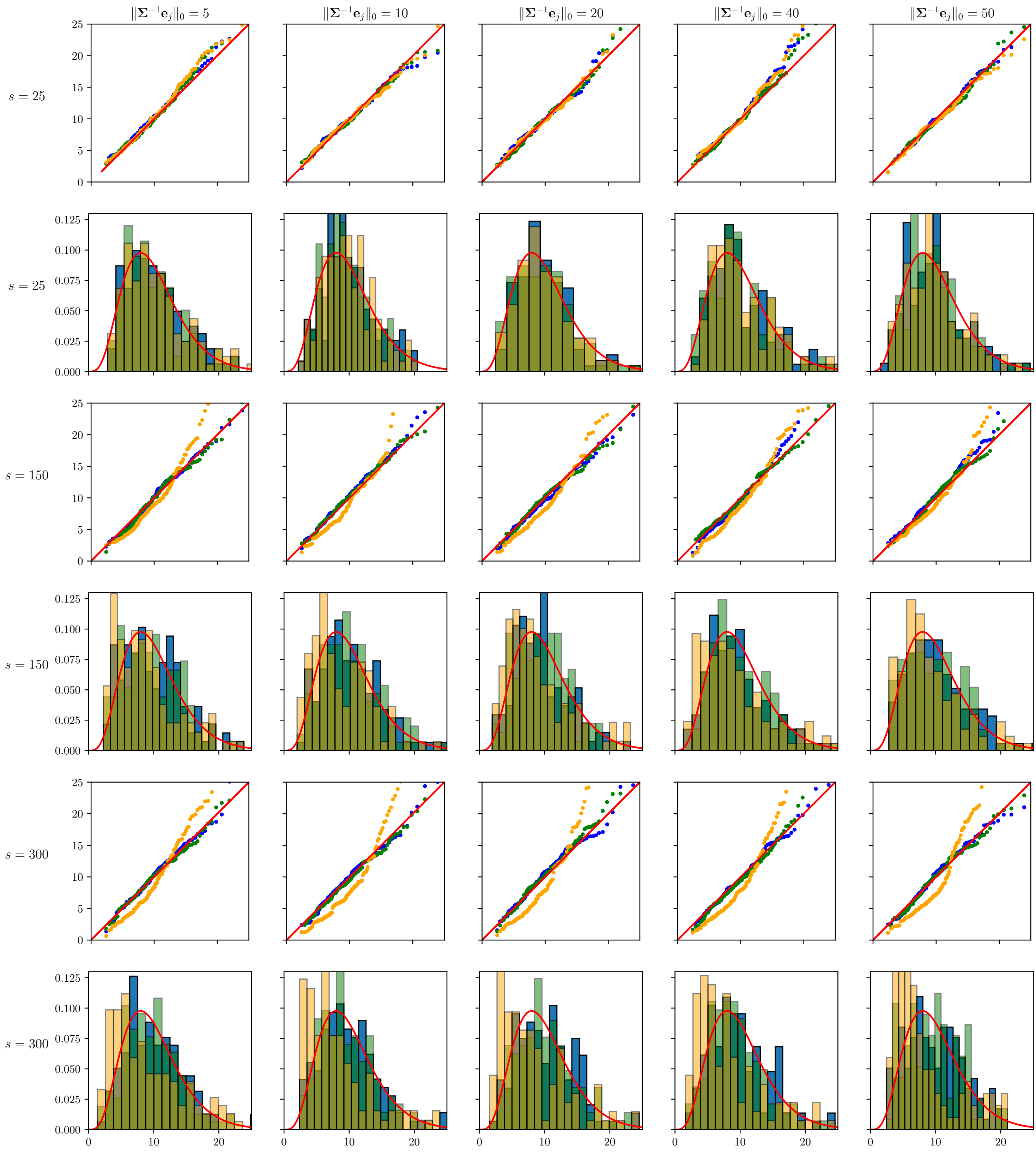

Figure 3.4: QQ-plots and histograms in the favorable setting (ii) for pivotal quantities in Theorem 3.6 (blue), Theorem 3.8 (green), Theorem 3.10 (orange).

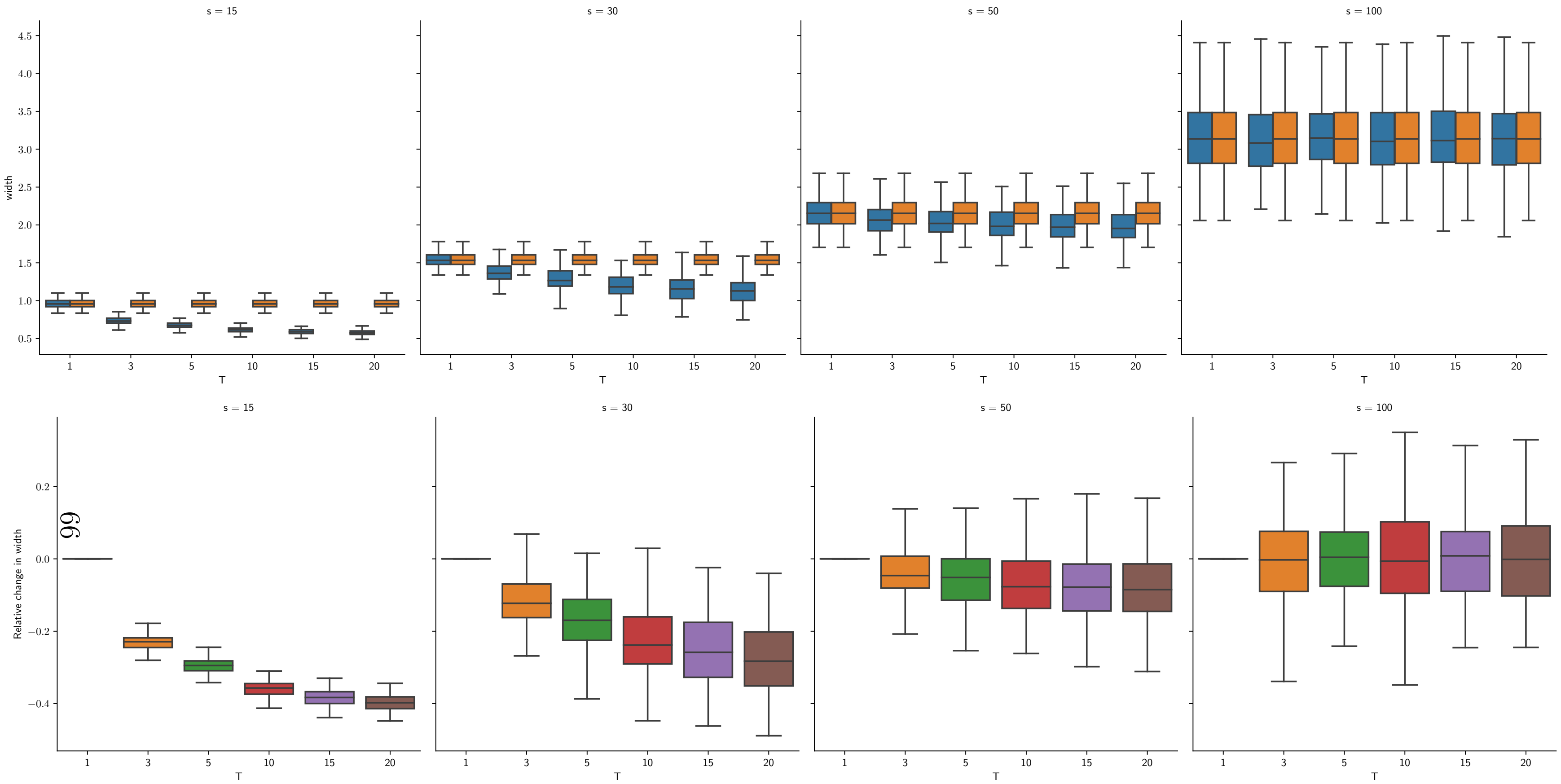

Figure 3.5: Top: boxplots for lengths of 95% confidence intervals using multi-task Lasso (blue) and single-task Lasso on the first task (orange).
Bottom: boxplots for relative change in length with single-task as reference. Only the no-overlap setting (ii) is shown and $\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_1\|_0$ is set to 5.

SUPPLEMENT

3.7 Intuition

Let us give some rationale behind the pivotal quantities stated in the main theorems. In this paragraph and only in this paragraph for the sake of providing some intuition, we assume that $\boldsymbol{a} = \boldsymbol{e}_j$ for some canonical basis vector in $\mathbb{R}^p$ and that $\boldsymbol{\Sigma} = \boldsymbol{I}_{p\times p}$ so that entries of $\boldsymbol{X}$ are i.i.d. $\mathcal{N}(0,1)$. In this setting, the random vector $\boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{e}_j$ has i.i.d. $\mathcal{N}(0,1)$ entries and is independent of $\boldsymbol{X}_{-j}$, the matrix $\boldsymbol{X}$ with j-th column removed. Since $\boldsymbol{z} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_{n\times n})$, Stein's formula [240, 241] states that $\mathbb{E}[\boldsymbol{z}_0^\top \mathbf{f}(\boldsymbol{z}_0)] = \mathbb{E}[\sum_{i=1}^n (\partial/\partial z_{0i}) f_i(\boldsymbol{z}_0)]$ for any differentiable vector field $\mathbf{f} = (f_1, ..., f_n)$ with $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n$, under integrability conditions. For the sake of the current informal argument, assume that Stein's formula provides reasonable approximation. Then applying Stein's formula to $\mathbf{f} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_t$ for each task $t = 1, ..., T$ (here, $\boldsymbol{e}_t$ is the t-th canonical basis vector in $\mathbb{R}^T$), by nontrivial computations that are made rigorous in the proofs given in the supplement, the approximations

$$
\begin{aligned}
\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_1 &\approx n\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_1 - \sum\nolimits_{t=1}^T \widehat{\mathbf{A}}_{1t}\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t,\\
\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_2 &\approx n\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_2 - \sum\nolimits_{t=1}^T \widehat{\mathbf{A}}_{2t}\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t,\\
&\vdots\\
\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_T &\approx n\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_T - \sum\nolimits_{t=1}^T \widehat{\mathbf{A}}_{Tt}\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t
\end{aligned}
\tag{3.50}
$$

hold up to smaller order terms, where $\widehat{\mathbf{A}}$ is the interaction matrix in Equation (3.18). By viewing (3.50) as a linear system with T equations and the T unknowns $(\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t)_{t=1,...,T}$, and assuming that solving the linear system maintains the approximations, we obtain that

$$
\begin{pmatrix}
\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_1\\
\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_2\\
\vdots\\
\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_T
\end{pmatrix}
\approx \left(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}\right)^{-1}
\begin{pmatrix}
\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_1\\
\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_2\\
\vdots\\
\boldsymbol{z}_0^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})\boldsymbol{e}_T
\end{pmatrix}
$$

or equivalently $(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top\boldsymbol{a} = \left(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}\right)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{z}_0$. Thus the matrix product of $(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})^{-1}$ times the residuals projected onto $\boldsymbol{z}_0$ provides us with estimates of the bias of $\widehat{\boldsymbol{B}}$ on the direction $\boldsymbol{a} \in \mathbb{R}^p$. This informal argument is the crux of the rigorous methodology developed in the next subsections. In the sequel, we drop the assumption that $\boldsymbol{\Sigma} = \boldsymbol{I}_{p\times p}$. When $\boldsymbol{\Sigma} \neq \boldsymbol{I}_{p\times p}$ is known as in Section 3.3.1, the score vector $\boldsymbol{z}_0$ in (3.50) has to be replaced by a random vector proportional to $\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$. When $\boldsymbol{\Sigma}$ is unknown as in Section 3.3.2, the score vector has to be estimated.

3.8 Proof of Proposition 3.2

We restate the proposition for convenience.

Proposition 3.2. *Let* $\widehat{\mathbf{A}}$ *be defined by* (3.18). *Then*
(i) $\widehat{\mathbf{A}}$ *is symmetric and positive semi-definite.*
(ii) If $\boldsymbol{X}_{\hat{S}}$ *is rank* $|\hat{S}|$ *then the spectral norm of* $\widehat{\mathbf{A}}$ *is bounded from above as* $\|\widehat{\mathbf{A}}\|_{op} \leq |\hat{S}|$.
(iii) If $\boldsymbol{X}_{\hat{S}}$ *is rank* $|\hat{S}|$ *and* $|\hat{S}|/n < 1$ *then* $\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n$ *is positive-definite and* $\|\boldsymbol{I}_{T\times T} - (\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op} \leq (|\hat{S}|/n)/(1 - |\hat{S}|/n)$.

Proof. (i) We have the following equalities:

$$\begin{aligned}
\boldsymbol{u}^\top \widehat{\mathbf{A}} \boldsymbol{v} &\overset{(i)}{=} \operatorname{Tr}\left[(\boldsymbol{u}^\top \otimes \boldsymbol{X}_{\hat{S}})[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger (\boldsymbol{v} \otimes (\boldsymbol{X}_{\hat{S}})^\top)\right] \\
&\overset{(ii)}{=} \operatorname{Tr}\left[(\boldsymbol{v}\boldsymbol{u}^\top \otimes (\boldsymbol{X}_{\hat{S}})^\top \boldsymbol{X}_{\hat{S}})[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger\right] \\
&\overset{(iii)}{=} \operatorname{Tr}\left[[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^{\dagger^\top} (\boldsymbol{v}\boldsymbol{u}^\top \otimes (\boldsymbol{X}_{\hat{S}})^\top \boldsymbol{X}_{\hat{S}})^\top\right] \\
&\overset{(iv)}{=} \operatorname{Tr}\left[[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger (\boldsymbol{u}\boldsymbol{v}^\top \otimes (\boldsymbol{X}_{\hat{S}})^\top \boldsymbol{X}_{\hat{S}})\right] \\
&\overset{(v)}{=} \operatorname{Tr}\left[(\boldsymbol{u}\boldsymbol{v}^\top \otimes (\boldsymbol{X}_{\hat{S}})^\top \boldsymbol{X}_{\hat{S}})[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger\right] = \boldsymbol{v}^\top \widehat{\mathbf{A}} \boldsymbol{u}
\end{aligned}$$

where (i) follows from (3.19), (ii) is a consequence of $\operatorname{Tr}[\boldsymbol{M}_1\boldsymbol{M}_2] = \operatorname{Tr}[\boldsymbol{M}_2\boldsymbol{M}_1]$ and the mixed product property (3.13), (iii) and (v) follow from $\operatorname{Tr}[\boldsymbol{M}] = \operatorname{Tr}[\boldsymbol{M}^\top]$, (iv) holds because the pseudoinverse preserves symmetry.

This proves that $\widehat{\mathbf{A}}$ is symmetric. Since the pseudoinverse of a positive semi-definite matrix is positive semi-definite as well, we also have

$$\boldsymbol{u}^\top \widehat{\mathbf{A}} \boldsymbol{u} = \|([\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger)^{1/2} (\boldsymbol{u} \otimes \boldsymbol{X}_{\hat{S}}^\top)\|_F^2 \geq 0 \tag{3.51}$$

so that $\widehat{\mathbf{A}}$ is positive semi-definite.

(ii) Recall that $\tilde{\boldsymbol{X}} = \boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{\hat{S}}$. By properties of Gram matrices, $\operatorname{rank}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) = \operatorname{rank}(\tilde{\boldsymbol{X}}) = |\hat{S}|T$, hence by the rank-nullity theorem, $\ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})$ has dimension $(p - |\hat{S}|)T$. By definition of $\boldsymbol{X}_{\hat{S}}$, each vector $\boldsymbol{e}_t \otimes \boldsymbol{e}_j$ is in the kernel of $\tilde{\boldsymbol{X}}$ for $j \notin \hat{S}$ and $t \in [T]$, hence in $\ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})$. These $(p - |\hat{S}|)T$ vectors are linearly independent, so they form a basis of $\ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})$.
Besides, since $\tilde{\boldsymbol{H}} = \sum_{k\in\hat{S}} \boldsymbol{H}^{(k)} \otimes (\boldsymbol{e}_k \boldsymbol{e}_k^\top)$, the mixed product property of Kronecker products (3.13) implies that $\tilde{\boldsymbol{H}}(\boldsymbol{e}_t \otimes \boldsymbol{e}_j) = \mathbf{0}$ for $j \notin \hat{S}$ and $t \in [T]$, hence $\ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) \subset \ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})$. Since these matrices are positive semi-definite, it is easy to check that the reverse inclusion holds as well, so that $\ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}) = \ker(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})$.

Since $\tilde{\boldsymbol{H}}$ is positive semi-definite, $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \preceq \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}$ holds in the sense of the positive semi-definite order, and

$$(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})^\dagger \preceq (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^\dagger \tag{3.52}$$

holds because the two matrices have the same kernel, see [143]. Next, using (3.51),

$$\begin{aligned}
\boldsymbol{u}^\top \widehat{\mathbf{A}} \boldsymbol{u} &= \|([\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}]^\dagger)^{1/2} (\boldsymbol{u} \otimes (\boldsymbol{X}_{\hat{S}})^\top)\|_F^2 + \operatorname{Tr}[(\boldsymbol{u}^\top \otimes \boldsymbol{X}_{\hat{S}})\{[\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}]^\dagger - [\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}]^\dagger\}(\boldsymbol{u} \otimes (\boldsymbol{X}_{\hat{S}})^\top)] \\
&\leq \|([\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}]^\dagger)^{1/2} (\boldsymbol{u} \otimes (\boldsymbol{X}_{\hat{S}})^\top)\|_F^2 \\
&= \operatorname{Tr}[(\boldsymbol{u}^\top \otimes \boldsymbol{X}_{\hat{S}})(\boldsymbol{I}_{T\times T} \otimes (\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}})^\dagger)(\boldsymbol{u} \otimes (\boldsymbol{X}_{\hat{S}})^\top)] \\
&= (\boldsymbol{u}^\top \boldsymbol{u}) \operatorname{Tr}[\boldsymbol{X}_{\hat{S}} (\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}})^\dagger (\boldsymbol{X}_{\hat{S}})^\top] \\
&= \|\boldsymbol{u}\|^2 |\hat{S}|,
\end{aligned}$$

where the first inequality follows from (3.52) and the third and fourth line follow respectively from $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \boldsymbol{I}_{T\times T} \otimes \boldsymbol{X}_{\hat{S}}$ and the mixed product property (3.13). The last line stems from the fact that $\boldsymbol{X}_{\hat{S}}(\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}})^\dagger \boldsymbol{X}_{\hat{S}}^\top$ is a projection matrix of rank $|\hat{S}|$ when $\mathrm{rank}(\boldsymbol{X}_{\hat{S}}) = |\hat{S}|$.

(iii) Since $\boldsymbol{X}_{\hat{S}}$ has rank $|\hat{S}|$, we have by (ii) that $\|\widehat{\mathbf{A}}\|_{op} \leq |\hat{S}| < n$. Since $\widehat{\mathbf{A}}$ is positive-semi definite, its spectral norm is its largest eigenvalue, hence all the eigenvalues of $\widehat{\mathbf{A}}/n$ are < 1, and $\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n$ is positive definite. For any $\boldsymbol{M} \in \mathbb{R}^{T\times T}$ with $\|\boldsymbol{M}\|_{op} < 1$ we have $(\boldsymbol{I}_{T\times T} - \boldsymbol{M})^{-1} = \sum_{k=0}^{\infty} \boldsymbol{M}^k$. By the triangle inequality and the submultiplicativity of the operator norm,

$$\|(\boldsymbol{I}_{T\times T} - \boldsymbol{M})^{-1} - \boldsymbol{I}_{T\times T}\|_{op} \leq \|\boldsymbol{M}\|_{op} \sum_{k=1}^{\infty} \|\boldsymbol{M}\|_{op}^{k-1} = \|\boldsymbol{M}\|_{op}/(1 - \|\boldsymbol{M}\|_{op}).$$

□

3.9 Preliminaries

In this section we develop a series of technical lemmas that will be used for proving Sections 3.3 and 3.4. We consider model (3.2) and the estimator $\widehat{\boldsymbol{B}}$ from (3.4). Let $\eta_1 > 0$, $\eta_2 \geq 2$, $\eta_3, \eta_4 \in (0,1)$ and set λ, λ_0 as in (3.8). Define the sparsity level

$$\bar{s} = s\Big(1/T + \frac{4\|\boldsymbol{\Sigma}\|_{op}(1+\eta_4)^2}{\kappa^2}(2 + \eta_2 + 1/\sqrt{T})^2\Big)\frac{2}{(\lambda/\lambda_0 - 1)^2} \tag{3.53}$$

and note that $\bar{s}$ is of the same order as s when the spectrum of $\boldsymbol{\Sigma}$ is bounded away from 0 and infinity as in Assumption (A1). Let $\mathcal{C} = \{\boldsymbol{U} \in \mathbb{R}^{p\times T} : \|\boldsymbol{U}\|_{2,1} \leq 3\sqrt{\bar{s}}\|\boldsymbol{U}\|_F\}$, $\kappa = (1-\eta_3)\phi_{\min}(\boldsymbol{\Sigma})^{1/2}$ and define the events

$$\Omega_1 = \Big\{\max_{\boldsymbol{U}\in\mathcal{C},\boldsymbol{U}\neq 0}\Big|\frac{\|\boldsymbol{X}\boldsymbol{U}\|_F}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{U}\|_F\sqrt{n}} - 1\Big| < \eta_3\Big\}, \quad \Omega_2 = \Big\{\sum_{j=1}^{p}(\|\boldsymbol{E}^T\boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+^2 < sn^2T\lambda_0^2\Big\},$$

$$\Omega_3 = \Big\{\max_{B\subset[p]:|B|\leq s+2\bar{s}+1}\Big(\max_{\boldsymbol{v}\in\mathbb{R}^p:\mathrm{supp}(\boldsymbol{v})\subset B}\Big|\frac{\|\boldsymbol{X}\boldsymbol{v}\|_2}{\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2} - 1\Big|\Big) < \eta_4\Big\}, \quad \Omega_4 = \Big\{\|\boldsymbol{E}\|_{op} < \sigma(2\sqrt{n} + \sqrt{T})\Big\}$$

as well as

$$\Omega_* = \Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_4. \tag{3.54}$$

Since the only randomness is with respect to $(\boldsymbol{X}, \boldsymbol{E})$, we view the underlying probability space as $\Omega = (\mathbb{R}^{n\times p}) \times (\mathbb{R}^{n\times T})$ and $\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_*$ as subsets of Ω so that Ω_i occurs if and only if $(\boldsymbol{X}, \boldsymbol{E}) \in \Omega_i$ for each $i = 1, 2, 3$.

Lemma 3.12. *Let Assumption (A1) be fulfilled. Then* $\mathbb{P}(\Omega_*) \to 1$.

Lemma 3.13. *On* Ω_* *we have:*

(i) $\widehat{\boldsymbol{B}} - \boldsymbol{B}^* \in \mathcal{C}$,

(ii) $n^{-1/2}\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \le (1-\eta_3)\bar{R}$,

(iii) $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \le \bar{R}$,

(iv) $\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_{2,1} \le 3\sqrt{s}\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_F \le 3\sqrt{s}\phi_{\min}(\boldsymbol{\Sigma})^{-1/2}\bar{R}$,

(v) $\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_F^2 \le 8\sigma^2 nT + 2(1-\eta_3)^2 n\bar{R}^2$,

where

$$\bar{R} \coloneqq (1-\eta_3)^{-1}\kappa^{-1}2(1+\eta_1)(3+\eta_2)\sigma \max_j \boldsymbol{\Sigma}_{jj}^{1/2}\sqrt{sT/n}\left(1+\sqrt{(2/T)\log(p/s)}\right).$$

Moreover, $\bar{R} \xrightarrow[n\to\infty]{} 0$ *under Assumption (A1).*

Lemma 3.14. *On* Ω_**, inequality* $|\hat{S}| \le \bar{s}$ *holds with* $\bar{s}$ *in* (3.53).

Lemma 3.15. *On* Ω_* *we have* $\operatorname{rank}(\boldsymbol{X}_{\hat{S}}) = |\hat{S}|$.

Lemma 3.16. *For almost every* $(\boldsymbol{X},\boldsymbol{E})$*, the KKT conditions of* $\widehat{\boldsymbol{B}}$ *in* (3.4) *hold strictly in the sense that* $\mathbb{P}(\max_{j\notin\hat{S}}\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{X}\boldsymbol{e}_j\|_2 < nT\lambda) = 1$.

Lemma 3.17. *Given the noise matrix* $\boldsymbol{E}$ *and two design matrices* $\boldsymbol{X},\overline{\boldsymbol{X}}$ *define* $\widehat{\boldsymbol{B}}$ *in* (3.4) *and* $\overline{\boldsymbol{B}}$ *by*

$$\overline{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\left(\tfrac{1}{2nT}\|\boldsymbol{E}+\overline{\boldsymbol{X}}(\boldsymbol{B}^*-\boldsymbol{B})\|_F^2 + \lambda\|\boldsymbol{B}\|_{2,1}\right).$$

If $\boldsymbol{X},\overline{\boldsymbol{X}},\boldsymbol{E}$ *are such that both* $\{(\boldsymbol{X},\boldsymbol{E}),(\overline{\boldsymbol{X}},\boldsymbol{E})\}\subset\Omega_*$ *then*

$$n^{1/2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F \le C_1(\eta_4)(\bar{R}+\|\boldsymbol{E}\|_{op}n^{-1/2})\|(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F,$$
$$\|\overline{\boldsymbol{X}}(\overline{\boldsymbol{B}}-\boldsymbol{B}^*)-\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \le C_2(\eta_4)(\bar{R}+\|\boldsymbol{E}\|_{op}n^{-1/2})\|(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F$$

for some constants that depend on η_4 *only and* $\bar{R}$ *is defined in Lemma 3.13.*

Lemma 3.18. *For almost every* $(\boldsymbol{X},\boldsymbol{E})$ *in the open set* $\Omega_1\cap\Omega_2\cap\Omega_3$*,* $\widehat{\boldsymbol{B}}$ *is a Fréchet differentiable function of* $\boldsymbol{X}$*. For almost every* $(\boldsymbol{X},\boldsymbol{E})$ *in* $\Omega_1\cap\Omega_2\cap\Omega_3$*, if*

$$\widehat{\boldsymbol{B}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\left(\tfrac{1}{2nT}\|\boldsymbol{E}+(\boldsymbol{X}+\boldsymbol{w}\boldsymbol{a}^\top)(\boldsymbol{B}^*-\boldsymbol{B})\|_F^2 + \lambda\|\boldsymbol{B}\|_{2,1}\right)$$

is the estimate (3.4) *with* $\boldsymbol{X}$ *replaced by the perturbed design* $\boldsymbol{X}+\boldsymbol{w}\boldsymbol{a}^\top$*, then for any* $\boldsymbol{b}\in\mathbb{R}^T$

$$\Big((\boldsymbol{X}+\boldsymbol{w}\boldsymbol{a}^\top)(\widehat{\boldsymbol{B}}(\boldsymbol{w})-\boldsymbol{B}^*)\Big)\boldsymbol{b} - \Big(\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\Big)\boldsymbol{b} = \big(\mathsf{D}(\boldsymbol{b})\big)\boldsymbol{w} + o(\|\boldsymbol{w}\|)$$

as $\|\boldsymbol{w}\| \to 0$*, where* $\mathsf{D} : \mathbb{R}^T \to \mathbb{R}^{n\times n}$ *is a linear map given by* $\mathsf{D}(\boldsymbol{b}) = \mathsf{D}^*(\boldsymbol{b}) + \mathsf{D}^{**}(\boldsymbol{b})$ *with*

$$
\begin{aligned}
\mathsf{D}^*(\boldsymbol{b}) &= (\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b})\boldsymbol{I}_{n\times n} - (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\big)^\dagger\big(((\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top\boldsymbol{a}) \otimes \boldsymbol{X}_{\hat{S}}^\top\big)\\
&= (\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b})\boldsymbol{I}_{n\times n} - (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\right)^\dagger\begin{pmatrix}\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_1\boldsymbol{X}_{\hat{S}}^\top\\ \vdots\\ \boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_T\boldsymbol{X}_{\hat{S}}^\top\end{pmatrix},\\
\mathsf{D}^{**}(\boldsymbol{b}) &= (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\big)^\dagger((\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \otimes \boldsymbol{a}_{\hat{S}})\\
&= (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\right)^\dagger\begin{pmatrix}\boldsymbol{a}_{\hat{S}}\boldsymbol{e}_1^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\\ \vdots\\ \boldsymbol{a}_{\hat{S}}\boldsymbol{e}_T^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\end{pmatrix}
\end{aligned}
$$

for all $\boldsymbol{b} \in \mathbb{R}^T$ *and* $\boldsymbol{w} \in \mathbb{R}^n$*. Note that* D, D^* *and* D^{**} *implicitly depend on* $(\boldsymbol{X}, \boldsymbol{E})$*. Hence the matrix* $\mathsf{D}(\boldsymbol{b})$ *of size* $n\times n$ *is the Jacobian of the map* $\boldsymbol{w} \mapsto (\boldsymbol{X}+\boldsymbol{w}\boldsymbol{a}^\top)(\widehat{\boldsymbol{B}}(\boldsymbol{w})-\boldsymbol{B}^*)\boldsymbol{b}$ *at* $\boldsymbol{w} = \boldsymbol{0}$*.*

Lemma 3.19. *For any* $\boldsymbol{b} \in \mathbb{R}^T$ *we have on* Ω_*

$$
\operatorname{Tr}[\mathsf{D}^*(\boldsymbol{b})] = \boldsymbol{b}^\top(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top\boldsymbol{a}, \tag{3.55}
$$

$$
\sum_{t=1}^T \left(\operatorname{Tr}[\mathsf{D}^{**}(\boldsymbol{e}_t)]\right)^2 \le C_3(\boldsymbol{\Sigma})\sigma^2 sT \tag{3.56}
$$

for some constant depending on $\boldsymbol{\Sigma}$ *and* $\eta_1, ..., \eta_4$ *only.*

Lemma 3.20. *Under Assumption (A1), as* $n, p \to +\infty$ *we have*

$$
\frac{1}{\sigma^2 n}\mathbb{E}\Big[I\{\Omega_*\}\sum_{t=1}^T\Big(\boldsymbol{z}_0^\top\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t - \operatorname{Tr}[\mathsf{D}(\boldsymbol{e}_t)]\Big)^2\Big] \longrightarrow 0.
$$

Since Ω_* *has probability approaching one, this implies that* $\frac{1}{\sigma^2 n}\sum_{t=1}^T(\boldsymbol{z}_0^\top\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\boldsymbol{e}_t - \operatorname{Tr}[\mathsf{D}(\boldsymbol{e}_t)])^2$ *converges to 0 in probability.*

We now prove each lemma. The lemmas are restated before their proofs for convenience.

Lemma 3.12. *Let Assumption (A1) be fulfilled. Then* $\mathbb{P}(\Omega_*) \to 1$.

Proof of Lemma 3.12. The fact that $\Omega_* = \Omega_1\cap\Omega_2\cap\Omega_3\cap\Omega_4$ has probability approaching one under Assumption (A1) follows from the propositions in Section 3.10: Proposition 3.21 *(iii)* with $k = 9s$ and $x = \log n$, Proposition 3.22, Proposition 3.23 applied with $k = s + 2\bar{s} + 1$, and $\mathbb{P}(\Omega_4) \ge 1 - e^{-n/2}$ by [71, Theorem II.13]. □

Lemma 3.13. *On* Ω_* *we have:*

(i) $\widehat{\boldsymbol{B}} - \boldsymbol{B}^* \in \mathcal{C}$,

(ii) $n^{-1/2}\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F \leq (1-\eta_3)\bar{R}$,

(iii) $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F \leq \bar{R}$,

(iv) $\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1} \leq 3\sqrt{s}\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F \leq 3\sqrt{s}\phi_{\min}(\boldsymbol{\Sigma})^{-1/2}\bar{R}$,

(v) $\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\|_F^2 \leq 8\sigma^2 nT + 2(1-\eta_3)^2 n\bar{R}^2$,

where

$$\bar{R} := (1-\eta_3)^{-1}\kappa^{-1}2(1+\eta_1)(3+\eta_2)\sigma \max_j \boldsymbol{\Sigma}_{jj}^{1/2}\sqrt{sT/n}\left(1 + \sqrt{(2/T)\log(p/s)}\right).$$

Moreover, $\bar{R} \xrightarrow[n\to\infty]{} 0$ *under Assumption (A1).*

Proof of Lemma 3.13. In the whole proof we place ourselves on the event Ω_*. We prove first that $\widehat{\boldsymbol{B}} - \boldsymbol{B}^* \in \mathcal{C}$.
By the definition of $\widehat{\boldsymbol{B}}$, $\frac{1}{nT}\|\boldsymbol{X}\widehat{\boldsymbol{B}} - \boldsymbol{Y}\|_F^2 + 2\lambda\|\widehat{\boldsymbol{B}}\|_{2,1} \leq \frac{1}{nT}\|\boldsymbol{X}\boldsymbol{B}^* - \boldsymbol{Y}\|_F^2 + 2\lambda\|\boldsymbol{B}^*\|_{2,1}$.
Rewriting the LHS as $\frac{1}{nT}\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) + (\boldsymbol{X}\boldsymbol{B}^* - \boldsymbol{Y})\|_F^2 + 2\lambda\|\widehat{\boldsymbol{B}}\|_{2,1}$ and expanding the square yields $\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F^2 \leq 2\langle \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*), \boldsymbol{E}\rangle_F + 2nT\lambda(\|\boldsymbol{B}^*\|_{2,1} - \|\widehat{\boldsymbol{B}}\|_{2,1})$.
The following chain of inequalities holds

$$\begin{aligned}
\langle \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*), \boldsymbol{E}\rangle_F &\overset{(i)}{\leq} \sum_{j=1}^p \|(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j\|_2 \cdot \|(\boldsymbol{X}^T\boldsymbol{E})^\top \boldsymbol{e}_j\|_2 \\
&\overset{(ii)}{\leq} \sum_{j=1}^p \|(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j\|_2 \left[(\|\boldsymbol{E}^T\boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+ + nT\lambda_0\right] \\
&\overset{(iii)}{\leq} \|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F \left(\sum_{j=1}^p (\|\boldsymbol{E}^T\boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+^2\right)^{1/2} + nT\lambda_0\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1} \\
&\overset{(iv)}{\leq} \|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F \sqrt{s} n\sqrt{T}\lambda_0 + nT\lambda_0\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1}.
\end{aligned}$$

(i) and (iii) follow from Cauchy-Schwarz inequality, (ii) stems from the inequality $a \leq (a-b)_+ + b$ and (iv) holds on Ω_2. Thus

$$\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F^2 \leq 2\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F\sqrt{s}nT\lambda_0 + 2nT\left[\lambda_0\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1} + \lambda(\|\boldsymbol{B}^*\|_{2,1} - \|\widehat{\boldsymbol{B}}\|_{2,1})\right]. \tag{3.57}$$

Besides, the quantity inside the bracket on the right hand side satisfies

$$\begin{aligned}
&\lambda_0\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1} + \lambda(\|\boldsymbol{B}^*\|_{2,1} - \|\widehat{\boldsymbol{B}}\|_{2,1}) \\
&\overset{(i)}{=} \lambda_0 \sum_{j\in S} \|(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{e}_j\|_2 + \lambda_0 \sum_{j\notin S} \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2 + \lambda\Big(\sum_{j\in S} \|\boldsymbol{B}^{*\top}\boldsymbol{e}_j\|_2 - \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2\Big) - \lambda \sum_{j\notin S} \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2 \\
&\overset{(ii)}{\leq} \lambda_0\sqrt{s}\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F + \lambda_0 \sum_{j\notin S} \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2 + \lambda \sum_{j\in S} \|(\boldsymbol{B}^* - \widehat{\boldsymbol{B}})^\top \boldsymbol{e}_j\|_2 - \lambda \sum_{j\notin S} \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2 \\
&\overset{(iii)}{\leq} (\lambda_0 + \lambda)\sqrt{s}\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_F + (\lambda_0 - \lambda)\sum_{j\notin S} \|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2,
\end{aligned}$$

where (ii) follows from Cauchy-Schwarz and the reverse triangle inequality applied respectively on the first and third summands of (i), whereas (iii) is a consequence of Cauchy-Schwarz. Combining this bound with (3.57) and plugging in the value $\lambda = (1+\eta_2)\lambda_0$ yields

$$\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F^2 \leq 2nT\sqrt{s}\lambda_0\Big(2+\eta_2+\frac{1}{\sqrt{T}}\Big)\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_F - 2nT\eta_2\lambda_0\sum_{j\notin S}\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2. \tag{3.58}$$

Non-negativity of the LHS, the equality $\sum_{j\notin S}\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2 = \|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_{2,1} - \sum_{j\in S}\|(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top \boldsymbol{e}_j\|_2$ and Cauchy-Schwarz lead to $\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_{2,1} \leq (1+\frac{3}{\eta_2}+\frac{1}{\eta_2\sqrt{T}})\sqrt{s}\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_F$. Since $T\geq 1$ and $\eta_2 \geq 2$, we get $\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_{2,1} \leq 3\sqrt{s}\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_F$, that is $\widehat{\boldsymbol{B}}-\boldsymbol{B}^* \in \mathcal{C}$.

The inequality

$$\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_F \leq \|\boldsymbol{\Sigma}^{-1/2}\|_{op}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F = \phi_{\min}(\boldsymbol{\Sigma})^{-1/2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \tag{3.59}$$

combined with (3.58) and the event Ω_1 yields

$$\begin{aligned}\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F &\leq 2\kappa^{-1}\Big(2+\eta_2+T^{-1/2}\Big)\sqrt{n}T\lambda_0\sqrt{s}\\ &\leq 2\kappa^{-1}(1+\eta_1)(3+\eta_2)\sigma\max_j \boldsymbol{\Sigma}_{jj}^{1/2}\sqrt{sT}\Big(1+\sqrt{(2/T)\log(p/s)}\Big)\\ &= \sqrt{n}(1-\eta_3)\bar{R}.\end{aligned}$$

Reusing Ω_1, we obtain $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \leq (\sqrt{n}(1-\eta_3))^{-1}\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \leq \bar{R}$. Combining this last bound with (3.59) yields $\|\widehat{\boldsymbol{B}}-\boldsymbol{B}^*\|_F \leq \phi_{\min}(\boldsymbol{\Sigma})^{-1/2}\bar{R}$, hence (iv). For inequality (v), using Ω_4, $(a+b)^2 \leq 2a^2+2b^2$ and $\|\boldsymbol{E}\|_F^2 \leq \operatorname{rank}(\boldsymbol{E})\|\boldsymbol{E}\|_{op}^2$ we have

$$\begin{aligned}\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_F^2 &\leq 2\|\boldsymbol{E}\|_F^2 + 2\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F^2\\ &\leq 2\operatorname{rank}(\boldsymbol{E})\|\boldsymbol{E}\|_{op}^2 + 2n(1-\eta_3)^2\bar{R}^2\\ &\leq 2\min(n,T)(2\sigma\max(\sqrt{n},\sqrt{T}))^2 + 2n(1-\eta_3)^2\bar{R}^2\\ &\leq 8\sigma^2 nT + 2(1-\eta_3)^2 n\bar{R}^2.\end{aligned}$$

Regarding the limit of $\bar{R}$, note that $\bar{R} \propto \big(\frac{sT}{n}\big)^{1/2} + \big(\frac{s}{n}\log(\frac{p}{s})\big)^{1/2}$. By Assumption (A1), each summand goes to 0 as n goes to ∞. □

Lemma 3.14. *On Ω_*, inequality $|\hat{S}| \leq \bar{s}$ holds with $\bar{s}$ in (3.53).*

Proof of Lemma 3.14. The KKT conditions of (3.4) are given by

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{X}\boldsymbol{e}_j &= nT\lambda\|\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j\|_2^{-1}\widehat{\boldsymbol{B}}^\top \boldsymbol{e}_j \qquad \text{for all } j\in\hat{S},\\ \|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{X}\boldsymbol{e}_j\|_2 &\leq nT\lambda \qquad \text{for all } j\notin\hat{S}.\end{aligned} \tag{3.60}$$

This implies that $\forall j\in\hat{S}$, $\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{X}\boldsymbol{e}_j\|_2 = nT\lambda$. Since $\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{X}\boldsymbol{e}_j\|_2 \leq \|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{e}_j\|_2 + \|(\boldsymbol{X}(\boldsymbol{B}^*-\widehat{\boldsymbol{B}}))^\top \boldsymbol{X}\boldsymbol{e}_j\|_2$ by the triangle inequality, we have for any $j\in\hat{S}$,

$$nT\lambda \leq (\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+ + nT\lambda_0 + \|(\boldsymbol{X}(\boldsymbol{B}^*-\widehat{\boldsymbol{B}}))^\top \boldsymbol{X}\boldsymbol{e}_j\|_2, \tag{3.61}$$

$$nT(\lambda-\lambda_0) \leq (\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+ + \|(\boldsymbol{X}(\boldsymbol{B}^*-\widehat{\boldsymbol{B}}))^\top \boldsymbol{X}\boldsymbol{e}_j\|_2. \tag{3.62}$$

Summing the squares of the above inequalities for a subset $B \subset \hat{S}$ and using $(a+b)^2 \leq 2a^2 + 2b^2$, we get

$$\frac{|B|n^2(\lambda-\lambda_0)^2T^2}{2} \leq \sum_{j\in B}(\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+^2 + \mathrm{Tr}(\{\boldsymbol{X}(\boldsymbol{B}^*-\widehat{\boldsymbol{B}})\}^\top\Big\{\sum_{j\in B}\boldsymbol{X}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{X}^\top\Big\}\{\boldsymbol{X}(\boldsymbol{B}^*-\widehat{\boldsymbol{B}})\}).$$

The first term is bounded from above by $sn^2T\lambda_0^2$ on the event Ω_2. Dividing by n^2T^2 we find

$$\frac{|B|(\lambda-\lambda_0)^2}{2} \leq s\lambda_0^2/T + \frac{1}{nT^2}\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F^2\psi_{\max}(B)$$

where $\psi_{\max}(B)$ is the largest eigenvalue of $\frac{1}{n}\sum_{j\in B}\boldsymbol{X}\boldsymbol{e}_j\boldsymbol{e}_j^\top\boldsymbol{X}^\top$, or equivalently the largest eigenvalue of $\frac{1}{n}(\boldsymbol{X}_B\boldsymbol{X}_B^\top)$, which is also the largest eigenvalue of $\frac{1}{n}(\boldsymbol{X}_B^\top\boldsymbol{X}_B)$. On the event of Ω_*, we obtain

$$\frac{|B|(\lambda-\lambda_0)^2}{2} \leq s\lambda_0^2/T + \frac{\psi_{\max}(B)}{nT^2}\frac{4}{\kappa^2}(2+\eta_2+1/\sqrt{T})^2nT^2\lambda_0^2s,$$

or equivalently

$$\frac{|B|(\lambda/\lambda_0-1)^2}{2} \leq s\Big(1/T + \frac{4\psi_{\max}(B)}{\kappa^2}(2+\eta_2+1/\sqrt{T})^2\Big).$$

Let $\bar{s}$ be as in (3.53) and assume that $|\hat{S}| \leq \bar{s}$ is violated on Ω_*. Then on Ω_3, any $B \subset \hat{S}$ with size $|B| = \lfloor\bar{s}\rfloor + 1$ satisfies $\forall \boldsymbol{v} \in \mathbb{R}^p$, $\|\boldsymbol{X}_B\boldsymbol{v}_B\|_2 \leq (1+\eta_4)\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\|_{op}\|\boldsymbol{v}_B\|_2$. Squaring yields $\psi_{\max}(B) \leq \|\boldsymbol{\Sigma}\|_{op}(1+\eta_4)^2$. Then

$$\frac{|B|(\lambda/\lambda_0-1)^2}{2} \leq s\Big(1/T + \frac{4\|\boldsymbol{\Sigma}\|_{op}(1+\eta_4)^2}{\kappa^2}(2+\eta_2+1/\sqrt{T})^2\Big)$$

which shows that $|B| \leq \bar{s}$ by definition of $\bar{s}$, a contradiction. □

Lemma 3.15. *On* Ω_* *we have* $\mathrm{rank}(\boldsymbol{X}_{\hat{S}}) = |\hat{S}|$.

Proof. By Lemma 3.14, we have $|\hat{S}| \leq \bar{s}$ on Ω_*. Since $s \leq s + 2\bar{s} + 1$, the event Ω_3 yields $\forall \boldsymbol{v} \in \mathbb{R}^p$, $\mathrm{supp}(\boldsymbol{v}) \subset \hat{S} \implies (1-\eta_4)\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2 \leq \|\boldsymbol{X}_{\hat{S}}\boldsymbol{v}\|_2$. If $\boldsymbol{v}$ is such that $\mathrm{supp}(\boldsymbol{v}) \subset \hat{S}$ and $\boldsymbol{X}_{\hat{S}}\boldsymbol{v} = \boldsymbol{0}$, then we must have $\boldsymbol{v} = \boldsymbol{0}$. Equivalently, the linear span of $(\boldsymbol{e}_j)_{j\in\hat{S}}$ has intersection $\{\boldsymbol{0}\}$ with $\ker(\boldsymbol{X}_{\hat{S}})$, hence $\ker(\boldsymbol{X}_{\hat{S}})$ must be contained in the span of $(\boldsymbol{e}_j)_{j\notin\hat{S}}$. Thus $\dim\ker(\boldsymbol{X}_{\hat{S}}) \leq p - |\hat{S}|$ and by the rank-nullity theorem, $\mathrm{rank}(\boldsymbol{X}_{\hat{S}}) \geq |\hat{S}|$. By definition of $\boldsymbol{X}_{\hat{S}}$, it is also clear that $\mathrm{rank}(\boldsymbol{X}_{\hat{S}}) \leq |\hat{S}|$, hence the conclusion. □

Lemma 3.16. *For almost every* $(\boldsymbol{X}, \boldsymbol{E})$*, the KKT conditions of* $\widehat{\boldsymbol{B}}$ *in* (3.4) *hold strictly in the sense that* $\mathbb{P}(\max_{j\notin\hat{S}}\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{X}\boldsymbol{e}_j\|_2 < nT\lambda) = 1$.

Proof of Lemma 3.16. This follows from the argument in Lemma 6.4 of [26, arXiv version v1, 24 Feb 2019]. □

Lemma 3.17. *Given the noise matrix* $\boldsymbol{E}$ *and two design matrices* $\boldsymbol{X}, \overline{\boldsymbol{X}}$ *define* $\widehat{\boldsymbol{B}}$ *in* (3.4) *and* $\overline{\boldsymbol{B}}$ *by*

$$\overline{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\left(\tfrac{1}{2nT}\|\boldsymbol{E}+\overline{\boldsymbol{X}}(\boldsymbol{B}^*-\boldsymbol{B})\|_F^2 + \lambda\|\boldsymbol{B}\|_{2,1}\right).$$

If $\boldsymbol{X}, \overline{\boldsymbol{X}}, \boldsymbol{E}$ *are such that both* $\{(\boldsymbol{X},\boldsymbol{E}), (\overline{\boldsymbol{X}},\boldsymbol{E})\}\subset\Omega_*$ *then*

$$n^{1/2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F \le C_4(\eta_4)(\bar{R}+\|\boldsymbol{E}\|_{op}n^{-1/2})\|(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F,$$
$$\|\overline{\boldsymbol{X}}(\overline{\boldsymbol{B}}-\boldsymbol{B}^*)-\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F \le C_5(\eta_4)(\bar{R}+\|\boldsymbol{E}\|_{op}n^{-1/2})\|(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F$$

for some constants that depend on η_4 *only and* $\bar{R}$ *is defined in Lemma 3.13.*

Proof of Lemma 3.17. By Lemma 3.14, $\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}}$ has at most $2\bar{s}$ non-zero rows. Ω_3 applied on each column of $\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})$ gives $(1-\eta_4)^2 n\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2 \le \|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2$. Similarly, using Ω_3 with $\overline{\boldsymbol{X}}$ and summing the resulting inequality with the previous one yields

$$2(1-\eta_4)^2 n\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2 \le \|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2 + \|\overline{\boldsymbol{X}}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2.$$

Define $\varphi : \boldsymbol{B}\mapsto \frac{1}{2nT}\|\boldsymbol{E}+\boldsymbol{X}(\boldsymbol{B}^*-\boldsymbol{B})\|_F^2+\lambda\|\boldsymbol{B}\|_{2,1}$, $\psi : \boldsymbol{B}\mapsto \frac{1}{2nT}\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B})\|_F^2$ and $\gamma : \boldsymbol{B}\mapsto\varphi(\boldsymbol{B})-\psi(\boldsymbol{B})$. When expanding the squares, it is clear that γ is the sum of a linear function and of the convex penalty, thus γ is convex. Additivity of subdifferentials yields $\partial\varphi(\widehat{\boldsymbol{B}})=\partial\gamma(\widehat{\boldsymbol{B}})+\partial\psi(\widehat{\boldsymbol{B}})=\partial\gamma(\widehat{\boldsymbol{B}})$. By optimality of $\widehat{\boldsymbol{B}}$ we have $\boldsymbol{0}_{p\times T}\in\partial\varphi(\widehat{\boldsymbol{B}})$, thus $\boldsymbol{0}_{p\times T}\in\partial\gamma(\widehat{\boldsymbol{B}})$. This implies $\gamma(\widehat{\boldsymbol{B}})\le\gamma(\overline{\boldsymbol{B}})$. Letting $\overline{\boldsymbol{H}}=\overline{\boldsymbol{B}}-\boldsymbol{B}^*$ and $\boldsymbol{H}=\widehat{\boldsymbol{B}}-\boldsymbol{B}^*$, the last inequality rewrites as

$$\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2 \le \|\boldsymbol{E}-\boldsymbol{X}\overline{\boldsymbol{H}}\|_F^2-\|\boldsymbol{E}-\boldsymbol{X}\boldsymbol{H}\|_F^2+g(\overline{\boldsymbol{B}})-g(\widehat{\boldsymbol{B}}).$$

Summing the similar inequality obtained by replacing $\boldsymbol{X}$ with $\overline{\boldsymbol{X}}$ yields

$$\|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2+\|\overline{\boldsymbol{X}}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2 \le \|\boldsymbol{E}-\boldsymbol{X}\overline{\boldsymbol{H}}\|_F^2-\|\boldsymbol{E}-\boldsymbol{X}\boldsymbol{H}\|_F^2+\|\boldsymbol{E}-\overline{\boldsymbol{X}}\boldsymbol{H}\|_F^2-\|\boldsymbol{E}-\overline{\boldsymbol{X}}\overline{\boldsymbol{H}}\|_F^2.$$

Combining the above displays, we obtain

$$\begin{aligned}
&2(1-\eta_4)^2 n\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2\\
&\le \|\boldsymbol{X}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2+\|\overline{\boldsymbol{X}}(\widehat{\boldsymbol{B}}-\overline{\boldsymbol{B}})\|_F^2\\
&\le \|\boldsymbol{E}-\boldsymbol{X}\overline{\boldsymbol{H}}\|_F^2-\|\boldsymbol{E}-\boldsymbol{X}\boldsymbol{H}\|_F^2+\|\boldsymbol{E}-\overline{\boldsymbol{X}}\boldsymbol{H}\|_F^2-\|\boldsymbol{E}-\overline{\boldsymbol{X}}\overline{\boldsymbol{H}}\|_F^2\\
&= \langle\boldsymbol{X}(\boldsymbol{H}-\overline{\boldsymbol{H}}), 2\boldsymbol{E}-\boldsymbol{X}(\boldsymbol{H}+\overline{\boldsymbol{H}})\rangle_F+\langle\overline{\boldsymbol{X}}(\overline{\boldsymbol{H}}-\boldsymbol{H}), 2\boldsymbol{E}-\overline{\boldsymbol{X}}(\boldsymbol{H}+\overline{\boldsymbol{H}})\rangle_F \qquad \text{thanks to } \langle a-b,a+b\rangle_F=\|a\|_F^2-\|b\|_F^2\\
&= \langle(\boldsymbol{X}-\overline{\boldsymbol{X}})(\boldsymbol{H}-\overline{\boldsymbol{H}}), 2\boldsymbol{E}\rangle_F+\langle\boldsymbol{H}-\overline{\boldsymbol{H}}, (\overline{\boldsymbol{X}}^\top\overline{\boldsymbol{X}}-\boldsymbol{X}^\top\boldsymbol{X})(\boldsymbol{H}+\overline{\boldsymbol{H}})\rangle_F\\
&= \langle\boldsymbol{H}-\overline{\boldsymbol{H}}, 2(\boldsymbol{X}-\overline{\boldsymbol{X}})^\top\boldsymbol{E}\rangle_F+\langle\boldsymbol{H}-\overline{\boldsymbol{H}}, \left[\overline{\boldsymbol{X}}^\top(\overline{\boldsymbol{X}}-\boldsymbol{X})+(\overline{\boldsymbol{X}}-\boldsymbol{X})^\top\boldsymbol{X}\right](\boldsymbol{H}+\overline{\boldsymbol{H}})\rangle_F.
\end{aligned}$$

The second summand rewrites as $\langle\overline{\boldsymbol{X}}(\boldsymbol{H}-\overline{\boldsymbol{H}}), (\overline{\boldsymbol{X}}-\boldsymbol{X})(\boldsymbol{H}+\overline{\boldsymbol{H}})\rangle_F+\langle(\overline{\boldsymbol{X}}-\boldsymbol{X})(\boldsymbol{H}-\overline{\boldsymbol{H}}), \boldsymbol{X}(\boldsymbol{H}+\overline{\boldsymbol{H}})\rangle_F$. By Cauchy-Schwarz and the submultiplicativity of the Frobenius norm, the second summand is bounded above by

$$\|\overline{\boldsymbol{X}}(\boldsymbol{H}-\overline{\boldsymbol{H}})\|_F\|(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{H}+\overline{\boldsymbol{H}})\|_F+\|(\boldsymbol{X}-\overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{H}-\overline{\boldsymbol{H}})\|_F\|\boldsymbol{X}(\boldsymbol{H}+\overline{\boldsymbol{H}})\|_F.$$

Since $\boldsymbol{H} - \overline{\boldsymbol{H}} = \widehat{\boldsymbol{B}} - \overline{\boldsymbol{B}}$ and $\boldsymbol{H} + \overline{\boldsymbol{H}} = \widehat{\boldsymbol{B}} + \overline{\boldsymbol{B}} - 2\boldsymbol{B}^*$ have respectively at most $2\bar{s}$ and $2\bar{s}+s$ non-zero rows, using Ω_3 twice gives the following bound on the second summand:

$$2(1+\eta_4)\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \overline{\boldsymbol{B}})\|_F \|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} + \overline{\boldsymbol{B}} - 2\boldsymbol{B}^*)\|_F.$$

Combining the above displays, we find

$$\begin{aligned}&2(1-\eta_4)^2 n\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \overline{\boldsymbol{B}})\|_F^2\\ \le{}& 2\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \overline{\boldsymbol{B}})\|_F \|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F \big(\|\boldsymbol{E}\|_{op} + (1+\eta_4)\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} + \overline{\boldsymbol{B}} - 2\boldsymbol{B}^*)\|_F\big).\end{aligned}$$

Thanks to Lemma 3.13 we have $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F \le \bar{R}$ and $\|\boldsymbol{\Sigma}^{1/2}(\overline{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F \le \bar{R}$, this shows that $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \overline{\boldsymbol{B}})\|_F \le \|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F \big(\|\boldsymbol{E}\|_{op} n^{-1} + 2(1+\eta_4)n^{-1/2}\bar{R}\big)(1-\eta_4)^{-2}$, hence

$$n^{1/2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \overline{\boldsymbol{B}})\|_F \le 2(1+\eta_4)(1-\eta_4)^{-2}(\bar{R} + \|\boldsymbol{E}\|_{op} n^{-1/2})\|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F.$$

We also have by the triangle inequality

$$\begin{aligned}&\|\overline{\boldsymbol{X}}(\overline{\boldsymbol{B}} - \boldsymbol{B}^*) - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F\\ &\le \|\overline{\boldsymbol{X}}(\overline{\boldsymbol{B}} - \widehat{\boldsymbol{B}})\|_F + \|(\overline{\boldsymbol{X}} - \boldsymbol{X})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F\\ &\le (1+\eta_4)n^{1/2}\|\boldsymbol{\Sigma}^{1/2}(\overline{\boldsymbol{B}} - \widehat{\boldsymbol{B}})\|_F + \|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F\\ &\le \|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F \Big[2(1+\eta_4)(1-\eta_4)^{-2}(\bar{R} + \|\boldsymbol{E}\|_{op} n^{-1/2}) + \bar{R}\Big]\\ &\le 4(1+\eta_4)(1-\eta_4)^{-2}\big(\bar{R} + \|\boldsymbol{E}\|_{op} n^{-1/2}\big)\|(\boldsymbol{X} - \overline{\boldsymbol{X}})\boldsymbol{\Sigma}^{-1/2}\|_F,\end{aligned}$$

where the last line follows from the inequality $2(1+\eta_4)(1-\eta_4)^{-2} \ge 2$ for $\eta_4 \in (0,1)$. □

Lemma 3.18. *For almost every $(\boldsymbol{X}, \boldsymbol{E})$ in the open set $\Omega_1 \cap \Omega_2 \cap \Omega_3$, $\widehat{\boldsymbol{B}}$ is a Fréchet differentiable function of $\boldsymbol{X}$. For almost every $(\boldsymbol{X}, \boldsymbol{E})$ in $\Omega_1 \cap \Omega_2 \cap \Omega_3$, if*

$$\widehat{\boldsymbol{B}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\Big(\tfrac{1}{2nT}\|\boldsymbol{E} + (\boldsymbol{X} + \boldsymbol{w}\boldsymbol{a}^\top)(\boldsymbol{B}^* - \boldsymbol{B})\|_F^2 + \lambda\|\boldsymbol{B}\|_{2,1}\Big)$$

is the estimate (3.4) with $\boldsymbol{X}$ replaced by the perturbed design $\boldsymbol{X} + \boldsymbol{w}\boldsymbol{a}^\top$, then for any $\boldsymbol{b} \in \mathbb{R}^T$

$$\Big((\boldsymbol{X} + \boldsymbol{w}\boldsymbol{a}^\top)(\widehat{\boldsymbol{B}}(\boldsymbol{w}) - \boldsymbol{B}^*)\Big)\boldsymbol{b} - \Big(\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\Big)\boldsymbol{b} = \big(\mathsf{D}(\boldsymbol{b})\big)\boldsymbol{w} + o(\|\boldsymbol{w}\|)$$

as $\|\boldsymbol{w}\| \to 0$, where $\mathsf{D} : \mathbb{R}^T \to \mathbb{R}^{n\times n}$ is a linear map given by $\mathsf{D}(\boldsymbol{b}) = \mathsf{D}^*(\boldsymbol{b}) + \mathsf{D}^{**}(\boldsymbol{b})$ with

$$\begin{aligned}\mathsf{D}^*(\boldsymbol{b}) &= (\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b})\boldsymbol{I}_{n\times n} - (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\big)^\dagger\big(((\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top\boldsymbol{a}) \otimes \boldsymbol{X}_{\hat{S}}^\top\big)\\ &= (\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b})\boldsymbol{I}_{n\times n} - (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\right)^\dagger \begin{pmatrix}\boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_1\boldsymbol{X}_{\hat{S}}^\top\\ \vdots\\ \boldsymbol{a}^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_T\boldsymbol{X}_{\hat{S}}^\top\end{pmatrix},\\ \mathsf{D}^{**}(\boldsymbol{b}) &= (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\big)^\dagger((\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \otimes \boldsymbol{a}_{\hat{S}})\\ &= (\boldsymbol{b}^\top \otimes \boldsymbol{X}_{\hat{S}})\left(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\right)^\dagger \begin{pmatrix}\boldsymbol{a}_{\hat{S}}\boldsymbol{e}_1^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\\ \vdots\\ \boldsymbol{a}_{\hat{S}}\boldsymbol{e}_T^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\end{pmatrix}\end{aligned}$$

for all $\boldsymbol{b} \in \mathbb{R}^T$ *and* $\boldsymbol{w} \in \mathbb{R}^n$. *Note that* D, D^* *and* D^{**} *implicitly depend on* $(\boldsymbol{X}, \boldsymbol{E})$. *Hence the matrix* $\mathsf{D}(\boldsymbol{b})$ *of size* $n \times n$ *is the Jacobian of the map* $\boldsymbol{w} \mapsto (\boldsymbol{X} + \boldsymbol{w}\boldsymbol{a}^\top)(\widehat{\boldsymbol{B}}(\boldsymbol{w}) - \boldsymbol{B}^*)\boldsymbol{b}$ *at* $\boldsymbol{w} = \boldsymbol{0}$.

Proof of Lemma 3.18. By Lemma 3.17 and Rademacher's theorem, we know that the Fréchet derivative of $\widehat{\boldsymbol{B}}$ with respect to $\boldsymbol{X}$ exists almost everywhere, so that $\mathsf{D}(\boldsymbol{b})$ exists for almost every $(\boldsymbol{X}, \boldsymbol{E}) \in \Omega_*$. By Lemma 3.16, we also have that for almost every $(\boldsymbol{X}, \boldsymbol{E})$, the KKT conditions are strict in the sense given in Lemma 3.16. In the following, we consider $(\boldsymbol{X}, \boldsymbol{E}) \in \Omega_*$ such that $\mathsf{D}(\boldsymbol{b})$ exists and such that the KKT conditions are strict; almost every $(\boldsymbol{X}, \boldsymbol{E}) \in \Omega_*$ satisfy these two conditions.

Since we know that the Jacobian $\mathsf{D}(\boldsymbol{b})$ exists by Rademacher's theorem, it is enough to characterize its value, for instance by computing the directional derivative in any fixed direction $\boldsymbol{w} \in \mathbb{R}^n$. To this end, for a real u in a neighborhood of 0, let $\boldsymbol{X}(u) = \boldsymbol{X} + u\boldsymbol{w}\boldsymbol{a}^\top$ and $\boldsymbol{B}(u) = \widehat{\boldsymbol{B}}(u\boldsymbol{w})$. Define the active set $\hat{S}(u) = \{j \in [p] : \|\boldsymbol{B}(u)^\top \boldsymbol{e}_j\|_2 > 0\}$. We also write $\dot{\boldsymbol{X}} = (d/du)\boldsymbol{X}\big|_{u=0} = \boldsymbol{w}\boldsymbol{a}^\top$, and $\dot{\boldsymbol{B}} = (d/du)\boldsymbol{B}(u)\big|_{u=0}$. At 0, we have $\boldsymbol{X}(0) = \boldsymbol{X}$ and $\boldsymbol{B}(0) = \widehat{\boldsymbol{B}}$ is the estimator computed at $(\boldsymbol{X}, \boldsymbol{Y})$ with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}$.

As in (3.60), the KKT conditions for $\boldsymbol{B}(u)$ read, for $j \in \hat{S}(u)$ (i.e., $\boldsymbol{e}_j^\top \boldsymbol{B}(u) \neq \boldsymbol{0}$),

$$\boldsymbol{e}_j^\top \boldsymbol{X}(u)^\top \big[\boldsymbol{E} - \boldsymbol{X}(u)(\boldsymbol{B}(u) - \boldsymbol{B}^*)\big] = \frac{nT\lambda}{\|\boldsymbol{B}(u)^\top \boldsymbol{e}_j\|_2} \boldsymbol{e}_j^\top \boldsymbol{B}(u) \qquad \in \mathbb{R}^{1\times T}$$

and for $j \notin \hat{S}(u)$ (i.e., $\boldsymbol{e}_j^\top \boldsymbol{B}(u) = \boldsymbol{0}$),

$$\|\boldsymbol{e}_j^\top \boldsymbol{X}(u)^\top \big[\boldsymbol{E} - \boldsymbol{X}(u)(\boldsymbol{B}(u) - \boldsymbol{B}^*)\big]\|_2 < nT\lambda.$$

By Lipschitz continuity of $u \mapsto \boldsymbol{B}(u)$ established in Lemma 3.17, the set $\hat{S}(u)$ is constant in a neighborhood of 0 because the KKT conditions on $\hat{S}(u)^c$ are bounded away from $nT\lambda$ on a neighborhood of 0 by continuity, and because the nonzero rows of $\boldsymbol{B}(u)$ are bounded away from $\boldsymbol{0}$ in a neighborhood of 0 again by continuity of $\boldsymbol{B}(u)$. Differentiation of the above display for $j \in \hat{S}(u)$ at $u = 0$ and the product rule yield

$$\boldsymbol{e}_j^\top \Big[\dot{\boldsymbol{X}}^\top (\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)) - \boldsymbol{X}^\top (\dot{\boldsymbol{X}}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) + \boldsymbol{X}\dot{\boldsymbol{B}})\Big] = nT\boldsymbol{e}_j^\top \dot{\boldsymbol{B}}\boldsymbol{H}^{(j)}$$

with $\boldsymbol{H}^{(j)}$ in (3.17). Rearranging and using $\dot{\boldsymbol{X}} = \boldsymbol{w}\boldsymbol{a}^\top$,

$$\boldsymbol{e}_j^\top \Big[\boldsymbol{a}\boldsymbol{w}^\top (\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)) - \boldsymbol{X}^\top (\boldsymbol{w}\boldsymbol{a}^\top (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*))\Big] = \boldsymbol{e}_j^\top \Big[nT\dot{\boldsymbol{B}}\boldsymbol{H}^{(j)} + \boldsymbol{X}^\top \boldsymbol{X}\dot{\boldsymbol{B}}\Big] \qquad \in \mathbb{R}^{1\times T}.$$

Let $\boldsymbol{P}_{\hat{S}} = \sum_{j\in\hat{S}} \boldsymbol{e}_j \boldsymbol{e}_j^\top \in \mathbb{R}^{p\times p}$. Multiplying by $\boldsymbol{e}_j$ to the left and summing over $j \in \hat{S}$, we obtain

$$\boldsymbol{P}_{\hat{S}}\Big[\boldsymbol{a}\boldsymbol{w}^\top (\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)) - \boldsymbol{X}^\top (\boldsymbol{w}\boldsymbol{a}^\top (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*))\Big] = \boldsymbol{P}_{\hat{S}}\Big[nT\dot{\boldsymbol{B}}\boldsymbol{H}^{(j)} + \boldsymbol{X}^\top \boldsymbol{X}\dot{\boldsymbol{B}}\Big] \qquad \in \mathbb{R}^{p\times T}.$$

Since $\hat{S}(u)$ is locally constant for u in a neighborhood of 0, we have $\boldsymbol{P}_{\hat{S}}\dot{\boldsymbol{B}} = \dot{\boldsymbol{B}}$ thus $\boldsymbol{X}\dot{\boldsymbol{B}} = \boldsymbol{X}_{\hat{S}}\dot{\boldsymbol{B}}$, hence

$$\boldsymbol{a}_{\hat{S}}\boldsymbol{w}^\top (\boldsymbol{E} - \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)) - \boldsymbol{X}_{\hat{S}}^\top (\boldsymbol{w}\boldsymbol{a}^\top (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)) = nT\Big[\sum_{j\in\hat{S}} \boldsymbol{e}_j \boldsymbol{e}_j^\top \dot{\boldsymbol{B}}\boldsymbol{H}^{(j)}\Big] + \boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}}\dot{\boldsymbol{B}}\boldsymbol{I}_{T\times T} \qquad \in \mathbb{R}^{p\times T}.$$

We now use the relationship between vectorization and Kronecker product (3.16). Applying (3.14) to the previous display for each term, we find

$$\begin{aligned}
&((\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top \otimes \boldsymbol{a}_{\hat{S}})\,\mathbf{vec}(\boldsymbol{w}^\top) - \Big(((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top \boldsymbol{a}) \otimes \boldsymbol{X}_{\hat{S}}^\top\Big)\,\mathbf{vec}(\boldsymbol{w}) \\
&= \Big(\Big[nT\sum_{j\in\hat{S}}(\boldsymbol{H}^{(j)}\otimes \boldsymbol{e}_j\boldsymbol{e}_j^\top)\Big] + \boldsymbol{I}_{T\times T}\otimes \boldsymbol{X}_{\hat{S}}^\top\boldsymbol{X}_{\hat{S}}\Big)\,\mathbf{vec}(\dot{\boldsymbol{B}}) \\
&= \Big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\Big)\,\mathbf{vec}(\dot{\boldsymbol{B}}).
\end{aligned}$$

Since $\mathbf{vec}(\cdot)$ is always a column vector, $\mathbf{vec}(\boldsymbol{w}^\top) = \mathbf{vec}(\boldsymbol{w}) = \boldsymbol{w}$. Finally, we have again using (3.16) and the chain rule, for any fixed $\boldsymbol{b}\in\mathbb{R}^T$,

$$\begin{aligned}
\mathsf{D}(\boldsymbol{b})\boldsymbol{w} &= \tfrac{d}{du}\boldsymbol{X}(u)(\boldsymbol{B}(u)-\boldsymbol{B}^*)\boldsymbol{b}\,\big|_{u=0} \\
&= \boldsymbol{w}\boldsymbol{a}^\top(\boldsymbol{B}(0)-\boldsymbol{B}^*)\boldsymbol{b} + \boldsymbol{X}\dot{\boldsymbol{B}}\boldsymbol{b} \\
&= \boldsymbol{w}\boldsymbol{a}^\top(\boldsymbol{B}(0)-\boldsymbol{B}^*)\boldsymbol{b} + \boldsymbol{X}_{\hat{S}}\dot{\boldsymbol{B}}\boldsymbol{b}.
\end{aligned}$$

By Lemma 3.15, $\mathrm{rank}(\boldsymbol{X}_{\hat{S}}) = |\hat{S}|$. The argument developed in the proof of Proposition 3.2 (ii) shows that the nullspace of the matrix $\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}$ is exactly the linear span of $\{\boldsymbol{e}_t\otimes\boldsymbol{e}_j,\ (j,t)\in\hat{S}^c\times[T]\}$. Because $\boldsymbol{P}_{\hat{S}}\dot{\boldsymbol{B}} = \dot{\boldsymbol{B}}$, $\mathbf{vec}(\dot{\boldsymbol{B}})$ is in $\ker(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})^\perp = \mathrm{range}(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})$. Since for any symmetric matrix $\boldsymbol{M}$, $\boldsymbol{M}^\dagger\boldsymbol{M}$ is the orthogonal projection on the range of $\boldsymbol{M}$, we have $\big(nT\tilde{\boldsymbol{H}} + \tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\big)^\dagger\big(nT\tilde{\boldsymbol{H}} + \tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\big)\,\mathbf{vec}(\dot{\boldsymbol{B}}) = \mathbf{vec}(\dot{\boldsymbol{B}})$. Since $\boldsymbol{X}_{\hat{S}}\dot{\boldsymbol{B}}\boldsymbol{b}$ is a column vector, using (3.16) again,

$$\begin{aligned}
\boldsymbol{X}_{\hat{S}}\dot{\boldsymbol{B}}\boldsymbol{b} &= \mathbf{vec}(\boldsymbol{X}_{\hat{S}}\dot{\boldsymbol{B}}\boldsymbol{b}) \\
&= (\boldsymbol{b}^\top\otimes\boldsymbol{X}_{\hat{S}})\,\mathbf{vec}(\dot{\boldsymbol{B}}) \\
&= (\boldsymbol{b}^\top\otimes\boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\big)^\dagger\Big[((\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\otimes\boldsymbol{a}_{\hat{S}}) - \Big(((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a})\otimes\boldsymbol{X}_{\hat{S}}^\top\Big)\Big]\boldsymbol{w}.
\end{aligned}$$

Since this holds for all $\boldsymbol{w}$, this provides the desired expression for $\mathsf{D}(\boldsymbol{b})$ for all $\boldsymbol{b}$. □

Lemma 3.19. *For any $\boldsymbol{b}\in\mathbb{R}^T$ we have on Ω_**

$$\mathrm{Tr}[\mathsf{D}^*(\boldsymbol{b})] = \boldsymbol{b}^\top(n\boldsymbol{I}_{T\times T} - \hat{\mathbf{A}})(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a}, \tag{3.55}$$

$$\sum_{t=1}^{T}\Big(\mathrm{Tr}[\mathsf{D}^{**}(\boldsymbol{e}_t)]\Big)^2 \le C_6(\boldsymbol{\Sigma})\sigma^2 sT \tag{3.56}$$

for some constant depending on $\boldsymbol{\Sigma}$ and $\eta_1,\dots,\eta_4$ only.

Proof of Lemma 3.19. For the first equality,

$$\mathrm{Tr}[\mathsf{D}^*(\boldsymbol{b})] = \mathrm{Tr}[(\boldsymbol{a}^\top(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\boldsymbol{b})\boldsymbol{I}_{n\times n} - (\boldsymbol{b}^\top\otimes\boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}}\big)^\dagger(((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a})\otimes\boldsymbol{X}_{\hat{S}}^\top)]$$

and the conclusion follows from (3.19).

For (3.56), the following bounds will be useful. Inequality $\|\boldsymbol{X}_{\hat S}\|_{op}^2 \le \|\boldsymbol{\Sigma}\|_{op}(1+\eta_4)^2 n$ holds on Ω_*. Furthermore since $\ker \boldsymbol{N} = \ker \boldsymbol{N}^\dagger$ for all symmetric matrices $\boldsymbol{N}$ and since $\tilde{\boldsymbol{H}}$ is positive semi-definite, on Ω_* we find

$$
\begin{aligned}
\|(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})^\dagger\|_{op}^2 &= \big[\min_{\boldsymbol{u}\in\mathbb{R}^{np}:\boldsymbol{u}\in\ker(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}})^\perp} \boldsymbol{u}^\top(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}})\boldsymbol{u}\big]^{-2}\\
&\le \big[\min_{\boldsymbol{u}\in\mathbb{R}^{np}:\boldsymbol{u}\in\ker(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}})^\perp} \boldsymbol{u}^\top(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}})\boldsymbol{u}\big]^{-2}\\
&\le \phi_{\min}(\boldsymbol{\Sigma})^{-2}(1-\eta_4)^{-4}n^{-2}. \qquad (3.63)
\end{aligned}
$$

We now work on $\sum_{t=1}^T \mathrm{Tr}[\mathsf{D}^{**}(\boldsymbol{e}_t)]^2 = \|\boldsymbol{v}\|^2$, the left hand side of (3.56). For brevity, define $\boldsymbol{M} = (\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + nT\tilde{\boldsymbol{H}})^\dagger(\boldsymbol{I}_{T\times T}\otimes \boldsymbol{a}_{\hat S})(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top$. Then if $\boldsymbol{e}_t \in \mathbb{R}^T$ and $\boldsymbol{e}_i \in \mathbb{R}^n$ denote canonical basis vectors, $\sum_{t=1}^T \mathrm{Tr}[\mathsf{D}^{**}(\boldsymbol{e}_t)]^2 = \|\boldsymbol{v}\|^2$ where $\boldsymbol{v}\in\mathbb{R}^T$ has components $\boldsymbol{v}_t = \mathrm{Tr}[\mathsf{D}^{**}(\boldsymbol{e}_t)]$ so that

$$
\boldsymbol{v}_t = \sum_{i=1}^n \boldsymbol{e}_i^\top[\mathsf{D}^{**}(\boldsymbol{e}_t)]\boldsymbol{e}_i = \sum_{i=1}^n \boldsymbol{e}_i^\top(\boldsymbol{e}_t^\top\otimes\boldsymbol{X}_{\hat S})\boldsymbol{M}\boldsymbol{e}_i = \boldsymbol{e}_t^\top\sum_{i=1}^n(\boldsymbol{I}_{T\times T}\otimes(\boldsymbol{e}_i^\top\boldsymbol{X}_{\hat S}))\boldsymbol{M}\boldsymbol{e}_i,
$$

where the last equality stems from two applications of the mixed product property (3.13):

$$
\begin{aligned}
\boldsymbol{e}_i^\top(\boldsymbol{e}_t^\top\otimes\boldsymbol{X}_{\hat S}) = (1\otimes\boldsymbol{e}_i^\top)(\boldsymbol{e}_t^\top\otimes\boldsymbol{X}_{\hat S}) = (\boldsymbol{e}_t^\top)\otimes(\boldsymbol{e}_i^\top\boldsymbol{X}_{\hat S}) &= (\boldsymbol{e}_t^\top\boldsymbol{I}_{T\times T})\otimes(1(\boldsymbol{e}_i^\top\boldsymbol{X}_{\hat S}))\\
&= \boldsymbol{e}_t^\top(\boldsymbol{I}_{T\times T}\otimes(\boldsymbol{e}_i^\top\boldsymbol{X}_{\hat S})).
\end{aligned}
$$

Thus $\boldsymbol{v} = \sum_{i=1}^n(\boldsymbol{I}_{T\times T}\otimes(\boldsymbol{e}_i^\top\boldsymbol{X}_{\hat S}))\boldsymbol{M}\boldsymbol{e}_i$ and since $\|\boldsymbol{v}\|_2^2 = \boldsymbol{v}^\top\boldsymbol{v} = (\boldsymbol{v}^\top\otimes 1)\boldsymbol{v}$, it follows that $\|\boldsymbol{v}\|^2 = \sum_{i=1}^n(\boldsymbol{v}^\top\otimes(\boldsymbol{e}_i^\top\boldsymbol{X}_{\hat S}))\boldsymbol{M}\boldsymbol{e}_i = \mathrm{Tr}[(\boldsymbol{v}^\top\otimes\boldsymbol{X}_{\hat S})\boldsymbol{M}]$ by (3.13).

By the definition of $\boldsymbol{M}$, using the commutation property of the trace we have

$$
\|\boldsymbol{v}\|_2^2 = \mathrm{Tr}\big[(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top(\boldsymbol{v}^\top\otimes\boldsymbol{X}_{\hat S})(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}})^\dagger(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{a}_{\hat S})\big].
$$

By the Cauchy-Schwarz inequality for $\langle\cdot,\cdot\rangle_F$ and using $\|\boldsymbol{U}\boldsymbol{V}\|_F \le \|\boldsymbol{U}\|_{op}\|\boldsymbol{V}\|_F$ twice, we find

$$
\begin{aligned}
\|\boldsymbol{v}\|_2^2 &\le \|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top(\boldsymbol{v}^\top\otimes\boldsymbol{X}_{\hat S})\|_F\|(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}})^\dagger(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{a}_{\hat S})\|_F\\
&\le \|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{op}\|\boldsymbol{v}^\top\otimes\boldsymbol{X}_{\hat S}\|_F\|(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}})^\dagger\|_{op}\|(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{a}_{\hat S})\|_F
\end{aligned}
$$

and the second factor equals $\|\boldsymbol{v}^\top\otimes\boldsymbol{X}_{\hat S}\|_F = \|\boldsymbol{v}\|_2\|\boldsymbol{X}_{\hat S}\|_F$ by (3.15) for the Frobenius norm.

We introduce the notation $\lesssim$ to denote an inequality up to a constant that depends on $\eta_1,...,\eta_4$ and $\phi_{\min}(\boldsymbol{\Sigma}),\phi_{\max}(\boldsymbol{\Sigma})$ only. On Ω_* we have the operator norm bound (3.63), the bound $\|\boldsymbol{X}_{\hat S}\|_F \le |\hat S|^{1/2}\|\boldsymbol{X}_{\hat S}\|_{op} \lesssim (|\hat S|n)^{1/2}$ as well as $\|(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{a}_{\hat S})\|_F = \sqrt{T}\|\boldsymbol{a}_{\hat S}\|_2 \lesssim \sqrt{T}$ so that

$$
\|\boldsymbol{v}\|_2 \lesssim \|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{op}\sqrt{ns}n^{-1}\sqrt{T}
$$

and $\|\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}\|_{op} \le \|\boldsymbol{E}\|_{op} + \|\boldsymbol{X}(\boldsymbol{B}^*-\widehat{\boldsymbol{B}})\|_F \le \sigma(\sqrt{T}+2\sqrt{n}) + \sqrt{n}\bar R$ thanks to Ω_4 and Lemma 3.13. Since $T\le n$ and $\bar R \lesssim 1$ under Assumption (A1), we have proved that $\|\boldsymbol{v}\|_2 \lesssim \sigma\sqrt{sT}$ holds on Ω_* which is exactly the desired bound (3.56).

□

Lemma 3.20. *Under Assumption (A1), as $n, p \to +\infty$ we have*

$$\frac{1}{\sigma^2 n}\mathbb{E}\Big[I\{\Omega_*\}\sum_{t=1}^{T}\Big(\boldsymbol{z}_0^\top \boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\boldsymbol{e}_t - \mathrm{Tr}[\mathsf{D}(\boldsymbol{e}_t)]\Big)^2\Big] \longrightarrow 0.$$

Since Ω_* has probability approaching one, this implies that $\frac{1}{\sigma^2 n}\sum_{t=1}^{T}(\boldsymbol{z}_0^\top \boldsymbol{X}(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\boldsymbol{e}_t - \mathrm{Tr}[\mathsf{D}(\boldsymbol{e}_t)])^2$ converges to 0 in probability.

Proof of Lemma 3.20. Recall that we assume the normalization $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|^2 = 1$. Following the notation in [26] we define the quantities:

$$\boldsymbol{u}_0 = \boldsymbol{\Sigma}^{-1}\boldsymbol{a}, \quad \boldsymbol{z}_0 = \boldsymbol{X}\boldsymbol{u}_0, \quad \boldsymbol{Q}_0 = \boldsymbol{I}_{p\times p} - \boldsymbol{u}_0\boldsymbol{a}^\top.$$

We have the decomposition $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{Q}_0 + \boldsymbol{z}_0\boldsymbol{a}^\top$, the vector $\boldsymbol{z}_0$ is independent of $\boldsymbol{X}\boldsymbol{Q}_0$ and $\boldsymbol{z}_0$ has distribution $\mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$. Given a value of $(\boldsymbol{E}, \boldsymbol{X}\boldsymbol{Q}_0)$, define the open set

$$U_0 = \{\boldsymbol{z}_0 \in \mathbb{R}^n : (\boldsymbol{E}, \boldsymbol{X}\boldsymbol{Q}_0 + \boldsymbol{z}_0\boldsymbol{a}^\top) \in \Omega_*\} \subset \mathbb{R}^n.$$

Since Ω_* is open , so is the set U_0. Given a value of $(\boldsymbol{E}, \boldsymbol{X}\boldsymbol{Q}_0)$ we also define the function $U_0 \to \mathbb{R}^{p\times T}$ given by

$$\widehat{\boldsymbol{B}}(\boldsymbol{z}_0) = \arg\min_{\boldsymbol{B}\in\mathbb{R}^{p\times T}}\Big(\tfrac{1}{2nT}\|\boldsymbol{E} + (\boldsymbol{X}\boldsymbol{Q}_0 + \boldsymbol{z}_0\boldsymbol{a}^\top)(\boldsymbol{B}^* - \boldsymbol{B})\|_F^2 + \lambda\|\boldsymbol{B}\|_{2,1}\Big)$$

as well as

$$\boldsymbol{F} : U_0 \to \mathbb{R}^{n\times T}, \qquad \boldsymbol{F} : \boldsymbol{z}_0 \mapsto (\boldsymbol{X}\boldsymbol{Q}_0 + \boldsymbol{z}_0\boldsymbol{a}^\top)(\widehat{\boldsymbol{B}}(\boldsymbol{z}_0) - \boldsymbol{B}^*).$$

Since $\bar{R} \to 0$ under Assumption (A1) and $\|\boldsymbol{E}\|_{op}n^{-1/2}$ is bounded by an absolute constant on Ω_4 when $T \le n$, Lemma 3.17 shows that $\boldsymbol{F}$ is L-Lipschitz for some constant L of the form $L = \sigma C_7(\eta_1, ..., \eta_4, \boldsymbol{\Sigma})$ where the constant depends only on $\eta_1, ..., \eta_4$ and the minimal and maximal eigenvalues of $\boldsymbol{\Sigma}$. By Kirszbraun's Theorem, there exists an L-Lipschitz function $\tilde{\boldsymbol{F}} : \mathbb{R}^n \to \mathbb{R}^{n\times T}$ which is an extension of $\boldsymbol{F}$, i.e., it satisfies $\boldsymbol{F}(\boldsymbol{z}_0) = \tilde{\boldsymbol{F}}(\boldsymbol{z}_0)$ for all $\boldsymbol{z}_0 \in U_0$. Since $\boldsymbol{F}(\boldsymbol{z}_0)$ is bounded from above by $n^{1/2}(1-\eta_3)\bar{R}$ in U_0 by Lemma 3.13, we define the function $\bar{\boldsymbol{F}} : \mathbb{R}^n \to \mathbb{R}^{n\times T}$ by

$$\bar{\boldsymbol{F}}(\boldsymbol{z}_0) = \boldsymbol{\Pi} \circ \tilde{\boldsymbol{F}}(\boldsymbol{z}_0)$$

where $\boldsymbol{\Pi} : \mathbb{R}^{n\times T} \to \mathbb{R}^{n\times T}$ is the convex projection onto the Frobenius ball of radius $n^{1/2}\bar{R}$ in $\mathbb{R}^{n\times T}$. Since convex projections are 1-Lipschitz functions, the function $\bar{\boldsymbol{F}}$ is also an L-Lipschitz extension of $\boldsymbol{F}$.

If $\bar{\mathsf{D}}(\boldsymbol{b})$ denotes the Jacobian such that $\bar{\boldsymbol{F}}(\boldsymbol{w})\boldsymbol{b} - \bar{\boldsymbol{F}}(\boldsymbol{0})\boldsymbol{b} = \bar{\mathsf{D}}(\boldsymbol{b})\boldsymbol{w} + o(\|\boldsymbol{w}\|)$ for all $\boldsymbol{b} \in \mathbb{R}^T$, then $\bar{\mathsf{D}}(\boldsymbol{b}) = \mathsf{D}(\boldsymbol{b})$ on U_0 because two functions that coincide on an open set have the same gradient on this open set. This implies

$$\begin{aligned}\mathbb{E}\Big[I\{\Omega_*\}\sum_{t=1}^{T}\Big(\boldsymbol{z}_0^\top\boldsymbol{F}(\boldsymbol{z}_0)\boldsymbol{e}_t - \mathrm{Tr}[\mathsf{D}(\boldsymbol{e}_t)]\Big)^2\Big] &= \mathbb{E}\Big[I\{\Omega_*\}\sum_{t=1}^{T}\Big(\boldsymbol{z}_0^\top\bar{\boldsymbol{F}}(\boldsymbol{z}_0)\boldsymbol{e}_t - \mathrm{Tr}[\bar{\mathsf{D}}(\boldsymbol{e}_t)]\Big)^2\Big] \\ &\le \mathbb{E}\Big[\sum_{t=1}^{T}\Big(\boldsymbol{z}_0^\top\bar{\boldsymbol{F}}(\boldsymbol{z}_0)\boldsymbol{e}_t - \mathrm{Tr}[\bar{\mathsf{D}}(\boldsymbol{e}_t)]\Big)^2\Big]\end{aligned}$$

where the second display simply follows from $I\{\Omega_*\} \le 1$. By the main result of [25] we find

$$\mathbb{E}\left[(\boldsymbol{z}_0^\top \bar{\boldsymbol{F}}(\boldsymbol{z}_0)\boldsymbol{e}_t - \mathrm{Tr}[\bar{\mathsf{D}}(\boldsymbol{e}_t)])^2\right] = \mathbb{E}\left[\|\bar{\boldsymbol{F}}(\boldsymbol{z}_0)\boldsymbol{e}_t\|_2^2 + \mathrm{Tr}(\{\bar{\mathsf{D}}(\boldsymbol{e}_t)\}^2)\right]$$
$$\le \mathbb{E}\left[\|\bar{\boldsymbol{F}}(\boldsymbol{z}_0)\boldsymbol{e}_t\|_2^2 + \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2\right]$$

for each $t = 1, ..., T$ since $\boldsymbol{z}_0 \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_{n\times n})$. Summing this inequality over $t = 1, ..., T$ yields

$$\begin{aligned}
&\frac{1}{n\sigma^2}\mathbb{E}\Big[\sum_{t=1}^T \Big(\boldsymbol{z}_0^\top \bar{\boldsymbol{F}}(\boldsymbol{z}_0)\boldsymbol{e}_t - \mathrm{Tr}[\bar{\mathsf{D}}(\boldsymbol{e}_t)]\Big)^2\Big] \\
&\le \frac{1}{n\sigma^2}\mathbb{E}\Big[\|\bar{\boldsymbol{F}}(\boldsymbol{z}_0)\|_F^2 + \sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2\Big] \\
&\le \frac{\bar{R}^2}{\sigma^2} + \frac{1}{n\sigma^2}\mathbb{E}\Big[\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2\Big] \\
&= \frac{\bar{R}^2}{\sigma^2} + \frac{1}{n\sigma^2}\mathbb{E}\Big[I\{\Omega_*\}\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2\Big] + \frac{1}{n\sigma^2}\mathbb{E}\Big[I\{\Omega_*^c\}\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2\Big].
\end{aligned}$$

Note that the first term, $\bar{R}^2/\sigma^2$, converges to 0, as stated in Lemma 3.13. We now bound the third term, on Ω_*^c. The quantity $\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2$ is exactly the squared Frobenius norm of the Jacobian of the map $\bar{\boldsymbol{F}} : \mathbb{R}^n \to \mathbb{R}^{n\times T}$ (this Jacobian has dimensions $(nT) \times n$ but we do not need to write it explicitly or choose a specific vectorization of $\mathbb{R}^{n\times T}$ into $\mathbb{R}^{nT}$). Since $\bar{\boldsymbol{F}}$ is L-Lipschitz, the operator norm of the Jacobian is at most L. Since the rank of the Jacobian of a map from $\mathbb{R}^n$ to any other linear space is at most n, the rank of the Jacobian is at most n. If follows from $\|\boldsymbol{J}\|_F^2 \le \mathrm{rank}(\boldsymbol{J})\|\boldsymbol{J}\|_{op}^2$ with $\boldsymbol{J} \in \mathbb{R}^{(nT)\times n}$ the Jacobian of $\bar{\boldsymbol{F}}$ that

$$\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2 = \|\boldsymbol{J}\|_F^2 \le nL^2$$

so that $\frac{1}{n\sigma^2}\mathbb{E}[I\{\Omega_*^c\}\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2] \le \mathbb{P}(\Omega_*^c)L^2/\sigma^2$ which converges to 0 under Assumption (A1) thanks to $\mathbb{P}(\Omega_*) \to 1$ in Lemma 3.12.

It remains to show that $\frac{1}{n\sigma^2}\mathbb{E}[I\{\Omega_*\}\sum_{t=1}^T \|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2]$ converges to 0. This quantity is equal to $\frac{1}{n\sigma^2}\mathbb{E}[I\{\Omega_*\}\sum_{t=1}^T \|\mathsf{D}(\boldsymbol{e}_t)\|_F^2]$ since the derivatives of $\bar{\boldsymbol{F}}$ and $\boldsymbol{F}$ coincide on U_0. To bound this quantity, we use the explicit formulae obtained in Lemma 3.18 with $\|\mathsf{D}(\boldsymbol{e}_t)\|_F^2 \le 2\|\mathsf{D}^*(\boldsymbol{e}_t)\|_F^2 + 2\|\mathsf{D}^{**}(\boldsymbol{e}_t)\|_F^2$. We can use the following property of Kronecker products. If $\boldsymbol{M}, \boldsymbol{Q}$ are two matrices, and $\boldsymbol{e}_t$ is the t-th canonical basis vector in $\mathbb{R}^T$, then by the mixed product property (3.13)

$$\begin{aligned}\sum_{t=1}^{T}\|(\boldsymbol{e}_t^\top\otimes\boldsymbol{M})\boldsymbol{Q}\|_F^2 &= \sum_{t=1}^{T}\mathrm{Tr}[\boldsymbol{Q}^\top(\boldsymbol{e}_t\otimes\boldsymbol{M}^\top)(\boldsymbol{e}_t^\top\otimes\boldsymbol{M})\boldsymbol{Q}]\\ &= \mathrm{Tr}[\boldsymbol{Q}^\top\sum_{t=1}^{T}\Big[(\boldsymbol{e}_t\otimes\boldsymbol{M}^\top)(\boldsymbol{e}_t^\top\otimes\boldsymbol{M})\Big]\boldsymbol{Q}]\\ &= \mathrm{Tr}[\boldsymbol{Q}^\top(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{M}^\top\boldsymbol{M})\Big]\boldsymbol{Q}]\\ &= \|(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{M})\boldsymbol{Q}\|_F^2. \qquad (3.64)\end{aligned}$$

Since $\|\mathsf{D}^*(\boldsymbol{e}_t)\|_F^2 \le 2(\boldsymbol{a}^\top(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\boldsymbol{e}_t)^2\|\boldsymbol{I}_{n\times n}\|_F^2 + 2\|(\boldsymbol{e}_t^\top\otimes\boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger\big(((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a})\otimes\boldsymbol{X}_{\hat{S}}^\top\big)\|_F^2$, thanks to (3.64) with $\boldsymbol{M}=\boldsymbol{X}_{\hat{S}}$ and $\boldsymbol{Q}=\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger\big(((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a})\otimes\boldsymbol{X}_{\hat{S}}^\top\big)$ for the second term we find

$$\sum_{t=1}^{T}\|\mathsf{D}^*(\boldsymbol{e}_t)\|_F^2 \le 2n\|(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a}\|_2^2+2\|(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger\big(((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a})\otimes\boldsymbol{X}_{\hat{S}}^\top\big)\|_F^2.$$

The first summand is bounded by $2n\|(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F^2\ \|\boldsymbol{a}\|_2^2 \le 2n\phi_{\min}(\boldsymbol{\Sigma})^{-1/2}\bar{R}\ \phi_{\max}(\boldsymbol{\Sigma})$ and the second summand by

$$\begin{aligned}&\overset{\text{(i)}}{\le} 2\|(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{X}_{\hat{S}})\|_{op}^2\ \|\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger\|_{op}^2\ \|((\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a})\otimes\boldsymbol{X}_{\hat{S}}^\top\|_F^2\\ &\overset{\text{(ii)}}{\le} 2\|\boldsymbol{X}_{\hat{S}}\|_{op}^2\ \|\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger\|_{op}^2\ \|(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)^\top\boldsymbol{a}\|_F^2\ \|\boldsymbol{X}_{\hat{S}}^\top\|_F^2\\ &\le 2\|\boldsymbol{X}_{\hat{S}}\|_{op}^2\ \|\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger\|_{op}^2\ \|(\widehat{\boldsymbol{B}}-\boldsymbol{B}^*)\|_F^2\ \|\boldsymbol{a}\|_2^2\ \mathrm{rank}(\boldsymbol{X}_{\hat{S}})\|\boldsymbol{X}_{\hat{S}}\|_{op}^2\\ &\overset{\text{(iii)}}{\le} 2(\phi_{\max}(\boldsymbol{\Sigma})(1+\eta_4)^2 n)^2(\phi_{\min}(\boldsymbol{\Sigma})^{-2}(1-\eta_4)^{-4}n^{-2})(\phi_{\min}(\boldsymbol{\Sigma})^{-1/2}\bar{R}\ \phi_{\max}(\boldsymbol{\Sigma}))\bar{s}\\ &= 2\phi_{\max}(\boldsymbol{\Sigma})^3\phi_{\min}(\boldsymbol{\Sigma})^{-5/2}\bar{s}\bar{R}.\end{aligned}$$

Above, (i) follows from $\|\boldsymbol{MNU}\|_F \le \|\boldsymbol{M}\|_{op}\|\boldsymbol{N}\|_{op}\|\boldsymbol{U}\|_F$, (ii) is a consequence of (3.15) and (iii) holds on Ω_*. Thus $\sum_{t=1}^{T}\|\mathsf{D}^*(\boldsymbol{e}_t)\|_F^2 \lesssim n\bar{R}$.

Likewise,

$$\begin{aligned}\sum_{t=1}^{T}\|\mathsf{D}^{**}(\boldsymbol{e}_t)\|_F^2 &\le \|(\boldsymbol{I}_{T\times T}\otimes\boldsymbol{X}_{\hat{S}})\big(\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}+nT\tilde{\boldsymbol{H}}\big)^\dagger((\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top\otimes\boldsymbol{a}_{\hat{S}})\|_F^2\\ &\le (\phi_{\max}(\boldsymbol{\Sigma})(1+\eta_4)^2 n)(\phi_{\min}(\boldsymbol{\Sigma})^{-2}(1-\eta_4)^{-4}n^{-2})(8\sigma^2 nT+2(1-\eta_3)^2 n\bar{R}^2)\phi_{\max}(\boldsymbol{\Sigma})\\ &\lesssim \sigma^2 T\end{aligned}$$

Thus $\frac{1}{n\sigma^2}\mathbb{E}[I\{\Omega_*\}\sum_{t=1}^{T}\|\bar{\mathsf{D}}(\boldsymbol{e}_t)\|_F^2] \lesssim \frac{\bar{R}}{\sigma^2}+\frac{T}{n}$ and the right hand side converges to 0 under Assumption (A1). □

3.10 Proof that $\mathbb{P}(\Omega_*) \to 1$

3.10.1 Ω_1: Restricted Eigenvalues for random matrices in multi-task learning

Proposition 3.21. *Let $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ be a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries and let $\mathcal{A}$ be a subset of $\mathbb{R}^{p\times T}$ with $\|\boldsymbol{B}\|_F = 1$ for all $\boldsymbol{B} \in \mathcal{A}$.*

(i) For any two $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{A}$, $\mathbb{P}(|\ \|\boldsymbol{GA}\|_F - \|\boldsymbol{GB}\|_F| \geq C_8\sqrt{x}\|\boldsymbol{B}-\boldsymbol{A}\|_F) \leq 6e^{-x}$ for all $x > 0$.

(ii) $\sup_{\boldsymbol{A},\boldsymbol{B}\in\mathcal{A}} |\ \|\boldsymbol{GA}\|_F - \|\boldsymbol{GB}\|_F| \leq C_9\mathbb{E}\sup_{\boldsymbol{B}\in\mathcal{A}} |\operatorname{Tr}[\boldsymbol{B}^\top \boldsymbol{G}']| + C_{10}\sqrt{x}$ with probability at least $1-e^{-x}$, where $\boldsymbol{G}' \in \mathbb{R}^{p\times T}$ has i.i.d. $\mathcal{N}(0,1)$ entries.

$\sup_{\boldsymbol{A}\in\mathcal{A}} |\ \|\boldsymbol{GA}\|_F - \sqrt{n}| \leq C_{11}\mathbb{E}\sup_{\boldsymbol{B}\in\mathcal{A}} |\operatorname{Tr}[\boldsymbol{B}^\top \boldsymbol{G}']| + C_{12}\sqrt{x}$ also holds with probability at least $1-3e^{-x}$.

(iii) If $\boldsymbol{X}$ has i.i.d. $\mathcal{N}_p(\boldsymbol{0},\boldsymbol{\Sigma})$ rows with $\max_{j\in[p]} \boldsymbol{\Sigma}_{jj} \leq 1$ and

$$\mathcal{C} = \{\boldsymbol{A} \in \mathbb{R}^{p\times T} : \|\boldsymbol{A}\|_{2,1} \leq \sqrt{k}\|\boldsymbol{A}\|_F\}, \tag{3.65}$$

then with probability at least $1-3e^{-x}$,

$$\begin{aligned}
&\sup_{\boldsymbol{A}\in\mathcal{C}:\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\|_F=1} \left|n^{-1/2}\|\boldsymbol{XA}\|_F - 1\right| = \sup_{\boldsymbol{B}\in\mathbb{R}^{p\times T}:\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\in\mathcal{C},\|\boldsymbol{B}\|_F=1} \left|n^{-1/2}\|\boldsymbol{X\Sigma}^{-1/2}\boldsymbol{B}\|_F - 1\right| \\
&\leq C_{13}\sqrt{x/n} + C_{14}n^{-1/2}\mathbb{E} \sup_{\boldsymbol{B}\in\mathbb{R}^{p\times T}:\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\in\mathcal{C},\|\boldsymbol{B}\|_F=1} |\operatorname{Tr}[\boldsymbol{B}^\top\boldsymbol{G}']| \\
&\leq C_{15}\sqrt{x/n} + C_{16}\sqrt{[kT + k\log(p/k)]/(\phi_{\min}(\boldsymbol{\Sigma})n)}
\end{aligned}$$

This implies that for any constant $\eta_3 \in (0,1)$, if $\{kT + k\log(p/k)\}/(n\phi_{\min}(\boldsymbol{\Sigma})) \to 0$ then $\mathbb{P}(\max_{\boldsymbol{A}\in\mathcal{C}:\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\|_F=1} |n^{-1/2}\|\boldsymbol{XA}\|_F - 1| \leq \eta_3) \to 1$.

The proof follows the argument from [172], adapted to the multi-task setting.

Proof of (i). We distinguish two cases.
Case (a): $\sqrt{xn} > n/4$. In this case we use that

$$\|\boldsymbol{GA}\|_F - \|\boldsymbol{GB}\|_F \leq \|\boldsymbol{G}(\boldsymbol{A}-\boldsymbol{B})\|_F = \Big(\sum_{i=1}^n \|(\boldsymbol{A}-\boldsymbol{B})^\top\boldsymbol{G}^\top\boldsymbol{e}_i\|_2^2\Big)^{1/2}$$

and we apply [263, Theorem 6.3.2] to the vector $\mathbf{vec}(\boldsymbol{G}^\top) \in \mathbb{R}^{np\times 1}$ and the block diagonal matrix with n blocks, each block being $(\boldsymbol{A}-\boldsymbol{B})^\top$. This yields

$$\mathbb{P}(|\ \|\boldsymbol{GA}\|_F - \|\boldsymbol{GB}\|_F| \geq \sqrt{x}\|\boldsymbol{B}-\boldsymbol{A}\|_{op} + \sqrt{n}\|\boldsymbol{B}-\boldsymbol{A}\|_F) \leq 2e^{-C_{17}x}.$$

Here, $\sqrt{n} \leq 4\sqrt{x}$ and we can bound from above the first term to obtain the desired bound.

Case (b): $\sqrt{xn} \leq n/4$. Write $\|\boldsymbol{G}\boldsymbol{A}\|_F - \|\boldsymbol{G}\boldsymbol{B}\|_F = \frac{\|\boldsymbol{G}\boldsymbol{A}\|_F^2 - \|\boldsymbol{G}\boldsymbol{B}\|_F^2}{\|\boldsymbol{G}\boldsymbol{A}\|_F + \|\boldsymbol{G}\boldsymbol{B}\|_F}$. We will use repeatedly the following concentration bounds: if $\boldsymbol{z} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{I}_{q\times q})$ and $\boldsymbol{M} \in \mathbb{R}^{q\times q}$ is symmetric positive semi-definite, then

$$\mathbb{P}\left(\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z} - \operatorname{Tr}\boldsymbol{M} < 2\sqrt{x}\|\boldsymbol{M}\|_F\right) \leq e^{-x}. \tag{3.66}$$

This is a straightforward consequence of [163, Lemma 1] after diagonalizing the symmetric positive semi-definite matrix $\boldsymbol{M}$. Furthermore, for any $\boldsymbol{M} \in \mathbb{R}^{q\times q}$,

$$\mathbb{P}\left(\boldsymbol{z}^\top \boldsymbol{M}\boldsymbol{z} - \operatorname{Tr}\boldsymbol{M} > 2\sqrt{x}\|\boldsymbol{M}\|_F + 2x\|\boldsymbol{M}\|_{op}\right) \leq e^{-x} \tag{3.67}$$

see for instance [41, Example 2.12] or [22, Lemma 3.1].

If $\boldsymbol{g}_1^\top, ..., \boldsymbol{g}_n^\top$ are the rows of $\boldsymbol{G}$ then $\|\boldsymbol{G}\boldsymbol{A}\|_F^2 = \sum_{i=1}^n \boldsymbol{g}_i^\top \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{g}_i$ is of the above form with $q = np$ and $\boldsymbol{M}$ is block diagonal with n blocks equal to $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{p\times p}$. Thus $\|\boldsymbol{G}\boldsymbol{A}\|_F^2 \geq n\|\boldsymbol{A}\|_F^2 - 2\sqrt{xn}\|\boldsymbol{A}\boldsymbol{A}^\top\|_F \geq n - 2\sqrt{xn}$ with probability at least $1 - e^{-x}$ by (3.66) and thanks to $\|\boldsymbol{A}\|_F = 1$. The same holds for a lower bound on $\|\boldsymbol{G}\boldsymbol{B}\|_F^2$. For the numerator, thanks to (3.67), with probability at least $1 - e^{-x}$:

$$\begin{aligned}
\|\boldsymbol{G}\boldsymbol{A}\|_F^2 - \|\boldsymbol{G}\boldsymbol{B}\|_F^2 &= \sum_{i=1}^n \boldsymbol{g}_i^\top (\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}+\boldsymbol{B})^\top \boldsymbol{g}_i \\
&\leq 2\sqrt{xn}\|(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}+\boldsymbol{B})^\top\|_F + 2x\|(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}+\boldsymbol{B})^\top\|_{op}.
\end{aligned}$$

By the union bound, with probability at least $1 - 3e^{-x}$,

$$\|\boldsymbol{G}\boldsymbol{A}\|_F - \|\boldsymbol{G}\boldsymbol{B}\|_F \leq \frac{2\sqrt{xn}\|(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}+\boldsymbol{B})^\top\|_F + 2x\|(\boldsymbol{A}-\boldsymbol{B})(\boldsymbol{A}+\boldsymbol{B})^\top\|_{op}}{2(n-2\sqrt{xn})_+^{1/2}}.$$

Since here $\sqrt{xn} \leq n/4$, the denominator is at least $2(n/2)^{1/2}$ and using the submultiplicativity of the Frobenius norm with $\|\boldsymbol{A}+\boldsymbol{B}\|_F \leq 2$ for the numerator we find

$$\frac{\|\boldsymbol{G}\boldsymbol{A}\|_F - \|\boldsymbol{G}\boldsymbol{B}\|_F}{\|\boldsymbol{A}-\boldsymbol{B}\|_F} \leq 2\frac{\sqrt{xn}+x}{(n/2)^{1/2}} \leq C_{18}\sqrt{x}.$$

□

Proof of (ii). Since (i) proves that the process $Z_{\boldsymbol{A}} = \|\boldsymbol{G}\boldsymbol{A}\|_F$ has subgaussian increment with respect to the Frobenius norm, (ii) follows by Talagrand Majorizing Measure theorem, for example as stated in [172, Theorem 4.1].

The second statement follows by taking a fixed $\boldsymbol{B} \in \mathcal{A}$ and using $|\sqrt{n} - \|\boldsymbol{G}\boldsymbol{B}\|_F| \leq C_{19}\sqrt{x}$ with probability at least $1 - 2e^{-x}$ by [263, Theorem 6.3.2] applied to the block diagonal matrix with n blocks, each block being $\boldsymbol{B}^\top$. □

Proof of (iii). Recall that $\boldsymbol{G}' \in \mathbb{R}^{p\times T}$ has i.i.d. $\mathcal{N}(0,1)$ entries. By application of (ii), it is sufficient to control the Gaussian width

$$\mathbb{E} \sup_{\boldsymbol{B}\in\mathbb{R}^{p\times T}: \boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\in\mathcal{C}, \|\boldsymbol{B}\|_F=1} |\operatorname{Tr}[\boldsymbol{B}^\top \boldsymbol{G}']| = \mathbb{E} \sup_{\boldsymbol{A}\in\mathcal{C}: \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\|_F=1} |\operatorname{Tr}[\boldsymbol{A}^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{G}']|. \tag{3.68}$$

Let $\boldsymbol{A} \in \mathcal{C}$ and let $\boldsymbol{g}_1^\top, ..., \boldsymbol{g}_p^\top$ be the rows of $\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}'$. For any fixed $j \in [p]$, the random vector $\boldsymbol{g}_j \in \mathbb{R}^{T\times 1}$ has $\mathcal{N}_T(\boldsymbol{0}_{1\times T}, \boldsymbol{\Sigma}_{jj}\boldsymbol{I}_{T\times T})$ distribution. By the triangle inequality and the Cauchy-Schwarz inequality we have for some $m, t > 0$

$$
\begin{aligned}
|\operatorname{Tr}[\boldsymbol{A}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}']| &\le \sum_{j=1}^p \|\boldsymbol{A}^\top\boldsymbol{e}_j\|_2\|\boldsymbol{g}_j\|_2 = \|\boldsymbol{A}\|_{2,1}(m+t) + \sum_{j=1}^p \|\boldsymbol{A}^\top\boldsymbol{e}_j\|_2(\|\boldsymbol{g}_j\|_2 - m - t) \\
&\le \|\boldsymbol{A}\|_F\sqrt{k}(m+t) + \|\boldsymbol{A}\|_F\Big(\sum_{j=1}^p(\|\boldsymbol{g}_j\|_2 - m - t)_+^2\Big)^{1/2}
\end{aligned}
$$

where for the second line we used that $\boldsymbol{A} \in \mathcal{C}$. We have $\|\boldsymbol{A}\|_F \le \|\boldsymbol{\Sigma}^{-1/2}\|_{op}$ if $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{A}\|_F = 1$. Next, we now define m such that m^2 is the median of the χ_T^2 distribution, and $t = \sqrt{2\log(p/k)}$. As explained in the proof of Proposition 3.22 around (3.70) we have $m \le \sqrt{T}$ [190] as well as $\mathbb{E}\sum_{j=1}^p(\|\boldsymbol{g}_j\|_2 - m - t)_+^2 \le k$. By the inequality $\sqrt{a} + \sqrt{b} \le \sqrt{2(a+b)}$, (3.68) is bounded from above by $\|\boldsymbol{\Sigma}^{-1/2}\|_{op}(\sqrt{2k(T + 2\log(p/k))} + \sqrt{k}) \le \|\boldsymbol{\Sigma}^{-1/2}\|_{op}\sqrt{8k(T + \log(p/k))}$ and the proof is complete.

□

3.10.2 Ω_2: Control of the noise

Proposition 3.22. *Let $a_+ = \max(0, a)$. If $\boldsymbol{E} \in \mathbb{R}^{n\times T}$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ rows independent of $\boldsymbol{E}$, then*

$$
\begin{aligned}
\sum_{j=1}^p \Big(\frac{\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{e}_j\|_2}{\sigma(1+\eta_1)\sqrt{n\boldsymbol{\Sigma}_{jj}}} - \sqrt{T} - \sqrt{2\log(p/s)}\Big)_+^2 &\le \sum_{j=1}^p \Big(\frac{\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{e}_j\|_2}{\sigma\|\boldsymbol{X}\boldsymbol{e}_j\|_2} - \sqrt{T} - \sqrt{2\log(p/s)}\Big)_+^2 \\
&\le s \qquad (3.69)
\end{aligned}
$$

with probability at least $1 - 4/\{(2\log(p/s)+2)(4\pi\log(p/s)+4)^{1/2}\} - pe^{-n\eta_1^2/2}$. Consequently, on the same event with

$$
\lambda_0 = \Big(\max_{j=1,\dots,p}\boldsymbol{\Sigma}_{jj}^{1/2}\Big)\frac{\sigma(1+\eta_1)}{\sqrt{nT}}\Big(1 + \sqrt{(2/T)\log(p/s)}\Big)
$$

we have $\sum_{j=1}^p(\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{e}_j\|_2 - nT\lambda_0)_+^2 \le \sigma^2(1+\eta_1)^2 n\max_j\boldsymbol{\Sigma}_{jj}s \le sn^2T\lambda_0^2$.

Proof. Since $\boldsymbol{X}\boldsymbol{e}_j$ has i.i.d. $\mathcal{N}(0, \boldsymbol{\Sigma}_{jj})$ entries, $\mathbb{P}(\|\boldsymbol{X}\boldsymbol{e}_j\|_2 \ge \boldsymbol{\Sigma}_{jj}^{1/2}(\sqrt{n}+t)) \le e^{-t^2/2}$ holds by standard bounds on χ_n^2 random variables, e.g., as a consequence of [41, Theorem 5.5]. The choice $t = \eta_1\sqrt{n}$ and the union bound over $\{1, ..., p\}$ provides the first inequality in (3.69).

Since $\boldsymbol{E}$ is independent of $\boldsymbol{X}$, conditionally on $\boldsymbol{X}$ the random variable $\boldsymbol{g}_j := \boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{e}_j/(\sigma\|\boldsymbol{X}\boldsymbol{e}_j\|_2)$ has standard normal distribution $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{I}_{T\times T})$. Since the conditional distribution does not depend on $\boldsymbol{X}$, the unconditional distribution of $\boldsymbol{g}_j$ is also $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{I}_{T\times T})$. By [41, Theorem 10.17] applied to the 1-Lipschitz function $\boldsymbol{g}_j \mapsto \|\boldsymbol{g}_j\|_2$, inequality $\mathbb{P}(\|\boldsymbol{g}_j\|_2 \ge m_j + t) \le \mathbb{P}(Z_j \ge t)$ holds, where $Z_j \sim \mathcal{N}(0,1)$ and m_j is the median of the random variable $\|\boldsymbol{g}_j\|_2$. It follows that for any $t > 0$

$$W := \sum_{j=1}^{p}(\|\boldsymbol{g}_j\|_2 - m_j - t)_+^2 \quad \text{satisfies} \quad \mathbb{E}[W] \leq \mathbb{E}\sum_{j=1}^{p}(Z_j - t)_+^2 \leq \frac{4pe^{-t^2/2}}{(t^2+2)(2\pi t^2+4)^{1/2}}, \tag{3.70}$$

where the second inequality follows from [25, Lemma G.1]. By the argument in [190], the median of the χ_T^2 distribution is smaller than T so that $m_j \leq \sqrt{T}$. Furthermore, for $t = (2\log(p/s))^{1/2}$ we have $\mathbb{E}[W] \leq sq$ where $q^{-1} = (t^2+2)(2\pi t^2+4)^{1/2}/4 > 1$. The second inequality in (3.69) thus holds with probability at least $1 - q$ by Markov's inequality $\mathbb{P}(W > \mathbb{E}[W]q^{-1}) \leq q$. □

3.10.3 Ω_3: Restricted Isometry Properties

The following bound is well known in the literature on the RIP property for Gaussian matrices, as a consequence of Gordon's Lemma, see, e.g., [279]. We provide the argument here for completeness.

Proposition 3.23 (Bound on upper sparse eigenvalues of random matrices, Gordon's lemma)**.** *Let $p \geq n$. If $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}(0, \boldsymbol{\Sigma})$ rows, then*

(i) for any set $B \subset [p]$ we have

$$\mathbb{P}\Big(\max_{\boldsymbol{v}\in\mathbb{R}^p:\operatorname{supp}(\boldsymbol{v})\subset B}\Big|\frac{\|\boldsymbol{X}\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|} - 1\Big| \leq \sqrt{|B|/n} + t\Big) \geq 1 - 2e^{-nt^2/2}$$

by Gordon's escape through the mesh theorem and its consequence, cf. for instance in [71, Theorem II.13] applied to the Gaussian matrix $\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$ and the intersection of the unit ball with the $|B|$ dimensional linear span of $\{\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_j, j \in B\}$.

(ii) Let $\eta_4 \in (0,1)$ be a constant. If k is such that $\sqrt{k/n} \leq \eta_4/2$ and $k\log(ep/k)/n \leq \eta_4^2/16$, then simultaneously for all B with $|B| \leq k$

$$\mathbb{P}\left(\max_{B\subset[p]:|B|\leq k}\left(\max_{\boldsymbol{v}\in\mathbb{R}^p:\operatorname{supp}(\boldsymbol{v})\subset B}\Big|\frac{\|\boldsymbol{X}\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|} - 1\Big|\right) > \eta_4\right) \leq 2\exp(-n\eta_4^2/16).$$

Proof. For (ii), by the union bound with $t = \eta_4/2$ we have

$$\mathbb{P}\left(\max_{B\subset[p]:|B|\leq k}\left(\max_{\boldsymbol{v}\in\mathbb{R}^p:\operatorname{supp}(\boldsymbol{v})\subset B}\Big|\frac{\|\boldsymbol{X}\boldsymbol{v}\|}{\sqrt{n}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|} - 1\Big|\right) > \eta_4\right) \leq 2\binom{p}{k}e^{-n\eta_4^2/8}.$$

Since $\log\binom{p}{k} \leq k\log(ep/k)$, the right hand side is bounded from above by $2\exp(-n\eta_4^2/16)$ by assumption on k. □

3.11 Proof of Theorems 3.3 and 3.4

Proof. By replacing $\boldsymbol{b}$ by $\boldsymbol{b}/\|\boldsymbol{b}\|_2$ if necessary, we assume that $\|\boldsymbol{b}\|_2 = 1$ without loss of generality. The proof is based on the decomposition

$$\begin{aligned}&(n\sigma^2)^{-1/2}\big(n\boldsymbol{a}^T(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b} + \boldsymbol{z}_0^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\big)\\ &= (n\sigma^2)^{-1/2}\boldsymbol{z}_0^T\boldsymbol{E}\boldsymbol{b} + \boldsymbol{r}^\top\boldsymbol{b} + \widetilde{\boldsymbol{r}}^\top\boldsymbol{b}\end{aligned}$$

with the remainder terms $\boldsymbol{r}^\top \boldsymbol{b}$ and $\widetilde{\boldsymbol{r}}^\top \boldsymbol{b}$ defined by the random vectors $\boldsymbol{r}, \widetilde{\boldsymbol{r}} \in \mathbb{R}^T$

$$\boldsymbol{r}^\top = (n\sigma^2)^{-1/2} \boldsymbol{z}_0^T \boldsymbol{E}\left[(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - (\boldsymbol{I}_{T\times T})\right],$$
$$\widetilde{\boldsymbol{r}}^\top = (n\sigma^2)^{-1/2}\left[\boldsymbol{a}^T(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}) - \boldsymbol{z}_0^T \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\right](\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}.$$

Since $\boldsymbol{z}_0 \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{I}_{n\times n})$ is independent of $\boldsymbol{E}\boldsymbol{b} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n\times n})$ we have $\boldsymbol{z}_0^\top \boldsymbol{E}\boldsymbol{b}/\|\boldsymbol{z}_0\|_2 \sim \mathcal{N}(0, \sigma^2)$. Since $\|\boldsymbol{z}_0\|_2^2 n^{-1} \xrightarrow{\mathbb{P}} 1$ by the law of large numbers, we obtain that $(n\sigma^2)^{-1/2}\boldsymbol{z}_0^\top \boldsymbol{E}\boldsymbol{b} \xrightarrow{d} \mathcal{N}(0,1)$ by Slutsky's theorem. To conclude with another application of Slutsky's theorem, it remains to prove that $\|\boldsymbol{r}\|_2$ and $\|\widetilde{\boldsymbol{r}}\|_2$ both converge to 0 in probability, and to prove that for the denominator, $(n\sigma^2)^{-1/2}\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2 \xrightarrow{\mathbb{P}} 1$.

For $\boldsymbol{r}$, on Ω_* we have $\|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - (\boldsymbol{I}_{T\times T})\|_{op} \le \bar{s}/(n-\bar{s})$ by Proposition 3.2(iii) and Lemma 3.14. It follows that

$$\begin{aligned}
\mathbb{E}[\min(1, \|\boldsymbol{r}\|_2)] &\le \mathbb{P}(\Omega_*^c) + \mathbb{E}\big[I\{\Omega_*\}(n\sigma^2)^{-1/2}\|\boldsymbol{E}^\top \boldsymbol{z}_0\|_2 \|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - (\boldsymbol{I}_{T\times T})\|_{op}\big] \\
&\le \mathbb{P}(\Omega_*^c) + (n\sigma^2)^{-1/2}\big(\bar{s}/(n-\bar{s})\big)\mathbb{E}\big[\|\boldsymbol{E}^\top \boldsymbol{z}_0\|_2\big] \\
&\le \mathbb{P}(\Omega_*^c) + (n\sigma^2)^{-1/2}\big(\bar{s}/(n-\bar{s})\big)\sqrt{nT\sigma^2} \\
&= \mathbb{P}(\Omega_*^c) + \big(\bar{s}/(n-\bar{s})\big)\sqrt{T}
\end{aligned}$$

by Jensen's inequality and $\mathbb{E}[\|\boldsymbol{E}^\top \boldsymbol{z}_0\|_2^2] = nT\sigma^2$. The last line converges to 0 by Lemma 3.12 and Assumption (A1). Since $W_n \xrightarrow{\mathbb{P}} 0$ if and only if $\mathbb{E}[\min(1, |W_n|)] \to 0$, this proves the convergence $\|\boldsymbol{r}\|_2 \xrightarrow{\mathbb{P}} 0$.

For $\widetilde{\boldsymbol{r}}$, we use that

$$\mathbb{E}[\min(1, \|\widetilde{\boldsymbol{r}}\|_2)] \le \mathbb{P}(\Omega_*^c) + \mathbb{E}[I\{\Omega_*\}\|\widetilde{\boldsymbol{r}}\|_2]$$

with $\mathbb{P}(\Omega_*^c) \to 0$ as above. For the second term, on Ω_* we have $\|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op} \le \|\boldsymbol{I}_{T\times T}\|_{op} + \|\boldsymbol{I}_{T\times T} - (\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op} \le 1 + \bar{s}/(n-\bar{s}) = (1-\bar{s}/n)^{-1}$ by Proposition 3.2 and Lemma 3.14. It follows that

$$\begin{aligned}
I\{\Omega_*\}\|\widetilde{\boldsymbol{r}}\|_2 &\le I\{\Omega_*\}\tfrac{1}{\sigma\sqrt{n}}(1 - \tfrac{\bar{s}}{n})^{-1}\|(n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{a} - (\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{X}^\top \boldsymbol{z}_0\|_2 \\
&= I\{\Omega_*\}\tfrac{1}{\sigma\sqrt{n}}(1 - \tfrac{\bar{s}}{n})^{-1}\Big[\sum_{t=1}^T \Big(\mathrm{Tr}[\mathsf{D}^*(\boldsymbol{e}_t)] - \boldsymbol{z}_0^\top \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{e}_t\Big)^2\Big]^{1/2}.
\end{aligned}$$

where the equality is a consequence of Lemma 3.19. Since $\mathsf{D}^* = \mathsf{D} - \mathsf{D}^{**}$, and using the inequalities $(a+b)^2 \le 2a^2 + 2b^2$ and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$,

$$\begin{aligned}
&\mathbb{E}\Big[I\{\Omega_*\}\sum_{t=1}^T \Big(\boldsymbol{z}_0^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{X}\boldsymbol{e}_t - \mathrm{Tr}[\mathsf{D}^*(\boldsymbol{e}_t)]\Big)^2\Big]^{1/2} \\
&\le \Big[2\mathbb{E}\Big(I\{\Omega_*\}\sum_{t=1}^T \big[\boldsymbol{z}_0^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{X}\boldsymbol{e}_t - \mathrm{Tr}[\mathsf{D}(\boldsymbol{e}_t)]\big]^2\Big) + 2\mathbb{E}\Big(I\{\Omega_*\}\sum_{t=1}^T \mathrm{Tr}[\mathsf{D}^{**}(\boldsymbol{e}_t)]^2\Big)\Big]^{1/2} \\
&\le o((n\sigma^2)^{1/2}) + O\big(\sigma \min(T, (sT)^{1/2})\big)
\end{aligned}$$

by Lemma 3.20 and inequality (3.56) in Lemma 3.19. Combining the above displays yields

$$\mathbb{E}[\min(1, \|\widetilde{\boldsymbol{r}}\|_2)] \le \mathbb{P}(\Omega_*^c) + (n\sigma^2)^{-1/2}(1 - \tfrac{\bar{s}}{n})^{-1}\big[o((n\sigma^2)^{1/2}) + O\big(\sigma(sT)^{1/2}\big)\big] = o(1),$$

or equivalently $\|\widetilde{\boldsymbol{r}}\|_2 \xrightarrow{\mathbb{P}} 0$.

Let us prove Theorem 3.4, that is $(n\sigma^2)^{-1/2}\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2 \xrightarrow{\mathbb{P}} 1$. By the law of large numbers, we have $\|\boldsymbol{E}\boldsymbol{b}\|_2^2/(n\sigma^2) \xrightarrow{\mathbb{P}} 1$, so it suffices to show that

$$(n\sigma^2)^{-1/2}\|\boldsymbol{E}\big[(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - \boldsymbol{I}_{T\times T}\big]\boldsymbol{b} - \boldsymbol{X}(\boldsymbol{B}^* - \widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2 \xrightarrow{\mathbb{P}} 0.$$

Techniques similar to those above show that $(n\sigma^2)^{-1/2}\|\boldsymbol{E}\big[(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1} - \boldsymbol{I}_{T\times T}\big]\boldsymbol{b}\|_2 \xrightarrow{\mathbb{P}} 0$ by Proposition 3.2(iii), and that $(n\sigma^2)^{-1/2}\|\boldsymbol{X}(\boldsymbol{B}^* - \widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2 \xrightarrow{\mathbb{P}} 0$ by Lemma 3.13 and $\bar{R} \to 0$.

An application of Slutsky's lemma completes the proof of Theorem 3.3. $\square$

3.12 Proof for χ^2_T limits, and confidence ellipsoid with nominal coverage

Lemma 3.24 (Differentiation with respect to $\boldsymbol{E}$)**.** *Here, we consider differentiation with respect to $\boldsymbol{E}$ for fixed $\boldsymbol{X}$. We have*

$$\mathbb{E}\Big[I\{\Omega_*\}\|\boldsymbol{E}^\top\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) - \sigma^2\widehat{\mathbf{A}}\|_F^2\Big] \le \sigma^2 nT\bar{R}^2 + \sigma^4 nT.$$

Proof. Let $\boldsymbol{F}: \mathbb{R}^{n\times T} \to \mathbb{R}^{n\times T}$ be the function $\boldsymbol{F}: \boldsymbol{E} \mapsto \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)$. The function $\boldsymbol{F}$ is 1-Lipschitz by [24, Proposition 3.1]. Furthermore, $\|\boldsymbol{F}\|_F \le \sqrt{n}\bar{R}$ on Ω_* by Lemma 3.13, so that if $\boldsymbol{\Pi}: \mathbb{R}^{n\times T} \to \mathbb{R}^{n\times T}$ is the convex projection onto the Frobenius ball of radius $\sqrt{n}\bar{R}$, the composition $\bar{\boldsymbol{F}} = \boldsymbol{\Pi} \circ \boldsymbol{F}$ coincides with $\boldsymbol{F}$ on Ω_*. The function $\bar{\boldsymbol{F}}$ is also 1-Lipschitz by composition of two 1-Lipschitz functions, and since Ω_* is open, the derivatives of $\bar{\boldsymbol{F}}$ and $\boldsymbol{F}$ with respect to $\boldsymbol{E}$ coincide in Ω_* where the derivatives exist (this existence of the derivatives is granted almost everywhere by Rademacher's theorem).

For any $t, t' \in [T]$, by the main result of [25] applied to the function $\boldsymbol{E}\boldsymbol{e}_{t'} \mapsto \bar{\boldsymbol{F}}\boldsymbol{e}_t$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(\boldsymbol{e}_{t'}^\top\boldsymbol{E}^\top\bar{\boldsymbol{F}}\boldsymbol{e}_t - \sigma^2\sum_{i=1}^n \frac{\partial \boldsymbol{e}_i^\top\bar{\boldsymbol{F}}\boldsymbol{e}_t}{\partial E_{it'}}\big)^2\Big] \\
&= \sigma^2\mathbb{E}\left[\|\bar{\boldsymbol{F}}\boldsymbol{e}_t\|_2^2\right] + \sigma^4\mathbb{E}\Big[\sum_{i=1}^n\sum_{i'=1}^n \Big(\frac{\partial}{\partial E_{i't'}}\boldsymbol{e}_i^\top\bar{\boldsymbol{F}}\boldsymbol{e}_t\Big)\Big(\frac{\partial}{\partial E_{it'}}\boldsymbol{e}_{i'}^\top\bar{\boldsymbol{F}}\boldsymbol{e}_t\Big)\Big] \\
&\le \sigma^2\mathbb{E}\left[\|\bar{\boldsymbol{F}}\boldsymbol{e}_t\|_2^2\right] + \sigma^4\mathbb{E}\Big[\sum_{i=1}^n\sum_{i'=1}^n \Big(\frac{\partial}{\partial E_{i't'}}\boldsymbol{e}_i^\top\bar{\boldsymbol{F}}\boldsymbol{e}_t\Big)^2\Big].
\end{aligned}$$

We now sum the above inequalities for all $t, t' \in [T]$ to find

$$\begin{aligned}
&\sum_{t=1}^T \sum_{t'=1}^T \mathbb{E}\Big[\big(\boldsymbol{e}_{t'}^\top \boldsymbol{E}^\top \bar{\boldsymbol{F}} \boldsymbol{e}_t - \sigma^2 \sum_{i=1}^n \frac{\partial \boldsymbol{e}_i^\top \bar{\boldsymbol{F}} \boldsymbol{e}_t}{\partial E_{it'}}\big)^2\Big] \\
&\le \sigma^2 T \mathbb{E}[\|\bar{\boldsymbol{F}}\|_F^2] + \sigma^4 \mathbb{E}\Big[\sum_{t=1}^T \sum_{t'=1}^T \sum_{i=1}^n \sum_{i'=1}^n \Big(\frac{\partial}{\partial E_{i't'}} \boldsymbol{e}_i^\top \bar{\boldsymbol{F}} \boldsymbol{e}_t\Big)^2\Big] \\
&\le \sigma^2 T n \bar{R}^2 + \sigma^4 n T,
\end{aligned}$$

where for the last inequality we used that $\|\bar{\boldsymbol{F}}\|_F \le \bar{R}\sqrt{n}$ by construction of $\bar{\boldsymbol{F}}$ and that $\bar{\boldsymbol{F}} : \mathbb{R}^{n\times T} \to \mathbb{R}^{n\times T}$ is 1-Lipschitz, so that the Frobenius norm of the Jacobian of $\bar{\boldsymbol{F}}$ (which is a matrix of size $(nT) \times (nT)$) is at most $\sqrt{nT}$. Finally, on Ω_* we have $\boldsymbol{F} = \bar{\boldsymbol{F}}$ and their derivatives coincide, and by differentiating the KKT conditions of $\widehat{\boldsymbol{B}}$ we find $\sum_{i=1}^n \frac{\partial \boldsymbol{e}_i^\top \boldsymbol{F} \boldsymbol{e}_t}{\partial E_{it'}} = \widehat{\mathbf{A}}_{tt'}$ on Ω_* for $\boldsymbol{F} = \boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)$. This completes the proof. ☐

Theorem 3.25. *Let $\boldsymbol{a} \in \mathbb{R}^p$ with $\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{a}\|_2 = 1$. Let $\boldsymbol{\xi}$ be defined in (3.37) and $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}\|_F^2/(nT)$. Then under Assumption (A1), $|\hat{\sigma}/\sigma - 1| = o_{\mathbb{P}}(T^{-1/2})$ as well as*

$$\max\{(\sigma^2 n)^{-1/2}, (\hat{\sigma}^2 n)^{-1/2}\}\big\|\boldsymbol{\xi} - \sqrt{n}\boldsymbol{E}^\top \boldsymbol{z}_0 \|\boldsymbol{z}_0\|_2^{-1}\big\|_2 = o_{\mathbb{P}}(1). \tag{3.71}$$

Proof of Theorem 3.25. By definition of $\boldsymbol{\xi}$ we have

$$(n\sigma^2)^{-1/2}\|\boldsymbol{\xi} - \boldsymbol{E}^\top \boldsymbol{z}_0\|_2 = (n\sigma^2)^{-1/2}\|(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{X}^\top \boldsymbol{z}_0 - (n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{a}\|_2$$

which converges to 0 in probability by Lemma 3.20. Next, with $\chi_T^2 = \sigma^{-2}\big\|\boldsymbol{E}^\top \boldsymbol{z}_0 \|\boldsymbol{z}_0\|_2^{-1}\big\|_2^2$,

$$(n\sigma^2)^{-1/2}\big\|\sqrt{n}\boldsymbol{E}^\top \boldsymbol{z}_0\|\boldsymbol{z}_0\|_2^{-1} - \boldsymbol{E}^\top \boldsymbol{z}_0\big\|_2 = (\chi_T^2)^{1/2}\big|1 - n^{-1/2}\|\boldsymbol{z}_0\|_2\big|. \tag{3.72}$$

By the Cauchy-Schwarz inequality we have $\mathbb{E}[(\chi_T^2)^{1/2}\big|1-n^{-1/2}\|\boldsymbol{z}_0\|_2\big|] \le \sqrt{T/n}\,\mathbb{E}[(\|\boldsymbol{z}_0\|_2 - \sqrt{n})]^{1/2}$. Combining Theorem 3.1.1 and Equation 2.15 in [263] yields $\mathbb{E}[(\|\boldsymbol{z}_0\|_2 - \sqrt{n})]^{1/2} \le C$ for some absolute constant C. Thus, by Assumption (A1) we have $T/n \to 0$ so that (3.72) converges to 0 in L^1, hence in probability. This proves $(\sigma^2 n)^{-1/2}\big\|\boldsymbol{\xi} - \sqrt{n}\boldsymbol{E}^\top \boldsymbol{z}_0\|\boldsymbol{z}_0\|_2^{-1}\big\|_2 = o_{\mathbb{P}}(1)$.

We now prove the same bound with $\sigma^2 n$ replaced by $\hat{\sigma}^2 n$. Let $\Omega_8 = \{|\|\boldsymbol{E}\|_F/\sigma - \sqrt{nT}| \le \sqrt{\log n}\}$. Then $\mathbb{P}(\Omega_8) \to 1$ by [263, Theorem 3.1.1] and

$$\begin{aligned}
I\{\Omega_8 \cap \Omega_*\}|\hat{\sigma}/\sigma - 1| &\le I\{\Omega_*\}\|\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_F(\sigma^2 nT)^{-1/2} + I\{\Omega_8\}|\sqrt{nT} - \|\boldsymbol{E}\|_F/\sigma|(nT)^{-1/2} \\
&\le (1-\eta_3)\bar{R}/\sqrt{\sigma^2 T} + (nT)^{-1/2}\sqrt{\log n}
\end{aligned} \tag{3.73}$$

by Lemma 3.13 for the first term. This proves that $|\hat{\sigma}/\sigma - 1| = o_{\mathbb{P}}(T^{-1/2})$. under Assumption (A1) so that using $\frac{1}{2}|\frac{1}{u} - 1| \le |u - 1|$ for $u \in [\frac{1}{2}, \frac{3}{2}]$ we obtain for n large enough

$$(1/2)I\{\Omega_8 \cap \Omega_*\}|\sigma/\hat{\sigma} - 1| \le I\{\Omega_8 \cap \Omega_*\}|\hat{\sigma}/\sigma - 1| \le (3.73).$$

Hence $\sigma/\hat{\sigma} = 1 + o_{\mathbb{P}}(1)$, thus

$$\begin{aligned}
(n\hat{\sigma}^2)^{-1/2}\big\|\sqrt{n}\boldsymbol{E}^\top \boldsymbol{z}_0\|\boldsymbol{z}_0\|_2^{-1} - \boldsymbol{E}^\top \boldsymbol{z}_0\big\|_2 &= (\sigma/\hat{\sigma})(n\sigma^2)^{-1/2}\big\|\sqrt{n}\boldsymbol{E}^\top \boldsymbol{z}_0\|\boldsymbol{z}_0\|_2^{-1} - \boldsymbol{E}^\top \boldsymbol{z}_0\big\|_2 \\
&= (1 + o_{\mathbb{P}}(1))o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).
\end{aligned}$$

☐

Theorem 3.6. *Define the observable positive semi-definite matrix* $\widehat{\mathbf{\Gamma}} = (\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})^\top(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}}) \in \mathbb{R}^{T\times T}$ *as well as*

$$\boldsymbol{\xi} = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{z}_0 + (n\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \boldsymbol{a}. \tag{3.37}$$

Then under Assumption (A1), there exists a random variable χ^2_T *with chi-square distribution with* T *degrees of freedom such that*

$$\sqrt{1-\tfrac{T}{n}}\Big\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\Big\|_2 - \sqrt{\chi^2_T} \le o_{\mathbb{P}}(1) + O_{\mathbb{P}}\Big(\min\Big\{\frac{T}{\sqrt{n}}, \frac{s^2\log^2(p/s)}{n\sqrt{T}}\Big\}\Big)$$

as well as

$$-o_{\mathbb{P}}(1) - O_{\mathbb{P}}\Big(\frac{T}{\sqrt{n}} + \frac{sT + s\log(p/s)}{n}\sqrt{T}\Big) \le \sqrt{1-\tfrac{T}{n}}\Big\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\Big\|_2 - \sqrt{\chi^2_T}.$$

Consequently,

(i) $(1-\frac{T}{n})^{\frac12}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 - (\chi^2_T)^{1/2} \le o_{\mathbb{P}}(1)$ *holds if additionally* $\min\{\frac{T^2}{n}, \frac{\log^8 p}{n}\} \to 0$, *and*

(ii) $(1-\frac{T}{n})^{\frac12}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 - (\chi^2_T)^{1/2} \ge o_{\mathbb{P}}(1)$ *holds if additionally* $\frac{T^2}{n} + \frac{sT+s\log(p/s)}{n}\sqrt{T} \to 0$.

Proof of Theorem 3.6. Theorem 3.25 applied with $\boldsymbol{z} = \boldsymbol{z}_0\|\boldsymbol{z}_0\|_2^{-1}$ yields the bound $(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi} - \sqrt{n}\boldsymbol{E}^\top\boldsymbol{z}\|_2 = o_{\mathbb{P}}(1)$. The proof then follows from Lemma 3.26. □

Lemma 3.26. *Let Assumption (A1) be fulfilled. Let* $\boldsymbol{z}, \boldsymbol{\xi}$ *be random vectors valued in* $\mathbb{R}^n$. *Assume that* $\boldsymbol{z}$ *is a measurable function of* $\boldsymbol{X}$ *with* $\mathbb{P}(\|\boldsymbol{z}\|_2 = 1) = 1$ *and let* $\boldsymbol{P}_{\boldsymbol{z}}^\perp = \boldsymbol{I}_n - \boldsymbol{z}\boldsymbol{z}^\top$. *Then the random variable* $F_{T,n-T} = \frac{n-T}{T}\|(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2^2$ *has the* F *distribution with degrees of freedom* T *and* $n-T$, *and the following holds:*

(i) $\sqrt{TF_{T,n-T}} = \sqrt{\chi^2_T} + o_{\mathbb{P}}(1)$ *as* $n \to +\infty$ *when* $T/n \to 0$ *where* χ^2_T *is a random variable with chi-square distribution with* T *degrees of freedom,*

(ii) $\mathbb{P}(\lambda_{\min}(\widehat{\mathbf{\Gamma}}) \ge n\sigma^2/2) \to 1$,

(iii) $\sqrt{n-T}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \sqrt{TF_{T,n-T}} \le o_{\mathbb{P}}(1) + O_{\mathbb{P}}(\frac{T}{\sqrt{n}})$,

(iv) $\sqrt{n-T}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \sqrt{TF_{T,n-T}} \ge -o_{\mathbb{P}}(1) - O_{\mathbb{P}}(\frac{T}{\sqrt{n}} + \frac{sT+s\log(p/s)}{n}\sqrt{T})$,

(v) $\sqrt{n-T}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \sqrt{TF_{T,n-T}} \le o_{\mathbb{P}}(1) + O_{\mathbb{P}}(\frac{s(s+T)\log^2(p/s)}{n\sqrt{T}})$.

Consequently, if $(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi} - \sqrt{n}\boldsymbol{E}^\top\boldsymbol{z}\|_2 = o_{\mathbb{P}}(1)$ *then*

$$(1-\tfrac{T}{n})^{1/2}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 \le (\chi^2_T)^{1/2} + o_{\mathbb{P}}(1) + O_{\mathbb{P}}\big(\min\{\tfrac{T}{\sqrt{n}}, \tfrac{\log^2(p/s)}{n^{1/4}}\}\big) \tag{3.74}$$

$$(1-\tfrac{T}{n})^{1/2}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2 \ge (\chi^2_T)^{1/2} - o_{\mathbb{P}}(1) - O_{\mathbb{P}}\big(\tfrac{T}{\sqrt{n}} + \tfrac{(sT+s\log\frac{p}{s})\sqrt{T}}{n}\big). \tag{3.75}$$

Proof of Lemma 3.26. For (i), we introduce the quantity

$$H := (n-1)\|(\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1/2}\boldsymbol{E}^\top \boldsymbol{z}\|_2^2 = (n-1)\boldsymbol{g}^\top \boldsymbol{W}^{-1}\boldsymbol{g} \tag{3.76}$$

where $\boldsymbol{g} = \sigma^{-1}\boldsymbol{E}^\top \boldsymbol{z}$ and $\boldsymbol{W} = \sigma^{-2}\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E}$. Since $\boldsymbol{E}$ and $\boldsymbol{z}$ are independent and since $\|\boldsymbol{z}\|_2 = 1$, $\boldsymbol{g}$ has distribution $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{I}_{T\times T})$. $\boldsymbol{P}_{\boldsymbol{z}}^\perp$ can be orthogonally diagonalized as $\boldsymbol{Q}\big(\sum_{i=1}^{n-1}\boldsymbol{e}_i\boldsymbol{e}_i{}^\top\big)\boldsymbol{Q}^\top$ where $\boldsymbol{Q}$ is an $n\times n$ orthogonal matrix, thus $\boldsymbol{W} = \sum_{i=1}^{n-1}\boldsymbol{n}_i\boldsymbol{n}_i^\top$ where the random vectors $\boldsymbol{n}_i = \sigma^{-1}\boldsymbol{E}^\top\boldsymbol{Q}\boldsymbol{e}_i$ are iid with standard normal $\mathcal{N}_T(\boldsymbol{0}, \boldsymbol{I}_{T\times T})$ distribution. Therefore $\boldsymbol{W}$ has the Wishart distribution with identity covariance and $n-1$ degrees-of-freedom. Since $\boldsymbol{E}^\top\boldsymbol{z}$ and $\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp$ are independent, so are $\boldsymbol{E}^\top\boldsymbol{z}$ and $(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp)(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp)^\top = \boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E}$, thus $\boldsymbol{g}$ and $\boldsymbol{W}$ are independent. By [113, Theorem 5.8] H has the Hotelling distribution with parameters $T, n-1$, and

$$\frac{n-1-T+1}{T}\frac{H}{n-1} \sim F_{T,n-1-T+1} = F_{T,n-T}$$

where the right-hand side is the F distribution with degrees-of-freedom T and $n-T$. Furthermore, since $F_{T,n-T} = \frac{\chi^2_T/T}{\chi^2_{n-T}/(n-T)}$ for some random variables having chi-square distributions with respective parameter T and $n-T$, we have

$$|\sqrt{TF_{T,n-T}} - \sqrt{\chi^2_T}| = |\sqrt{\chi^2_T/(\chi^2_{n-T}/(n-T))} - \sqrt{\chi^2_T}| = O_{\mathbb{P}}(\sqrt{T})|1-\sqrt{\chi^2_{n-T}/(n-T)}|$$

where the last equality follows from $\mathbb{E}[(\chi^2_T)^{1/2}] \le \mathbb{E}[\chi^2_T]^{1/2} = \sqrt{T}$ and the a.s. convergence of $\chi^2_{n-T}/(n-T)$ to 1. Furthermore $|1-|a|| \le |1-a^2|$ and the Central Limit Theorem yield

$$|1-\sqrt{\chi^2_{n-T}/(n-T)}| \le \frac{|\chi^2_{n-T}-(n-T)|}{n-T} = O_{\mathbb{P}}((n-T)^{-1/2}).$$

Thus $\sqrt{TF_{T,n-T}} = (\chi^2_T)^{1/2} + O_{\mathbb{P}}((\frac{n}{T}-1)^{-1/2})$, and since $\frac{n}{T}\to 0$ we have $\sqrt{TF_{T,n-T}} = (\chi^2_T)^{1/2} + o_{\mathbb{P}}(1)$ and $\mathbb{P}(\sqrt{TF_{T,n-T}} \le q_{T,\alpha}) \to 1-\alpha$ by Proposition 3.7. This proves (i).

Next we exhibit a lower bound on the eigenvalues of $\widehat{\boldsymbol{\Gamma}}$. Let $\boldsymbol{H} = \widehat{\boldsymbol{B}} - \boldsymbol{B}^*$ and consider the decomposition

$$\widehat{\boldsymbol{\Gamma}} = \boldsymbol{E}^\top\boldsymbol{E} + (\boldsymbol{X}\boldsymbol{H})^\top(\boldsymbol{X}\boldsymbol{H}) - [\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H} + (\boldsymbol{X}\boldsymbol{H})^\top\boldsymbol{E}]. \tag{3.77}$$

Since $(\boldsymbol{X}\boldsymbol{H})^\top(\boldsymbol{X}\boldsymbol{H})$ is positive semidefinite we have

$$\widehat{\boldsymbol{\Gamma}} \ \succeq \ \boldsymbol{E}^\top\boldsymbol{E} - 2\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H}\|_{op}\boldsymbol{I}_{T\times T}. \tag{3.78}$$

Since $\boldsymbol{E}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, if $s_{\min}(\boldsymbol{E})$ and $s_{\max}(\boldsymbol{E})$ denote the smallest and greatest singular values of $\boldsymbol{E}$ we have $\sigma(\sqrt{n}-\sqrt{T}) \le \mathbb{E}[s_{\min}(\boldsymbol{E})] \le \mathbb{E}[s_{\max}(\boldsymbol{E})] \le \sigma(\sqrt{n}+\sqrt{T})$ by [71, Theorem II.13]. Since $s_{\min}(\boldsymbol{E})$ and $s_{\max}(\boldsymbol{E})$ are 1-Lipschitz functions of $\boldsymbol{E}$ when considered as a vector in $\mathbb{R}^{nT}$, Gaussian concentration as stated in [104, Theorem B.6] yields the existence of exponential random variables $Z_1, Z_2 \sim \text{Exp}(1)$ such that almost surely

$$\sigma(\sqrt{n}-\sqrt{T}-\sqrt{2Z_1}) \le s_{\min}(\boldsymbol{E}) \le s_{\max}(\boldsymbol{E}) \le \sigma(\sqrt{n}+\sqrt{T}+\sqrt{2Z_2}).$$

Letting $Z = 2\max(Z_1, Z_2)$, we have

$$\sigma^2(\sqrt{n} - \sqrt{T} - \sqrt{Z})_+^2 \boldsymbol{I}_{T\times T} \preceq \boldsymbol{E}^\top \boldsymbol{E} \preceq \sigma^2(\sqrt{n} + \sqrt{T} + \sqrt{Z})^2 \boldsymbol{I}_{T\times T}.$$

Thanks to (3.78) and the inequality $(1-x)_+^2 \geq 1 - 2x$ for $x \geq 0$ we have

$$\widehat{\boldsymbol{\Gamma}} \succeq \sigma^2 n[1 - 2(\sqrt{T/n} + \sqrt{Z/n}) - 2\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}/(\sigma^2 n)]\boldsymbol{I}_{T\times T}. \tag{3.79}$$

On the event $\Omega_9 = \{1 - 2(\sqrt{T/n} + \sqrt{Z/n}) - 2\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}/(\sigma^2 n) > 1/2\}$ we have $\lambda_{\min}(\widehat{\boldsymbol{\Gamma}}) \geq \lambda_{\min}(\boldsymbol{E}^\top \boldsymbol{E} - 2\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}\boldsymbol{I}_{T\times T}) \geq \sigma^2 n/2$. We now proceed to show that $\mathbb{P}(\Omega_9) \to 1$. We have by the triangle inequality for the norm $\mathbb{E}[(\cdot)^2]^{1/2}$ that

$$\begin{aligned} &\mathbb{E}\Big[I\{\Omega_*\}\Big(\sqrt{T/n} + \sqrt{Z/n} + \|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}/(\sigma^2 n)\Big)^2\Big]^{1/2} \\ &\leq \sqrt{T/n} + \mathbb{E}[Z]^{1/2}/\sqrt{n} + \bar{s}/n + \mathbb{E}[I\{\Omega_*\}\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H} - \sigma^2\widehat{\mathbf{A}}\|_{op}^2/(\sigma^2 n)^2]^{1/2} \\ &\leq \sqrt{T/n} + \mathbb{E}[Z]^{1/2}/\sqrt{n} + \bar{s}/n + [(T/n)(1 + \bar{R}^2/\sigma^2)]^{1/2} \end{aligned} \tag{3.80}$$

where we used Proposition 3.2(ii) and Lemma 3.14 to bound $\|\widehat{\mathbf{A}}\|_{op}$ from above by $\bar{s}$ on Ω_* for the first inequality, and Lemma 3.24 the second inequality. Hence under Assumption (A1), the previous display converges to 0. Next, $\mathbb{P}(\Omega_9^c) = \mathbb{P}(\Omega_9^c \cap \Omega_*^c) + \mathbb{P}(\Omega_9^c \cap \Omega_*)$, Markov's inequality and an application of Jensen's inequality yield

$$\begin{aligned} \mathbb{P}(\Omega_9^c) &= \mathbb{P}(\Omega_*^c \cap \Omega_9^c) + \mathbb{P}\Big(1/4 \leq I\{\Omega_*\}\big(\sqrt{T/n} + \sqrt{Z/n} + \|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}/(\sigma^2 n)\big)\Big) \\ &\leq \mathbb{P}(\Omega_*^c) + 4\mathbb{E}\Big[I\{\Omega_*\}\big(\sqrt{T/n} + \sqrt{Z/n} + \|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}/(\sigma^2 n)\big)\Big] \leq \mathbb{P}(\Omega_*^c) + 4(3.80) \end{aligned}$$

where 4(3.80) refers to four times the quantity (3.80) which converges to 0. Thus the event Ω_9 has probability approaching one and claim (ii) follows.

We now prove (iii)-(v). Let $\Omega(n)$ be a sequence of events with $\mathbb{P}(\Omega(n)) \to 1$, V_n be any sequence of random variables and a_n be any deterministic sequence of real numbers. It is easily seen that $I\{\Omega(n)\}V_n = o_{\mathbb{P}}(a_n)$ implies $V_n = o_{\mathbb{P}}(a_n)$ and $I\{\Omega(n)\}V_n = O_{\mathbb{P}}(a_n)$ implies $V_n = O_{\mathbb{P}}(a_n)$. This observation will allow us to transition seamlessly from bounds on $I\{\Omega(n)\}V_n$ to bounds on V_n by choosing, e.g., $\Omega(n) = \Omega_* \cap \Omega_9$ or other events of probability approaching one in our problem. It will be useful to note that by the same argument as above $\phi_{\min}(\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E}) \geq \sigma^2(\sqrt{n-1} - \sqrt{T} - \sqrt{2Z_3})_+^2$ where $Z_3 \sim \text{Exp}(1)$, so that $\phi_{\min}(\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E}) \geq \sigma^2 n/2$ on an event Ω_8 of probability approaching one. We will use the following fact: if $\boldsymbol{M}, \boldsymbol{N}$ are two positive definite matrices with eigenvalues at least 1/2 then

$$\|\boldsymbol{M}^{-1/2} - \boldsymbol{N}^{-1/2}\|_{op} \leq 2\|\boldsymbol{M}^{1/2} - \boldsymbol{N}^{1/2}\|_{op} \leq \sqrt{2}\|\boldsymbol{M} - \boldsymbol{N}\|_{op} \tag{3.81}$$

using the resolvent identity $\boldsymbol{M}^{-1/2} - \boldsymbol{N}^{-1/2} = \boldsymbol{N}^{-1/2}(\boldsymbol{N}^{1/2} - \boldsymbol{M}^{1/2})\boldsymbol{M}^{-1/2}$ for the first inequality and [144] for the second. To prove (iii), we apply (3.81) to $\boldsymbol{M} = (\sigma^2 n)^{-1}[\boldsymbol{E}^\top \boldsymbol{E} - 2\|\boldsymbol{E}^\top \boldsymbol{X}\boldsymbol{H}\|_{op}\boldsymbol{I}_{T\times T}]$ and $\boldsymbol{N} = (\sigma^2 n)^{-1}\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E}$, both matrices having eigenvalues at least 1/2 on $\Omega_9 \cap \Omega_8$. Rewriting (3.78) as $\widehat{\boldsymbol{\Gamma}}^{-1/2} \preceq (\sigma^2 n)^{-1/2}\boldsymbol{M}^{-1/2}$, applying the triangle inequality and (3.81), we have on $\Omega_8 \cap \Omega_9$

$$
\begin{aligned}
\Delta &:= \sqrt{n-T}\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \sqrt{TF_{T,n-T}} \\
&= \sqrt{n-T}(\|\widehat{\mathbf{\Gamma}}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \|(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2) \\
&\le \sqrt{(n-T)/(\sigma^2 n)}\|\boldsymbol{M}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \sqrt{(n-T)/(\sigma^2 n)}\|\boldsymbol{N}^{-1/2}\boldsymbol{E}^\top\boldsymbol{z}\|_2 \\
&\le \sqrt{(n-T)/(\sigma^2 n)}\|(\boldsymbol{M}^{-1/2} - \boldsymbol{N}^{-1/2})\boldsymbol{E}^\top\boldsymbol{z}\|_2 \\
&\le \sqrt{1-T/n}\sqrt{2}\Big\|(\sigma^2 n)^{-1}[\boldsymbol{E}^\top\boldsymbol{z}\boldsymbol{z}^\top\boldsymbol{E} - 2\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H}\|_{op}\boldsymbol{I}_{T\times T}]\Big\|_{op}\|\boldsymbol{E}^\top\boldsymbol{z}\|_2\sigma^{-1}.
\end{aligned}
\tag{3.82}
$$

The bounds used in (3.80) yield $I\{\Omega_*\}\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H} - \sigma^2\widehat{\mathbf{A}}\|_{op}(\sigma^2 n)^{-1} = O_{\mathbb{P}}(\sqrt{T/n})$ and $I\{\Omega_*\}\|\widehat{\mathbf{A}}\|_{op} = O_{\mathbb{P}}(\bar{s})$, hence $\frac{\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H}\|_{op}}{\sigma^2 n} \le \frac{\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H}-\sigma^2\widehat{\mathbf{A}}\|_{op}}{\sigma^2 n} + \frac{\|\widehat{\mathbf{A}}\|_{op}}{n} = O_{\mathbb{P}}(\frac{\sqrt{T}}{\sqrt{n}}) + O_{\mathbb{P}}(\frac{\bar{s}}{n})$. Furthermore $\|\boldsymbol{E}^\top\boldsymbol{z}\|_2^2/\sigma^2$ has χ^2_T distribution, thus $\|\boldsymbol{E}^\top\boldsymbol{z}\|_2^2/\sigma^2 = O_{\mathbb{P}}(T)$ and we obtain

$$
\Delta \le \sqrt{1-T/n}\Big(O_{\mathbb{P}}(\tfrac{T}{n}) + O_{\mathbb{P}}(\tfrac{\sqrt{T}}{\sqrt{n}}) + O_{\mathbb{P}}(\tfrac{\bar{s}}{n})\Big)O_{\mathbb{P}}(\sqrt{T}).
$$

Since $\frac{T}{n} \to 0$, the right-hand side of the equality is $O_{\mathbb{P}}(\frac{T}{\sqrt{n}}) + O_{\mathbb{P}}(\frac{s\sqrt{T}}{n}) = O_{\mathbb{P}}(\frac{T}{\sqrt{n}}) + o_{\mathbb{P}}(1)$.

For claim (iv), with Δ defined in (3.82) a similar argument yields

$$
\begin{aligned}
|\Delta| &\le \sqrt{n-T}\|\big(\widehat{\mathbf{\Gamma}}^{-1/2} - (\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-1/2}\big)\boldsymbol{E}^\top\boldsymbol{z}\|_2 \\
&\le \sqrt{2}\sqrt{1-T/n}(\sigma^2 n)^{-1}\big[\|\boldsymbol{E}^\top\boldsymbol{z}\|_{op}^2 + 2\|\boldsymbol{E}^\top\boldsymbol{X}\boldsymbol{H}\|_{op} + \|\boldsymbol{X}\boldsymbol{H}\|_{op}^2\big]\|\boldsymbol{E}^\top\boldsymbol{z}\|_2/\sigma
\end{aligned}
$$

on $\Omega_8 \cap \Omega_9$, thus $|\Delta| \le \sqrt{1-T/n}\big(O_{\mathbb{P}}(\frac{T}{n}) + O_{\mathbb{P}}(\frac{\sqrt{T}}{\sqrt{n}}) + O_{\mathbb{P}}(\frac{\bar{s}}{n}) + O_{\mathbb{P}}(\bar{R}^2)\big)O_{\mathbb{P}}(\sqrt{T})$ thanks to Lemma 3.13(ii) for the term $\|\boldsymbol{X}\boldsymbol{H}\|_{op}/(\sigma^2 n)$. This proves (iv).

It remains to prove (v), for which we need a more subtle argument. The important remark is that on the one hand $\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\top$ is independent of $\boldsymbol{E}^\top\boldsymbol{z}$ because $\boldsymbol{E}$ has iid $\mathcal{N}(0,\sigma^2)$ entries, while on the other hand $\widehat{\mathbf{\Gamma}}$ is not independent of $\boldsymbol{E}^\top\boldsymbol{z}$. To overcome this lack of independence, we bound $\widehat{\mathbf{\Gamma}}$ from below by a positive definite matrix independent of $\boldsymbol{E}^\top\boldsymbol{z}$, as follows. For a fixed subset $J \subset [p]$, let $\boldsymbol{P}_J$ be the orthogonal projection matrix onto the linear span of $\{\boldsymbol{z}\} \cup \{\boldsymbol{X}\boldsymbol{e}_j, j \in J\}$ so that the rank of $\boldsymbol{P}_J$ is at most $|J|+1$. Set $\boldsymbol{P}_J^\perp = \boldsymbol{I}_{n\times n} - \boldsymbol{P}_J$. Then in the event

$$
\hat{S} \cup \mathrm{supp}(\boldsymbol{B}^*) \subset J, \tag{3.83}
$$

we have $\boldsymbol{P}_J^\perp\boldsymbol{X}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*) = \boldsymbol{0}$, hence $\widehat{\mathbf{\Gamma}} \succeq (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top\boldsymbol{P}_J^\perp(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}}) = \boldsymbol{E}^\top\boldsymbol{P}_J^\perp\boldsymbol{E}$, thus

$$
\sqrt{n-T}\|\widehat{\mathbf{\Gamma}}^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2 \le \sqrt{n-T}\|(\boldsymbol{E}^\top\boldsymbol{P}_J^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2.
$$

For a fixed J and in the event $\hat{S} \cup \mathrm{supp}\,\boldsymbol{B}^* \subset J$, we can bound from above Δ in (3.82) as

$$
\begin{aligned}
\Delta &\le \sqrt{n-T}\big[\|(\boldsymbol{E}^\top\boldsymbol{P}_J^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2 - \|(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2\big] \\
&\le \frac{\sqrt{n-T}}{\|(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2}\Big[\|(\boldsymbol{E}^\top\boldsymbol{P}_J^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2^2 - \|(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2^2\Big]_+ \\
&= \frac{\sqrt{n-T}}{\|(\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-\frac{1}{2}}\boldsymbol{E}^\top\boldsymbol{z}\|_2}\Big[\boldsymbol{g}^\top\Big\{(\boldsymbol{E}^\top\boldsymbol{P}_J^\perp\boldsymbol{E})^{-1} - (\boldsymbol{E}^\top\boldsymbol{P}_{\boldsymbol{z}}^\perp\boldsymbol{E})^{-1}\Big\}\boldsymbol{g}\Big]_+
\end{aligned}
\tag{3.84}
$$

where $\boldsymbol{g} = \boldsymbol{E}^\top \boldsymbol{z} \sim \mathcal{N}_T(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{T\times T})$ as before, the first inequality follows from $\widehat{\boldsymbol{\Gamma}}^{-\frac{1}{2}} \preceq (\boldsymbol{E}^\top \boldsymbol{P}_J^\perp \boldsymbol{E})^{-\frac{1}{2}}$ and the second from $\sqrt{a} - \sqrt{b} \le (a-b)_+/\sqrt{b}$. For any $J \subset [p]$, the null space inclusion $\ker \boldsymbol{P}_J \subset \ker \boldsymbol{z}\boldsymbol{z}^\top$ holds and the matrix $\boldsymbol{P}_{\boldsymbol{z}}^\perp - \boldsymbol{P}_J^\perp$ is an orthogonal projection matrix with rank $r \le |J|$ so that $\boldsymbol{P}_{\boldsymbol{z}}^\perp - \boldsymbol{P}_J^\perp = \boldsymbol{Q}_J \boldsymbol{Q}_J^\top$ for the matrix $\boldsymbol{Q}_J \in \mathbb{R}^{n\times r}$ with orthonormal columns given by $\boldsymbol{Q}_J = \sum_{k=1}^r \boldsymbol{u}_k \boldsymbol{e}_k^\top$ where $\boldsymbol{u}_k \in \mathbb{R}^n$ are orthonormal eigenvectors of $\boldsymbol{P}_{\boldsymbol{z}}^\perp - \boldsymbol{P}_J^\perp$ corresponding to the non-zero eigenvalues and $\boldsymbol{e}_k$ are canonical basis vectors in $\mathbb{R}^r$. By the Sherman-Morrison-Woodbury identity, the matrix in curly brackets is equal to

$$\boldsymbol{M}_J := (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1} \boldsymbol{E}^\top \boldsymbol{Q}_J \Big(\boldsymbol{I}_{r\times r} - \boldsymbol{Q}_J^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1} \boldsymbol{E}^\top \boldsymbol{Q}_J \Big)^{-1} \boldsymbol{Q}_J^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1}.$$

Applying [71, Theorem II.13] to the Gaussian matrices $\boldsymbol{Q}_J^\top \boldsymbol{E}$ and $\boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E}$, we find

$$\begin{aligned} \mathbb{P}\big(\|\boldsymbol{E}^\top \boldsymbol{Q}_J\|_{op} \ge \sigma(\sqrt{T} + \sqrt{|J|} + t) \big) &\le e^{-t^2/2}, \\ \mathbb{P}\big(\phi_{\min}(\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E}) \le \sigma^2 (\sqrt{n-1} - \sqrt{T} - t)_+^2 \big) &\le e^{-t^2/2} \end{aligned} \tag{3.85}$$

for all $t > 0$. As long as $\frac{1}{2} \ge (\frac{\sqrt{T} + \sqrt{|J|} + t}{\sqrt{n-1} - \sqrt{T} - t})^2$ we have

$$\boldsymbol{I}_{r\times r} - \boldsymbol{Q}_J^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1} \boldsymbol{E}^\top \boldsymbol{Q}_J \succeq \boldsymbol{I}_{r\times r}/2$$

and thus $\boldsymbol{g}^\top \boldsymbol{M}_j \boldsymbol{g} \le 2\|\boldsymbol{Q}_j^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1} \boldsymbol{g}\|_2^2$. Applying Theorem 6.3.2 in [263] and because $\boldsymbol{g}$ is independent of $(\boldsymbol{Q}_J^\top \boldsymbol{E}, \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})$, we find

$$\mathbb{P}(\|\boldsymbol{Q}_j^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1} \boldsymbol{g}/\sigma\|_2 \ge \|\boldsymbol{Q}_j^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1}\|_F + Ct \|\boldsymbol{Q}_j^\top \boldsymbol{E} (\boldsymbol{E}^\top \boldsymbol{P}_{\boldsymbol{z}}^\perp \boldsymbol{E})^{-1}\|_{op}) \le 2e^{-t^2/2}$$

for some absolute constant $C > 0$. Combined with (3.85) and the union bound,

$$\mathbb{P}\Big[\boldsymbol{g}^\top \boldsymbol{M}_J \boldsymbol{g} \ge 2 \Big(\frac{\sigma^{-1} \|\boldsymbol{Q}_J^\top \boldsymbol{E}\|_F}{(\sqrt{n-1} - \sqrt{T} - t)_+^2} + Ct \frac{\sqrt{T} + \sqrt{|J|} + t}{(\sqrt{n-1} - \sqrt{T} - t)_+^2} \Big)^2 \Big] \le 4e^{-t^2/2}.$$

By concentration of chi-square distributed random variables with Tr degrees of freedom (e.g., Theorem 5.6 in [41]), we also have $\mathbb{P}(\sigma^{-1}\|\boldsymbol{Q}_J^\top \boldsymbol{E}\|_F \ge \sqrt{T|J|} + t) \le e^{-t^2/2}$ since $Tr \le T|J|$. Let $s_* = \bar{s} + s$ and note that for $t \ge 0$,

$$\begin{aligned} &\mathbb{P}\Big(\Big\{ \Delta \ge \frac{n-T}{\sqrt{TF_{T,n-T}}} \Big[\frac{\sqrt{Ts_*} + t + Ct(\sqrt{T} + \sqrt{s_*} + t)}{(\sqrt{n-1} - \sqrt{T} - t)_+^2} \Big]^2 \Big\} \cap \Omega_* \Big) \\ &\le \mathbb{P}\Big(\bigcup_{\substack{J\subset[p] \\ |J|=s_*}} \Big\{ \Delta \ge \frac{n-T}{\sqrt{TF_{T,n-T}}} \Big[\frac{\sqrt{T|J|} + t + Ct(\sqrt{T} + \sqrt{|J|} + t)}{(\sqrt{n-1} - \sqrt{T} - t)_+^2} \Big]^2 \Big\} \cap \{\hat{S} \cup \operatorname{supp}(\boldsymbol{B}^*) \subset J\} \Big) \\ &\le \sum_{\substack{J\subset[p] \\ |J|=s_*}} \mathbb{P}\Big(\Big\{ \Delta \ge \frac{n-T}{\sqrt{TF_{T,n-T}}} \Big[\frac{\sqrt{T|J|} + t + Ct(\sqrt{T} + \sqrt{|J|} + t)}{(\sqrt{n-1} - \sqrt{T} - t)_+^2} \Big]^2 \Big\} \cap \{\hat{S} \cup \operatorname{supp}(\boldsymbol{B}^*) \subset J\} \Big) \\ &\le 5 \binom{p}{s_*} e^{-t^2/2}, \end{aligned}$$

where the first inequality holds because $|\hat{S} \cup \mathrm{supp}(\boldsymbol{B}^*)| \le s_*$ on Ω_* by Lemma 3.14. and the last one is obtained by putting together the previous concentration bounds. Setting $t = x + (2\log\binom{p}{s_*})^{1/2}$, we find that Δ is smaller than

$$\frac{n-T}{\sqrt{TF_{T,n-T}}}\Big[\frac{\sqrt{Ts_*} + \sqrt{2\log\binom{p}{s_*}} + x + C(\sqrt{2\log\binom{p}{s_*}} + x)(\sqrt{T} + \sqrt{s_*} + \sqrt{2\log\binom{p}{s_*}} + x)}{(\sqrt{n-1} - \sqrt{T} - \sqrt{2\log\binom{p}{s_*}} - x)_+^2}\Big]^2$$

with probability at least $1 - 5e^{-x^2/2} - \mathbb{P}(\Omega_*^c)$. Since $\mathbb{E}[F_{T,n-T}^{-1}] = T/(T-2)$, we have the estimate $F_{T,n-T}^{-1} = O_{\mathbb{P}}(1)$. Under Assumption (A1)(iv) to control the denominator, and by the bound $\log\binom{p}{s_*} \le s_* \log(\frac{ep}{s_*})$ the above display is thus

$$\begin{aligned} &O_{\mathbb{P}}\Big(\frac{n}{\sqrt{T}}\Big[\frac{Ts + s\log(p/s) + sT\log(p/s) + s^2\log^2(p/s)}{n^2}\Big]\Big) \\ &= O_{\mathbb{P}}\Big(\frac{Ts + s\log(p/s)}{n\sqrt{T}}\Big) + O_{\mathbb{P}}\Big(\frac{sT\log(p/s)}{n\sqrt{T}}\Big) + O_{\mathbb{P}}\Big(\frac{s^2\log^2(p/s)}{n\sqrt{T}}\Big). \end{aligned}$$

In the right-hand side, the first term is $o_{\mathbb{P}}(1)$ thanks to Assumption (A1)(iv). For n large enough $\log(p/s) \ge 1$ holds, thus the second and third term are smaller than $O_{\mathbb{P}}(\frac{s(s+T)\log^2(p/s)}{n\sqrt{T}})$. This proves (v).

In order to deduce the upper bound (3.74) from (iii) and (v), it is sufficient to show that

$$\min\{T/\sqrt{n}, s(s+T)\log^2(p/s)/(n\sqrt{T})\} = o(1) + o(\log^2(p/s)n^{-1/4}) \tag{3.86}$$

holds under Assumption (A1). Let $u_n = sT/n$ and note that $u_n \to 0$ by Assumption (A1). On the one hand, if $T \le \max\{\sqrt{n}u_n, s\}$ then $T/\sqrt{n} \le \max\{u_n, \sqrt{sT/n}\} = o(1)$. On the other hand, if $T > \max\{\sqrt{n}u_n, s\}$ then

$$\frac{s(s+T)\log^2(p/s)}{n\sqrt{T}} \le \frac{2sT\log^2(p/s)}{n\sqrt{T}} = \frac{2u_n\log^2(p/s)}{\sqrt{T}} \le \frac{2u_n^{1/2}\log^2(p/s)}{n^{1/4}} = o\Big(\frac{\log^2(p/s)}{n^{1/4}}\Big).$$

This proves (3.86) and completes the proof. □

Proposition 3.7. *Let $(W_n)_{n\ge1}$ be a sequence of random random variables and χ_T^2 a sequence of random variables with chi-square distribution with T degrees-of-freedom, where $T = T_n$ is function of n (in particular, $T \to +\infty$ as $n \to +\infty$ is allowed). If $\alpha \in (0,1)$ is a fixed constant not depending on n, T and $q_{T,\alpha} > 0$ is the quantile defined by $\mathbb{P}((\chi_T^2)^{1/2} \le q_{T,\alpha}) = 1 - \alpha$ then*

(i) $W_n - (\chi_T^2)^{1/2} \le o_{\mathbb{P}}(1)$ implies that $\mathbb{P}(W_n \le q_{T,\alpha}) \ge 1 - \alpha - o(1)$ and

(ii) $W_n - (\chi_T^2)^{1/2} \ge -o_{\mathbb{P}}(1)$ implies that $\mathbb{P}(W_n \le q_{T,\alpha}) \le 1 - \alpha + o(1)$.

Proof of Proposition 3.7. We first prove case (i). Then by definition of $q_{T,\alpha}$ and the union bound, for any constant $\delta > 0$ not depending on n, T,

$$\begin{aligned} \mathbb{P}(W_n > q_{T,\alpha}) &\le \mathbb{P}(o_{\mathbb{P}}(1) > \delta) + \mathbb{P}((\chi_T^2)^{1/2} > q_{T,\alpha} - \delta) \\ &= \mathbb{P}(o_{\mathbb{P}}(1) > \delta) + \alpha + \mathbb{P}\big((\chi_T^2)^{1/2} \in [q_{T,\alpha} - \delta, q_{T,\alpha}]\big). \end{aligned}$$

We now bound the third term. Let $f_T : [0,+\infty) \to [0,\infty)$ be the probability density function of $(\chi^2_T)^{1/2}$, which admits the closed form $f_T(x) = (2^{T/2-1}\Gamma(T/2))^{-1}x^{T-1}e^{-x^2/2}$ for $x \geq 0$. Then $\mathbb{P}((\chi^2_T)^{1/2} \in [q_{T,\alpha} - \delta, q_{T,\alpha}]) \leq \delta \sup_{x>0} f_T(x)$. The supremum $\sup_{x>0} f_T(x)$ is attained at $x = \sqrt{T-1}$, the mode of the chi distribution with T degrees of freedom, so that

$$\sup_{x>0} f_T(x) = (2^{T/2-1}\Gamma(T/2))^{-1}(T-1)^{(T-1)/2}e^{-(T-1)/2} \xrightarrow[T\to+\infty]{} 1/\sqrt{\pi}$$

by Stirling's formula. Hence there exists an absolute constant $C_0 > 0$ such that

$$\mathbb{P}(W_n > q_{T,\alpha}) \leq \mathbb{P}(o_{\mathbb{P}}(1) > \delta) + \alpha + \delta C_0.$$

For any $\epsilon > 0$, let $\delta = \epsilon/C_0$. Using by the definition of convergence in probability, for n large enough we have $\mathbb{P}(o_{\mathbb{P}}(1) > \delta) \leq \epsilon$ so that $\mathbb{P}(W_n > q_{T,\alpha}) - \alpha \leq 2\epsilon$. Since this holds for any $\epsilon > 0$, the claim is proved. The same argument can be applied in case (ii) by reversing the inequalities.

□

Proof of (3.38). The convergence in distribution

$$\sqrt{2}\big((\chi^2_T)^{1/2} - \sqrt{T}\big) = \frac{(\sqrt{2T})^{-1}(\chi^2_T - T)}{(\chi^2_T/T)^{1/2}/2 + 1/2} \to^d \mathcal{N}(0,1) \tag{3.87}$$

holds by the Central Limit Theorem for $(\sqrt{2T})^{-1}(\chi^2_T - T) \to^d \mathcal{N}(0,1)$, the weak law of large numbers for $(\chi^2_T/T)^{1/2} \to^{\mathbb{P}} 1$ and Slutsky's theorem. If $\Phi(u) = \mathbb{P}(\mathcal{N}(0,1) \leq u)$ is the standard normal cdf, for any subsequence $(a_{T'})_{T'}$ of $a_T = \Phi\big(\sqrt{2}(q_{T,\alpha} - \sqrt{T})\big)$ converging to an accumulation point L, we have for any $\epsilon > 0$ and T' large enough

$$\mathbb{P}\big[\Phi\big(\sqrt{2}((\chi^2_{T'})^{1/2} - \sqrt{T'})\big) \leq L - \epsilon\big] \leq 1 - \alpha \leq \mathbb{P}\big[\Phi\big(\sqrt{2}((\chi^2_{T'})^{1/2} - \sqrt{T'})\big) \leq L + \epsilon\big]$$

so that $L - \epsilon \leq 1 - \alpha + o(1)$ and $1 - \alpha \leq L + \epsilon + o(1)$ by the weak convergence (3.87). It follows that $L = 1 - \alpha$ is the only accumulation point and $q_{T,\alpha} - \sqrt{T} \to z_\alpha/\sqrt{2}$, as desired. □

3.13 Proofs for unknown covariance

3.13.1 Asymptotic normality

Proof of Theorem 3.5 under assumption (3.34). We will use throughout the proof the notation defined after (3.31) for $\tau_j, \boldsymbol{\gamma}^{(j)}$ and $\boldsymbol{\varepsilon}^{(j)}$. Define the direction $\widetilde{\boldsymbol{a}}_j = \boldsymbol{e}_j(\boldsymbol{\Sigma}^{-1})_{jj}^{-1/2} = \tau_j \boldsymbol{e}_j$ normalized such that $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{a}}_j\|_2 = 1$ by construction, as well as $\widetilde{\boldsymbol{z}}_j = \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{a}}_j \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Next, define $\boldsymbol{\xi}_j, \widehat{\boldsymbol{\xi}}_j \in \mathbb{R}^T$ by

$$\boldsymbol{\xi}_j = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \widetilde{\boldsymbol{z}}_j + (n\boldsymbol{I}_{T\times T} - \widehat{\mathrm{A}})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \widetilde{\boldsymbol{a}}_j, \tag{3.88}$$

$$\widehat{\boldsymbol{\xi}}_j = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \widehat{\boldsymbol{z}}_j \left[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\right] \tau_j + (n\boldsymbol{I}_{T\times T} - \widehat{\mathrm{A}})(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \widetilde{\boldsymbol{a}}_j \tag{3.89}$$

so that $\boldsymbol{\xi}_j$ coincides with (3.37) for the normalized direction $\widetilde{\boldsymbol{a}}_j$. Since the second term in $\boldsymbol{\xi}_j$ is the same as the second term in $\widehat{\boldsymbol{\xi}}_j$,

$$\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2 = \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \Big\{ \widehat{\boldsymbol{z}}_j \tau_j^{-1} \Big[n\tau_j^2 (\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1} \Big] - \widetilde{\boldsymbol{z}}_j \Big\} \|_2. \tag{3.90}$$

Since $\boldsymbol{\gamma}^{(j)} = -(\boldsymbol{I}_p - \boldsymbol{e}_j\boldsymbol{e}_j^\top)\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j(\boldsymbol{\Sigma}^{-1})_{jj}^{-1}$ in (3.31), or equivalently $\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)} = \tau_j^2 \boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j$, we have

$$\widetilde{\boldsymbol{z}}_j = \tau_j \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j = \tau_j^{-1}\boldsymbol{X}(\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)}).$$

Next, $\widehat{\boldsymbol{z}}_j = \boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}^{(j)} = \boldsymbol{X}[\boldsymbol{e}_j - \widehat{\boldsymbol{\gamma}}^{(j)}]$ since by definition of $\widehat{\boldsymbol{\gamma}}^{(j)}$, the j-th coordinate of $\widehat{\boldsymbol{\gamma}}^{(j)}$ is zero, so that $\boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}^{(j)} = \boldsymbol{X}\widehat{\boldsymbol{\gamma}}^{(j)}$. By inserting $\boldsymbol{I}_{p\times p} = \sum_{k=1}^p \boldsymbol{e}_k\boldsymbol{e}_k{}^\top$ in (3.90), using that the KKT conditions of $\widehat{\boldsymbol{B}}$ imply that $\max_{k\in[p]} \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{X}\boldsymbol{e}_k\|_2 \le nT\lambda$ and the triangle inequality, we find

$$\begin{aligned}
\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2 &= \tau_j^{-1}\Big\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})^\top \boldsymbol{X}\sum_{k=1}^p \boldsymbol{e}_k\boldsymbol{e}_k^\top \Big\{ (\boldsymbol{e}_j - \widehat{\boldsymbol{\gamma}}^{(j)}) \Big[n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big] - (\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)}) \Big\} \Big\|_2 \\
&\le \tau_j^{-1} nT\lambda \sum_{k=1}^p \Big| \boldsymbol{e}_k^\top \Big\{ (\boldsymbol{e}_j - \widehat{\boldsymbol{\gamma}}^{(j)}) \Big[n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big] - (\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)}) \Big\} \Big| \qquad (3.91) \\
&= \tau_j^{-1} nT\lambda \Big\| \Big\{ (\boldsymbol{e}_j - \widehat{\boldsymbol{\gamma}}^{(j)}) \Big[n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big] - (\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)}) \Big\} \Big\|_1.
\end{aligned}$$

There are two errors at this point: the estimation error $\|\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma}^{(j)}\|_1$ and the estimation error $|n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1} - 1|$, which corresponds to the relative error of the estimation of the variance τ_j^2 by $n^{-1}\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j$ in the linear model (3.31). Keeping these two errors in mind, by the triangle inequality the previous display yields

$$\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2 \le \frac{\tau_j^{-1} nT\lambda}{(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)/(n\tau_j^2)} \Big(\Big\|\boldsymbol{\gamma}^{(j)} - \widehat{\boldsymbol{\gamma}}^{(j)}\Big\|_1 + \Big\|\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)}\Big\|_1 \Big| 1 - \frac{\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j}{n\tau_j^2}\Big| \Big). \tag{3.92}$$

For the first term in the parenthesis, inequality (3.33) holds: this is the usual ℓ_1 estimation rate for the Lasso estimate $\widehat{\boldsymbol{\gamma}}^{(j)}$ for the sparse estimation target $\boldsymbol{\gamma}^{(j)}$ in the linear model (3.31) with noise variance τ_j^2. For the second term, inequality

$$\tau_j^{-1}\|\boldsymbol{e}_j - \boldsymbol{\gamma}^{(j)}\|_1 = \tau_j\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_1 \le \tau_j\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0^{1/2}\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_2 \le \|\boldsymbol{\Sigma}^{-1/2}\|_{op}\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0^{1/2} \tag{3.93}$$

holds thanks to the Cauchy-Schwarz inequality and $\tau_j = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}_j\|_2^{-1}$. Furthermore, by the triangle inequality, we have

$$\Big|1 - \frac{\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j}{n\tau_j^2}\Big| \le \Big|1 - \frac{\|\boldsymbol{\varepsilon}^{(j)}\|_2^2}{n\tau_j^2}\Big| + \Big|\frac{(\boldsymbol{\varepsilon}^{(j)} - \widehat{\boldsymbol{z}}_j)^\top\boldsymbol{\varepsilon}^{(j)}}{n\tau_j^2} - \frac{\widehat{\boldsymbol{z}}_j^\top(\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{\varepsilon}^{(j)})}{n\tau_j^2}\Big|. \tag{3.94}$$

As $\|\boldsymbol{\varepsilon}^{(j)}\|_2^2/\tau_j^2$ has χ_n^2 distribution, the first term is $O(n^{-1/2})$ by the Central Limit Theorem. For the next term, we use again the triangle inequality. To bound the next term, notice that by Hölder's inequality,

$$|(\boldsymbol{\varepsilon}^{(j)} - \widehat{\boldsymbol{z}}_j)^\top\boldsymbol{\varepsilon}^{(j)}| = |(\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma}^{(j)})^\top\boldsymbol{X}_{-j}^\top\boldsymbol{\varepsilon}^{(j)}| \le \|\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma}^{(j)}\|_1\|\boldsymbol{X}_{-j}{}^\top\boldsymbol{\varepsilon}^{(j)}\|_\infty.$$

Each factor in the right hand side is bounded from above as follows: $\|\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma}^{(j)}\|_1 = \tau_j\|\boldsymbol{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0 O_{\mathbb{P}}\big(\sqrt{n^{-1}\log p}\big)$ thanks to (3.33) and $\|\boldsymbol{X}_{-j}{}^\top\boldsymbol{\varepsilon}^{(j)}\|_\infty = \tau_j O_{\mathbb{P}}(\sqrt{n\log p})$ because $\boldsymbol{X}_{-j}$ is independent of $\boldsymbol{\varepsilon}^{(j)}$ and $\max_{k\in[p]\setminus\{j\}}\boldsymbol{\Sigma}_{kk}\le 1$. This proves that $|(\boldsymbol{\varepsilon}^{(j)} - \widehat{\boldsymbol{z}}_j)^\top\boldsymbol{\varepsilon}^{(j)}|/(n\tau_j^2) \le \|\boldsymbol{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0 O_{\mathbb{P}}(n^{-1}\log p)$. We also have

$$|\widehat{\boldsymbol{z}}_j^\top(\boldsymbol{X}\boldsymbol{e}_j - \boldsymbol{\varepsilon}^{(j)})| = |\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}_{-j}\boldsymbol{\gamma}^{(j)}| \le \|\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}_{-j}\|_\infty\|\boldsymbol{\gamma}^{(j)}\|_1 \le O_{\mathbb{P}}(\tau_j\sqrt{n\log p})\|\boldsymbol{\gamma}^{(j)}\|_1$$

thanks to Hölder's inequality and the KKT conditions for $\widehat{\boldsymbol{\gamma}}^{(j)}$ in (3.32) to bound the ℓ_∞ norm. We have $\|\boldsymbol{\gamma}^{(j)}\|_1 \le \|\boldsymbol{\gamma}^{(j)}\|_0^{1/2}\|\boldsymbol{\gamma}^{(j)}\|_2$ and $\|\boldsymbol{\gamma}^{(j)}\|_2 \le \tau_j^2\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_2 \le \tau_j\|\boldsymbol{\Sigma}^{-1/2}\|_{op}$ by definition of $\boldsymbol{\gamma}^{(j)}$ and the Cauchy-Schwarz inequality. Combining these bounds provide an upper bound on the right hand side of (3.94), so that

$$\begin{aligned}\Big|1 - \frac{\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j}{n\tau_j^2}\Big| &\le O_{\mathbb{P}}\Big(\frac{1}{\sqrt{n}}\Big) + \|\boldsymbol{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0 O_{\mathbb{P}}\Big(\frac{\log p}{n}\Big) + \|\boldsymbol{\Sigma}^{-1/2}\|_{op}\Big(\frac{\|\boldsymbol{\gamma}^{(j)}\|_0\log p}{n}\Big)^{1/2} \\ &\le \|\boldsymbol{\Sigma}^{-1}\|_{op}\big(\|\boldsymbol{\gamma}^{(j)}\|_0\log(p)/n\big)^{1/2}O_{\mathbb{P}}(1) \qquad (3.95)\end{aligned}$$

where the second line follows by bounding from above the first two terms thanks to assumption (3.34) and $\|\boldsymbol{\Sigma}^{-1}\|_{op} \ge 1$ (this is a consequence of $\boldsymbol{\Sigma}_{jj} \le 1$ in Assumption (A1)). The bound (3.95) also provides $\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j/(n\tau_j^2)\xrightarrow{\mathbb{P}}1$ and thus $n\tau_j^2/(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j) = O_{\mathbb{P}}(1)$. Using (3.33), (3.93) and (3.95) to bound from above the right hand side of (3.92) we find

$$\begin{aligned}\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2 \le n T\lambda\Big(&\|\boldsymbol{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0 O_{\mathbb{P}}(\sqrt{n^{-1}\log p}) \\ &+ \|\boldsymbol{\Sigma}^{-1/2}\|_{op}\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0^{1/2}\|\boldsymbol{\Sigma}^{-1}\|_{op}\|\boldsymbol{\gamma}^{(j)}\|_0^{1/2}O_{\mathbb{P}}(\sqrt{n^{-1}\log p})\Big).\end{aligned}$$

Since $\|\boldsymbol{\gamma}^{(j)}\|_0 = \|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0 - 1$, this implies $\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2 \le nT\lambda\|\boldsymbol{\Sigma}^{-1}\|_{op}^{3/2}\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0 O_{\mathbb{P}}(\sqrt{n^{-1}\log p})$. Thanks to $\lambda = O\big(\sigma(nT)^{-1/2}\big)(1+\sqrt{\log(p/s)/T})\big)$ by definition of λ, we eventually obtain

$$(\sigma^2 n)^{-1/2}\|\widehat{\boldsymbol{\xi}}_j - \boldsymbol{\xi}_j\|_2 = O_{\mathbb{P}}\big([\sqrt{T} + \sqrt{\log(p/s)}]\|\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_j\|_0\sqrt{\log(p)/n}\big) \qquad (3.96)$$

which converges to 0 in probability thanks to assumption (3.34).

To complete the proof of Theorem 3.5 and prove asymptotic normality for some fixed $\boldsymbol{b}\in\mathbb{R}^T$ with $\|\boldsymbol{b}\|_2 = 1$, notice that

$$\zeta_j := \frac{n\widetilde{\boldsymbol{a}}_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b} + \widetilde{\boldsymbol{z}}_j^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}}{\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2}$$

satisfies $\zeta_j\xrightarrow{d}\mathcal{N}(0,1)$ by Theorem 3.3 applied to the normalized direction $\widetilde{\boldsymbol{a}}_j$. Furthermore,

$$\begin{aligned}&\Big|\zeta_j - \frac{n\boldsymbol{e}_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b} + n(\widehat{\boldsymbol{z}}_j^\top\boldsymbol{X}\boldsymbol{e}_j)^{-1}\widehat{\boldsymbol{z}}_j^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}}{(\tau_j)^{-1}\ \|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2}\Big| \\ &= \frac{|(\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j)^\top(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}|}{\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2} \\ &\le (\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2\|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op}\Big(\|(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2^{-1}(\sigma^2 n)^{1/2}\Big).\end{aligned}$$

In the above display, $(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}_j - \widehat{\boldsymbol{\xi}}_j\|_2 \xrightarrow{\mathbb{P}} 0$ when (3.34) holds, $\|(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\|_{op} \xrightarrow{\mathbb{P}} 1$ by Proposition 3.2(iii) and Lemma 3.14, and the rightmost factor converges to 1 in probability by Theorem 3.4.

Since $\tau_j = (\boldsymbol{\Sigma}^{-1})_{jj}^{-1/2}$, the last claim follows by $\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j/(n\tau_j^2) \xrightarrow{\mathbb{P}} 1$ by (3.95) and Slutsky's theorem. We also have $\|\widehat{\boldsymbol{z}}_j\|_2/(\tau_j\sqrt{n}) \xrightarrow{\mathbb{P}} 1$ since, using (3.32) and the triangle inequality,

$$
\begin{aligned}
(\tau_j^2 n)^{-1}\big|\|\widehat{\boldsymbol{z}}_j\|_2^2 - \widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j\big| &= (\tau_j^2 n)^{-1}|\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}_{-j}\widehat{\boldsymbol{\gamma}}^{(j)}| \\
&\le (\tau_j^2 n)^{-1} O_{\mathbb{P}}(1)\tau_j\sqrt{n\log p}\|\widehat{\boldsymbol{\gamma}}^{(j)}\|_1 \\
&\le O_{\mathbb{P}}(1)\sqrt{\log(p)/n}\big[\|\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma}^{(j)}\|_1 + \|\boldsymbol{\gamma}^{(j)}\|_1\big]/\tau_j \\
&\le O_{\mathbb{P}}(1)\sqrt{\log(p)/n}\big[\|\boldsymbol{\gamma}^{(j)}\|_0\sqrt{\log(p)/n} + \|\boldsymbol{\gamma}^{(j)}\|_0^{1/2}\big] \quad (3.97)\\
&= o_{\mathbb{P}}(1)
\end{aligned}
$$

thanks to (3.33) for the first term and the Cauchy-Schwarz inequality for the second. The convergence to 0 in probability in the last line follows from (3.34). $\square$

Proof of Theorem 3.5 under assumption (3.35). With $\widehat{\boldsymbol{\xi}}_j$ in (3.89) and $\widetilde{\boldsymbol{\xi}}_j := \boldsymbol{E}^\top \widehat{\boldsymbol{z}}_j[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}]\tau_j$ we have

$$
\begin{aligned}
\|\widehat{\boldsymbol{\xi}}_j - \widetilde{\boldsymbol{\xi}}_j\|_2 &= \| - \widehat{\mathbf{A}}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)^\top \widetilde{\boldsymbol{a}}_j + \tau_j(\boldsymbol{B}^* - \widehat{\boldsymbol{B}})^\top \boldsymbol{X}_{-j}^\top \widehat{\boldsymbol{z}}_j[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}]\|_2 \\
&\le \|\widehat{\mathbf{A}}\|_{op}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{op} + \tau_j\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1}\|\boldsymbol{X}_{-j}^\top \widehat{\boldsymbol{z}}_j\|_\infty |[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)]^{-1}| \quad (3.98)
\end{aligned}
$$

thanks to $\|\boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{a}}_j\|_2 = 1$ for the first term and Hölder's inequality for the second term. Thanks to Lemma 3.13(iii), Lemma 3.14 and Proposition 3.2 we find $\|\widehat{\mathbf{A}}\|_{op}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_{op} = O_{\mathbb{P}}(\bar{s}\bar{R})$. For the second term, thanks to (3.32) and Lemma 3.13(iv) we have

$$
\tau_j\|\widehat{\boldsymbol{B}} - \boldsymbol{B}^*\|_{2,1}\|\boldsymbol{X}_{-j}^\top \widehat{\boldsymbol{z}}_j\|_\infty |[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)]^{-1}| \le O_{\mathbb{P}}(\sqrt{s}\bar{R})\sqrt{n\log p}\big|[n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)]^{-1}\big|
$$

and the bound (3.95) grants $\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j/(n\tau_j^2) \xrightarrow{\mathbb{P}} 1$ thanks to the leftmost assumption in (3.35). In summary, $(\sigma^2 n)^{-1/2}\|\widehat{\boldsymbol{\xi}}_j - \widetilde{\boldsymbol{\xi}}_j\|_2 = O_{\mathbb{P}}(n^{-1/2}s\bar{R} + \sqrt{s}\bar{R}\sqrt{\log p}) = O_{\mathbb{P}}(\sqrt{s}\bar{R}\sqrt{\log p})$ thanks to $n^{-1/2}\sqrt{s} \le 1$. Hence due to the rightmost assumption in (3.35),

$$
(\sigma^2 n)^{-1/2}\|\widehat{\boldsymbol{\xi}}_j - \widetilde{\boldsymbol{\xi}}_j\|_2 \xrightarrow{\mathbb{P}} 0. \quad (3.99)
$$

Next, assume without loss of generality that $\|\boldsymbol{b}\|_2 = 1$. By definition of $\widehat{\boldsymbol{\xi}}_j$ in (3.89),

$$
\begin{aligned}
&\frac{n\boldsymbol{e}_j^\top(\widehat{\boldsymbol{B}} - \boldsymbol{B}^*)\boldsymbol{b} + n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\widehat{\boldsymbol{z}}_j^\top(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}}{(\boldsymbol{\Sigma}^{-1})_{jj}^{1/2}\ \sigma\sqrt{n}} - \frac{\widetilde{\boldsymbol{\xi}}_j^\top \boldsymbol{b}}{\sigma\sqrt{n}} \\
&= \frac{(\widehat{\boldsymbol{\xi}}_j - \widetilde{\boldsymbol{\xi}}_j)^\top(\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}}{\sigma\sqrt{n}} - \frac{\widetilde{\boldsymbol{\xi}}_j^\top(\boldsymbol{I}_{T\times T} - (\boldsymbol{I}_{T\times T} - \widehat{\mathbf{A}}/n)^{-1})\boldsymbol{b}}{\sigma\sqrt{n}}. \quad (3.100)
\end{aligned}
$$

The first term converges to 0 in probability thanks to the previous paragraph, while the second term is $O_{\mathbb{P}}(s/n)\|\widetilde{\boldsymbol{\xi}}_j\|_2(\sigma^2 n)^{-1/2}$ by Proposition 3.2 and Lemma 3.14. $\boldsymbol{\zeta}_j :=$

$[n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}]^{-1}\tau_j\|\widehat{\boldsymbol{z}}_j\|_2^{-1}\widetilde{\boldsymbol{\xi}}_j$ has $\mathcal{N}_T(\mathbf{0},\sigma^2\boldsymbol{I}_{T\times T})$ distribution by independence of $\boldsymbol{E}$ and $\boldsymbol{X}$. Next, $\|\widetilde{\boldsymbol{\xi}}_j\|_2 = \|\widehat{\boldsymbol{z}}_j\|_2[n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}]\|\boldsymbol{\zeta}_j\|_2$ and $\|\boldsymbol{\zeta}_j\|_2 = O_{\mathbb{P}}(\sqrt{T})$ since $\mathbb{E}[\|\boldsymbol{\zeta}_j\|_2^2] = T$. Furthermore, $n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\xrightarrow{\mathbb{P}}1$ by (3.95). We also have $\tau_j\sqrt{n}\|\widehat{\boldsymbol{z}}_j\|_2^{-1}\xrightarrow{\mathbb{P}}1$ by (3.97), thanks to the leftmost assumption in (3.35) for the last line in (3.97). This shows that $\|\widetilde{\boldsymbol{\xi}}_j\|_2/(\sqrt{n}\|\boldsymbol{\zeta}_j\|_2)\xrightarrow{\mathbb{P}}1$ and that the second term in (3.100) is $O_{\mathbb{P}}(s/n)$ and converges to 0 in probability. We conclude by observing that $\widetilde{\boldsymbol{\xi}}_j^\top \boldsymbol{b}/(\sigma\sqrt{n})\xrightarrow{d}\mathcal{N}(0,1)$ by Slutsky's theorem thanks to $n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\xrightarrow{\mathbb{P}}1$ and $\tau_j\sqrt{n}\|\widehat{\boldsymbol{z}}_j\|_2^{-1}\xrightarrow{\mathbb{P}}1$. In the denominator, $\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2$ and $\sigma\sqrt{n}$ can be used interchangeably, again by Slutsky's theorem, since $\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{B}})(\boldsymbol{I}_{T\times T}-\widehat{\mathbf{A}}/n)^{-1}\boldsymbol{b}\|_2/(\sigma\sqrt{n})\xrightarrow{\mathbb{P}}1$ by Theorem 3.4. □

3.13.2 Asymptotic χ_T^2 distribution

Proof of Theorem 3.8 under assumption (3.35) *.* Let $\widehat{\boldsymbol{\xi}}_j$ and $\widetilde{\boldsymbol{\xi}}_j$ be defined respectively in (3.89) and in the sentence preceding (3.98). Notice that the quantity in the left hand side of (3.43) is equal to $(1-T/n)^{1/2}\|\widehat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\xi}\|_2$ where

$$\boldsymbol{\xi} = \Big(\tau_j^{-1}\frac{\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\big]^{-1}\Big)\widehat{\boldsymbol{\xi}}_j. \tag{3.101}$$

Set $\boldsymbol{z} = \widehat{\boldsymbol{z}}_j/\|\widehat{\boldsymbol{z}}_j\|_2$. For these values of $\boldsymbol{\xi}$ and $\boldsymbol{z}$, we have

$$\begin{aligned}(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}-\sqrt{n}\boldsymbol{E}^\top\boldsymbol{z}\|_2 &= (\sigma^2 n)^{-1/2}\Big\|\widehat{\boldsymbol{\xi}}_j\tau_j^{-1}\frac{\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} - \boldsymbol{E}^\top\widehat{\boldsymbol{z}}_j\frac{\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big\|_2 \\ &= (\sigma^2 n)^{-1/2}\Big\|\widehat{\boldsymbol{\xi}}_j-\widetilde{\boldsymbol{\xi}}_j\Big\|_2\tau_j^{-1}\frac{\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}]^{-1}.\end{aligned}$$

Hence the above is $o_{\mathbb{P}}(1)$ by combining (3.99) with (3.95) and (3.97). An application of Lemma 3.26 for these values of $\boldsymbol{z}$ and $\boldsymbol{\xi}$ yields (3.74) which completes the proof. □

Proof of Theorem 3.8 under assumption (3.34) *.* Let $\boldsymbol{\xi}_j$ be defined in (3.88) Since $\widetilde{\boldsymbol{a}}_j, \widetilde{\boldsymbol{z}}_j$ defined in the proof of Theorem 3.5 satisfy the assumptions of Theorem 3.6, we have already established that $(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}_j - \sqrt{n}\boldsymbol{E}^\top\widetilde{\boldsymbol{z}}_j\|\widetilde{\boldsymbol{z}}_j\|_2^{-1}\|_2 = o_{\mathbb{P}}(1)$, cf. (3.71) with $\boldsymbol{a} = \widetilde{\boldsymbol{a}}_j$ and $\boldsymbol{z}_0 = \widetilde{\boldsymbol{z}}_j$.

We now proceed to show that $(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}-\boldsymbol{\xi}_j\|_2 = o_{\mathbb{P}}(1)$ for $\boldsymbol{\xi}$ defined in (3.101). By the triangle inequality and since $\widehat{\boldsymbol{\xi}}_j$ in (3.101) is proportional to $\widehat{\boldsymbol{\xi}}_j$, we have

$$\begin{aligned}&(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}-\boldsymbol{\xi}_j\|_2 \\ &= \frac{1}{\sigma\sqrt{n}}\Big\|_2\widehat{\boldsymbol{\xi}}_j\frac{\tau_j^{-1}\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} - \boldsymbol{\xi}_j\Big\|_2 \\ &= \frac{1}{\sigma\sqrt{n}}\Big\|(\widehat{\boldsymbol{\xi}}_j-\boldsymbol{\xi}_j)\frac{\tau_j^{-1}\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} + \boldsymbol{\xi}_j\Big(\frac{\tau_j^{-1}\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} - 1\Big)\Big\|_2 \\ &\le \frac{1}{\sigma\sqrt{n}}\Big\|(\widehat{\boldsymbol{\xi}}_j-\boldsymbol{\xi}_j)\Big\|_2\frac{\tau_j^{-1}\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} + \Big\|\boldsymbol{\xi}_j\Big\|_2\Big|\frac{\tau_j^{-1}\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} - 1\Big|.\end{aligned}$$

For the first term, by (3.96) we already have $(\sigma^2 n)^{-1/2}\|\widetilde{\boldsymbol{\xi}}_j - \boldsymbol{\xi}_j\|_2 = o_{\mathbb{P}}(1)$. Combined with $\|\widehat{\boldsymbol{z}}_j\|_2/(\tau_j\sqrt{n})\xrightarrow{\mathbb{P}}1$ and $n\tau_j^2(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\xrightarrow{\mathbb{P}}1$ (see (3.95) and (3.97)), this proves that the first term above is $o_{\mathbb{P}}(1)$ For the remaining terms, $(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi}_j\|_2 = O_{\mathbb{P}}(\sqrt{T})$ by (3.71), and the question is whether

$$O_{\mathbb{P}}(\sqrt{T})\Big|\frac{\tau_j^{-1}\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} - 1\Big| \tag{3.102}$$

converges to 0 using (3.95) and (3.97). With $a_j = \|\widehat{\boldsymbol{z}}_j\|_2^2/(\tau_j^2 n)$ and $b_j = \widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j/(\tau_j^2 n)$ for brevity,

$$\begin{aligned}\Big|\tau_j^{-1}\frac{\sqrt{n}}{\|\widehat{\boldsymbol{z}}_j\|_2}\Big[n(\widehat{\boldsymbol{z}}_j^\top \boldsymbol{X}\boldsymbol{e}_j)^{-1}\Big]^{-1} - 1\Big| &= a_j^{-1/2}|b_j - a^{1/2}| \\ &\le a_j^{-1/2}(|b_j - 1| + |1 - a_j^{1/2}|) \\ &= a_j^{-1/2}(|b_j - 1| + |1 - a_j|(1 + a_j)^{-1}).\end{aligned}$$

We have $|a_j - 1| + |b_j - 1| = \sqrt{\|\boldsymbol{\gamma}^{(j)}\|_0 \log(p)/n}\; O_{\mathbb{P}}(1)$ thanks to (3.95) and (3.97). Hence thanks to (3.34), quantity (3.102) is $o_{\mathbb{P}}(1)$. Combining all the pieces, we have proved that

$$(\sigma^2 n)^{-1/2}\|\boldsymbol{\xi} - \sqrt{n}\boldsymbol{E}^\top \widetilde{\boldsymbol{z}}_j\|\widetilde{\boldsymbol{z}}_j\|_2^{-1}\|_2 \le (\sigma^2 n)^{-1/2}\|\boldsymbol{\xi} - \boldsymbol{\xi}_j\|_2 + o_{\mathbb{P}}(1) \le o_{\mathbb{P}}(1).$$

Applying Lemma 3.26 to $\boldsymbol{\xi}$ in (3.101) and $\boldsymbol{z} = \widetilde{\boldsymbol{z}}_j\|\widetilde{\boldsymbol{z}}_j\|_2^{-1}$, conclusion (3.74) completes the proof. □

Chapter 4

Statistical properties of approximate geometric quantiles in infinite-dimensional Banach spaces

4.1 Introduction

Data samples are sometimes modeled as points living in a metric space [37, 92, 96], or more specifically in a manifold [206, 209], or in an infinite-dimensional normed space [69, 216, 217]. The practioner is often interested in a measure of central tendency, i.e., a point in the space that is most representative of the whole sample. Once such a measure is defined, it is worth investigating its statistical properties: as the sample size grows to infinity, does this measure approach the central tendency of the population, and if so, at which rate? Means and medians are classical measures of central tendency in a Euclidean space; by viewing them as solutions to optimization problems they have been generalized to the aforementioned non-Euclidean settings. Such extensions have been termed "Fréchet means" and "Fréchet medians", and they are defined for a finite sample or more generally for a probability measure [95, 147, 243, 12, 274]. Their statistical properties have attracted much attention recently [230, 5, 167, 138].

Regarding normed spaces, the Fréchet median was introduced in the two-dimensional Euclidean setting by Weber [272] and was later reintroduced in the same setting by Gini and Galvani [103, 225] as well as Haldane [110], who referred to it as a "geometric" or "geometrical" median. Throughout this chapter, we adopt Haldane's terminology of geometric median, but "spatial median" and "L^1 median" are also common names in the literature [46, 239]. Valadier [253, 254] extended the concept to any reflexive Banach space and Kemperman [148] performed a systematic study of existence and uniqueness in general Banach spaces, as well as statistical properties in finite dimension. Chaudhuri [65] and Koltchinskii [156, 157] defined geometric quantiles in Banach spaces by slightly changing the objective function of the minimization problem. Note that geometric quantiles include the geometric median as a special case.

Infinite-dimensional normed spaces play a major role in kernel methods [121, 281, 198] and in functional data analysis [93, 124, 127], since they are an appropriate setting for the modeling of curves (e.g., radar waveforms, spectrometric data, electricity consumption, stock prices). Functional data is mostly modeled in the Hilbert space L^2, however there is recent interest in other functional spaces such as the space of continuous functions [74, 73]. Moreover, the non-Hilbertian infinite-dimensional setting is relevant when working with operators [182, 158].

From a statistical standpoint, a geometric quantile is a location parameter that fits the framework of M-estimation. Replacing the objective function with its empirical counterpart naturally yields an estimator, usually called empirical (or sample) geometric quantile. In a Euclidean space, consistency and asymptotic normality of empirical geometric quantiles are easily obtained by applying general results from the theory of M-estimation [132, 108]. In infinite dimension, technical challenges arise. First, the normed space E can be equipped with the weak, the weak$*$ (if E is a dual space) or the norm topology. These topologies give very different meanings to convergence in the space, hence also to consistency. Consistency in the norm topology is the most desirable mode of convergence and it is also the most difficult to establish. Second, the non-compactness of spheres and closed balls in infinite-dimensional spaces invalidates many reasonings commonly used in finite-dimensional M-estimation; different techniques are therefore required. The recent paper [238] aims to develop a general theory of M-estimation in Hilbert spaces, with an emphasis on the infinite-dimensional function space L^2. As noted by the authors, their consistency result in the norm topology [238, Theorem 3.4] covers only finite-dimensional spaces.

For a given geometric quantile, the estimator we consider here is an approximate empirical version, in the sense that it minimizes the empirical objective function up to some (possibly random) additive precision ϵ_n. By setting $\epsilon_n = 0$ we recover the exact empirical geometric quantile studied in [53, 100, 64]. Such a relaxation is standard in M-estimation [132, 116, 259, 10], as it is more realistic and covers estimators obtained by iterative optimization methods like gradient descent.

Some statistical results are known for the infinite-dimensional exact median ($\ell = 0$ and $\epsilon_n = 0$): Cadre [53] proved that the empirical geometric median is consistent in the weak$*$ topology when E is the dual of separable Banach space (thus also in the weak topology when E is reflexive) and Gervini [100] obtained a similar result for the space $E = L^2$. Chakraborty and Chaudhuri [64] proved consistency with respect to the norm topology in separable Hilbert spaces. Notably, their result [64, Theorem 4.2.2] has distributional assumptions that are superfluous in the finite-dimensional case, which suggests that they are also unnecessary in infinite dimension. Regarding asymptotic normality, the only result in infinite dimension that we are aware of is the central limit theorem [100, Theorem 6] which holds for the space $E = L^2$ and under a very specific assumption: to the probability measure on L^2 corresponds a stochastic process X, and the Karhunen–Loève decomposition of X is assumed to have only a finite number of summands.

Recently, other estimators of geometric quantiles and median have been proposed [57, 63, 105, 58] which have good computational and statistical properties in infinite dimension. Most related to our work is [63], which explores the properties of a sieved M-estimator [261, Chapter 3.4] by carrying optimization over finite-dimensional subspaces. Their proof techniques are incompatible with the study of our estimator (see Remark 4.75 below for more details). Some works [192, 205] have exploited empirical geometric quantiles and median as auxiliary tools in the construction of robust estimators.

Contributions and outline

The main goal of this chapter is to investigate the fundamental large-sample properties of the approximate empirical geometric quantile in infinite-dimensional Banach spaces. We describe below how the chapter is organized, and we give a brief overview of our contributions. In the body of the chapter, immediately before or after each result, we explain in detail how it relates to or improves on the existing literature.

- In Section 4.2, we recall the definition of a geometric quantile and we address the issues of existence and uniqueness. Proposition 4.8 provides a new condition for existence of a geometric median, and we see in Corollary 4.13 that it ensures existence in a wide variety of L^1 spaces, thus extending a result by Kemperman [148, Corollary 3.2]. In Proposition 4.19 we characterize the set of geometric medians when the space is strictly convex and the measure is supported on some affine line.

- In Section 4.3, we introduce our estimation setting and the approximate empirical geometric quantile. This estimator is defined in an implicit fashion, which opens the door for measurability issues. After detailing our treatment of measurability woes, we consider the adjacent topic of measurable selections. Sinova et al. [238, Proposition 3.3] state a selection result for generic M-estimators that is valid only in σ-compact Hilbert spaces, i.e., finite-dimensional Hilbert spaces. In contrast, our selection Theorems 4.25 and 4.26 cover a wide range of infinite-dimensional Banach spaces. Finally, in Theorem 4.30 we provide an asymptotic uniqueness result for empirical quantiles.

- In Section 4.4.1 we examine convergence of the estimator in the setting where there might be multiple population quantiles. We leverage the theory of variational convergence to obtain Theorems 4.37, 4.42 and 4.43, which are asymptotic statements on subsequences in the weak topology.

- In Section 4.4.2 we switch to the setting of a unique population quantile and we study the consistency of our estimator.
  - Section 4.4.2 is dedicated to consistency in the weak topology. As an immediate consequence of the results developed in Section 4.4.1 we obtain the consistency Theorem 4.45, which is a minor generalization of [53, Theorem 1 (i)] and [100, Theorem 2].
  - In Section 4.4.2 we turn to consistency in the norm topology. Theorems 4.54 and 4.55 provide consistency in separable, reflexive, strictly convex spaces that satisfy the Radon–Riesz property, hence as a special case in separable, uniformly convex spaces (e.g., separable Hilbert spaces, L^p, $W^{k,p}$ with $p \in (1, \infty)$). Our findings holds under minimal assumptions that match those of

the finite-dimensional case. They are a significant improvement on the result by Chakraborty and Chaudhuri [64, Theorem 4.2.2], which is only valid in separable Hilbert spaces and requires extra distributional assumptions.

- In Section 4.5 we study asymptotic normality of the estimator in separable Hilbert spaces. Theorem 4.73 provides weak Bahadur–Kiefer representations of the empirical quantile, which generalize results by Niemiro [199, Theorem 5], Arcones [11, Proposition 4.1] and Van der Vaart [259, Theorem 5.1] to infinite dimension. As an immediate consequence, Theorem 4.77 states the asymptotic normality of the empirical quantile, under distributional assumptions that exactly match those of the finite-dimensional case. This improves significantly on Gervini's normality result [100, Theorem 6]. This is the first central limit theorem for geometric quantiles that holds in a generic Hilbert space and under minimal assumptions.

The setting considered in this chapter is quite general compared to the existing literature: we consider geometric quantiles instead of geometric medians, general Banach spaces instead of Hilbert spaces, and our estimator is based on approximate minimization instead of exact minimization. However, the novelty of our contributions is not based solely on this generality. Indeed, our results on consistency in the norm topology and asymptotic normality improve the state of the art even in the special case where the parameter is the geometric median ($\ell = 0$), E is a Hilbert space, and the estimator is the exact empirical median ($\epsilon_n = 0$).

Proofs are deferred to appendices. For the reader's convenience, we provide precise references whenever we invoke a technical result from topology, functional analysis or measure theory.

4.2 Geometric quantiles, existence and uniqueness

4.2.1 Setting

Definition 4.1. Let $(E, \|\cdot\|)$ be a real normed vector space and let $(E^*, \|\cdot\|_*)$ denote its continuous dual space. Let $\ell \in E^*$ be such that $\|\ell\|_* < 1$ and μ be a Borel probability measure on E. We define the *objective function* ϕ_ℓ as follows:

$$\begin{aligned}\phi_\ell \colon E &\to \mathbb{R}\\ \alpha &\mapsto \int_E (\|\alpha - x\| - \|x\|) d\mu(x) - \ell(\alpha).\end{aligned}$$

We let X be a random element from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(E, \mathcal{B}(E))$ such that X has distribution μ, i.e., the corresponding pushforward measure is equal to μ. With this notation, the objective function rewrites as $\phi_\ell : \alpha \mapsto \mathbb{E}[\|\alpha - X\| - \|X\|] - \ell(\alpha)$.

The following proposition gives basic properties of ϕ_ℓ. Its proof is in Section 4.7.1.

Proposition 4.2. *1. ϕ_ℓ is well-defined, $(1 + \|\ell\|_*)$-Lipschitz and convex.*

2. $\lim_{\|\alpha\| \to \infty} \frac{\phi_0(\alpha)}{\|\alpha\|} = 1$ and $\lim_{\|\alpha\| \to \infty} \phi_\ell(\alpha) = \infty$.

3. *ϕ_ℓ is bounded below.*

Definition 4.3. We consider the following minimization problem:

$$\inf_{\alpha \in E} \phi_\ell(\alpha). \tag{4.1}$$

We let $\mathsf{Quant}_\ell(\mu)$ denote the subset of E where the infimum in (4.1) is attained. The elements of $\mathsf{Quant}_\ell(\mu)$ are called *geometric ℓ-quantiles* of the measure μ. When $\ell = 0$ we speak of *geometric medians* and we write $\mathsf{Med}(\mu)$ instead of $\mathsf{Quant}_0(\mu)$.

Remark 4.4. When no ambiguity arises, we will drop the ℓ-subscripts and write ϕ, $\mathsf{Quant}(\mu)$ for the sake of legibility.

The infimum in (4.1) is finite by Proposition 4.2. The set $\mathsf{Quant}(\mu)$ may be empty, a singleton or contain several elements. Some conditions for the existence and the uniqueness of minimizers are given in the next subsections.

4.2.2 The univariate case

We start with the univariate setting where $E = \mathbb{R}$ with the absolute value as norm. We identify ℓ with the corresponding scalar in $(-1, 1)$, so that $\ell(\alpha) = \ell \cdot \alpha$ and we define $p = (1+\ell)/2$ which lies in $(0, 1)$. We show, as is well-known, that the notion of geometric ℓ-quantile coincides with the usual definition of p-th quantile in one dimension: α must satisfy both $\mathbb{P}(X \leq \alpha) \geq p$ and $\mathbb{P}(X \geq \alpha) \geq 1 - p$.

Proposition 4.5. *We write F_X for the cdf of X.*

1. *Let $M_1 = \{\alpha \in \mathbb{R} : \mathbb{P}(X \leq \alpha) \geq p\} = \{\alpha \in \mathbb{R} : F_X(\alpha) \geq p\}$,*
   $M_2 = \{\alpha \in \mathbb{R} : \mathbb{P}(X \geq \alpha) \geq 1 - p\} = \{\alpha \in \mathbb{R} : F_X(\alpha^-) \leq p\}$.
   Then M_1 is an interval of the form $[\min(M_1), \infty)$, and M_2 is an interval of the form $(-\infty, \max(M_2)]$.

2. *The inequality $\min(M_1) \leq \max(M_2)$ holds and $\mathsf{Quant}(\mu)$ is the nonempty closed bounded interval $M_1 \cap M_2 = [\min(M_1), \max(M_2)]$.*

The statements of this subsection are all proved in Section 4.7.2. Existence is therefore guaranteed and uniqueness reduces to a standard problem. For the sake of completeness, the following corollary provides conditions for uniqueness of ℓ-quantiles in the univariate case, which we expect are already known. The first condition is stated in terms of the measure μ and the second in terms of the cdf F_X. The third item gives more details about the set $F_X^{-1}(\{p\})$ when there is more than one quantile.

Corollary 4.6. *With the notation of Proposition 4.5,*

1. *μ has at least two ℓ-quantiles if and only if there exist real numbers $\alpha_1 < \alpha_2$ such that $\mu((-\infty, \alpha_1]) = p$ and $\mu([\alpha_2, \infty)) = 1 - p$.*

2. *μ has a unique ℓ-quantile if and only if the set $F_X^{-1}(\{p\})$ is empty or a singleton.*

3. *If μ has at least two ℓ-quantiles, then $F_X < p$ over $(-\infty, \min(M_1))$, $F_X = p$ over $[\min(M_1), \max(M_2))$ and $F_X > p$ over $(\max(M_2), \infty)$.*

Remark 4.7. In particular, if a univariate measure has more than one ℓ-quantile it is possible to split its mass between two disjoint subsets (more precisely, between two disjoint half-lines). Besides, if $\alpha_1 < \alpha_2$ are as in the first item of Corollary 4.6, then the open interval (α_1, α_2) has measure 0, hence $\text{supp}(\mu) \cap (\alpha_1, \alpha_2) = \emptyset$ and the support of μ is disconnected. These two observations give convenient sufficient conditions on the measure μ that ensure the uniqueness of quantiles in one dimension. If the cdf of the associated random variable X is known, then uniqueness can be assessed simply by considering the preimage of p by F_X.

4.2.3 Existence in the general case

We turn now to the existence of geometric quantiles in the general setting of Definition 4.1. First we list some concepts and notations from topology and functional analysis that we will use below. Let $(F, \|\cdot\|)$ be a normed vector space over $\mathbb{R}$ or C_{20}. We recall that F^* and F^{**} denote respectively the topological dual and second dual of F. These two vector spaces are equipped with their dual norms, which we write respectively $\|\cdot\|_*$ and $\|\cdot\|_{**}$. Let $J : F \to F^{**}$ be the canonical linear isometry from F into F^{**}. F is said to be *reflexive* if J is surjective. We will say that the subspace $J(F)$ is 1-*complemented* in F^{**} if there is a linear projection operator $P : F^{**} \to F^{**}$ with range equal to $J(F)$ and operator norm $\|P\|_{op}$ equal to 1. F is said to be *separable* if it contains a countable dense subset. We give proofs for the statements of this subsection in Section 4.7.3.

The next proposition provides three sufficient conditions for existence, which involve only topological properties of the space E, independently of the measure μ.

Proposition 4.8. *The measure μ has at least one geometric ℓ-quantile in any of the following cases:*

1. *[253, 254] E is a reflexive space.*
2. *[148] There is an isometric isomorphism I between E and F^*, where F is a separable normed space and $\ell \circ I^{-1} \in J(F)$.*
3. *E is separable, $J(E)$ is 1-complemented in E^{**} and $\ell = 0$.*

Remark 4.9. For medians ($\ell = 0$), the first condition of Proposition 4.8 was given by Valadier [253, 254], and the second was stated in less generality by Kemperman [148]. These two conditions already cover a large number of spaces that are used in applications, with the notable exception of L^1 spaces. Kemperman states that "medians always exist for many L^1 spaces", but he only proves it in the very special case $L^1(S, \mathcal{P}(S), \nu)$ where S is a countable set and ν is a measure supported on a subset of S [148, Corollary 3.2].

Remark 4.10. Our third condition, which is new, is more intricate as it exploits complementability in the second dual: Lemma 4.81 and the proof of Proposition 4.8 in

Section 4.7.3 reveal a link between geometric medians of μ and geometric medians of some pushforward of μ in E^{**}. This third condition covers separable L^1 spaces, and separability is verified in a number of usual settings, as seen in the first item of Corollary 4.13. The proof of existence in this case is technical. The requirement that $\ell = 0$ seems to be an artifact of our proof technique.

Remark 4.11. Items 1. and 2. in the proposition clearly imply that E is a Banach space. For the third, letting $P : E^{**} \to E^{**}$ denote a bounded linear projection with range $J(E)$, we have $J(E) = \ker(\mathrm{Id} - P)$. Therefore, $J(E)$ is a closed subspace of E^{**}, so the space $(J(E), \|\cdot\|_{**})$ is Banach and so is $(E, \|\cdot\|)$. It is unclear to us if there is sufficient condition that would not require the completeness of E.

The following corollaries provide a list of spaces in which ℓ-quantiles or medians are guaranteed to exist. Among these, some can originally be defined as vector spaces over the field C_{21}; this is especially the case when F is a complex Hilbert space or when F is the space of Schatten p-class operators on a complex Hilbert space. In such circumstances we put $E = F_{\mathbb{R}}$, the real vector space obtained by restriction of the scalar multiplication to $\mathbb{R} \times F$. To avoid notational overburden we keep these subscripts implicit in the statements of the corollaries.

Corollary 4.12. *The measure μ has at least one ℓ-quantile in any of the following cases (as explained in the previous paragraph each space is viewed as a real vector space):*

1. *E is finite-dimensional and equipped with any norm,*
2. *E is a Hilbert space equipped with its Hilbert norm,*
3. *$E = L^p(S, \mathcal{A}, \nu)$ equipped with the L^p norm, where $1 < p < \infty$ and $(S, \mathcal{A}, \nu)$ is any measure space,*
4. *$E = W^{k,p}(\Omega)$ a Sobolev space with the Sobolev norm $\|u\|_{k,p} = (\sum_{|\alpha| \le k} \|D^\alpha u\|^p_{L^p(\Omega)})^{1/p}$, where Ω is an open subset of $\mathbb{R}^n$, k and n are positive integers and $1 < p < \infty$,*
5. *$E = L^\Phi(S, \mathcal{A}, \nu)$ an Orlicz space equipped with its Orlicz norm or its gauge (Luxemburg) norm, where $(S, \mathcal{A}, \nu)$ is any measure space, Φ is a Young function such that Φ and its complementary function Ψ both satisfy the Δ_2 condition (see [218] for terminology),*
6. *$E = S_p(H)$ the space of Schatten p-class operators equipped with the Schatten p-norm, where $1 < p < \infty$ and H is a Hilbert space.*

Corollary 4.13. *The measure μ has at least one geometric median in any of the following cases (each space is viewed as a real vector space):*

1. *$E = L^p(S, \mathcal{A}, \nu)$ equipped with the L^p norm, where $p \in \{1, \infty\}$, $(S, \mathcal{A}, \nu)$ is a sigma-finite measure space and $\mathcal{A}$ is countably generated. This includes the case where $(S, \mathcal{A})$ is a separable metric space with its Borel sigma-algebra and the case where $(S, \mathcal{A})$ is a countable space with its discrete sigma-algebra.*
2. *$E = BV(\Omega)$ the space of functions of bounded variation equipped with the BV norm, where Ω is an open subset of $\mathbb{R}^n$.*

3. *$E = S_1(H)$ the space of trace-class operators equipped with the trace norm, where H is a separable Hilbert space.*

4. *$E = B(H)$ the space of bounded operators on a separable Hilbert space H, equipped with the operator norm.*

Remark 4.14. The nonexistence of geometric quantiles is a possibility. In [170] the authors consider the Banach space c_0 of real sequences that converge to 0, equipped with the supremum norm, and they construct a Borel probability measure μ such that $\mathsf{Med}(\mu) = \emptyset$.

4.2.4 Uniqueness in the general case

Now that we have shed some light on the existence of geometric quantiles, we turn to the question of uniqueness. Proofs for this subsection are given in Section 4.7.4. Contrary to the univariate case, in general spaces the set of minimizers $\mathsf{Med}(\mu)$ may be empty. Consequently when we speak of uniqueness in this section, we mean the situation where a measure has at most one ℓ-quantile.

Unlike existence, the study of uniqueness involves geometric properties of both the space E and the measure μ.

Definition 4.15. Let $(F, \|\cdot\|)$ be a normed space over $\mathbb{R}$ or C_{22}.

1. F is *strictly convex* (or *strictly rotund*) if for every distinct unit vectors $x, y \in F$, the inequality $\|x + y\| < 2$ holds. Equivalently, the unit sphere of F contains no nontrivial line segments.

2. F is *uniformly convex* (or *uniformly rotund*) if

$$\forall \epsilon > 0, \exists \delta > 0, \forall (x, y) \in E^2, [\|x\| = \|y\| = 1 \text{ and } \|x-y\| \geq \epsilon] \implies \|\tfrac{1}{2}(x+y)\| \leq 1-\delta.$$

Definition 4.16. We let $\mathcal{M}_\sim$ denote the set of Borel probability measures μ on E that are not concentrated on a line, i.e., $\mu(L) < 1$ for every affine line L. We write $\mathcal{M}_-$ its complement.

The following proposition gives a sufficient condition for $\mathsf{Quant}(\mu)$ to contain at most one element; this condition already appears for medians in [148, 191] and for quantiles in [65]. As a novel result, we provide a converse statement that exhibits an interplay between uniqueness and the geometry of the space E.

Proposition 4.17.

1. *[148, 191, 65] If E is strictly convex and $\mu \in \mathcal{M}_\sim$, then μ has at most one ℓ-quantile.*

2. *If any of these two conditions is dropped, μ may have more than one ℓ-quantile.*

3. *Suppose that every $\mu \in \mathcal{M}_\sim$ has at most one median. Then E is strictly convex.*

As a consequence, when $\mu \in \mathcal{M}_\sim$ we obtain the following list of spaces for which $\mathsf{Quant}(\mu)$ is a singleton, i.e., there is both existence and uniqueness of a median. Remarkably, all these spaces are reflexive. Besides, the list includes every uniformly convex Banach space. Uniform convexity is a strong condition, since it implies both strict convexity and reflexivity. As in Corollary 4.12, complex vector spaces are regarded as real vector spaces and we also make $\mathbb{R}$-subscripts implicit in the following statement.

Corollary 4.18. *Let $\mu \in \mathcal{M}_\sim$ be a measure on E. The existence and uniqueness of a geometric ℓ-quantile for μ is guaranteed in any of the following cases (each space is viewed as a real vector space):*

1. *E is a uniformly convex Banach space, e.g.,*
   - (a) *E is finite-dimensional and strictly convex,*
   - (b) *E is a Hilbert space equipped with its Hilbert norm,*
   - (c) *$E = L^p(S, \mathcal{A}, \nu)$ as in Corollary 4.12,*
   - (d) *$E = W^{k,p}(\Omega)$ as in Corollary 4.12,*
   - (e) *$E = S_p(H)$ as in Corollary 4.12,*
2. *$E = L^\Phi(S, \mathcal{A}, \nu)$ an Orlicz space equipped with its Orlicz norm, where $(S, \mathcal{A}, \nu)$ is a sigma-finite measure space and ν is diffuse, Φ is a strictly convex N-function such that both Φ and its complementary function Ψ satisfy the Δ_2 condition,*
3. *$E = L^\Phi(S, \mathcal{A}, \nu)$ an Orlicz space equipped with its gauge (Luxemburg) norm, where $(S, \mathcal{A}, \nu)$ is a measure space and ν is diffuse on some set of positive measure, Φ is a finite strictly convex Young function such that both Φ and its complementary function Ψ satisfy the Δ_2 condition.*

So far we have tackled uniqueness for measures in $\mathcal{M}_\sim$. In the next proposition we consider members of $\mathcal{M}_-$, i.e., measures that are concentrated on some affine line. We show that if E is strictly convex, any geometric median must lie on the supporting line. The problem thus becomes completely univariate so Proposition 4.5 applies: $\mathsf{Med}(\mu)$ is a nonempty closed line segment and uniqueness can be addressed with Corollary 4.6. Strikingly, this does not hold for ℓ-quantiles with $\ell \neq 0$. As in Proposition 4.17 we provide a converse that illustrates the interconnection between medians of measures and geometric features of the space. To the best of our knowledge, these results are new.

Proposition 4.19. *Let $\mu \in \mathcal{M}_-$ and let L denote an affine line such that $\mu(L) = 1$.*

1. *If E is strictly convex, then the set of medians $\mathsf{Med}(\mu)$ is a nonempty closed line segment included in L.*
2. *Without the strict convexity hypothesis, the conclusion of 1. may not be true.*
3. *In 1., $\mathsf{Med}(\mu)$ cannot be replaced with $\mathsf{Quant}_\ell(\mu)$ for arbitrary $\ell \neq 0$.*
4. *Suppose that for every $\mu \in \mathcal{M}_-$, $\mathsf{Med}(\mu)$ is a closed line segment included in the affine line supporting μ. Then E is strictly convex.*

4.3 Empirical geometric quantiles: measurability, selections and uniqueness

4.3.1 Estimation setting

Estimating geometric medians fits the general framework of M-estimation [133, Section 6.2], which we quickly recall in its simplest form. There is a parameter space Θ, a probability space $(\mathcal{X}, \mathcal{A}, \mu)$, a contrast function $\varphi : \mathcal{X} \times \Theta \to \mathbb{R}$ that is integrable in the first argument, giving rise to the objective function ϕ

$$\phi : \theta \mapsto \int_{\mathcal{X}} \varphi(x, \theta) d\mu(x).$$

Given an i.i.d. sample $X_1, X_2 \ldots \sim \mu$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, an estimator is obtained by approximate minimization of the empirical objective function $\widehat{\phi}_n$

$$\widehat{\phi}_n : \theta \mapsto \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i, \theta).$$

In our case, $\Theta = \mathcal{X} = E$ a real normed vector space, $\mathcal{A} = \mathcal{B}(E)$ its Borel σ-algebra, μ is a fixed Borel probability measure on E, and φ is the function

$$\varphi : (x, \alpha) \mapsto \|\alpha - x\| - \|x\| - \ell(\alpha).$$

The following definition gives the precise setting and provides additional terminology.

Definition 4.20. Let $(X_n)_{n \geq 1}$ be a sequence of i.i.d. E-valued Borel random elements defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, each with distribution μ. Additionally let $(\epsilon_n)_{n \geq 1}$ be a sequence of (not necessarily measurable) maps from Ω to $\mathbb{R}_{\geq 0}$. For every $n \geq 1$ we define the *empirical measure* $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$, the *empirical objective function* $\widehat{\phi}_n : \alpha \mapsto \frac{1}{n} \sum_{i=1}^{n} (\|\alpha - X_i\| - \|X_i\|) - \ell(\alpha)$ and the *set of ϵ_n-empirical ℓ-quantiles*

$$\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) = \{\alpha \in E : \widehat{\phi}_n(\alpha) \leq \inf(\widehat{\phi}_n) + \epsilon_n\}.$$

We say that $(\widehat{\alpha}_n)_{n \geq 1}$ is a *sequence of ϵ_n-empirical ℓ-quantiles* if for all $n \geq 1$, $\widehat{\alpha}_n$ is a (not necessarily measurable) map from Ω to E such that $\widehat{\alpha}_n \in \epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$.

The quantities X_n, ϵ_n, $\widehat{\mu}_n$, $\widehat{\phi}_n$, $\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$, $\widehat{\alpha}_n$ all depend on $\omega \in \Omega$; when needed, the dependence will be indicated with a superscript, e.g., X_n^ω, ϵ_n^ω, $\widehat{\mu}_n^\omega$, $\widehat{\phi}_n^\omega$, $\widehat{\alpha}_n^\omega$. In the definition of $(\widehat{\alpha}_n)$ above, we mean more precisely that

$$\forall n \geq 1, \forall \omega \in \Omega, \, \widehat{\alpha}_n^\omega \in \epsilon_n^\omega\text{-}\mathsf{Quant}(\widehat{\mu}_n^\omega).$$

As is customary in the theory of M-estimation, we consider approximate minimizers of $\widehat{\phi}_n$, i.e., elements of E that are ϵ_n-optimal. When $\epsilon_n = 0$ we recover exact empirical ℓ-quantiles and in that case we write $\mathsf{Quant}(\widehat{\mu}_n)$ instead of $0\text{-}\mathsf{Quant}(\widehat{\mu}_n)$. For each n, the function $\widehat{\phi}_n$ is obtained by replacing the measure μ in Definition 4.1 with the empirical measure $\widehat{\mu}_n$. Consequently the existence and uniqueness results developed in Section 4.2 apply equally well to each $\widehat{\phi}_n$. By Proposition 4.2 the infimum of $\widehat{\phi}_n$ is finite, thus the set $\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$ is nonempty whenever ϵ_n is positive. However the following assumption is needed to cover the case $\epsilon_n = 0$.

Assumption (A2). *E is a separable Banach space that verifies any of the conditions in Proposition 4.8.*

Under this assumption a sequence of exact empirical ℓ-quantiles is always guaranteed to exist. Even though the completeness of the normed space E is a byproduct of Proposition 4.8, we add it to the assumption for the sake of clarity. Moreover the separability of E is a natural hypothesis, as it will be needed for crucial facts, such as the equality between σ-algebras $\mathcal{B}(E^2) = \mathcal{B}(E) \otimes \mathcal{B}(E)$ and the weak convergence of $\widehat{\mu}_n$ to μ with $\mathbb{P}$-probability 1.

4.3.2 Measurability

Measurability issues arise naturally in this chapter, especially because we will state asymptotic results valid for any sequence $(\widehat{\alpha}_n)_{n\geq 1}$ of ϵ_n-empirical ℓ-quantiles, regardless of whether each $\widehat{\alpha}_n$ is measurable. Therefore, we will repeatedly want to evaluate the probability of subsets of Ω that may not be in the σ-algebra $\mathcal{F}$. Besides, in order to match the generality of the M-estimation works by Huber [132], Perlman [210] and Dudley [83], we do not require that the maps ϵ_n of Definition 4.20 be measurable. And yet, we will often assume that the sequence $(\epsilon_n)_{n\geq 1}$ converges stochastically to 0 in some way.

To resolve these measurability difficulties we will employ the notions of outer and inner probability $\mathbb{P}^*, \mathbb{P}_*$. For any subset B of Ω they are defined respectively as $\mathbb{P}^*(B) = \inf\{\mathbb{P}(A) : A \in \mathcal{F}, B \subset A\}$ and $\mathbb{P}_*(B) = 1 - \mathbb{P}^*(B^c)$. Some useful properties of $\mathbb{P}^*$ and $\mathbb{P}_*$ are stated in Lemma 4.85 of Section 4.8.1. Further properties can be found in Chapter 1.2 of Van der Vaart and Wellner [261].

We make use of the adjective "stochastic" to designate any object or notion that depends on $\omega \in \Omega$ and is subject to a lack of measurability. Conversely we reserve the adjective "random" for quantities that are measurable. We will say that a stochastic property holds $\mathbb{P}_*$-*almost surely* if the subset of Ω where the property is verified has inner probability 1.

A theory of stochastic convergence for arbitrary maps can be found in Chapter 1.9 of [261]. We recall the definitions of three modes of convergence that will be needed in this chapter, as well as some asymptotic notations.

Definition 4.21. Let $Y, Y_1, Y_2, \ldots$ be maps from Ω to some topological space F.

1. $(Y_n)_{n\geq 1}$ converges $\mathbb{P}_*$-*almost surely* to Y if $\mathbb{P}_*(\{\omega : Y_n^\omega \to Y^\omega\}) = 1$.

We assume next that F is a metric space with metric d.

2. $(Y_n)_{n\geq 1}$ converges *in outer probability* to Y if $\mathbb{P}^*\big(d(Y_n, Y) > \epsilon\big) \to 0$ for each $\epsilon > 0$.

3. $(Y_n)_{n\geq 1}$ converges *outer almost surely* to Y if there exist random variables $\Delta_1, \Delta_2, \ldots$ such that $d(Y_n, Y) \leq \Delta_n$ for each n and $(\Delta_n)_{n\geq 1}$ converges $\mathbb{P}$-almost surely to 0.

Remark 4.22. Van der Vaart and Wellner refer to the first mode as "convergence almost surely". However, for clarity and consistency with the terminology of the preceding paragraph, when measurability is not guaranteed we prefer the terminology "convergence $\mathbb{P}_*$-almost surely".

Definition 4.23. Let $Y_1, Y_2, \ldots$ be maps from Ω to some normed space $(F, \|\cdot\|)$ and $(a_n)_{n\geq 1}$ be a sequence of nonzero real numbers.

1. We write $Y_n = o_{\mathbb{P}^*}(a_n)$ to signify that $a_n^{-1} Y_n$ converges in outer probability to 0.
2. We write $Y_n = O_{\mathbb{P}^*}(a_n)$ when for every $\varepsilon > 0$, there exists $M > 0$ such that

$$\forall n \geq 1,\, \mathbb{P}^*(\|a_n^{-1} Y_n\| > M) < \varepsilon.$$

4.3.3 Measurable selections

While we have the tools to deal with non-measurability, it is generally more convenient and less technical to work with measurable quantities. This is why statisticians often look for measurable selections, which we define next.

Definition 4.24. Let $n \geq 1$. We say that the map $\widehat{\alpha}_n : \Omega \to E$ is a *measurable selection* from the set ϵ_n-$\mathsf{Quant}(\widehat{\mu}_n)$ if it is $(\mathcal{F}, \mathcal{B}(E))$-measurable and $\widehat{\alpha}_n^\omega$ belongs to ϵ_n^ω-$\mathsf{Quant}(\widehat{\mu}_n^\omega)$ for each $\omega \in \Omega$.

When such a selection is found for each $n \geq 1$, one considers the resulting sequence of Borel random elements $(\widehat{\alpha}_n)_{n\geq 1}$, the analysis of which involves fewer technicalities compared with a non-measurable sequence.

Previous works related to M-estimation (e.g., [8, 108, 234, 2, 36, 238]) have relied on [142, Lemma 2; 212, Theorem 1.9; 47, Corollary 1] to obtain measurable selections. Lemma 2 in [142] is only suited to the finite-dimensional case. The statement of [212, Theorem 1.9] (resp., of [47, Corollary 1]) has a local compactness (resp., σ-compactness) assumption, which in our setting requires that E be locally compact (resp., σ-compact). Either of these conditions excludes infinite-dimensional Banach spaces and is therefore too restrictive for our purposes. Another classical reference for measurable selections is the survey by Wagner [266] and its update [267]. In Section 9 of [266] selection results are listed for the setting of optimization problems. In most of the references it is assumed that "F is compact-valued", which means for us that E is compact. One exception is [169, Proposition 14.8] but their "Suslin operation" assumption does not hold here. The second exception is Theorem 1 in Schäl [228] which is applicable to ϵ_n-empirical ℓ-quantiles under a mild assumption on ϵ_n. We obtain as a consequence the following theorem, however we prove it via a different and simpler technique in Section 4.8.2.

Theorem 4.25. *Let $n \geq 1$. Assume that E is separable and ϵ_n is a positive random variable. Then a measurable selection from ϵ_n-$\mathsf{Quant}(\widehat{\mu}_n)$ exists.*

This selection theorem is not applicable when the map ϵ_n is allowed to take the value 0. In that situation the existence of empirical ℓ-quantiles is no longer automatic and we will need Assumption (A2). Using Theorem 2 (ii) in Brown and Purves [47] or Proposition 4.2 (c) in Hess [116], we obtain a *universally measurable* selection from the set $\mathsf{Quant}(\widehat{\mu}_n) = 0$-$\mathsf{Quant}(\widehat{\mu}_n)$ (see Definition 4.86 of the Appendix). Universal measurability is a weaker concept than measurability and the following assumption is needed to obtain measurability in the usual sense.

Assumption (A3). $(\Omega, \mathcal{F}, \mathbb{P})$ *is a complete probability space, i.e.,*

$$\forall (S, N) \in \mathcal{P}(\Omega) \times \mathcal{F}, \quad [S \subset N \text{ and } \mathbb{P}(N) = 0] \implies S \in \mathcal{F}.$$

This is not a strong assumption since a probability space can always be uniquely completed. Besides, if $(\Omega, \mathcal{F}, \mathbb{P})$ is complete, subsets $A \subset \Omega$ such that $\mathbb{P}^*(A) = 0$ or $\mathbb{P}_*(A) = 1$ are automatically in $\mathcal{F}$ (see Problem 10 in [261, Chapter 1.2]). We can now state a second selection theorem, which holds irrespective of the measurability of ϵ_n.

Theorem 4.26. *Let* $n \geq 1$*. Under Assumptions (A2) and (A3), a measurable selection from* $\mathsf{Quant}(\widehat{\mu}_n)$ *exists, hence from* ϵ_n-$\mathsf{Quant}(\widehat{\mu}_n)$ *as well.*

Remark 4.27. Sinova et al. [238, Proposition 3.3] state a selection result for generic M-estimators in separable Hilbert spaces. Their proof relies on [47, Corollary 1], which results in a σ-compactness requirement on the space and excludes the infinite-dimensional setting. Our selection Theorems 4.25 and 4.26 are an improvement in this regard.

4.3.4 Uniqueness of empirical quantiles

We close this section with a novel uniqueness result for empirical ℓ-quantiles in the case where E is strictly convex and μ is in $\mathcal{M}_\sim$, i.e., $\mu(L) < 1$ for every affine line L. In this setting, μ has at most one geometric ℓ-quantile by Proposition 4.17. First we show a stronger separating inequality for μ: the measure cannot get arbitrarily close to 1 on affine lines. Proofs for this subsection are in Section 4.8.3.

Proposition 4.28. *Each* $\mu \in \mathcal{M}_\sim$ *is separated away from* 1 *on affine lines: there exists* $\delta_\mu \in (0, 1]$ *such that for any affine line* L *we have the inequality* $\mu(L) \leq 1 - \delta_\mu$*.*

To show that empirical measures $\widehat{\mu}_n$ have at most one ℓ-quantile, it suffices to prove that they inherit the separation property of μ. To this end, we establish a Glivenko–Cantelli result for the class of affine lines. In fact we obtain one for the slightly larger class

$$\mathcal{C} = \{u + \mathbb{R}v : (u, v) \in E^2\}$$

of singletons and affine lines: this is no more difficult and actually removes some notational burden in the proof. We show that $\mathcal{C}$ is Vapnik–Červonenkis, then we use the theory of empirical processes.

The class $\mathcal{C}$ is neither countable, nor does it contain a countable subclass $\mathcal{C}_0$ verifying $\sup_{C \in \mathcal{C}} |\widehat{\mu}_n(C) - \mu(C)| = \sup_{C \in \mathcal{C}_0} |\widehat{\mu}_n(C) - \mu(C)|$, so the supremum of interest may not be measurable. Two standard workarounds are described in Section 2.2 of Ledoux and Talagrand [168]. The first approach is to focus instead on the essential (or lattice) supremum of the process $(|\widehat{\mu}_n(C) - \mu(C)|)_{C \in \mathcal{C}}$, which is equal to $\sup_{C \in \mathcal{C}_0} |\widehat{\mu}_n(C) - \mu(C)|$ for some countable $\mathcal{C}_0 \subset \mathcal{C}$ (see, e.g., [237, Definition and Lemma p.230]). However this restricted supremum is not suitable for obtaining the separation property, since we want an upper bound on $\widehat{\mu}_n(L)$ for every affine line L. The second approach is to replace the process with a separable version $(\Lambda_C)_{C \in \mathcal{C}}$. Then for fixed $C \in \mathcal{C}$, the equality $|\widehat{\mu}_n^\omega(C) - \mu(C)| = \Lambda_C^\omega$ holds for ω in the complement of a null set N_C. The uncountable union $\cup_{C\in\mathcal{C}} N_C$ may not be in $\mathcal{F}$ and it may not have outer probability 0 either; this is a major hindrance for our purpose. Other tools are therefore needed to deal with the supremum: we make use of the theory developed by Van der Vaart and Wellner [261].

We apply a Glivenko–Cantelli theorem based on random L_1-entropy and symmetrization: this forces $(\Omega, \mathcal{F}, \mathbb{P})$ in Definition 4.20 to be the countable product space $(E^{\mathbb{N}}, \mathcal{B}(E)^{\otimes\mathbb{N}}, \mu^{\mathbb{N}})$ and X_n to be the n-th coordinate map. This is not a strong requirement since it could be assumed without loss of generality. Additionally we have to verify that the corresponding class of indicators $\mathcal{F} = \{\mathbb{1}_C : C \in \mathcal{C}\}$ is μ-measurable (see [261, Definition 2.3.3]). For this, standard methods described in [261, Examples 2.3.4 and 2.3.5] are not applicable. To resolve that challenging technicality, our proof makes use of image admissible Suslin classes, a notion developed by Dudley [81, Section 10.3; 82, Section 5.3]. Now we can state the following proposition.

Proposition 4.29. *Assume E is separable, $(\Omega, \mathcal{F}, \mathbb{P})$ is the product probability space $(E^{\mathbb{N}}, \mathcal{B}(E)^{\otimes\mathbb{N}}, \mu^{\mathbb{N}})$ and $X_n : \Omega \to E$ is the n-th coordinate map for each $n \geq 1$. Let*

$$\mathcal{C} = \{u + \mathbb{R}v : (u, v) \in E^2\}$$

denote the class of singletons and affine lines in E. Then the stochastic quantity

$$\sup_{C\in\mathcal{C}} |\widehat{\mu}_n(C) - \mu(C)|$$

converges to 0 outer almost surely.

Combining Propositions 4.28 and 4.29 yields the following asymptotic uniqueness theorem, which is new.

Theorem 4.30. *We require the assumptions of Proposition 4.29, strict convexity of E and $\mu \in \mathcal{M}_{\sim}$. Then the following holds $\mathbb{P}_*$-almost surely: for large enough n the empirical measure $\widehat{\mu}_n$ has at most one geometric ℓ-quantile.*

In other words, with $\mathbb{P}_*$-probability 1 the set $\mathsf{Quant}(\widehat{\mu}_n)$ is empty or a singleton for large enough n. If we add the existence and the completeness assumptions (A2), (A3) then Theorem 4.26 applies and there exists a measurable selection from $\mathsf{Quant}(\widehat{\mu}_n)$ for each n. In that case, $\mathbb{P}$-almost surely, the selection is trivial for large enough n.

4.4 Convergence of approximate empirical quantiles

In this section we investigate the asymptotic behaviour of the set-valued stochastic sequence $(\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n))_{n\geq 1}$. The analysis can be carried out in two different settings:

(a) μ is allowed to have multiple ℓ-quantiles,

(b) μ has a unique ℓ-quantile, say $\alpha_\star$: $\mathsf{Quant}(\mu) = \{\alpha_\star\}$.

In setting (a) we look for any kind of statement that may indicate closeness of $\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$ to $\mathsf{Quant}(\mu)$ for large n. Setting (b) fits the usual framework of estimation theory: we consider for each n an element $\widehat{\alpha}_n$ from the set $\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$, and we are interested in the convergence of the stochastic sequence $(\widehat{\alpha}_n)_{n\geq 1}$ to the unknown parameter $\alpha_\star$.

4.4.1 The case of multiple true quantiles

Variational convergence

Asymptotic statements about empirical quantiles, even in the absence of uniqueness, can be obtained using the theory of variational convergence [14; 20; 129, Section 7.5; 40, Section 6.2], which introduces several different but related ways in which sequences of sets and functions converge. We will be interested in two such kinds of convergence: the first one is epi-convergence, also known as Kuratowski–Painlevé convergence [40] or as Γ-convergence [70, 43]; and the second is Mosco-convergence [196, 21, 20, 40]. The following definitions are given in the context of lower semicontinuous proper convex functions on a normed space. This setting, which is well suited for our purposes but not the most general, allows for simpler definitions.

Definition 4.31. Let F be a real normed space and $f, (f_n)_{n\geq 1}$ be lower semicontinuous convex functions defined on F with values in $\mathbb{R}$.

1. The sequence $(f_n)_{n\geq 1}$ epi-converges to f if for each $x \in E$ both of the following conditions hold:
   (i) $\liminf_n f_n(x_n) \geq f(x)$ for every sequence $(x_n)_{n\geq 1}$ that converges in the norm topology to x,
   (ii) $\limsup_n f_n(x_n) \leq f(x)$ for some sequence $(x_n)_{n\geq 1}$ that converges in the norm topology to x.
2. The sequence $(f_n)_{n\geq 1}$ Mosco-converges to f if for each $x \in E$ both of the following conditions hold:
   (i) $\liminf_n f_n(x_n) \geq f(x)$ for every sequence $(x_n)_{n\geq 1}$ that converges to x in the weak topology of E,
   (ii) $\limsup_n f_n(x_n) \leq f(x)$ for some sequence $(x_n)_{n\geq 1}$ that converges in the norm topology to x.

Since convergence in the norm topology implies convergence in the weak topology, Mosco-convergence implies epi-convergence. If F is finite-dimensional, they are equivalent.

Various works in statistics (e.g., [84, 116, 271, 29, 102, 76, 226, 231]) and stochastic optimization (e.g., [146, 222, 152, 176, 277, 117]) have leveraged variational convergence to study the consistency of estimators defined through a minimization procedure. Indeed, the following proposition shows that epi- or Mosco-convergence of (f_n) to f results in some kind of asymptotic closeness between the convex sets $\varepsilon_n\text{-}\arg\min f_n = \{x : f_n(x) \leq \inf(f) + \varepsilon_n\}$ and $\arg\min f$. Given $(\varepsilon_n)_{n\geq 1}$ a deterministic sequence of nonnegative real numbers, we say that $(x_n)_{n\geq 1}$ is a *sequence of ε_n-minimizers* if for all $n \geq 1$, $x_n \in \varepsilon_n\text{-}\arg\min f_n$.

Proposition 4.32. *Let F be a normed vector space and $f, (f_n)_{n\geq 1}$ be lower semicontinuous, proper convex functions defined on F with values in $\mathbb{R}$. Let $(\varepsilon_n)_{n\geq 1}$ be a sequence of nonnegative real numbers with $\lim_{n\to\infty} \varepsilon_n = 0$, and let $(x_n)_{n\geq 1}$ denote a sequence of*

ε_n-minimizers. Assume that $(f_n)_{n\geq 1}$ epi-converges (resp., Mosco-converges) to f. If some subsequence $(x_{n_k})_{k\geq 1}$ converges in the norm topology (resp., in the weak topology) to some $x \in E$, then $x \in \arg\min f$.

In other words if a sequence of ε_n-minimizers has a convergent subsequence, then the subsequential limit is itself a minimizer of f. An elementary proof of Proposition 4.32 is given in Section 4.9.1.

Application to geometric quantiles

Next we return to the setting of empirical ℓ-quantiles introduced in Section 4.3.1. Before diving into convergence results we state the following assumption, which controls the degree of optimality of the minimizers as n goes to infinity.

Assumption (A4). *$(\epsilon_n)_{n\geq 1}$ converges $\mathbb{P}_*$-almost surely to 0.*

As a first convergence statement we mention Kemperman's result of asymptotic closeness between the sets $\mathsf{Med}(\widehat{\mu}_n)$ and $\mathsf{Med}(\mu)$ in the special case where E is finite-dimensional. For any subset A of E and $\delta > 0$, we let $A^\delta = \{x \in E : \exists \alpha \in A, \|x - \alpha\| < \delta\}$ denote the δ-fattening (also known as δ-enlargement) of A.

Theorem 4.33 (Theorem 2.24 in [148])**.** *Suppose E is finite-dimensional. Then $\mathbb{P}_*$-almost surely:*

$$\forall \delta > 0, \exists N \geq 1, \forall n \geq N, \mathsf{Med}(\widehat{\mu}_n) \subset \mathsf{Med}(\mu)^\delta.$$

Remark 4.34. In words, for any δ and sufficently large n, each empirical median is at least δ-close to some true median. When E has finite dimension the weak topology coincides with the norm topology, hence Theorem 4.33 clearly implies with $\mathbb{P}_*$-probability 1 the conclusion of Proposition 4.32: for any sequence of empirical ℓ-medians $(\widehat{\alpha}_n)_{n\geq 1}$, if some subsequence $(\widehat{\alpha}_{n_k})_{k\geq 1}$ converges to some $\alpha \in E$ then α is an ℓ-median of μ. Kemperman's result relies crucially on the compactness (w.r.t. the norm topology) of closed balls, which is only valid in finite-dimensional spaces. As a straightforward consequence of our results below, we obtain in Corollary 4.44 a generalization of this theorem to ϵ_n-empirical ℓ-quantiles.

Remark 4.35. Kemperman's original theorem is *deterministic*, in the sense that he considers an arbitrary sequence of probability measures $(\mu_n)_{n\geq 1}$ that converges weakly (in the usual sense for measures) to μ, whereas we work with the random measure $\widehat{\mu}_n$. The result we state above is a slight generalization to the random setting.

In infinite dimension the study of empirical ℓ-quantiles is amenable to Mosco-convergence. We show indeed the stronger statement that $\mathbb{P}$-almost surely $(\widehat{\phi}_n)_{n\geq 1}$ converges uniformly on bounded sets to ϕ. Kemperman had proved [148, Section 2.19] in the deterministic setting (see Remark 4.35) that $(\widehat{\phi}_n)_{n\geq 1}$ converges uniformly on compact sets to ϕ, which is useful only when E is finite-dimensional. Our result is an improvement in this regard.

Proposition 4.36. *Assume E is separable. Then $\mathbb{P}$-almost surely, the sequence of functions $(\widehat{\phi}_n)_{n\geq 1}$ converges uniformly on bounded sets to ϕ.*

The proofs for results in this subsection are in Section 4.9.2. We show that the subset of Ω under consideration in the proposition belongs to $\mathcal{F}$, so there is no measurability hurdle here. Since uniform convergence on bounded sets implies Mosco-convergence (see [40, Theorem 6.2.14]), Proposition 4.32 applies. By combining it with Proposition 4.36, we obtain the following convergence theorem for empirical ℓ-quantiles. It relies on the weak topology of E, which is not metrizable or even first-countable in general, hence the need for inner probability.

Theorem 4.37. *Under Assumptions (A2) and (A4), the following statement holds $\mathbb{P}_*$-almost surely: for any sequence of ϵ_n-empirical ℓ-quantiles $(\widehat{\alpha}_n)_{n\geq 1}$, if some subsequence $(\widehat{\alpha}_{n_k})_{k\geq 1}$ converges in the weak topology of E to some $\alpha \in E$, then α is an ℓ-quantile of μ.*

Remark 4.38. Some works in stochastic optimization have applied the theory of variational convergence to study general optimization problems in which the objective function ϕ has the form $\phi : \alpha \mapsto \int_S g(\alpha, x)d\mu(x)$ where $g : E \times S \to \mathbb{R}$ is a generic integrand, E is a metric space and $(S, \mathcal{S}, \mu)$ is a probability space. In references [84, 152, 176, 116, 277] the authors assume various topological conditions on the spaces E, S and various regularity conditions on g. Additionally they consider a sequence of probability measures $(\mu_n)_{n\geq 1}$ (deterministic or random) that converges weakly to μ, from which they obtain approximating functions $\phi_n : \alpha \mapsto \int_S g(\alpha, x)d\mu_n(x)$, and they study epi- or Mosco-convergence of ϕ_n to ϕ. Epi-convergence (which is weaker than Mosco-) in the infinite-dimensional setting is proved in [116, Theorem 5.1] and [277, Theorem 2]; both of these theorems apply to quantiles. Mosco-convergence in the infinite-dimensional setting is established in [152, Theorem 2.4] and [176, Theorem 14]; however these results do not cover quantiles: regarding [152] the integrability condition on the subderivative is not verified, while in [176] the authors' Condition 4 is not met. Consequently, to our knowledge, our theorem cannot be obtained from existing results in the literature on variational convergence.

Remark 4.39. Theorem 4.37 does not mean that the closed convex sets $\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$ converge in the usual sense to $\mathsf{Quant}(\mu)$ (see [195, Definition 1.7.1 and Theorem 1.7.16] for the notion of convergence). Indeed, there may be an $\alpha_\star \in \mathsf{Quant}(\mu)$ whose distance to $\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n)$ remains bounded away from 0.

Theorem 4.37 is a statement on weakly convergent subsequences, the existence of which is only hypothesized. In reflexive spaces however, we show that at least one such subsequence exists, even under a weaker convergence mode for $(\epsilon_n)_{n\geq 1}$, hence the following assumptions.

Assumption (A5). *E is a separable reflexive Banach space.*

Assumption (A6). *$(\epsilon_n)_{n\geq 1}$ converges in outer probability to 0.*

Reflexive Banach spaces enjoy the property that closed balls are weakly compact. In order to leverage this fact for the existence of a convergent subsequence, we need some kind of boundedness result on ϵ_n-empirical ℓ-quantiles. This is the subject of the next proposition.

Proposition 4.40. *Under Assumptions (A2) and (A4):*

1. *there exists $R > 0$ such that the following holds $\mathbb{P}_*$-almost surely: for n large enough, ϵ_n-$\mathsf{Quant}(\widehat{\mu}_n)$ is contained in the closed ball $\bar{B}(0, R)$,*

2. *$\mathbb{P}_*$-almost surely, any sequence of ϵ_n-empirical ℓ-quantiles $(\widehat{\alpha}_n)_{n\geq 1}$ is bounded in norm, i.e., $\sup_n \|\widehat{\alpha}_n\| < \infty$.*

Under Assumptions (A2) and (A6):

3. *there exists $R > 0$ such that $\mathbb{P}_*\big(\{\omega \in \Omega : \epsilon_n^\omega\text{-}\mathsf{Quant}(\widehat{\mu}_n^\omega) \subset \bar{B}(0, R)\}\big) \xrightarrow[n\to\infty]{} 1$,*

4. *any sequence of ϵ_n-empirical ℓ-quantiles $(\widehat{\alpha}_n)_{n\geq 1}$ is stochastically bounded, i.e., $\widehat{\alpha}_n = O_{\mathbb{P}^*}(1)$.*

Remark 4.41. As is made clear in the proof, the constant R depends only on the measure μ. A deterministic version (see Remark 4.35) of the second item in the case of exact empirical medians (i.e., $\epsilon_n = 0$ and $\ell = 0$) was noted by Kemperman [148, p.221] and proved by Cadre [53, Lemma 2] with techniques different from ours.

By combining Theorem 4.37 and Proposition 4.40 we obtain the following corollary.

Theorem 4.42. *Under Assumptions (A4) and (A5), the following holds $\mathbb{P}_*$-almost surely: for any sequence of ϵ_n-empirical ℓ-quantiles $(\widehat{\alpha}_n)_{n\geq 1}$ and any of its subsequence $(\widehat{\alpha}_{n_k})_{k\geq 1}$, there exists a further subsequence $(\widehat{\alpha}_{n_{k_j}})_{j\geq 1}$ that converges in the weak topology of E to some ℓ-quantile $\alpha \in \mathsf{Quant}(\mu)$.*

Next, we change the convergence of $(\epsilon_n)_{n\geq 1}$ to convergence in outer probability. We still obtain a $\mathbb{P}_*$-almost sure statement, but it is weaker than Theorem 4.42.

Theorem 4.43. *Under Assumptions (A5), (A6) the following holds $\mathbb{P}_*$-almost surely: any sequence of ϵ_n-empirical ℓ-quantiles $(\widehat{\alpha}_n)_{n\geq 1}$ has a subsequence $(\widehat{\alpha}_{n_k})_{k\geq 1}$ that converges in the weak topology of E to some ℓ-quantile $\alpha \in \mathsf{Quant}(\mu)$.*

A direct consequence of Theorem 4.42 is the following generalization of Theorem 4.33 to ϵ_n-empirical ℓ-quantiles.

Corollary 4.44. *Suppose E is finite-dimensional and Assumption (A4) holds. Then $\mathbb{P}_*$-almost surely:*

$$\forall \delta > 0, \exists N \geq 1, \forall n \geq N, \epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) \subset \mathsf{Quant}(\mu)^\delta.$$

4.4.2 The case of a single true quantile

We turn our attention to the setting where the measure μ has a single geometric ℓ-quantile, denoted by $\alpha_\star$. This quantity is the parameter of location that we seek to estimate using the approximate empirical ℓ-quantiles from Definition 4.20. To guarantee that $\alpha_\star$ exists we will require Assumption (A2), or the more specific reflexivity Assumption (A5). The following assumption is a generic placeholder to ensure uniqueness of $\alpha_\star$. It is met for example when E is strictly convex and μ is in $\mathcal{M}_\sim$, as seen in Proposition 4.17.

Assumption (A7). *The measure μ has at most one ℓ-quantile.*

Next we state a stronger assumption that guarantees existence and uniqueness for measures in $\mathcal{M}_-$, as shown in Proposition 4.19.

Assumption (A8). *E is a separable, strictly convex normed space and $\ell = 0$. The measure μ is in $\mathcal{M}_-$ and has a unique median.*

A desired property for any sequence of approximate empirical ℓ-quantiles $(\widehat{\alpha}_n)_{n\geq 1}$ is some form of consistency, i.e., some kind of stochastic convergence of $(\widehat{\alpha}_n)$ to $\alpha_\star$. Our setting encompasses infinite-dimensional Banach spaces, which can be equipped with the weak, the weak$*$ (when considering a dual space) or the norm topology [7, Chapter 6]. Therefore, the topology must be specified before any of the convergence modes in Definition 4.21 is considered.

The finite-dimensional case is very special since this is precisely when the weak and norm topologies of E coincide. The consistency of approximate empirical quantiles in this context is well-understood: when E is finite-dimensional, Theorems 4.45, 4.54 and 4.55 established below are corollaries of general M-estimation results [132, Theorem 1; 83, Theorem 6.6; 108, Theorem 5.1; 199, Theorem 1]. The proofs in these references crucially exploit the compactness of closed balls and spheres, which holds only in finite-dimensional normed spaces. The reliance on compactness severely limits the generalization of these works to infinite-dimensional Banach spaces.

Consistency in the weak topology

As a direct consequence of Theorem 4.42 we obtain a consistency result in the weak topology when E is a separable, reflexive, strictly convex Banach space.

Theorem 4.45. *Under (A4), (A5) and (A7) the following holds $\mathbb{P}_*$-almost surely: any sequence of ϵ_n-empirical ℓ-quantiles converges in the weak topology of E to $\alpha_\star$.*

Proofs for this subsection are in Section 4.9.3.

Remark 4.46. Cadre obtained by other means a similar result [53, Theorem 1 (i)] for exact empirical medians in a deterministic setting (see Remark 4.35). His theorem covers the case where E is equal to the dual of a separable Banach space and the case where E is reflexive.

Remark 4.47. Gervini [100, 101] invoked Theorem 1 in Huber's seminal work [132] to obtain consistency in the weak topology of exact empirical medians for the space $E = L^2(T)$, where T is a closed interval of the real line. We show in Remark 4.88 that the argument Gervini uses to apply [132] is incorrect. In Remark 4.89 however, we see that Huber's theorem is indeed applicable; this provides another proof of Theorem 4.45.

Remark 4.48. Consistency in the weak topology is useful in practice since it is equivalent to convergence along linear functionals: for any f in E^*, the sequence of real numbers $(f(\widehat{\alpha}_n))$ converges to $f(\alpha_\star)$. This topology can however be counterintuitive: for example, if $(e_n)_{n\geq 1}$ is an orthonormal basis of a separable Hilbert space, then the sequence (e_n) converges weakly to 0 although these vectors have unit norm.

Consistency in the norm topology

Consistency results for the norm topology are much closer to the statistician's intuition, but they are also more challenging to obtain. To establish such results, a standard technique in M-estimation is to exploit uniform convergence of $(\widehat{\phi}_n)$ and the following condition on the minimizer of ϕ [259, Theorem 5.7; 261, Corollary 3.2.3 (i)].

Definition 4.49. We say that ϕ has a *well-separated minimizer* if it has a minimizer $\alpha_\star$ such that the inequality

$$\phi(\alpha_\star) < \inf_{\substack{\alpha \in E \\ \|\alpha - \alpha_\star\| \geq \epsilon}} \phi(\alpha)$$

holds for every $\epsilon > 0$.

In words, if α is separated away from $\alpha_\star$ then $\phi(\alpha)$ cannot get arbitrarily close to the minimum value of ϕ. We introduce an equivalent condition on the function ϕ: it is stated in terms of sequences, and verifying it is more convenient.

Definition 4.50. A deterministic sequence $(\alpha_n)_{n\geq 1}$ is a *minimizing sequence* if the sequence of real numbers $(\phi(\alpha_n))$ converges to $\phi(\alpha_\star)$. The function ϕ is *well-posed* if it has a unique minimizer $\alpha_\star$ and if any minimizing sequence $(\alpha_n)_{n\geq 1}$ converges in the norm topology of E to $\alpha_\star$.

The function ϕ has a well-separated minimizer if and only if it is well-posed; this is the subject of Lemma 4.90 in Section 4.9.4. Proving that ϕ is well-posed is the key technical hurdle before we obtain consistency results. If E is finite-dimensional and ψ is any continuous function with a unique minimizer, then this minimizer is automatically well-separated and ψ is well-posed. There is no such result in infinite dimension; our specific function ϕ requires a bespoke approach and we add the following topological assumption on the normed space E.

Assumption (A9). *E has the Radon–Riesz (or Kadec–Klee) property: for any sequence $(x_n)_{n\geq 1}$ and any $x \in E$, if simultaneously (x_n) converges weakly to x and the sequence of real numbers $(\|x_n\|)$ converges to $\|x\|$, then (x_n) converges in the norm topology to x.*

By exploiting this assumption, we obtain the following proposition. Its proof is technically challenging.

Proposition 4.51. *1. Under (A5), (A7) and (A9), the function ϕ is well-posed.*

2. Assume (A8) and let L denote an affine line such that $\mu(L) = 1$. Then any minimizing sequence lying on L converges in the norm topology of E to $\alpha_\star$.

Remark 4.52. Assumption (A7) is broad and typically fulfilled by requiring that E be strictly convex and μ be in $\mathcal{M}_\sim$. The second item of Proposition 4.51 covers measures in the complementary class $\mathcal{M}_-$; notably it requires neither reflexivity nor the Radon–Riesz assumption and it states a result weaker than well-posedness.

Remark 4.53. We justify next why the Radon–Riesz assumption (A9) seems necessary to obtain the first item of Proposition 4.51. A related minimization problem is that of *best approximation*: given C a closed convex subset of E, a best approximation of 0 in C is a minimizer of the function $\alpha \mapsto \|\alpha\|$ with the constraint that $\alpha \in C$. It is known in the literature [265; 77, Theorem 2 p.41; 175, Theorem 10.4.6] that the best approximation problem is well-posed for every closed convex C if and only if:

$$E \text{ is a reflexive, strictly convex Banach space having the Radon–Riesz property.} \tag{4.2}$$

The problem of geometric medians bears some resemblance to best approximation since the objective function ϕ_0 involves the norm, and the flexibility in the choice of C is paralleled by freedom in the choice of the measure μ. Consequently we conjecture that well-posedness of ϕ_0 for every $\mu \in \mathcal{M}_\sim$ is equivalent to (4.2). That (4.2) is sufficient follows from Proposition 4.51. Regarding necessity, we obtained strict convexity in the third item of Proposition 4.17; proving reflexivity and the Radon–Riesz property is an open problem.

In addition to well-posedness of ϕ, the sequence $(\widehat{\phi}_n)$ converges uniformly on bounded sets to ϕ with $\mathbb{P}$-probability 1 by Proposition 4.36. We can therefore adapt the M-estimation technique described above and obtain the following consistency results in the norm topology.

Theorem 4.54. *Under (A4), (A5), (A7), (A9) the following holds $\mathbb{P}_*$-almost surely: any sequence of ϵ_n-empirical ℓ-quantiles converges in the norm topology to $\alpha_\star$.*

Theorem 4.55. *Assume (A6), (A5), (A7), (A9). Any sequence $(\widehat{\alpha}_n)_{n\geq 1}$ of ϵ_n-empirical ℓ-quantiles converges in outer probability to $\alpha_\star$, i.e.,*

$$\forall \delta > 0, \quad \mathbb{P}^*(\|\widehat{\alpha}_n - \alpha_\star\| > \delta) \xrightarrow[n\to\infty]{} 0.$$

Remark 4.56. As in the M-estimation reference [132], the convergence mode of (ϵ_n) in each theorem is reflected in the convergence mode obtained for approximate empirical quantiles.

Next, we formulate consistency for measures in $\mathcal{M}_-$: the reflexivity and Radon–Riesz assumptions can be replaced with strict convexity (which implies neither of the previous two), however the precision ϵ_n must be set to 0.

Proposition 4.57. *The conclusions of Theorems 4.54 and 4.55 also hold under the combined assumptions $\epsilon_n = 0$ and (A8).*

Uniformly convex spaces (see Definition 4.15) are reflexive, strictly convex and they verify the Radon–Riesz property. As a consequence we obtain the following explicit list of spaces where the two previous consistency theorems hold.

Corollary 4.58. *Let $\mu \in \mathcal{M}_\sim$. Under Assumption (A4) (resp., (A6)), the conclusion of Theorem 4.54 (resp., 4.55) holds in any separable, uniformly convex Banach space, hence in any of the following spaces (each space is viewed as a real vector space, see Corollary 4.12):*

1. *E is finite-dimensional and strictly convex.*

2. *E is a separable Hilbert space equipped with its Hilbert norm,*

3. *$E = L^p(S, \mathcal{A}, \nu)$ equipped with the L^p norm, where $1 < p < \infty$, $(S, \mathcal{A}, \nu)$ is a sigma-finite measure space and $\mathcal{A}$ is countably generated. This includes the case where $(S, \mathcal{A})$ is a separable metric space with its Borel sigma-algebra and the case where $(S, \mathcal{A})$ is a countable space with its discrete sigma-algebra.*

4. *$E = W^{k,p}(\Omega)$ as in Corollary 4.12,*

5. *$E = S_p(H)$ as in Corollary 4.12 and H is separable.*

The same statement holds for $\mu \in \mathcal{M}_-$, if we assume additionally that μ has a unique ℓ-quantile.

Next we compare our results to those in the literature. Since geometric quantiles fit the framework of convex M-estimation, the following consistency result is already known and it holds when E is finite-dimensional.

Lemma 4.59 ([199, Theorem 1])**.** *Assume $\mu \in \mathcal{M}_\sim$, $E = \mathbb{R}^d$, $d \geq 2$, equipped with the Euclidean norm and $\epsilon_n = 0$. Given $(\widehat{\alpha}_n)_{n\geq 1}$ a sequence of measurable selections, we have*

$$\mathbb{P}(\|\widehat{\alpha}_n - \alpha_\star\| \xrightarrow[n\to\infty]{} 0) = 1.$$

Note that our Corollary 4.58 recovers this statement, with no additional assumption.

To our knowledge, the only preexisting consistency result in the norm topology for infinite-dimensional spaces is Theorem 4.2.2 in Chakraborty and Chaudhuri [64]. They state an almost-sure consistency theorem for exact empirical medians ($\ell = 0$, $\epsilon_n = 0$) in separable Hilbert spaces. Their statement covers measures in $\mathcal{M}_\sim$ verifying two additional assumptions on which we comment below:

$$\mathbb{E}[\|X - \alpha_\star\|^{-1}] < \infty \quad \text{and} \quad \mathbb{E}[\|X\|^2] < \infty, \tag{4.3}$$

where X is a random element with distribution μ. Their proof exploits properties of the Hessian of ϕ_0 to obtain the almost-sure inequality $\phi_0(\widehat{\alpha}_n) - \phi_0(\alpha_\star) \geq \frac{c}{2}\|\widehat{\alpha}_n - \alpha_\star\|^2$ for a positive constant c and large enough n. The conclusion then follows since $\phi_0(\widehat{\alpha}_n) \to \phi_0(\alpha_\star)$. For the Hessian at $\alpha_\star$ to exist, the condition $\mathbb{E}[\|X - \alpha_\star\|^{-1}] < \infty$ must be satisfied. It is restrictive, since it implies $0 = \mathbb{P}(X = \alpha_\star) = \mu(\{\alpha_\star\})$ and additionally μ cannot put too much mass around $\alpha_\star$. Furthermore, Chakraborty and Chaudhuri leverage a key property of the Hessian that is found in Proposition 2.1 of Cardot, Cénac and Zitt [57], which requires additionally the moment condition $\mathbb{E}[\|X\|^2] < \infty$ (we dedicate a paragraph on the next page to a further detailed assessment of [57, Proposition 2.1]). Unlike [64], because we do not rely on first- or second-order methods, our Theorems 4.54 and 4.55 are free from the extra distributional assumptions (4.3). Moreover, our theorems match exactly the minimal assumptions needed for consistency in the finite-dimensional setting, as seen in Lemma 4.59.

Besides, the authors of [57] crucially rely on the Hilbert structure to obtain their equality (6), which plays a key role in the proof of [57, Proposition 2.1] (and by extension, in the proof of [64, Theorem 4.2.2]). Chakraborty and Chaudhuri argue [64, p.38] that their proof technique extends to other Banach spaces by applying Proposition 1 in Asplund [13] with $f = \phi$, $a = 0$, $b = \alpha_\star$. They claim that the third clause of [13, Proposition 1] is true by [13, Theorem 3]. This reasoning is correct as long as the set G in Asplund's Theorem 3 contains 0, however this theorem has no such guarantee.

Our Theorems 4.54 and 4.55 hold for $\ell \neq 0$ and $\epsilon_n \neq 0$, in a large variety of Banach space. They do not require the distributional assumptions (4.3) and they match the assumptions needed for consistency in finite dimension. Therefore, they are a significant improvement on the existing literature.

Detailed assessment of [57, Proposition 2.1] The proof of Proposition 2.1 in [57] is quite terse; as we have spent quite some time figuring out the details when reading this paper, we deem it worthwile to provide an extended proof here.

Proof. We begin by repeating one of the key assumptions.

Assumption (A10). *For every $v \in E$, there exists some $w \in E$ satisfying both $\langle w, v\rangle = 0$ and $\mathbb{V}[\langle w, X\rangle] = 0$.*

Let $\mathcal{S} = \{K \subset E : K$ is a linear subspace and for every $x \in K$, $\langle x, X\rangle$ is a.s. constant$\}$. Note that $\mathcal{S}$ is partially ordered by inclusion and if $\mathcal{R}$ is a chain in $\mathcal{S}$, then $\bigcup_{K\in\mathcal{R}} K$ is in $\mathcal{S}$ and it is an upper bound of $\mathcal{R}$. By Zorn's lemma, $\mathcal{S}$ has a maximal element, say $F \in \mathcal{S}$.

We show that F is closed: let $(x_n)_{n\geq 1}$ denote a sequence in F that converges to some $x \in E$. For each $n \geq 1$, there exists an almost-sure event Ω_n and a constant c_n such that $\langle x_n, X\rangle = c_n$ on Ω_n. Since $\langle x, X\rangle = \lim_n \langle x_n, X\rangle$, by considering the almost-sure event $\cap_{n\geq 1}\Omega_n$, we obtain that $(c_n)_{n\geq 1}$ converges to some $c \in \mathbb{R}$ and $\langle x, X\rangle = c$ with probability 1, thus $x \in F$.

Next, we see that $\dim(F^\perp) \geq 2$. If $\dim(F^\perp) = 0$, then since F is closed we have $F = E$ and for every $x \in E$, the random variable $\langle x, X\rangle$ is a.s. constant which contradicts Assumption (A10). If $\dim(F^\perp) = 1$ there exists $v \neq 0$ such that $F^\perp = \mathbb{R}v$. By Assumption (A10), there is some $w \in (\mathbb{R}v)^\perp = F$ such that $\langle w, X\rangle$ is not a.s. constant. This is in contradiction with $F \in \mathcal{S}$.

Consider v_1, v_2 orthonormal vectors in $F^\perp$ and the mapping

$$\begin{aligned} \phi : [0, 2\pi] &\to \mathbb{R} \\ t &\mapsto \mathbb{V}[\langle \cos(t)v_1 + \sin(t)v_2, X\rangle], \end{aligned}$$

which reaches a minimum at some t_0. Assuming $\phi(t_0) = 0$ and letting $z = \cos(t_0)v_1 + \sin(t_0)v_2$, we have $z \in F^\perp \setminus \{0\}$, thus $F \oplus (\mathbb{R}z)$ is in $\mathcal{S}$ and contradicts the maximality of F. Consequently $\phi(t_0)$ is positive.

Fix some arbitrary vector $u \in E$ with unit norm for the rest of the proof. Elementary linear algebra shows that $(\mathbb{R}u)^\perp \cap \operatorname{span}(v_1, v_2) \neq \{0\}$; let v denote a unit vector in this intersection. Since $\mathbb{R}u \subset (\mathbb{R}v)^\perp$ we have $P_vP_u = 0$, i.e., $P_v = P_vP_{u^\perp}$, hence for every $y \in E$ the following inequality holds $\|P_v y\| \leq \|P_{u^\perp} y\|$, which rewrites as $\langle y, v\rangle^2 \leq \|P_{u^\perp} y\|^2$.

For any $K > 0$ we obtain the following estimate:

$$\begin{aligned}
\mathbb{E}\Big[\frac{\|P_{u^\perp}(X-\alpha)\|^2}{\|X-\alpha\|^3}\Big] &\geq \mathbb{E}\Big[\frac{\langle X-\alpha, v\rangle^2}{\|X-\alpha\|^3}\Big] \\
&\geq \mathbb{E}\Big[\frac{\langle X-\alpha, v\rangle^2}{(K+A)^3}\mathbb{1}_{\|X\|\leq K}\Big] \\
&\geq \frac{\mathbb{V}[\langle X-\alpha, v\rangle]}{(K+A)^3} - \frac{\mathbb{E}\big[\|X-\alpha\|^2\mathbb{1}_{\|X\|>K}\big]}{(K+A)^3} \\
&\geq \frac{\phi(t_0)}{(K+A)^3} - \frac{2}{(K+A)^3}\mathbb{E}\big[(\|X\|^2+A^2)\mathbb{1}_{\|X\|>K}\big].
\end{aligned}$$

This lower bound is asymptotically $\sim \frac{\phi(t_0)}{K^3}$ as $K \to \infty$, hence it is possible to choose K so large (and independent of α) in order that the lower bound be positive. □

While the definition of $\mathcal{S}$ and F does not require the moment assumption $\mathbb{E}[\|X\|^2] < \infty$, it seems inescapable that such an assumption is needed in the rest of the proof.

4.5 Asymptotic normality of approximate empirical quantiles

In the preceding section, approximate empirical quantiles were shown to converge in the norm topology to the true ℓ-quantile $\alpha_\star$ under mild assumptions on the space E, the measure μ and the precision ϵ_n. In order to perform more advanced inference on $\alpha_\star$ (e.g., developing confidence regions and hypothesis testing), it is necessary to determine the asymptotic distribution of these estimates.

When E is a Euclidean space, M-estimation results based on empirical processes [259, Theorem 5.23; 261, Example 3.2.22] yield a linear representation for $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$, from which the asymptotic normality of $\widehat{\alpha}_n$ follows. We identify the functional ℓ with the corresponding element of $\mathbb{R}^d$.

Theorem 4.60. *Assume that (i) E is a d-dimensional Euclidean space with $d \geq 2$, (ii) the moment condition $\mathbb{E}[\|X-\alpha_\star\|^{-1}] < \infty$ holds, (iii) μ is in $\mathcal{M}_\sim$ and (iv) $\epsilon_n = o_{\mathbb{P}}(1/n)$. Let $\widehat{\alpha}_n$ be a measurable selection from ϵ_n-$\mathsf{Quant}(\widehat{\mu}_n)$ for each $n \geq 1$ and let H, V be $d \times d$ symmetric matrices defined by*

$$\begin{aligned}
H &= \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}\frac{1}{\|\alpha_\star - X\|}\Big(I_d - \frac{(\alpha_\star - X)(\alpha_\star - X)^\top}{\|\alpha_\star - X\|^2}\Big)\Big], \\
V &= \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}\Big(\frac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell\Big)\Big(\frac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell\Big)^\top\Big].
\end{aligned}$$

Then H is positive-definite and

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) = -H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\mathbb{1}_{X_i\neq\alpha_\star}\frac{\alpha_\star - X_i}{\|\alpha_\star - X_i\|} - \ell\Big) + o_{\mathbb{P}}(1). \quad (4.4)$$

As a consequence, $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$ *converges in distribution to the multivariate normal* $\mathcal{N}_d(0, H^{-1}VH^{-1})$*.*

The moment assumption (ii) ensures that the function $\alpha \mapsto \|\alpha - x\| - \|x\|$ is differentiable at $\alpha_\star$ for μ-almost every x, that H is well-defined and that ϕ has the second-order Taylor expansion $\phi(\alpha_\star + h) = \phi(\alpha_\star) + \frac{1}{2}h^\top Hh + o(\|h\|^2)$. Assumption (iii) guarantees not only that μ has a unique ℓ-quantile, but also that H is invertible. These facts combined with assumption (iv) warrant the application of [259, Theorem 5.23], from which Theorem 4.60 is obtained. The proof of this theorem in [259] relies crucially on bounding the $L_2(\mu)$ bracketing number of the function class $\mathcal{F}_\delta = \{\varphi_\alpha - \varphi_{\alpha_\star} : \|\alpha - \alpha_\star\| \leq \delta\}$ as follows:

$$N_{[\,]}(\eta, \mathcal{F}_\delta, L_2(\mu)) \leq C\Big(\frac{\delta}{\eta}\Big)^d \quad \text{for every } \eta \in (0, \delta),$$

where φ_α denotes the function $x \mapsto \|\alpha - x\| - \|x\|$, δ is a positive real and C is a constant depending only on d and δ. That the dimension d appears in the right-hand side stems from standard volumetric arguments, which do not generalize to the infinite-dimensional case. Other M-estimation works [108, Theorem 6.1; 199, Theorem 4; 119, Theorem 2.1] that leverage convexity reach a weaker conclusion, namely the asymptotic normality of exact empirical ℓ-quantiles (i.e., when $\epsilon_n = 0$). The proofs in these works rely critically on the compactness of closed balls and spheres in the norm topology, which is characteristic of the finite-dimensional setting.

Theorem 4.60 will serve as a benchmark when we establish normality results that encompass infinite-dimensional spaces.

4.5.1 Asymptotic normality in Hilbert spaces

In this subsection E is a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and μ is in $\mathcal{M}_\sim$. By the Riesz representation theorem, there is a unique vector $l \in E$ such that ℓ coincides with the functional $\alpha \mapsto \langle l, \alpha \rangle$. Troughout we will identify l with ℓ and write $\langle \ell, \alpha \rangle$ for convenience. As an example,

$$\phi_\ell(\alpha) = \phi_0(\alpha) - \langle \ell, \alpha \rangle \text{ for every } \alpha \in E.$$

By Corollary 4.18, μ has a unique geometric median $\alpha_\star$.

Preliminaries

For convenience we will sometimes denote by N the norm function: $N(\alpha) = \|\alpha\| = \langle \alpha, \alpha \rangle^{1/2}$. We will leverage derivatives of N and ϕ, hence the following refresher for gradients and Hessians in Hilbert spaces.

Definition 4.61. Let $f : E \to \mathbb{R}$ and $\alpha \in E$.

1. If f is Fréchet differentiable at α, the *gradient* of f at α is the unique element of E denoted by $\nabla f(\alpha)$ such that the Fréchet derivative $Df(\alpha)$ is the linear functional $h \mapsto \langle \nabla f(\alpha), h \rangle$.

2. If f is twice Fréchet differentiable at α, the *Hessian* of f at α is the unique bounded operator denoted by $\nabla^2 f(\alpha)$ such that the second-order Fréchet derivative $D^2 f(\alpha)$ verifies

$$D^2 f(\alpha)(h_1, h_2) = \langle \nabla^2 f(\alpha) h_1, h_2 \rangle \quad \text{for every } (h_1, h_2) \in E^2.$$

By elementary differential calculus, the norm N is infinitely differentiable at each nonzero $\alpha \in E$ with gradient and Hessian given by

$$\nabla N(\alpha) = \frac{\alpha}{\|\alpha\|} \qquad \nabla^2 N(\alpha) = \frac{1}{\|\alpha\|}\Big(\mathrm{Id} - \frac{\alpha \otimes \alpha}{\|\alpha\|^2}\Big), \tag{4.5}$$

where Id is the identity operator and $\alpha \otimes \alpha$ denotes the operator $u \mapsto \langle u, \alpha \rangle \alpha$. The following lemma gives explicit error bounds for the second-order (resp., first-order) Taylor approximation of N (resp., ∇N). We let $a \wedge b$ denote the minimum of the real numbers a and b.

Lemma 4.62. *For any $\alpha \in E \setminus \{0\}$ and any $h \in E$ the following inequality holds:*

$$\Big| \|\alpha + h\| - \|\alpha\| - \langle \nabla N(\alpha), h \rangle - \frac{1}{2} \langle \nabla^2 N(\alpha) h, h \rangle \Big| \le \frac{1}{2} \Big(\frac{\|h\|^2}{\|\alpha\|} \wedge \frac{\|h\|^3}{\|\alpha\|^2} \Big). \tag{4.6}$$

Assuming additionally that $\alpha + h$ is nonzero,

$$\|\nabla N(\alpha + h) - \nabla N(\alpha) - \nabla^2 N(\alpha) h\| \le 2 \Big(\frac{\|h\|}{\|\alpha\|} \wedge \frac{\|h\|^2}{\|\alpha\|^2} \Big). \tag{4.7}$$

Proofs for this subsection are in Section 4.10.1.

Remark 4.63. In the case of linear dependence where $h = \lambda \alpha$ with $\lambda \in \mathbb{R}$ and $\|\alpha\| = 1$, the inequalities rewrite as statements involving only λ (with $\lambda \neq -1$ in the second):

$$\big| |1 + \lambda| - 1 - \lambda \big| \le \frac{1}{2} (\lambda^2 \wedge |\lambda|^3) \quad \text{and} \quad \Big| \frac{1 + \lambda}{|1 + \lambda|} - 1 \Big| \le 2 (|\lambda| \wedge \lambda^2). \tag{4.8}$$

It is easily seen that the constants $1/2$ and 2 in the right-hand sides of (4.8) are sharp, hence they are also optimal in inequalities (4.6) and (4.7). Remarkably, these constants are independent of the Hilbert space E.

By combining Equation (4.5) and Lemma 4.62 we obtain differentiability properties of the objective function ϕ. We say that $\alpha \in E$ is an *atom* of the measure μ if $\mu(\{\alpha\}) > 0$.

Proposition 4.64. *Let $\alpha \in E$.*

1. *ϕ is Fréchet differentiable at α if and only if α is not an atom of μ. In that case, the gradient of ϕ is given by*

$$\nabla \phi(\alpha) = \mathbb{E}\Big[\mathbb{1}_{X \neq \alpha} \frac{\alpha - X}{\|\alpha - X\|} \Big] - \ell. \tag{4.9}$$

2. *Assume that* $\mathbb{E}[\|X-\alpha\|^{-1}] < \infty$. *Then the operator on* E
$$H = \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha - X\|}\Big(\mathrm{Id} - \frac{(\alpha - X)\otimes(\alpha - X)}{\|\alpha - X\|^2}\Big)\Big] \tag{4.10}$$
is well-defined, bounded, self-adjoint and nonnegative, i.e.,
$$\langle Hh_1, h_2\rangle = \langle h_1, Hh_2\rangle \quad \text{and} \quad \langle Hh_1, h_1\rangle \geq 0 \quad \text{for every } (h_1, h_2) \in E^2.$$
Moreover the following second-order Taylor expansion holds:
$$\phi(\alpha + h) = \phi(\alpha) + \langle \nabla\phi(\alpha), h\rangle + \tfrac{1}{2}\langle Hh, h\rangle + o(\|h\|^2).$$

3. *Under the additional assumption that* μ *is in* $\mathcal{M}_\sim$, *the operator* H *is invertible, its inverse is bounded, self-adjoint, nonnegative and* $\inf_{\|h\|=1}\langle Hh, h\rangle > 0$.

4. *Assume that* $\mathbb{E}[\|X-\alpha\|^{-1}] < \infty$ *and* α *has a neighborhood without any atom. Then* ϕ *is twice Fréchet differentiable at* α *with Hessian* $\nabla^2\phi(\alpha) = H$.

Remark 4.65. Expliciting $\nabla\phi(\alpha)$ and H requires integrating functions with values respectively in the Hilbert space E and the Banach space $B(E)$ of bounded operators on E (equipped with the operator norm). The expectations in (4.9) and (4.10) are understood as Bochner integrals (see, e.g., [75, Section II.2]). Since $B(E)$ is not separable when E is infinite-dimensional, there are measurability issues that we address in the proof of Proposition 4.64.

Remark 4.66. Cardot, Cénac and Zitt [57] also obtain by other means that H is invertible and $\inf_{\|h\|=1}\langle Hh, h\rangle > 0$. However their proof requires the extra assumption $\mathbb{E}[\|X\|^2] < \infty$.

Remark 4.67. The quantity $\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{\alpha - X}{\|\alpha - X\|}\Big] - \ell$ is always well-defined, regardless of the differentiability of ϕ at α. In fact, it is a subgradient of the convex function ϕ at α. For convenience, we will use the notation $\nabla\phi(\alpha)$ as a shorthand instead, even when ϕ is not differentiable. Similarly, the operator H is well-defined as soon as $\mathbb{E}[\|X-\alpha\|^{-1}] < \infty$. When this condition is met we will write $\nabla^2\phi(\alpha)$ to denote the aforementioned operator, even when ϕ need not be twice differentiable.

Weak Bahadur–Kiefer representations and asymptotic normality

Given $(\widehat{\alpha}_n)$ a sequence of ϵ_n-empirical ℓ-quantiles, we wish to establish convergence in distribution of the sequence $(\sqrt{n}(\widehat{\alpha}_n - \alpha_\star))$. For each $n \geq 1$, we note that $\widehat{\alpha}_n$ is an ϵ_n-minimizer of the empirical objective function $\widehat{\phi}_n$ if and only if $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$ is an ϵ_n-minimizer of the function $\widehat{\psi}_n$ defined next. We will derive limiting statements on $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$ by approximating $\widehat{\psi}_n$ with a quadratic function $\widehat{\Psi}_n$ that resembles the second-order Taylor expansion of $\widehat{\psi}_n$.

Definition 4.68. We let $\widehat{\psi}_n$ denote the shifted and rescaled empirical objective function
$$\widehat{\psi}_n : \beta \mapsto \widehat{\phi}_n\Big(\alpha_\star + \tfrac{\beta}{\sqrt{n}}\Big)$$
and we define the quadratic function
$$\widehat{\Psi}_n : \beta \mapsto \widehat{\phi}_n(\alpha_\star) + \langle \nabla\widehat{\phi}_n(\alpha_\star), \tfrac{\beta}{\sqrt{n}}\rangle + \tfrac{1}{2}\langle \nabla^2\phi(\alpha_\star)\tfrac{\beta}{\sqrt{n}}, \tfrac{\beta}{\sqrt{n}}\rangle.$$

Remark 4.69. In Definition 4.68, the abuse of notation described in Remark 4.67 occurs when we write $\nabla^2\phi(\alpha_\star)$. The following analogous abuse is performed here and later:

$$\nabla\widehat{\phi}_n(\alpha) = \tfrac{1}{n}\textstyle\sum_{i=1}^n(\mathbb{1}_{X_i\neq\alpha}\frac{\alpha-X_i}{\|\alpha-X_i\|} - \ell),$$
$$\nabla^2\widehat{\phi}_n(\alpha) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathbb{1}_{X_i\neq\alpha}\frac{1}{\|\alpha-X_i\|}\Big(\mathrm{Id} - \frac{(\alpha-X_i)\otimes(\alpha-X_i)}{\|\alpha-X_i\|^2}\Big).$$

Note that $\nabla\widehat{\phi}_n(\alpha)$ is always a subgradient of $\widehat{\phi}_n$ at α, regardless of the differentiability of $\widehat{\phi}_n$.

The quadratic function $\widehat{\Psi}_n$ has a unique minimizer $\widehat{\beta}_n$ which is easy to write in closed form: $\widehat{\beta}_n = -\sqrt{n}[\nabla^2\phi(\alpha_\star)^{-1}]\nabla\widehat{\phi}_n(\alpha_\star)$. By the central limit theorem for Hilbert spaces, $(\widehat{\beta}_n)_{n\geq 1}$ converges in distribution to a Gaussian measure (see, e.g., [161, Section I.2] for Gaussian measures on Hilbert spaces and [168, Chapter 10] for central limit theorems). This is part of the following proposition.

Proposition 4.70. *Assume that* $\mathbb{E}[\|X-\alpha_\star\|^{-1}] < \infty$ *and* $\mu \in \mathcal{M}_\sim$*.*

1. *For each* $n \geq 1$*, the function* $\widehat{\Psi}_n$ *is convex, with unique minimizer*
$$\widehat{\beta}_n = -\sqrt{n}[\nabla^2\phi(\alpha_\star)^{-1}]\nabla\widehat{\phi}_n(\alpha_\star).$$
2. *The sequence* $(\widehat{\beta}_n)_{n\geq 1}$ *converges in distribution to the centered Gaussian measure with covariance operator* $\nabla^2\phi(\alpha_\star)^{-1}\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}(\frac{\alpha_\star-X}{\|\alpha_\star-X\|}-\ell)\otimes(\frac{\alpha_\star-X}{\|\alpha_\star-X\|}-\ell)\Big]\nabla^2\phi(\alpha_\star)^{-1}$*. As a consequence,* $\widehat{\beta}_n = O_{\mathbb{P}}(1)$*.*
3. *Letting* $\kappa = \inf_{\|h\|=1}\langle\nabla^2\phi(\alpha_\star)h,h\rangle$*,* $\widehat{\Psi}_n$ *is* $\frac{\kappa}{n}$*-strongly convex and the following bound holds:*
$$\widehat{\Psi}_n(\beta) \geq \widehat{\Psi}_n(\widehat{\beta}_n) + \frac{\kappa}{2n}\|\beta-\widehat{\beta}_n\|^2 \quad \textit{for every } \beta \in E.$$

Proofs for this subsection are in Section 4.10.2

The next proposition shows that the quadratic function $\widehat{\Psi}_n$ is uniformly close to $\widehat{\psi}_n$ on bounded sets.

Proposition 4.71. *Let* $R > 0$ *be fixed.*

1. *Assuming* $\mathbb{E}[\|X-\alpha_\star\|^{-1}] < \infty$*, the random sequence* $(n\cdot\sup_{\|\beta\|\leq R}|\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)|)_{n\geq 1}$ *converge* $\mathbb{P}$*-almost surely to* 0*, hence*
$$\sup_{\|\beta\|\leq R}|\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| = o_{\mathbb{P}}(n^{-1}).$$
2. *Assuming* $\mathbb{E}[\|X-\alpha_\star\|^{-2}] < \infty$*, we have*
$$\sup_{\|\beta\|\leq R}|\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| = O_{\mathbb{P}}(n^{-3/2}).$$
and
$$\sup_{\|\beta\|\leq R}|\nabla\widehat{\psi}_n(\beta) - \nabla\widehat{\Psi}_n(\beta)| = O_{\mathbb{P}}(n^{-3/2}).$$

Remark 4.72. When $\mathbb{E}[\|X-\alpha_\star\|^{-2}] < \infty$ we obtain the tigher bound $O_{\mathbb{P}}(n^{-3/2})$ instead of $o_{\mathbb{P}}(n^{-1})$. This second moment assumption is crucial in the proof because we apply the central limit theorem to the random variable $\mathbb{1}_{X_i \neq \alpha_\star} \|X_i - \alpha_\star\|^{-1}$ and to the random element $\mathbb{1}_{X_i\neq\alpha_\star} \frac{(\alpha_\star - X_i)\otimes(\alpha_\star - X_i)}{\|X_i-\alpha_\star\|^3}$ which takes values in the Hilbert space $S_2(E)$ of Hilbert–Schmidt operators.

Since the convex functions $\widehat{\psi}_n$ and $\widehat{\Psi}_n$ are close, it is expected that approximate minimizers of $\widehat{\psi}_n$ are close to $\widehat{\beta}_n$. The next theorem formalizes this idea. We obtain linear representations similar to (4.4), which we call *weak Bahadur–Kiefer representations* (see Remark 4.74 below). The first representation is sufficient to derive asymptotic normality later. With stronger assumptions we obtain two substantially refined representations.

Theorem 4.73. *Assume $\mu \in \mathcal{M}_\sim$ and let $(\widehat{\alpha}_n)_{n\geq 1}$ denote a sequence of ϵ_n-empirical ℓ-quantiles.*

1. *If $\mathbb{E}[\|X-\alpha_\star\|^{-1}] < \infty$ and $\epsilon_n = o_{\mathbb{P}^*}(n^{-1})$, we have*

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) = \widehat{\beta}_n + o_{\mathbb{P}^*}(1).$$

2. *If $\mathbb{E}[\|X-\alpha_\star\|^{-2}] < \infty$ and $\epsilon_n = o_{\mathbb{P}^*}(n^{-3/2})$, we have*

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) = \widehat{\beta}_n + O_{\mathbb{P}^*}(n^{-1/4}).$$

3. *If $\mathbb{E}[\|X-\alpha_\star\|^{-2}] < \infty$ and $\epsilon_n = o_{\mathbb{P}^*}(n^{-2})$, we have*

$$\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) = \widehat{\beta}_n + O_{\mathbb{P}^*}(n^{-1/2}).$$

Remark 4.74. The idea of approaching $\widehat{\psi}_n$ by a quadratic function and then leveraging the closeness of minimizers is not new. Niemiro used it in [199] to derive Bahadur–Kiefer representations for a wide range of M-estimators (see [16, 150] for the seminal works of Bahadur and Kiefer on one-dimensional quantiles). When applied to geometric quantiles, Niemiro's Theorem 5 yields the representation

$$\mathbb{P}\Big(\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) = \widehat{\beta}_n + O\big(n^{-(1+s)/4}(\log n)^{1/2}(\log\log n)^{(1+s)/4}\big)\Big) = 1$$

for any $s \in [0,1)$ under the following assumptions: $E = \mathbb{R}^d$ with $d \geq 2$, $\epsilon_n = 0$, $\mu \in \mathcal{M}_\sim$, μ has a neighborhood of $\alpha_\star$ without any atom and $\mathbb{E}[\|X-\alpha_\star\|^{-2}] < \infty$. Sharper representations were obtained in later works [155, 65, 11]. The sharpest is given by Arcones in [11, Proposition 4.1] where he obtains

$$\mathbb{P}\Big(\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) = \widehat{\beta}_n + O\big(n^{-1/2}\log\log n\big)\Big) = 1$$

with the additional assumption (relative to Niemiro) that $\nabla\phi$ has a second-order Taylor expansion. Arcones further derives a law of the iterated logarithm which shows that the rate $O\big(n^{-1/2}\log\log n\big)$ cannot be improved upon. The proofs in the aforementioned references all rely crucially on E being finite-dimensional (essentially to ensure that closed balls and spheres are compact in the norm topology, or to exploit results from empirical process theory) and there is a major technical hurdle in generalizing their techniques to the infinite-dimensional setting.

In comparison, our estimate $O_{\mathbb{P}^*}(n^{-1/2})$ has the correct order and it holds in infinite dimension under assumptions less stringent than those of Niemiro and Arcones. However our representation is weaker, in the sense that it is not almost sure.

Remark 4.75. Chakraborty and Chaudhuri [63] consider a different estimator of the population geometric quantile. In a separable Hilbert space with orthonormal basis $(e_n)_{n\geq 1}$, they define the nested finite-dimensional subspaces $\mathcal{Z}_n = \mathrm{span}(e_1, \dots, e_{d(n)})$ where $(d(n))_{n\geq 1}$ is a sequence of positive integers. For fixed $n \geq 1$, the data is projected orthogonally on $\mathcal{Z}_n$ and the linear functional ℓ is projected orthogonally on the corresponding subspace of the dual, thus defining the transformed data $X_1^{(n)}, \dots, X_n^{(n)}$ and functional $\ell^{(n)}$. Their estimator $\widetilde{\alpha}_n$ is then defined as a minimizer of

$$\alpha \mapsto \frac{1}{n}\sum_{i=1}^{n}(\|\alpha - X_i^{(n)}\| - \|X_i^{(n)}\|) - \ell^{(n)}(\alpha)$$

over the $d(n)$-dimensional subspace $\mathcal{Z}_n$. By projecting the random element X on $\mathcal{Z}_n$, they also define a population quantity $\alpha_\star^{(n)}$ as a minimizer of

$$\alpha \mapsto \mathbb{E}[\|\alpha - X^{(n)}\| - \|X^{(n)}\|] - \ell^{(n)}(\alpha)$$

over $\mathcal{Z}_n$. In [63, Theorem 3.3] they develop a Bahadur–Kiefer representation for the quantity $\sqrt{n}(\widetilde{\alpha}_n - \alpha_\star^{(n)})$ and later obtain as a consequence the asymptotic normality of their estimator $\widetilde{\alpha}_n$. In the proofs Chakraborty and Chaudhuri rely crucially on the finite dimensionality of the $\mathcal{Z}_n$, thus their work is not applicable to our estimator $\widehat{\alpha}_n$.

Remark 4.76. In Theorem 4.73, the second item is obtained by combining the estimate $\sup_{\|\beta\|\leq R}|\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| = O_{\mathbb{P}}(n^{-3/2})$ with the (κ/n)-strong convexity of $\widehat{\Psi}_n$. In the third item we only use closeness of the gradients: $\sup_{\|\beta\|\leq R}|\nabla\widehat{\psi}_n(\beta) - \nabla\widehat{\Psi}_n(\beta)| = O_{\mathbb{P}}(n^{-3/2})$, which allows us to improve the bound on $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) - \widehat{\beta}_n$ from $O_{\mathbb{P}^*}(n^{-1/4})$ to $O_{\mathbb{P}^*}(n^{-1/2})$. In exchange however, we must put an additional constraint on the precision ϵ_n.

A consequence of Theorem 4.73 is the asymptotic normality of $\widehat{\alpha}_n$.

Theorem 4.77. *Assume $\mu \in \mathcal{M}_\sim$, $\mathbb{E}[\|X - \alpha_\star\|^{-1}] < \infty$ and $\epsilon_n = o_{\mathbb{P}^*}(n^{-1})$. For any sequence $(\widehat{\alpha}_n)_{n\geq 1}$ of ϵ_n-empirical ℓ-quantiles, $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$ converges in distribution to the centered Gaussian measure with covariance operator*

$$\Sigma = \nabla^2\phi(\alpha_\star)^{-1}\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}\big(\tfrac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell\big) \otimes \big(\tfrac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell\big)\Big]\nabla^2\phi(\alpha_\star)^{-1}.$$

Remark 4.78. The functions $\widehat{\alpha}_n : \Omega \to E$ may not be measurable, as discussed in Section 4.3.2. To make sense of the convergence in Theorem 4.77, we adopt the theory developed by Van der Vaart and Wellner [261, Chapter 1.3]: letting γ denote the aforementioned Gaussian measure, for any continuous and bounded function $f : E \to \mathbb{R}$ the following convergence of outer expectations holds: $\mathbb{E}^*\big[f\big(\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)\big)\big] \xrightarrow[n\to\infty]{} \int_E f(x)d\gamma(x)$.

Remark 4.79. If E is finite-dimensional, Theorem 4.77 reduces exactly to Van der Vaart's result stated in Theorem 4.60.

Remark 4.80. The only normality result in infinite dimension that we are aware of is [100, Theorem 6]. Gervini considers L^2 spaces and associates to the measure μ a stochastic process X. The normality result is stated for exact medians ($\ell = 0$, $\epsilon_n = 0$), under the assumption that the Karhunen–Loève decomposition of X has only a finite number of summands. In contrast, our normality Theorem 4.77 is valid in a generic separable Hilbert space, and under minimal distributional assumptions that match those of the finite-dimensional case.

4.5.2 In other Banach spaces

In Corollary 4.58 we gave a list of Banach spaces where approximate empirical ℓ-quantiles are consistent in the norm topology. Among these spaces, Theorem 4.77 indicates that asymptotic normality holds in $L^2(S, \mathcal{A}, \nu)$, $W^{k,2}(\Omega)$ and $S_2(H)$ since they are separable Hilbert spaces.

In the rest of this subsection we justify why the technique used in Section 4.5.1 (i.e., approximation with the quadratic function $\widehat{\Psi}_n$) fails to provide normality in the spaces

$$L^p(S, \mathcal{A}, \nu),\ W^{k,p}(\Omega),\ S_p(H) \quad \text{with } p > 2. \tag{4.11}$$

We consider a general separable Banach space $(E, \|\cdot\|)$ such that its norm (which we write alternatively as N) is twice Fréchet differentiable on $E \setminus \{0\}$. This is not a strong assumption since it is known to be satisfied for each $L^p(S, \mathcal{A}, \nu)$ when $p \geq 2$ (see [246, Section 2.2]). We let $\langle \cdot, \cdot \rangle$ denotes the duality pairing: given $f \in E^*$ and $\alpha \in E$, $\langle f, \alpha \rangle = f(\alpha)$. The quadratic function $\widehat{\Psi}_n$ is defined as

$$\begin{aligned}\widehat{\Psi}_n : \beta \mapsto \widehat{\phi}_n(\alpha_\star) &+ \langle \tfrac{1}{n} \textstyle\sum_{i=1}^n \left(\mathbb{1}_{X_i \neq \alpha_\star} DN(\alpha_\star - X_i) - \ell \right), \tfrac{\beta}{\sqrt{n}} \rangle \\ &+ \tfrac{1}{2} \mathbb{E}[\mathbb{1}_{X \neq \alpha_\star} D^2 N(\alpha_\star - X)](\tfrac{\beta}{\sqrt{n}}, \tfrac{\beta}{\sqrt{n}}).\end{aligned} \tag{4.12}$$

For fixed $x \in E$, $D^2N(\alpha_\star - x)$ is a bounded symmetric bilinear form (see [59, Theorem 5.1.1]) and it is nonnegative because N is convex. Here, for the sake of the argument, we disregard measurability and integrability concerns related to the Bochner integral that appears in (4.12). The Fréchet derivative of $\widehat{\Psi}_n$ at β is the linear functional

$$D\widehat{\Psi}_n(\beta) = \frac{1}{n^{3/2}} \sum_{i=1}^n \left(\mathbb{1}_{X_i \neq \alpha_\star} DN(\alpha_\star - X_i) - \ell \right) + \frac{1}{n} \mathbb{E}[\mathbb{1}_{X \neq \alpha_\star} D^2 N(\alpha_\star - X)](\beta, \cdot).$$

We define the operator $T : E \to E^*$ such that $T\beta = \mathbb{E}[\mathbb{1}_{X \neq \alpha_\star} D^2 N(\alpha_\star - X)](\beta, \cdot)$ and it is easily seen that T is bounded, $\langle T\beta_1, \beta_2 \rangle = \langle T\beta_2, \beta_1 \rangle$ and $\langle T\beta, \beta \rangle \geq 0$. Since $\widehat{\Psi}_n$ is convex, β is a minimizer of $\widehat{\Psi}_n$ if and only if

$$T\beta = -\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\mathbb{1}_{X_i \neq \alpha_\star} DN(\alpha_\star - X_i) - \ell \right).$$

To identify a unique minimizer $\widehat{\beta}_n$ and apply the central limit theorem as in Section 4.5.1, we need T to be a bijection. Furthermore, it was crucial in the proofs that $\kappa = \inf_{\|h\|=1}\langle Th, h\rangle$ be positive. Assuming these properties of T hold, we define the bilinear form $[\cdot,\cdot]$ by $[\beta_1, \beta_2] = \langle T\beta_1, \beta_2\rangle$. It is symmetric and positive definite by assumption. We write $\|\cdot\|_T$ for the associated norm and we note that

$$\kappa\|\beta\| \leq \|\beta\|_T \leq \|T\|_{op}\|\beta\|,$$

hence $(E, \|\cdot\|_T)$ is complete and the identity operator is a linear isomorphism between $(E, \|\cdot\|)$ and the Hilbert space $(E, \|\cdot\|_T)$. Hilbert spaces have Rademacher cotype 2 (see [168, Section 9.2] for the definition). Since the cotype is isomorphic invariant [6, Remark 6.2.11 (f)], the space $(E, \|\cdot\|)$ has cotype 2 as well. It is known however that the best possible cotype for spaces in (4.11) is p (see [168] and [178]), which is > 2. We have reached an absurdity, this is why the approximation technique with $\widehat{\Psi}_n$ is unsuccessful in these spaces.

By [168, Theorem 10.7], for each space in (4.11) we can find a mean-zero Borel probability measure ν with finite second moment that does not satisfy the central limit theorem: if $(Y_n)_{n\geq 1}$ is a sequence of i.i.d. Borel random elements with distribution ν, the sequence $(n^{-1/2}\sum_{i=1}^n Y_i)$ does not converge in distribution. This suggests that, in these spaces, approximate empirical quantiles may not converge at the parametric rate $n^{1/2}$.

4.6 Concluding remarks

A natural question is whether one can develop a general theory of M-estimation in infinite dimension, or *a minima* whether our work can be transposed to other M-parameters. We list key technical ingredients of this chapter, the proofs of which were quantiles-specific:

- Proposition 4.36 on uniform convergence of $(\widehat{\phi}_n)$ to ϕ over bounded sets,
- Proposition 4.40 on asymptotic boundedness of the estimator,
- Proposition 4.51 on well-posedness of ϕ,
- Lemma 4.62 on errors bounds for the Taylor expansion of the norm.

Extending our work to other parameters requires either new approaches, or an adaptation of these points.

Other directions for future research include: showing almost-sure Bahadur–Kiefer representations, deriving the exact rate of convergence of the estimator in Banach spaces such as $L^p(S, \mathcal{A}, \nu)$, $W^{k,p}(\Omega)$, $S_p(H)$ for $p \neq 2$, and investigating nonasymptotic properties (e.g., concentration bounds).

4.7 Proofs for Section 4.2

4.7.1 Proofs for Section 4.2.1

Proof of Proposition 4.2. 1. Given $\alpha \in E$, the reverse triangle inequality yields integrability of $x \mapsto \|\alpha - x\| - \|x\|$, hence ϕ_ℓ is well-defined. Furthermore,

$$\forall (\alpha_1, \alpha_2) \in E^2, \quad |\phi_\ell(\alpha_1) - \phi_\ell(\alpha_2)| = \int_E (\|\alpha_1 - x\| - \|\alpha_2 - x\|) d\mu(x) - \ell(\alpha_1 - \alpha_2)$$
$$\leq (1 + \|\ell\|_*)\|\alpha_1 - \alpha_2\|.$$

Convexity of ϕ_0 is a consequence of the standard triangle inequality, hence ϕ_ℓ is convex as well.

2. We adapt Valadier's proof [253]. Let $(\alpha_n)_{n \geq 1}$ be a sequence of nonzero vectors such that $\|\alpha_n\| \to \infty$. Note that

$$\Big|\frac{\phi_0(\alpha_n)}{\|\alpha_n\|} - 1\Big| = \Big| \int_E \frac{\|\alpha_n - x\| - \|x\| - \|\alpha_n\|}{\|\alpha_n\|} d\mu(x) \Big| \leq \int_E \frac{\big|\|\alpha_n - x\| - \|x\| - \|\alpha_n\|\big|}{\|\alpha_n\|} d\mu(x)$$

and for each $x \in E$, $\big|\|\alpha_n - x\| - \|x\| - \|\alpha_n\|\big|/\|\alpha_n\|$ is less than $2\min(\|x\|/\|\alpha_n\|, 1)$ which is bounded by 2 and converges pointwise to 0. By the dominated convergence theorem $\frac{\phi_0(\alpha_n)}{\|\alpha_n\|} \to_n 1$, hence the claim. As a consequence, $\frac{\phi_0(\alpha)}{\|\alpha\|} - \|\ell\|_*$ is no less than $(1 - \|\ell\|_*)/2$ when $\|\alpha\|$ is large enough, and in that case we have the estimate

$$\phi_\ell(\alpha) \geq \|\alpha\| \Big(\frac{\phi_0(\alpha)}{\|\alpha\|} - \|\ell\|_* \Big) \geq \frac{1 - \|\ell\|_*}{2} \|\alpha\|,$$

hence $\lim_{\|\alpha\| \to \infty} \phi_\ell(\alpha) = \infty$.

3. By the second item, there is some $r \geq 0$ such that $\|\alpha\| \geq r \implies \phi_\ell(\alpha) \geq 0$. From Lipschitzness and $\phi_\ell(0) = 0$, it follows that $\|\alpha\| \leq r \implies |\phi_\ell(\alpha)| \leq (1 + \|\ell\|_*)r$, hence $\phi_\ell(\alpha) \geq -(1 + \|\ell\|_*)r$ for every $\alpha \in E$. □

4.7.2 Proofs for Section 4.2.2

Proof of Proposition 4.5. 1. Since $\lim_{\alpha \to -\infty} F_X(\alpha) = 0$ and $\lim_{\alpha \to \infty} F_X(\alpha) = 1$, M_1 is nonempty and bounded below, so it has an infimum $\inf(M_1)$. Let $(\alpha_n)_{n \geq 1}$ be a nonincreasing sequence of elements of M_1 which converges to $\inf(M_1)$. Since F_X is right-continuous and $\alpha_n \in M_1$, we have $F_X(\inf(M_1)) \geq p$, hence $\inf(M_1) \in M_1$ and $\inf(M_1)$ is actually the minimal element of M_1. Since F_X is nondecreasing, any $\alpha \geq \min(M_1)$ is an element of M_1, hence $M_1 = [\min(M_1), \infty)$.

Since $\alpha \in M_2 \iff \mathbb{P}(-X \leq -\alpha) \geq 1 - p$, by replacing X with $-X$ and using the result we just proved on M_1, we see that M_2 has a maximal element and $M_2 = (-\infty, \max(M_2)]$.

2. Let $(\alpha_n)_{n \geq 1}$ be such that $\forall n \geq 1$, $\alpha_n < \min(M_1)$ and $\alpha_n \to \min(M_1)$. Since $\alpha_n \notin M_1$, we have $F_X(\alpha) < p$ and letting n go to infinity, $F_X(\min(M_1)^-) \leq p$, i.e., $\min(M_1) \in M_2$, thus $\min(M_1) \leq \max(M_2)$.

ϕ being a convex function over $\mathbb{R}$, it has left and right derivatives at each $\alpha \in \mathbb{R}$. They can be computed using the left and right derivatives of the absolute value, followed by an application of the dominated convergence theorem. Thus

$$\phi'_-(\alpha) = \int_{\mathbb{R}} (\mathbb{1}_{x<\alpha} - \mathbb{1}_{x\geq\alpha}) d\mu(x) - (2p-1) = 2(\mathbb{P}(X < \alpha) - p)$$

and

$$\phi'_+(\alpha) = \int_{\mathbb{R}} (\mathbb{1}_{x\leq\alpha} - \mathbb{1}_{x>\alpha}) d\mu(x) - (2p-1) = 2(\mathbb{P}(X \leq \alpha) - p).$$

Since ϕ is convex, α is a minimizer of ϕ if and only if $0 \in \partial\phi(\alpha)$, i.e., 0 is in the interval $[2(\mathbb{P}(X < \alpha) - p), 2(\mathbb{P}(X \leq \alpha) - p)]$, or equivalently $\alpha \in M_1 \cap M_2$. Using the explicit forms of M_1 and M_2 proved above, we find $\mathsf{Med}(\mu) = [\min(M_1), \max(M_2)]$. □

Proof of Corollary 4.6. 1. Suppose that μ has at least two ℓ-quantiles. By the second item of Proposition 4.5, we must have $\min(M_1) < \max(M_2)$. By the definitions of M_1 and M_2 we have the chain of inequalities

$$p \leq \mathbb{P}(X \leq \min(M_1)) \leq \mathbb{P}(X < \max(M_2)) \leq p,$$

hence $\mu\big((-\infty, \min(M_1)]\big) = \mathbb{P}(X \leq \min(M_1)) = p$. Replacing X with $-X$ yields similarly $\mathbb{P}(X \geq \max(M_2)) = 1-p$. For the converse, if $\alpha_1 < \alpha_2$ verify $\mu((-\infty, \alpha_1]) = p$ and $\mu([\alpha_2, \infty)) = 1 - p$, then $\alpha_1 \in M_1$ and $\alpha_2 \in M_2$, hence $\min(M_1) \leq \alpha_1 < \alpha_2 \leq \max(M_2)$. Consequently, by the second item of Proposition 4.5 there are at least two ℓ-quantiles.

2. If $F_X(\alpha_1) = F_X(\alpha_2) = p$ with $\alpha_1 \neq \alpha_2$, since $F_X(\alpha_1^-) \leq F_X(\alpha_1) = p$, α_1 is an ℓ-quantile, and so is α_2. Conversely, if μ has at least two ℓ-quantiles we consider $\alpha \in (\min(M_1), \max(M_2))$. Then as in the proof of the first item, $p \leq F_X(\min(M_1)) \leq F_X(\alpha) \leq F_X(\max(M_2)^-) \leq p$, hence $F_X(\min(M_1)) = F_X(\alpha) = p$.

3. If $\alpha < \min(M_1)$, then $\alpha \notin M_1$ and $F_X(\alpha) < p$. If $\alpha \in (\min(M_1), \max(M_2))$, then as seen in the proof of the second item, $F_X(\min(M_1)) = F_X(\alpha) = p$ hence F_X is equal to p on the interval $[\min(M_1), \alpha]$, hence also over $[\min(M_1), \max(M_2))$. If $\alpha > \max(M_2)$, then $\alpha \notin M_2$ and $F_X(\alpha) \geq F_X(\alpha^-) > p$. □

4.7.3 Proofs for Section 4.2.3

In the following lemma we consider the situation where there is an isometry J from E into another vector space F. The measure μ is then naturally transported to a measure $\widetilde{\mu}$ on the image $J(E)$, and this gives rise to another objective function $\widetilde{\phi}$. Note that J need not be surjective.

Lemma 4.81. *Let $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ be normed spaces, $J : E \to F$ be a linear isometry such that $J(E)$ is a Borel subset of F, and μ be a Borel probability measure on E.*

1. *Let $\widetilde{\mu}$ be the measure induced by J on $J(E)$, $\widetilde{\ell} = \ell \circ J^{-1}$ and $\widetilde{\phi} : J(E) \to \mathbb{R}$ be the corresponding function given by Definition 4.1. Then $\|\widetilde{\ell}\|_* = \|\ell\|_* < 1$,*

$$\forall \alpha \in E,\ \phi(\alpha) = \widetilde{\phi}(J\alpha) \quad \text{and} \quad \forall \beta \in J(E),\ \widetilde{\phi}(\beta) = \phi(J^{-1}\beta).$$

2. *Let $F = E^{**}$ and J be the canonical isometry into the second dual of E. Further, let $\widehat{\mu} = J_\sharp\mu$ be the measure induced by J on E^{**}, and $\widehat{\phi}_0 : E^{**} \to \mathbb{R}$ be the function given by Definition 4.1. Then*

$$\forall\alpha \in E,\ \phi_0(\alpha) = \widehat{\phi}_0(J\alpha) \quad \text{and} \quad \forall\beta \in J(E),\ \widehat{\phi}_0(\beta) = \phi_0(J^{-1}\beta).$$

Proof of Lemma 4.81. Since J is a bounded operator, it is Borel measurable from E to F, hence the pushforward measure $J_\sharp\mu$ is a Borel probability measure on F, and $\widetilde{\mu}$ is its restriction to the Borel subset $J(E)$. We have

$$\|\widetilde{\ell}\|_* = \sup_{\substack{\beta\in J(E)\\ \|\beta\|_F=1}} \ell(J^{-1}(\beta)) = \sup_{\substack{\alpha\in E\\ \|\alpha\|_E=1}} \ell(\alpha) = \|\ell\|_*.$$

For any $\alpha \in E$,

$$\begin{aligned}
\phi(\alpha) &= \int_E (\|J^{-1}(J\alpha - Jx)\|_E - \|J^{-1}(Jx)\|_E) d\mu(x) - \ell(J^{-1}J\alpha) \\
&= \int_E (\|J\alpha - Jx\|_F - \|Jx\|_F) d\mu(x) - (\ell \circ J^{-1})(J\alpha) \\
&= \int_{J(E)} (\|J\alpha - y\|_F - \|y\|_F) d\widetilde{\mu}(y) - \widetilde{\ell}(J\alpha) = \widetilde{\phi}(J\alpha).
\end{aligned}$$

The second item is obtained by a similar computation with $\ell = 0$. □

Proof of Proposition 4.8. 1. By Proposition 4.2 ϕ is continuous, convex and coercive. Since E is reflexive, ϕ reaches its infimum by Theorem 2.11 and Remark 2.13 in [17].

2. In Section 3.9 of [148] Kemperman proves existence only when $\ell = 0$ and $E = F^*$ where F is a separable Banach space. It is easily seen from his work that the completeness assumption on F is superfluous. Since ℓ is a linear functional on F^*, it is in fact an element of F^{**}. By assumption ℓ is in $J(F)$, hence ℓ is simply an evaluation map. To extend Kemperman's proof, it suffices to note that $\ell(\alpha_n) \to_n \ell(\alpha)$ for any sequence (α_n) that converges in the weak∗ topology of E^* to a limit α.

Assume now that equality is replaced by the surjective isometry $I : E \to F^*$. Let $\widetilde{\mu}$, $\widetilde{\ell}$, $\widetilde{\phi}$ be as in Lemma 4.81. The assumption on ℓ rewrites as $\widetilde{\ell} \in J(F)$, thus by the previous paragraph $\widetilde{\mu}$ has at least one $\widetilde{\ell}$-quantile, say $\beta_\star \in \mathsf{Quant}(\widetilde{\mu})$. By Lemma 4.81, for any $\alpha \in E$ we have $\phi(\alpha) = \widetilde{\phi}(I\alpha) \geq \widetilde{\phi}(\beta_\star) = \phi(I^{-1}\beta_\star)$, hence $I^{-1}\beta_\star \in \mathsf{Quant}(\mu)$.

3. Let $P : E^{**} \to E^{**}$ denote the bounded linear projection with range $J(E)$, so that $J(E) = \ker(\mathrm{Id} - P)$ and $J(E)$ is a closed subspace of E^{**}, hence a Borel subset of E^{**} (this will be needed below).

Using the notations of Lemma 4.81, we prove first that $\widehat{\phi}_0$ has a minimum by exploiting the weak∗ compactness of closed balls in E^{**} and the weak∗ lower semicontinuity of the norm. Since $(E^{**}, \|\cdot\|_{**})$ may not be separable (this is typically the case when $E = L^1$), we must at times consider the separable subspace $J(E)$ in order to invoke some external results.

Lemma 4.82. *The objective function $\widehat{\phi}_0$ reaches its infimum over E^{**}.*

Proof of Lemma 4.82. By Proposition 4.2, $\widehat{\phi}_0$ has a finite infimum and it is coercive, so there is some $R > 0$ such that $\|\beta\|_{**} > R \implies \widehat{\phi}_0(\beta) > \inf(\widehat{\phi}_0)+1$. Since E^{**} is the dual space of E^*, we can equip E^{**} with the weak$*$ topology. Let $B = \{y \in E^{**} : \|y\|_{**} \leq R\}$ denote the closed ball with center 0 and radius R. By the Banach–Alaoglu theorem [7, Theorem 6.21], B is weak$*$ compact. If we prove that the restriction of $\widehat{\phi}_0$ to B is weak$*$ lower semicontinuous (i.e., lower semicontinuous w.r.t. the topology that B inherits from the weak$*$ topology), then by [7, Theorem 2.43] $\widehat{\phi}_0$ reaches its infimum over B, which coincides with its infimum over the whole space E^{**} by the definition of R.

Consequently, it remains to prove that $\widehat{\phi}_0|_B$ is lower semicontinuous with respect to $\mathcal{T}_B$, the relative topology induced on B by the weak$*$ topology of E^{**}. We fix $y_0 \in \mathbb{R}$ and we prove that the set $\{\beta \in B : \widehat{\phi}_0(\beta) > y_0\}$ is open w.r.t. the topology $\mathcal{T}_B$. Fix $\beta_0 \in B$ such that $\widehat{\phi}_0(\beta_0) > y_0$, and set $\epsilon = \widehat{\phi}_0(\beta_0) - y_0 > 0$. We consider $\widetilde{\mu}$, the restriction of $J_\sharp\mu$ to the Borel set $J(E)$. Note that $\widetilde{\mu}(J(E)) = \widehat{\mu}(J(E) \cap J(E)) = (J_\sharp\mu)(J(E)) = 1$, hence $\widetilde{\mu}$ is a probability measure on $J(E)$. Since $(E, \|\cdot\|)$ is separable, so is $(J(E), \|\cdot\|_{**})$. By Theorems 15.10 and 15.12 in [7], there is some sequence $(\widetilde{\mu}_n)_{n\geq 1}$ of finitely supported measures on $J(E)$ which converges weakly (i.e., in the usual sense for measures) to $\widetilde{\mu}$. For each $\beta \in B$ we define the function $f_\beta : J(E) \to \mathbb{R}$, $y \mapsto \|\beta - y\|_{**} - \|y\|_{**}$. The family $(f_\beta)_{\beta\in B}$ is uniformly bounded by R and pointwise equicontinuous. Theorem 3.1 in [219] yields the following convergence:

$$\sup_{\beta\in B} \left| \int_{J(E)} f_\beta(y) d\widetilde{\mu}_n(y) - \int_{J(E)} f_\beta(y) d\widetilde{\mu}(y) \right| \xrightarrow[n\to\infty]{} 0,$$

so there is some $n_0 \geq 1$ such that

$$\sup_{\beta\in B} \left| \int_{J(E)} f_\beta(y) d\widetilde{\mu}_{n_0}(y) - \int_{J(E)} f_\beta(y) d\widetilde{\mu}(y) \right| < \frac{\epsilon}{2}. \tag{4.13}$$

Let us write the measure $\widetilde{\mu}_{n_0}$ as $\sum_{i=1}^m p_i \delta_{y_i}$ with $y_i \in J(E)$. Note that

$$\int_{J(E)} f_\beta(y) d\widetilde{\mu}_{n_0}(y) = \sum_{i=1}^m p_i(\|\beta - y_i\|_{**} - \|y_i\|_{**})$$

and

$$\begin{aligned} \int_{J(E)} f_\beta(y) d\widetilde{\mu}(y) &= \int_{J(E)} (\|\beta - y\|_{**} - \|y\|_{**}) d\widetilde{\mu}(y) \\ &\overset{(i)}{=} \int_{J(E)} (\|\beta - y\|_{**} - \|y\|_{**}) d\widehat{\mu}(y) \\ &\overset{(ii)}{=} \int_{E^{**}} (\|\beta - y\|_{**} - \|y\|_{**}) d\widehat{\mu}(y) = \widehat{\phi}_0(\beta), \end{aligned}$$

where equality (i) follows from the definition of $\widetilde{\mu}$ and integration over $J(E)$, and (ii) from the fact that $\widehat{\mu}$ is concentrated on $J(E)$. Letting $\varphi : E^{**} \to \mathbb{R}$, $\beta \mapsto \sum_{i=1}^m p_i(\|\beta - y_i\|_{**} - \|y_i\|_{**})$, (4.13) rewrites as

$$\sup_{\beta\in B} |\varphi(\beta) - \widehat{\phi}_0(\beta)| < \epsilon/2. \tag{4.14}$$

Since $\|\cdot\|_{**}$ is weak$*$ lower semicontinuous (see [7, Lemma 6.22]), the function φ is weak$*$ lower semicontinuous, hence the restriction of φ to B is lower semicontinuous w.r.t. the topology $\mathcal{T}_B$. By inequality (4.14), $\varphi(\beta_0) > \widehat{\phi}_0(\beta_0) - \epsilon/2 = y_0 + \epsilon/2$. By lower semicontinuity of φ there is some open $U \in \mathcal{T}_B$ such that $\beta \in U \implies \varphi(\beta) > y_0 + \epsilon/2$. Finally, since $U \subset B$,

$$\forall \beta \in U, \quad \widehat{\phi}_0(\beta) = \varphi(\beta) + (\widehat{\phi}_0(\beta) - \varphi(\beta)) > (y_0 + \epsilon/2) - \epsilon/2 = y_0.$$

This shows the set $\{\beta \in B : \widehat{\phi}_0(\beta) > y_0\}$ is open in $\mathcal{T}_B$. Thus $\widetilde{\phi}_0|_B$ is lower semicontinuous with respect to $\mathcal{T}_B$. This ends the proof of Lemma 4.82. □

Let $\beta_\star \in E^{**}$ be a minimizer of $\widehat{\phi}_0$, which exists according to Lemma 4.82. Note that

$$\begin{aligned}
\phi_0(J^{-1}(P\beta_\star)) &= \widehat{\phi}_0(P\beta_\star) && (4.15)\\
&= \int_E (\|P\beta_\star - Jx\|_{**} - \|Jx\|_{**}) d\mu(x) \\
&= \int_E (\|P(\beta_\star - Jx)\|_{**} - \|Jx\|_{**}) d\mu(x) && (4.16)\\
&\le \int_E (\|\beta_\star - Jx\|_{**} - \|Jx\|_{**}) d\mu(x) && (4.17)\\
&= \widehat{\phi}_0(\beta_\star) \\
&\le \inf_{\beta \in J(E)} \widehat{\phi}_0(\beta) \\
&= \inf_{\alpha \in E} \phi_0(\alpha), && (4.18)
\end{aligned}$$

where (4.15) and (4.18) stem from Lemma 4.81, (4.16) follows from $P(Jx) = Jx$, and (4.17) is a consequence of $\|P\| = 1$. We obtained the bound $\phi_0(J^{-1}(P\beta_\star)) \le \inf_{\alpha \in E} \phi_0(\alpha)$, hence $J^{-1}(P\beta_\star) \in E$ is a minimizer of ϕ_0, i.e., a geometric median of μ. □

Let F be a normed vector space over C_{23}. The following lemma states connections between F and $F_{\mathbb{R}}$:

Lemma 4.83. *1. If F is reflexive then $F_{\mathbb{R}}$ is reflexive.*

2. If F is C_{24}-isometrically isomorphic to the dual of a separable complex normed space, then $F_{\mathbb{R}}$ is $\mathbb{R}$-isometrically isomorphic to the dual of a separable real normed space.

3. If F is separable and $J_{C_{25}}(F)$ is 1-complemented in F^{**} (where $J_{C_{26}}$ is the canonical C_{27}-linear isometry from F to F^{**}), then $F_{\mathbb{R}}$ is separable and $J(F)$ is 1-complemented in $(F_{\mathbb{R}})^{**}$.

Proof of Lemma 4.83. Item 1. follows from Proposition 1.13.1 in [186].

For the second item, suppose F is C_{28}-isometrically isomorphic to G^* where G is a complex separable normed space. Then $G_{\mathbb{R}}$ is separable and $F_{\mathbb{R}}$ is $\mathbb{R}$-isometrically isomorphic to $(G^*)_\mathbb{R}$. Since $(G^*)_\mathbb{R}$ is $\mathbb{R}$-isometrically isomorphic to $(G_\mathbb{R})^*$ (see the proof of [186, Proposition 1.13.1]), $F_\mathbb{R}$ is $\mathbb{R}$-isometrically isomorphic to $(G_\mathbb{R})^*$, which is the dual of a separable real normed space.

Item 3. follows from Remark c) in [114, p.101]. □

Proof of Corollary 4.12. By the first point of Proposition 4.8, it suffices to check that each space is reflexive. If E derives from a complex vector space F we show that F is reflexive and then we apply Lemma 4.83. If E is truly a real vector space we directly check that E is reflexive.

Reflexivity is a well-known fact for finite-dimensional, Hilbert and $L^p(S, \mathcal{A}, \nu)$ spaces. $W^{k,p}(\Omega)$ is reflexive when $1 < p < \infty$, see [1, Theorem 3.6]. Under the assumptions of the corollary, $L^\Phi(S, \mathcal{A}, \nu)$ is reflexive by [218, Theorem 10 p.112]. When $p \in (1, \infty)$, the space $S_p(H)$ is uniformly convex (see Definition 4.15 and [184] for the proof), hence reflexive (see [186, Theorem 5.2.15]). □

Proof of Corollary 4.13. Regarding the first item, the assumptions on $(S, \mathcal{A}, \nu)$ are given to ensure the separability of $L^1(S, \mathcal{A}, \nu)$ (see [229, Lemma 27.23]). Since $L^\infty(S, \mathcal{A}, \nu)$ is always isometric to the dual of $L^1(S, \mathcal{A}, \nu)$, existence of medians when $p = \infty$ follows. For $p = 1$, it suffices to note that $L^1(S, \mathcal{A}, \nu)$ is 1-complemented in its second dual: this is proved for $L^1([0,1], \mathcal{B}([0,1]), \lambda)$ in Proposition 6.3.10 of [6] and for any $L^1(S, \mathcal{A}, \nu)$ in Appendix B10 of [72].

For item 2., it is proven in [208, Proposition 2.4] (with $k = 1$ in their notation) that $BV(\Omega)$ is isometrically isomorphic to the dual of the quotient space $C_0(\Omega, \ell^n_\infty)/F$ where ℓ^n_∞ classically denotes $\mathbb{R}^n$ with the $\|\cdot\|_\infty$ norm , $C_0(\Omega, \ell^n_\infty)$ is the closure in the sup norm of $C_c(\Omega, \ell^n_\infty)$ (the space of ℓ^n_∞-valued continuous functions with compact supports in Ω), and F is a closed subspace of $C_0(\Omega, \ell^n_\infty)$. It is easily seen that $C_0(\Omega, \ell^n_\infty)$ is separable, hence the quotient above is separable as well.

For the third item, we let $C(H)$ denote the space of compact operators on H equipped with the operator norm. Since H is separable, so is $C(H)$ by [90, Proposition 7.5]. Besides, $S_1(H)$ is isometrically isomorphic to $C(H)^*$ (see [189, Proposition 16.24]), so existence is obtained for $S_1(H)$.

It is known that $B(H)$ is isometrically isomorphic to $S_1(H)^*$ (see [189, Proposition 16.26]). Since H is separable, $S_1(H)$ is separable w.r.t. the trace-class norm by [68, Theorem 18.11 (d)]. □

4.7.4 Proofs for Section 4.2.4

Proof of Proposition 4.17. 1. We show the stronger result that ϕ is a strictly convex function. Assume the contrary: there are some $\lambda \in (0, 1)$, $\alpha_1 \neq \alpha_2$ with

$$\phi((1-\lambda)\alpha_1 + \lambda\alpha_2) = (1-\lambda)\phi(\alpha_1) + \lambda\phi(\alpha_2).$$

Then the function

$$f : x \mapsto (1-\lambda)\|\alpha_1 - x\| + \lambda\|\alpha_2 - x\| - \|(1-\lambda)(\alpha_1 - x) + \lambda(\alpha_2 - x)\|$$

is nonnegative and has μ-integral zero. Let $A = \{x \in E : f(x) = 0\}$, so that $\mu(A) = 1$. Consider $x \in A$ and assume first that $x \notin \{\alpha_1, \alpha_2\}$. By the strict convexity of E, there is some $K_x > 0$ with $(1-\lambda)(\alpha_1 - x) = K_x\lambda(\alpha_2 - x)$. Since $\alpha_1 \neq \alpha_2$, we must have $1 - \lambda - \lambda K_x \neq 0$, and the previous equality yields

$$x = \alpha_1 + \frac{\lambda K_x}{1-\lambda-\lambda K_x}(\alpha_1 - \alpha_2),$$

hence

$$x \in \alpha_1 + \mathbb{R}(\alpha_1 - \alpha_2). \tag{4.19}$$

Since (4.19) holds as well in the case where $x \in \{\alpha_1, \alpha_2\}$, we obtain the inclusion $A \subset \alpha_1 + \mathbb{R}(\alpha_1 - \alpha_2)$ and μ gives full mass to the line $\alpha_1 + \mathbb{R}(\alpha_1 - \alpha_2)$. This contradicts $\mu \in \mathcal{M}_\sim$ hence ϕ is strictly convex, so it can have at most one minimizer.

2. First we drop the condition $\mu \in \mathcal{M}_\sim$. Consider $\mu = \frac{1}{2}(\delta_{x_1} + \delta_{x_2})$ where x_1, x_2 are any two distinct unit vectors in E. By the triangle inequality, $\phi_0(\alpha) \geq 1/2\|x_1 - x_2\| - 1$ holds for any $\alpha \in E$ and equality is attained over $[x_1, x_2]$, the closed line segment between x_1 and x_2. Hence $[x_1, x_2] \subset \mathsf{Med}(\mu)$. Note that the inclusion holds in any space E, regardless of strict convexity.

Next, we give an example of a space lacking strict convexity and a measure $\mu \in \mathcal{M}_\sim$ with more than one median. Consider $(\mathbb{R}^2, \|\cdot\|_\infty)$ and $\mu = \frac{1}{4}(\delta_{(-1,0)} + \delta_{(1,0)} + \delta_{(0,1)} + \delta_{(0,-1)})$ (this is Example 3.4 in [148]). This space is not strictly convex and μ is not concentrated on a line. Straightforward computations show that $\mathsf{Med}(\mu)$ is the convex hull of the four points that support μ.

3. We prove the contrapositive: we assume E is not strictly convex and we construct some $\mu \in \mathcal{M}_\sim$ with at least two medians. By assumption, it is possible to find two distinct points y_1, y_2 in the unit sphere such that the segment $[y_1, y_2]$ is a subset of the unit sphere as well. Let $x_1 = \frac{2}{3}y_1 + \frac{1}{3}y_2$, $x_2 = \frac{1}{3}y_1 + \frac{2}{3}y_2$ be on the segment and $\mu = \frac{1}{4}(\delta_{x_1} + \delta_{-x_1} + \delta_{x_2} + \delta_{-x_2})$. For every $\alpha \in E$, by the triangle inequality

$$\phi_0(\alpha) + 1 = \frac{1}{4}(\|\alpha - x_1\| + \|\alpha + x_1\| + \|\alpha - x_2\| + \|\alpha + x_2\|) \geq \frac{1}{4}(\|2x_1\| + \|2x_2\|) = 1,$$

hence ϕ_0 is bounded from below by 0. But $\phi_0(0) = 0$ and

$$\phi_0\big(1/6(y_1 - y_2)\big) + 1 = \frac{1}{4}\Big(\|\frac{1}{2}y_1 + \frac{1}{2}y_2\| + \|\frac{5}{6}y_1 + \frac{1}{6}y_2\| + \|\frac{5}{6}y_1 + \frac{1}{6}y_2\| + \|\frac{1}{2}y_1 + \frac{1}{2}y_2\|\Big) = 1$$

since by assumption the quantities inside each norm lie on the unit sphere. Consequently, 0 and $1/6(y_1 - y_2)$ are distinct geometric medians of μ. It remains to show that $\mu \in \mathcal{M}_\sim$. If it is not the case, x_2 must lie on the line $\mathbb{R}x_1$, hence y_1 and y_2 are linearly dependent. Since they are distinct unit vectors, this implies $y_1 = -y_2$. This contradicts the hypothesis that $[y_1, y_2]$ be a subset of the unit sphere, since the segment contains zero. □

Proof of Corollary 4.18. With the notations of Lemma 4.83, if F is a complex vector space such that $(F, \|\cdot\|)$ is strictly convex then $(F_{\mathbb{R}}, \|\cdot\|)$ is a strictly convex real vector space. Consequently, when E derives from a complex vector space F it suffices to check that F is strictly convex.

1. A uniformly convex Banach space, whether real or complex, is both reflexive [186, Theorem 5.2.15] and strictly convex [186, Proposition 5.2.6].

a) A strictly convex finite-dimensional space is uniformly convex [186, Proposition 5.2.14].

b) A Hilbert space is uniformly convex [91, p.430].

c) For $1 < p < \infty$, $L^p(S, \mathcal{A}, \nu)$ is uniformly convex [91, Theorem 9.3].

d) For $1 < p < \infty$, $W^{k,p}(\Omega)$ is uniformly convex [1, Theorem 3.6].

e) $S_p(H)$ with $1 < p < \infty$ is uniformly convex [184].

2. Under these assumptions, $L^\Phi(S, \mathcal{A}, \nu)$ is strictly convex when equipped with the Orlicz norm [218, Corollary 7 p.275]. Existence is already obtained from Corollary 4.12.
3. Under these assumptions, $L^\Phi(S, \mathcal{A}, \nu)$ is strictly convex when equipped with the gauge norm [218, Corollary 5 p.272]. Existence is already obtained from Corollary 4.12. □

In the proof of Proposition 4.19 below, we need the following common technical fact which we show here since we could not find a reference. With the terminology of Section 4.2.3, we show that L is 1-complemented in E.

Lemma 4.84. *Let L be a one-dimensional subspace of E. There exists a bounded linear projection $P : E \to E$ with range L and satisfying $\|P\| = 1$.*

Proof of Lemma 4.84. We write $L = \mathbb{R}v$ with $\|v\| = 1$. Let $f \in L^*$ be the bounded linear functional which maps $x = \lambda v$ to $f(x) = \lambda$. It is clear that f has dual norm 1 and $|f(v)| = 1$. By the Hahn-Banach theorem it has an extension $g \in E^*$ with dual norm 1. Let $P : E \to E$, $x \mapsto g(x)v$. P is linear, $P^2 = P$, the range of P is L, P is bounded and $\|P\| = 1$. □

Proof of Proposition 4.19. 1. We write the affine line as $L = u + \mathbb{R}v$ with $u, v \in E$ and $\|v\| = 1$. Let $\widetilde{\mu}$ be the shifted measure defined by $\widetilde{\mu}(A) = \mu(u + A)$, so that $\widetilde{\mu}$ is concentrated on the line $\mathbb{R}v$. We temporarily write ϕ_μ and $\phi_{\widetilde{\mu}}$ to clarify the measure we consider when integrating. It is easily seen that for any $\alpha \in E$, $\phi_{\widetilde{\mu}}(\alpha - u) = \phi_\mu(\alpha) - \phi_\mu(u)$, hence $\mathsf{Med}(\mu) = u + \mathsf{Med}(\widetilde{\mu})$. We can therefore suppose w.l.o.g. that $u = 0$, so that L is a linear subspace of dimension one which supports μ.

We show the inclusion $\mathsf{Med}(\mu) \subset L$. Suppose first that μ is degenerate, i.e., $\mu = \delta_x$ for some $x \in E$. Since $\mu(L) = 1$, x must lie on L. We also have $\mathsf{Med}(\mu) = \{x\}$, hence the claim. We can therefore assume that μ is nondegenerate in the rest of the paragraph. Suppose for the sake of contradiction that there is a minimizer $\alpha_\star$ that is not in L. Since μ is concentrated on L and using Lemma 4.84,

$$\begin{aligned}\phi_0(\alpha_\star) = \int_L (\|\alpha_\star - x\| - \|x\|) d\mu(x) &\geq \int_L (\|P(\alpha_\star - x)\| - \|x\|) d\mu(x) \\ &= \int_L (\|P(\alpha_\star) - x\| - \|x\|) d\mu(x) = \phi_0(P(\alpha_\star))\end{aligned}$$

hence $P(\alpha_\star)$ is a minimizer as well, which is distinct from $\alpha_\star$ by assumption. By the strict convexity of $(E, \|\cdot\|)$, following the proof of item 1. in Proposition 4.17 we obtain that μ is concentrated on the affine line $\alpha_\star + \mathbb{R}(\alpha_\star - P\alpha_\star)$, hence the intersection

$$L \cap \big(\alpha_\star + \mathbb{R}(\alpha_\star - P\alpha_\star)\big)$$

has probability 1, and is thus nonempty. Since $\alpha_\star$ is in $\alpha_\star + \mathbb{R}(\alpha_\star - P\alpha_\star)$ but not in L, the intersection of the lines must be some singleton $\{x\}$ which has mass 1, hence μ is degenerate, a contradiction. The inclusion $\mathsf{Med}(\mu) \subset L$ is proved.

Since any minimizer of ϕ must lie in L we can restrict our attention to this line. As in the proof of Lemma 4.84, let $f \in L^*$ be the isomorphism $f : \lambda v \mapsto \lambda$ and ν be the pushforward measure on $\mathbb{R}$ defined by $\nu = f_\sharp \mu$. For $\alpha \in L$, the equality $\alpha = f(\alpha)v$ holds, thus

$$\begin{aligned}\phi_\mu(\alpha) &= \int_L (\|f(\alpha)v - x\| - \|x\|)d\mu(x) = \int_L (\|f(\alpha)v - f(x)v\| - \|f(x)v\|)d\mu(x) \\ &= \int_{\mathbb{R}} (|f(\alpha) - \lambda| - |\lambda|)d\nu(\lambda) = \phi_\nu(f(\alpha)),\end{aligned}$$

where we used that v has norm 1. Let $m_{\min}$ and $m_{\max} \in \mathbb{R}$ denote the smallest and largest median of ν (see Proposition 4.5). From the previous paragraph, $\alpha \in \mathsf{Med}(\mu) \iff \alpha \in \mathsf{Med}(\mu) \cap L$ and from the last computation $\alpha \in \mathsf{Med}(\mu) \cap L \iff f(\alpha) \in \mathsf{Med}(\nu)$, so equivalently α is in the segment $[m_{\min}v, m_{\max}v] \subset L$.

2. Consider $(\mathbb{R}^2, \|\cdot\|_\infty)$ and $\mu = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)})$ (this is Example 3.4 in [148]). Then $\mu \in \mathcal{M}_-$ but $\mathsf{Med}(\mu)$ is the square with vertices $(-1, 0), (1, 0), (0, 1), (0, -1)$.

3. We consider $(\mathbb{R}^2, \|\cdot\|_2)$, $\mu = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)})$ and $\ell : (\alpha_1, \alpha_2) \mapsto \alpha_2/2$. The smallest value attained by ϕ on the supporting line $\mathbb{R} \times \{0\}$ is 0, while the global minimum value of ϕ is $\sqrt{3}/2 - 1 < 0$, which is attained at $\alpha = (0, 1/\sqrt{3})$.

4. The proof is similar to that of item 3. in Proposition 4.17. We proceed with the contrapositive: we assume that E is not strictly convex and we construct $\mu \in \mathcal{M}_-$ such that $\mathsf{Med}(\mu)$ is not a subset of the affine line supporting μ. It is possible to find two distinct points y_1, y_2 in the unit sphere such that the segment $[y_1, y_2]$ is a subset of the unit sphere as well. Let $x_1 = \frac{1}{2}y_1 + \frac{1}{2}y_2$ be on the segment and $\mu = \frac{1}{2}(\delta_{x_1} + \delta_{-x_1}) \in \mathcal{M}_-$. For every $\alpha \in E$ the triangle inequality yields $\phi(\alpha) \geq 0$, and

$$\phi\big(1/4(y_1 - y_2)\big) = \frac{1}{2}\Big(\|\frac{1}{4}y_1 + \frac{3}{4}y_2\| + \|\frac{3}{4}y_1 + \frac{1}{4}y_2\|\Big) - 1 = 0.$$

Consequently, $1/4(y_1 - y_2)$ is a median of μ and it remains to show that $1/4(y_1 - y_2) \notin \mathbb{R}x_1$, or equivalently that $y_1 - y_2$ is linearly independent of $y_1 + y_2$. It suffices to observe that y_1 and y_2 are linearly independent themselves. Otherwise, since they are distinct unit vectors we have $y_2 = -y_1$ and $x_1 = 0$ lies on the unit sphere, which is absurd.

□

4.8 Proofs for Section 4.3

4.8.1 Proofs for Section 4.3.2

The following lemma collects useful facts on outer and inner probabilities that we will use in proofs later.

Lemma 4.85. *For arbitrary subsets B_1, B_2 of Ω,*

1. *If B_1 is measurable, i.e., $B_1 \in \mathcal{F}$, then $\mathbb{P}^*(B_1) = \mathbb{P}_*(B_1) = \mathbb{P}(B_1)$.*
2. *If $B_1 \subset B_2$ then $\mathbb{P}^*(B_1) \leq \mathbb{P}^*(B_2)$ and $\mathbb{P}_*(B_1) \leq \mathbb{P}_*(B_2)$.*
3. *$\mathbb{P}^*(B_1 \cup B_2) \leq \mathbb{P}^*(B_1) + \mathbb{P}^*(B_2)$ and $\mathbb{P}_*(B_1 \cap B_2) \geq \mathbb{P}_*(B_1) + \mathbb{P}_*(B_2) - 1$.*
4. *If $\mathbb{P}_*(B_1) = \mathbb{P}_*(B_2) = 1$ then $\mathbb{P}_*(B_1 \cap B_2) = 1$.*

Let $(A_n)_{n\geq 1}$, $(B_n)_{n\geq 1}$ be sequences of subsets of Ω.

5. *If $\lim_n \mathbb{P}^*(A_n \cap B_n) = 0$ and $\lim_n \mathbb{P}_*(B_n) = 1$, then $\lim_n \mathbb{P}^*(A_n) = 0$.*

Proof of Lemma 4.85. The first and second item follow easily from the definitions of inner and outer probabilities. For the third item it suffices to prove $\mathbb{P}^*(B_1 \cup B_2) \leq \mathbb{P}^*(B_1) + \mathbb{P}^*(B_2)$: this inequality is a consequence of Problem 15 in [261, Chapter 1.2]. The fourth item follows immediately from the third. For the last item, we note the inclusion $A_n \subset (A_n \cap B_n) \cup B_n^c$, hence by subadditivity

$$\mathbb{P}^*(A_n) \leq \mathbb{P}^*(A_n \cap B_n) + \mathbb{P}^*(B_n^c) = \mathbb{P}^*(A_n \cap B_n) + 1 - \mathbb{P}_*(B_n),$$

and the claim follows. □

4.8.2 Addendum and proofs for Section 4.3.3

We introduce the weaker notion of universal measurability.

Definition 4.86. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be a measurable spaces.

1. A subset $C \subset X$ is called *universally measurable* in $(X, \mathcal{A})$ if C belongs to the ν-completion of $\mathcal{A}$ for each probability measure ν defined on $(X, \mathcal{A})$.
2. A map $f : X \to Y$ is called universally measurable if for each $D \in \mathcal{B}$, the set $f^{-1}(D)$ is *universally measurable* in $(X, \mathcal{A})$.

Proof of Theorem 4.25. Let $n \geq 1$ be fixed. We let f be the function

$$\begin{aligned} f \colon E^n \times \mathbb{R}_{>0} \times E &\to \mathbb{R} \\ (x_1, \ldots, x_n, \epsilon, \alpha) &\mapsto \frac{1}{n} \sum_{i=1}^{n} (\|\alpha - x_i\| - \|x_i\|) - \ell(\alpha). \end{aligned}$$

We show the existence of a Borel measurable $\psi : E^n \times \mathbb{R}_{>0} \to E$ such that

$$f\big(x_1, \ldots, x_n, \epsilon, \psi(x_1, \ldots, x_n, \epsilon)\big) \le \epsilon + \inf_{\alpha \in E} f(x_1, \ldots, x_n, \epsilon, \alpha)$$

for each $x_1, \ldots, x_n, \epsilon$. The infimum is finite by Proposition 4.2. We could make use of Theorem 1 in Schäl [228] by letting, with their notation, $S = E^n \times \mathbb{R}_{>0}$, $A = E$, $D : s \mapsto E$ and $\varepsilon : (x_1, \ldots, x_n, \epsilon) \mapsto \epsilon$. We give a direct proof instead that does not resort to Schäl's machinery. We strive for the greatest generality in our arguments, so that our method may be applied to other contexts.

Since E is separable it has some countable dense subset $\{e_p : p \ge 1\}$. We let τ be the function

$$\begin{aligned} \tau\colon \quad & E^n \times \mathbb{R}_{>0} \to \mathbb{N}_{>0} \\ & (x_1, \ldots, x_n, \epsilon) \mapsto \min\{p : f(x_1, \ldots, x_n, \epsilon, e_p) \le \inf_{\alpha \in E} f(x_1, \ldots, x_n, \epsilon, \alpha) + \epsilon\}. \end{aligned}$$

τ is well-defined because ϵ is positive and f is upper semicontinuous in its last argument. Regarding measurability of τ, we put the discrete σ-algebra on $\mathbb{N}_{>0}$ and we note that

$$\begin{aligned} \{\tau = p\} = & \bigcap_{i=1}^{p-1} \{(x_1, \ldots, x_n, \epsilon) : f(x_1, \ldots, x_n, \epsilon, e_i) > \inf_{\alpha \in E} f(x_1, \ldots, x_n, \epsilon, \alpha) + \epsilon\} \\ & \cap \{(x_1, \ldots, x_n, \epsilon) : f(x_1, \ldots, x_n, \epsilon, e_p) \le \inf_{\alpha \in E} f(x_1, \ldots, x_n, \epsilon, \alpha) + \epsilon\}. \end{aligned}$$

For each α, the function $(x_1, \ldots, x_n, \epsilon) \mapsto f(x_1, \ldots, x_n, \epsilon, \alpha)$ is upper semicontinuous, hence so is the function $(x_1, \ldots, x_n, \epsilon) \mapsto \inf_{\alpha \in E} f(x_1, \ldots, x_n, \epsilon, \alpha)$. Since upper semicontinuity implies Borel measurability, τ is Borel measurable. Finally we let $\psi = e_\tau$. The vector $(X_1, \ldots, X_n, \epsilon_n)$ is measurable between Ω and and $E^n \times \mathbb{R}_{>0}$, where the latter is equipped with the product σ-algebra $\mathcal{B}(E)^{\otimes n} \otimes \mathcal{B}(\mathbb{R}_{>0})$. Since E is separable, this last σ-algebra is equal to $\mathcal{B}(E^n \times \mathbb{R}_{>0})$. By composition $\psi(X_1, \ldots, X_n, \epsilon_n)$ is Borel measurable and by construction it is a selection from the set ϵ_n-$\mathsf{Quant}(\widehat{\mu}_n)$. ☐

Proof of Theorem 4.26. Let $n \ge 1$ be fixed and define f as

$$\begin{aligned} f\colon & E^n \times E \to \mathbb{R} \\ & (x_1, \ldots, x_n, \alpha) \mapsto \frac{1}{n} \sum_{i=1}^{n} (\|\alpha - x_i\| - \|x_i\|) - \ell(\alpha). \end{aligned}$$

By Theorem 2 (ii) in Brown and Purves [47] there is a universally Borel measurable $\psi : E^n \to E$ such that

$$f\big(x_1, \ldots, x_n, \psi(x_1, \ldots, x_n)\big) = \min_{\alpha \in E} f(x_1, \ldots, x_n, \alpha)$$

for each $x_1, \ldots, x_n$.

For notational convenience we let $Z = (X_1, \ldots, X_n)$ and we show that $\psi(Z)$ is Borel measurable: let $B \in \mathcal{B}(E)$ and note that

$$[\psi(Z)]^{-1}(B) = Z^{-1}(\psi^{-1}(B)).$$

Since ψ is universally measurable, the set $\psi^{-1}(B)$ is universally measurable in $(E^n, \mathcal{B}(E^n))$. Let $\mathbb{P}_Z$ denote the pushforward of $\mathbb{P}$ by the vector Z. This is a measure on the measurable space $(E^n, \mathcal{B}(E)^{\otimes n}) = (E^n, \mathcal{B}(E^n))$ where the equality is due to the separability of E. Consequently $\psi^{-1}(B)$ is in the $\mathbb{P}_Z$-completion of $\mathcal{B}(E^n)$: there exists two measurable sets $A, N \in \mathcal{B}(E^n)$ and a subset $M \subset N$ such that $\mathbb{P}_Z(N) = 0$ and $\psi^{-1}(B) = A \cup M$. We obtain therefore

$$[\psi(Z)]^{-1}(B) = Z^{-1}(A) \cup Z^{-1}(M).$$

The set $Z^{-1}(A)$ is in $\mathcal{F}$ by measurability of Z. Furthermore $Z^{-1}(M) \subset Z^{-1}(N)$ and $\mathbb{P}(Z^{-1}(N)) = \mathbb{P}_Z(N) = 0$. By the completeness assumption on $(\Omega, \mathcal{F}, \mathbb{P})$ the set $Z^{-1}(M)$ is in $\mathcal{F}$, hence so is $[\psi(Z)]^{-1}(B)$. We have proved that $\psi(Z)$ is a Borel measurable selection from $\mathsf{Quant}(\widehat{\mu}_n)$. □

4.8.3 Proofs for Section 4.3.4

Proof of Proposition 4.28. We fix $\mu \in \mathcal{M}_\sim$ and we proceed by contradiction: for each $\delta \in (0, 1]$ there is some affine line L with $\mu(L) > 1 - \delta$.

Let $n \geq 1$. Exploiting the hypothesis twice we find two affine lines L_n, L'_n verifying $\mu(L_n) > 1 - 1/2^{n+1}$ and $\mu(L'_n) > \mu(L_n)$. By the inequality $2(1 - 1/2^{n+1}) > 1$ the lines are neither disjoint, nor are they equal since $\mu(L'_n) \neq \mu(L_n)$. Consequently there exists $x_n \in E$ such that $L_n \cap L'_n = \{x_n\}$, thus

$$\mu(\{x_n\}) = \mu(L'_n) - \mu(L'_n \cap L_n^c) > 1 - 1/2^{n+1} - 1/2^{n+1} = 1 - 1/2^n.$$

Since for $n \geq 2$ we have $(1 - 1/2) + (1 - 1/2^n) > 1$, we obtain $x_n = x_1$ hence

$$\mu(\{x_1\}) > 1 - 1/2^n$$

for each $n \geq 1$, and finally $\mu(\{x_1\}) = 1$. Therefore μ gives mass 1 to any affine line going through x_1, which contradicts $\mu \in \mathcal{M}_\sim$. □

Proof of Proposition 4.29. We show first that the class $\mathcal{C}$ has VC dimension 2. Any set with two elements $\{x, y\} \subset E$ is clearly shattered by $\mathcal{C}$ and any subset with three elements $\{x, y, z\}$ cannot be shattered. Indeed, either w.l.o.g. $y \in [x, z]$ and we consider the labeling $x \mapsto 1$, $y \mapsto 0$, $z \mapsto 1$, or the three points are not on an affine line and we label $x \mapsto 1$, $y \mapsto 1$, $z \mapsto 1$.

Next we define the class of indicator functions $\mathcal{F} = \{\mathbb{1}_C : C \in \mathcal{C}\}$ and we verify that $\mathcal{F}$ is μ-measurable [261, Definition 2.3.3]: we fix some $n \geq 1$, $(e_1, \ldots, e_n) \in \mathbb{R}^n$ and we show that the function

$$\begin{aligned} E^n &\to \mathbb{R} \\ (x_1, \ldots, x_n) &\mapsto \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n e_i f(x_i) \right| \end{aligned}$$

is measurable, where the σ-algebra on E^n is the μ^n-completion of $\mathcal{B}(E)^{\otimes n} = \mathcal{B}(E^n)$.

To this end we let $\pi_i : E^n \to E$ denote the i-th projection map on the first coordinate, we define the diagonal class

$$\mathcal{F}_\Delta = \{(f \circ \pi_1, \ldots, f \circ \pi_n) : f \in \mathcal{F}\},$$

as well as the map γ

$$\begin{aligned}\gamma\colon \mathcal{F}_\Delta \times E^n &\to \mathbb{R}\\ (h_1,\dots,h_n,x_1,\dots,x_n) &\mapsto \sum_{i=1}^n e_i h_i(x_1,\dots,x_n)\end{aligned}$$

and the map T

$$\begin{aligned}T\colon E^2 &\to \mathcal{F}_\Delta\\ (u,v) &\mapsto (\mathbb{1}_{u+\mathbb{R}v}\circ\pi_1,\dots,\mathbb{1}_{u+\mathbb{R}v}\circ\pi_n).\end{aligned}$$

We show next that γ is image admissible Suslin via $(E^2,\mathcal{B}(E^2),T)$ (see [82, Section 5.3] for the definition). Since E is a separable Banach space, E^2 is a Suslin measurable space. T is clearly surjective and it remains to verify that the map

$$\begin{aligned}E^2\times E^n &\to \mathbb{R}\\ ((u,v),(w_1,\dots,w_n)) &\mapsto \sum_{i=1}^n e_i \mathbb{1}_{u+\mathbb{R}v}(w_i)\end{aligned}$$

is $(\mathcal{B}(E^2)\otimes\mathcal{B}(E^n),\mathcal{B}(\mathbb{R}))$-measurable. By composition and the separability of E it suffices more simply to show that ψ defined by

$$\begin{aligned}\psi\colon \quad E^3 &\to \mathbb{R}\\ (u,v,w) &\mapsto \mathbb{1}_{u+\mathbb{R}v}(w)\end{aligned}$$

is Borel measurable, i.e., that $\psi^{-1}(0)\in\mathcal{B}(E^3)$. We let $A=\{(u,v,w): v=0 \text{ and } u\neq w\}$ and $B=\{(u,v,w): u-w \text{ and } v \text{ are linearly independent}\}$, so that

$$\psi^{-1}(\{0\}) = A\cup B.$$

Since $A=\{(u,v,w): v=0\}\cap\{(u,v,w): u\neq w\}$, A is the intersection of a closed set and an open set, and is thus a Borel subset of E^3. It is a standard exercise in topology that $\{(x,y): x \text{ and } y \text{ linearly independent}\}$ is an open subset of E^2 (see, e.g., [97, Problem 5.5]). From this fact it easily follows that B is open in E^3, hence ψ is Borel measurable and γ is image admissible Suslin. Then by Corollary 5.25 in [82] the supremum function

$$\begin{aligned}E^n &\to \mathbb{R}\\ (x_1,\dots,x_n) &\mapsto \sup_{(h_1,\dots,h_n)\in\mathcal{F}_\Delta} |\gamma(h_1,\dots,h_n,x_1,\dots,x_n)|\end{aligned}$$

is universally measurable, hence measurable when E^n is endowed with the μ^n-completion of $\mathcal{B}(E^n)$. By construction, the equality of suprema

$$\sup_{(h_1,\dots,h_n)\in\mathcal{F}_\Delta} |\gamma(h_1,\dots,h_n,x_1,\dots,x_n)| = \sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^n e_i f(x_i)\Big|$$

holds for each $(x_1, \dots, x_n)$, therefore the proof of μ-measurability is complete.

An obvious envelope for $\mathcal{F}$ is the constant function 1. By the discussion closing Chapter 2.4 in [261] and the bound on covering numbers for VC classes [261, Theorem 2.6.4] we have all the ingredients needed to apply the Glivenko–Cantelli theorem [261, Theorem 2.4.3], and the claim is proved. □

Proof of Theorem 4.30. We use the notations of the previous proposition. Since μ is in $\mathcal{M}_\sim$, by Proposition 4.28 there exists $\delta \in (0, 1]$ such that for each affine line L, $\mu(L) \leq 1 - \delta$. By Proposition 4.29 and the definition of convergence outer almost surely (see Definition 4.21), there is a sequence of random variables $(\Delta_n)_{n\geq1}$ such that

$$\sup_{C\in\mathcal{C}} |\widehat{\mu}_n(C) - \mu(C)| \leq \Delta_n \tag{4.20}$$

for each n and $(\Delta_n)_{n\geq1}$ converges $\mathbb{P}$-almost surely to 0. We let $\Omega_0 = \{\omega \in \Omega : \lim_n \Delta_n^\omega = 0\}$ so that $\mathbb{P}(\Omega_0) = 1$, and we fix some $\omega \in \Omega_0$. There exists $N \geq 1$ such that $n \geq N \implies \Delta_n^\omega \leq \delta/2$. For $n \geq N$ and any affine line L, since $L \in \mathcal{C}$ we obtain by (4.20) that $\widehat{\mu}_n^\omega(L) \leq 1 - \delta/2$. The strict convexity of E combined with Proposition 4.17 implies that $\mathsf{Med}(\widehat{\mu}_n^\omega)$ is empty or a singleton whenever $n \geq N$.

Let Ω_1 denote the subset of Ω under consideration in Theorem 4.30. We have proved the inclusion $\Omega_0 \subset \Omega_1$, hence by items 1. and 2. in Lemma 4.85 we have $\mathbb{P}_*(\Omega_1) = 1$. □

4.9 Proofs for Section 4.4

4.9.1 Proofs for Section 4.4.1

Proof of Proposition 4.32. We prove the result for Mosco-convergence, the case of epi-convergence is similar. Let $(x_{n_k})_{k\geq1}$ be a subsequence that converges in the weak topology to some $x \in E$. We define the sequence $(\tilde{x}_n)_{n\geq1}$ as follows: if $n \in \{n_k, n_k + 1, \dots, n_{k+1} - 1\}$ then we let $\tilde{x}_n = x_{n_k}$, and we extend with $\tilde{x}_n = 0$ for $n < n_1$. By construction $(\tilde{x}_n)_{n\geq1}$ converges in the weak topology to x. Thus by Mosco-convergence of (f_n) we have

$$f(x) \leq \liminf_n f_n(\tilde{x}_n) \leq \liminf_k f_{n_k}(\tilde{x}_{n_k}) = \liminf_k f_{n_k}(x_{n_k}) \leq \limsup_k f_{n_k}(x_{n_k}). \tag{4.21}$$

By definition of (x_n), the inequality $f_{n_k}(x_{n_k}) \leq \inf(f_{n_k}) + \varepsilon_{n_k}$ holds for each $k \geq 1$, hence

$$\limsup_k f_{n_k}(x_{n_k}) \leq \limsup_k [\inf(f_{n_k})] \leq \limsup_n [\inf(f_n)]. \tag{4.22}$$

Next, we consider any $z \in E$ and we show $f(x) \leq f(z)$. By Mosco-convergence, there is some sequence $(z_n)_{n\geq1}$ that converges in the norm topology to z and such that $\limsup_n f_n(z_n) \leq f(z)$. Since $\inf(f_n) \leq f_n(z_n)$ for each $n \geq 1$, we obtain

$$\limsup_n [\inf(f_n)] \leq \limsup_n f_n(z_n) \leq f(z). \tag{4.23}$$

Combining inequalities (4.21), (4.22) and (4.23) yields $f(x) \leq f(z)$, hence $x \in \arg\min f$. Note that reflexivity of E is not needed for the claim to hold. □

4.9.2 Proofs for Section 4.4.1

Proof of Proposition 4.36. We begin by showing that the set

$$\{\omega \in \Omega : \forall B \text{ bounded}, \sup_{\alpha\in B} |\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)| \xrightarrow[n\to\infty]{} 0\} \tag{4.24}$$

is in $\mathcal{F}$. For fixed $\omega \in \Omega$, the sequence $(\widehat{\phi}_n^\omega)_{n\geq 1}$ converges uniformly on bounded sets to ϕ if and only if it converges uniformly on closed balls centered at 0 with rational radii, so it suffices to prove for any $r \in \mathbb{Q}_{>0}$ and any $n \geq 1$ that the function

$$A_n : \omega \mapsto \sup_{\alpha\in \bar{B}(0,r)} |\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)|$$

is measurable. Since E is separable, $\bar{B}(0,r)$ has a countable dense subset, say C. For any fixed $(x_1,\dots,x_n) \in E^n$, the function

$$\alpha \mapsto \frac{1}{n}\sum_{i=1}^n (\|\alpha - x_i\| - \|x_i\|) - \ell(\alpha) - \phi(\alpha)$$

is continuous, hence the index set in the supremum can be replaced with C, so that $A_n = \sup_{\alpha\in C} |\widehat{\phi}_n(\alpha) - \phi(\alpha)|$. For each $\alpha \in C$, the random variable $\omega \mapsto |\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)|$ is measurable, hence A_n is measurable as well. Consequently,

$$\bigcap_{r\in\mathbb{Q}_{>0}} \{\omega \in \Omega : \sup_{\alpha\in\bar{B}(0,r)} |\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)| \xrightarrow[n\to\infty]{} 0\} \in \mathcal{F},$$

hence the set (4.24) is in $\mathcal{F}$.

Since E is separable, by Varadarajan's theorem [262] the set $\Omega_0 := \{\omega : \widehat{\mu}_n^\omega \xrightarrow{\text{weakly}} \mu\}$ is in $\mathcal{F}$ and $\mathbb{P}(\Omega_0) = 1$. Fix some $\omega \in \Omega_0$ and let $B \subset E$ be bounded in norm by some $r \geq 0$. We show that $\widehat{\phi}_n^\omega$ converges uniformly to ϕ over B. For each $\alpha \in B$, we let $\varphi_\alpha : E \to \mathbb{R}$, $x \mapsto \|\alpha - x\| - \|x\|$. By the reverse triangle inequality, $\forall x \in E, |\varphi_\alpha(x)| \leq \|\alpha\| \leq r$, so the family $(\varphi_\alpha)_{\alpha\in B}$ is uniformly bounded. Fix some $x_0 \in E$ and note similarly that $|\varphi_\alpha(x) - \varphi_\alpha(x_0)| \leq 2\|x - x_0\|$, hence the family $(\varphi_\alpha)_{\alpha\in B}$ is pointwise equicontinuous. By Theorem 3.1 in [219],

$$\sup_{\alpha\in B}\left| \int_E \varphi_\alpha(x) d\widehat{\mu}_n^\omega(x) - \int_E \varphi_\alpha(x) d\mu(x)\right| \xrightarrow[n\to\infty]{} 0$$

which rewrites as

$$\sup_{\alpha\in B}\left|\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)\right| \xrightarrow[n\to\infty]{} 0.$$

Therefore the event (4.24) contains Ω_0, so it has probability 1. □

Proof of Theorem 4.37. Let Ω_0 be as in the proof of Proposition 4.36 and let $\Omega_1 = \{\omega : \lim_n \epsilon_n^\omega = 0\}$. By Proposition 4.36 and Theorem 6.2.14 in [40] we have the inclusion

$$\Omega_0 \subset \{\omega \in \Omega : \widehat{\phi}_n^\omega \xrightarrow[n\to\infty]{\text{Mosco}} \phi\}.$$

Let Ω_2 denote the subset of Ω considered in the statement of Theorem 4.37. The Proposition 4.32 yields the further inclusion $\Omega_0 \cap \Omega_1 \subset \Omega_2$. Since $\mathbb{P}(\Omega_0) = \mathbb{P}_*(\Omega_1) = 1$ we conclude by items 1., 4. and 2. in Lemma 4.85 that $\mathbb{P}_*(\Omega_2) = 1$. □

Proof of Proposition 4.40. We make use of the machinery developed by Kemperman in [148, Section 2]: he defines the function

$$h : \mathbb{R}_{>0} \to \mathbb{R}, r \mapsto \frac{1}{r}\int_0^r \mu(\{\alpha \in E : \|\alpha\| > u\})du,$$

and he proves that h is nonincreasing, as well as the following lower bound: $\phi_0(\alpha) \geq \|\alpha\|(1 - 2h(\|\alpha\|))$, hence $\phi(\alpha) \geq \|\alpha\|(1 - 2h(\|\alpha\|)) - \ell(\alpha)$ for every $\alpha \in E$. Similarly we define for each $\omega \in \Omega$ and $n \geq 1$ the functions h_n^ω by replacing μ with $\widehat{\mu}_n^\omega$; they are nonincreasing and verify the same lower bound:

$$\begin{aligned}\widehat{\phi}_n^\omega(\alpha) &\geq \|\alpha\|(1 - 2h_n^\omega(\|\alpha\|)) - \ell(\alpha) \\ &\geq \|\alpha\|(1 - \|\ell\|_* - 2h_n^\omega(\|\alpha\|)).\end{aligned} \tag{4.25}$$

For fixed r, n, ω we note that

$$\begin{aligned}h_n^\omega(r) &= \frac{1}{r}\int_0^r \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\|X_i^\omega\|>u}du = \frac{1}{n}\sum_{i=1}^n \left(\tfrac{\|X_i^\omega\|}{r}\mathbb{1}_{\|X_i^\omega\|\leq r} + \mathbb{1}_{\|X_i^\omega\|>r}\right) \\ &= \frac{1}{r}\frac{1}{n}\sum_{i=1}^n \left(\|X_i^\omega\|\mathbb{1}_{\|X_i^\omega\|\leq r}\right) + \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\|X_i^\omega\|>r}.\end{aligned} \tag{4.26}$$

By the dominated convergence theorem we obtain the limits

$$\frac{1}{r}\mathbb{E}[\|X\|\mathbb{1}_{\|X\|\leq r}] \xrightarrow[r\to\infty]{} 0 \quad \text{and} \quad \mathbb{E}[\mathbb{1}_{\|X\|>r}] \xrightarrow[r\to\infty]{} 0,$$

hence we can find some $R > 0$ such that

$$R^{-1}\mathbb{E}[\|X\|\mathbb{1}_{\|X\|\leq R}] + \mathbb{E}[\mathbb{1}_{\|X\|>R}] < (1 - \|\ell\|_*)/4.$$

By (4.26) and the strong law of large numbers, the measurable random variables $\omega \mapsto h_n^\omega(R)$ converge $\mathbb{P}$-almost surely to a constant strictly less than $(1 - \|\ell\|_*)/4$. Note that R depends solely on the distribution of X, i.e., on the measure μ.

We can now turn to the proof of the first item in the proposition. We let Ω_0 (resp., Ω_1) be the event (resp., the set) where the almost-sure convergence of $h_n(R)$ (resp., of ϵ_n) holds and we fix some $\omega \in \Omega_0 \cap \Omega_1$. Since $\omega \in \Omega_0$ (resp., $\omega \in \Omega_1$) the inequality

$$h_n^\omega(R) < (1 - \|\ell\|_*)/4 \quad (\text{resp., } \epsilon_n^\omega \leq R(1 - \|\ell\|_*)/2) \tag{4.27}$$

holds for sufficiently large n. Therefore there exists $N \geq 1$ such that for every $n \geq N$ and $\alpha \in E$ verifying $\|\alpha\| > R$, the following chain of inequalities is true:

$$\begin{aligned}\widehat{\phi}_n^\omega(\alpha) \geq \|\alpha\|(1 - \|\ell\|_* - 2h_n^\omega(\|\alpha\|)) &\geq \|\alpha\|(1 - \|\ell\|_* - 2h_n^\omega(R)) \\ &> R(1 - \|\ell\|_*)/2 \geq \epsilon_n^\omega \geq \inf(\widehat{\phi}_n^\omega) + \epsilon_n^\omega,\end{aligned}$$

where we used successively inequality (4.25), the monotonicity of h_n^ω, inequalities (4.27) and $\widehat{\phi}_n^\omega(0) = 0$. Thus, for each n larger than N the set $\epsilon_n^\omega\text{-}\mathsf{Quant}(\widehat{\mu}_n^\omega)$ is a subset of the closed ball $\bar{B}(0, R)$. Since $\mathbb{P}_*(\Omega_0 \cap \Omega_1) = 1$, the claim follows by item 2. of Lemma 4.85.

The second item of Proposition 4.40 is an immediate consequence of the first.

For the third item the proof is similar. Since $h_n(R)$ converges $\mathbb{P}$-almost surely, it converges in $\mathbb{P}$-probability as well, hence

$$\mathbb{P}\Big(\big|h_n(R) - (R^{-1}\mathbb{E}[\|X_1\|\mathbb{1}_{\|X_1\|\le R}] + \mathbb{E}[\mathbb{1}_{\|X_1\|>R}])\big| \le (1-\|\ell\|_*)/8\Big) \xrightarrow[n\to\infty]{} 1. \quad (4.28)$$

The following inclusions are obtained as above:

$$\Big\{\big|h_n(R) - (R^{-1}\mathbb{E}[\|X_1\|\mathbb{1}_{\|X_1\|\le R}] + \mathbb{E}[\mathbb{1}_{\|X_1\|>R}])\big| \le \tfrac{1-\|\ell\|_*}{8}\Big\} \cap \{\epsilon_n \le \tfrac{R(1-\|\ell\|_*)}{4}\} \quad (4.29)$$
$$\subset \{h_n(R) < 3(1-\|\ell\|_*)/8\} \cap \{\epsilon_n \le R(1-\|\ell\|_*)/4\}$$
$$\subset \{\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) \subset \bar{B}(0,R)\}.$$

By (4.28), by the convergence in outer probability of (ϵ_n) and item 3. of Lemma 4.85, the set (4.29) has $\mathbb{P}_*$-probability converging to 1, and the claim easily follows.

The last item is obtained directly from the third. □

Proof of Theorem 4.42. Let Ω_0 (resp., Ω_1) be the subset of Ω having inner probability 1 in the second item of Proposition 4.40 (resp., in Theorem 4.37) and fix some $\omega \in \Omega_0\cap\Omega_1$. We consider $(\widehat{\alpha}_n^\omega)_{n\ge1}$ a sequence of ϵ_n^ω-empirical ℓ-quantiles and we write $(\widehat{\alpha}_{n_k}^\omega)_{k\ge1}$ an arbitrary subsequence. Since $\omega \in \Omega_0$ there exists $R > 0$ such that all the $\widehat{\alpha}_n^\omega$ lie in the closed ball $\bar{B}(0,R)$. Since E is a reflexive Banach space, $\bar{B}(0,R)$ is weakly compact, i.e., compact in the weak topology of E (see [7, Theorem 6.25]). By the Eberlein–Šmulian theorem [7, Theorem 6.34], $\bar{B}(0,R)$ is weakly sequentially compact. Therefore $(\widehat{\alpha}_{n_k}^\omega)_{k\ge1}$ has a subsequence $(\widehat{\alpha}_{n_{k_j}}^\omega)_{j\ge1}$ that converges in the weak topology to some $\alpha \in E$. Since $(\widehat{\alpha}_{n_{k_j}}^\omega)_{j\ge1}$ is a subsequence of the original sequence $(\widehat{\alpha}_n^\omega)_{n\ge1}$ and since $\omega \in \Omega_1$ we have $\alpha \in \mathsf{Quant}(\mu)$.

Let Ω_2 be the subset of Ω under scrutiny in the statement of Theorem 4.42. We have proved the inclusion $\Omega_0 \cap \Omega_1 \subset \Omega_2$. Since $\mathbb{P}_*(\Omega_0) = \mathbb{P}_*(\Omega_1) = 1$, items 2. and 4. in Lemma 4.85 yield $\mathbb{P}_*(\Omega_2) = 1$. □

Proof of Theorem 4.43. Since $(\epsilon_n)_{n\ge1}$ converges in outer probability to 0, it has a subsequence $(\epsilon_{n_k})_{k\ge1}$ that converges outer almost surely to 0 (see [261, Lemma 1.9.2]). This convergence clearly implies $\mathbb{P}_*$-almost sure convergence to 0: we let $\Omega_0 = \{\omega : \lim_k \epsilon_{n_k}^\omega = 0\}$ so that $\mathbb{P}_*(\Omega_0) = 1$. Additionally we let I denote the set of integers $I = \{n_k : k \ge 1\}$. We define $(e_n)_{n\ge1}$ another sequence of nonnegative random variables as follows:

$$e_n^\omega = \begin{cases} \epsilon_n^\omega & \text{if } \omega \in \Omega_0 \text{ and } n \in I, \\ 0 & \text{otherwise,} \end{cases}$$

so that $(e_n)_{n\ge1}$ converges $\mathbb{P}_*$-almost surely to 0 and $\forall k \ge 1, \forall\omega \in \Omega_0,\ e_{n_k}^\omega = \epsilon_{n_k}^\omega$. We apply the second item of Proposition 4.40 and Theorem 4.37 with the sequence $(e_n)_{n\ge1}$ in lieu of $(\epsilon_n)_{n\ge1}$; let Ω_1 and Ω_2 denote the respective subsets of Ω that have inner probability 1.

Fix some $\omega \in \Omega_0\cap\Omega_1\cap\Omega_2$ and consider $(\widehat{\alpha}_n^\omega)$ a sequence of ϵ_n^ω-empirical ℓ-quantiles. We let $\hat{\beta}_n^\omega$ be defined for each $n \ge 1$ by

$$\hat{\beta}_n^\omega = \begin{cases} \widehat{\alpha}_n^\omega & \text{if } n \in I, \\ \text{any element of } \mathsf{Quant}(\widehat{\mu}_n^\omega) & \text{otherwise,} \end{cases}$$

so that $(\hat{\beta}_n^\omega)_{n\geq 1}$ is a sequence of e_n^ω-empirical ℓ-quantiles. Since $\omega \in \Omega_1$ the sequence $(\hat{\beta}_n^\omega)_{n\geq 1}$ is bounded in norm, hence so is the subsequence $(\hat{\beta}_{n_k}^\omega)_{k\geq 1}$. All these approximate minimizers lie in some closed ball $\bar{B}(0,R)$ with $R > 0$. Since E is reflexive, $\bar{B}(0,R)$ is weakly compact thus $(\hat{\beta}_{n_k}^\omega)_{k\geq 1}$ has a subsequence $(\hat{\beta}_{n_{k_j}}^\omega)_{j\geq 1}$ that converges in the weak topology to some $\alpha \in E$. Since ω is in Ω_2, Theorem 4.37 yields $\alpha \in \mathsf{Quant}(\mu)$.

But by definition $\hat{\beta}_{n_k}^\omega$ coincides with $\widehat{\alpha}_{n_k}^\omega$ for each $k \geq 1$, hence $(\widehat{\alpha}_{n_{k_j}}^\omega)_{j\geq 1}$ converges in the weak topology to $\alpha \in \mathsf{Quant}(\mu)$. Let Ω_3 be the subset of Ω under consideration in the statement of Theorem 4.43. We have proved the inclusion $\Omega_0 \cap \Omega_1 \cap \Omega_2 \subset \Omega_3$. Since $\mathbb{P}_*(\Omega_0) = \mathbb{P}_*(\Omega_1) = \mathbb{P}_*(\Omega_2) = 1$, we obtain $\mathbb{P}_*(\Omega_3) = 1$. □

Proof of Corollary 4.44. Let Ω_0 be the subset of inner probability 1 in the statement of Theorem 4.42. Fix some $\omega \in \Omega_0$ and suppose for the sake of contradiction that there is some $\delta > 0$ and increasing indexes n_k such that for all $k \geq 1$,

$$\epsilon_{n_k}^\omega\text{-}\mathsf{Quant}(\widehat{\mu}_{n_k}^\omega) \not\subset \mathsf{Quant}(\mu)^\delta.$$

Then for each $k \geq 1$ we can find some $\widehat{\alpha}_{n_k}^\omega \in \epsilon_{n_k}^\omega\text{-}\mathsf{Quant}(\widehat{\mu}_{n_k}^\omega) \setminus \mathsf{Quant}(\mu)^\delta$. Since the weak topology coincides with the norm topology in finite dimension, by Theorem 4.42 and taking a subsequence we may assume that $(\widehat{\alpha}_{n_k}^\omega)_k$ converges to some $\alpha \in \mathsf{Quant}(\mu)$. This contradicts $\widehat{\alpha}_{n_k}^\omega \notin \mathsf{Quant}(\mu)^\delta$, hence the inclusion

$$\Omega_0 \subset \{\omega : \forall \delta > 0, \exists N \geq 1, \forall n \geq N, \epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) \subset \mathsf{Quant}(\mu)^\delta\},$$

from which the claim follows. □

4.9.3 Proofs for Section 4.4.2

The following lemma gives a useful criterion for convergence of sequences in topological spaces.

Lemma 4.87. *Let $(G, \mathcal{T})$ be a topological space, $(x_n)_{n\geq 1}$ be a sequence in G and $x \in G$. If any subsequence $(x_{n_k})_{k\geq 1}$ has a further subsequence $(x_{n_{k_j}})_{j\geq 1}$ such that $\lim_j x_{n_{k_j}} = x$, then the sequence $(x_n)_{n\geq 1}$ converges to x.*

Proof of Lemma 4.87. Assume for the sake of contradiction that $(x_n)_{n\geq 1}$ does not converge to x. Then there exists a neighborhood U of x such that the set $\{n : x_n \notin U\}$ is infinite. Consequently we can find a subsequence $(x_{n_k})_{k\geq 1}$ with $x_{n_k} \notin U$ for each $k \geq 1$. By assumption, there is a further subsequence $(x_{n_{k_j}})_{j\geq 1}$ that converges to x: in particular there exists $j \geq 1$ with $x_{n_{k_j}} \in U$. This is a contradiction. □

Proof of Theorem 4.45. Let Ω_0 denote the subset of Ω that has inner probability 1 in Theorem 4.42. We fix $\omega \in \Omega_0$, we let $(\widehat{\alpha}_n^\omega)_{n\geq 1}$ be a sequence of ϵ_n^ω-empirical ℓ-quantiles and we consider an arbitrary subsequence $(\widehat{\alpha}_{n_k}^\omega)_{k\geq 1}$. By Theorem 4.42 there exists a further subsequence $(\widehat{\alpha}_{n_{k_j}}^\omega)_{j\geq 1}$ that converges weakly to some $\alpha \in \mathsf{Quant}(\mu)$. Because of Assumption (A7), $\mathsf{Quant}(\mu) = \{\alpha_\star\}$ hence $\alpha = \alpha_\star$. By Lemma 4.87 the sequence $(\widehat{\alpha}_n^\omega)_{n\geq 1}$ converges in the weak topology of E to $\alpha_\star$.

Let Ω_1 denote the subset of Ω considered in the statement of Theorem 4.45. We have established the inclusion $\Omega_0 \subset \Omega_1$. Since $\mathbb{P}_*(\Omega_0) = 1$, the proof is complete. □

Remark 4.88. Gervini [101, Proof of Theorem 2] argues that $L^2(T)$ equipped with its weak topology is locally compact. With this topology, $L^2(T)$ is a Hausdorff topological vector space. If such a space is locally compact, then it is finite-dimensional [7, Theorem 5.26]. Since $L^2(T)$ is infinite-dimensional, it is not locally compact when equipped with its weak topology.

Remark 4.89. By the first item in Proposition 4.40, as far as convergence is concerned we can assume w.l.o.g. that any sequence of ϵ_n-empirical ℓ-quantiles is contained in the closed ball $\bar{B}(0, R)$. This set is compact in the weak topology of E [7, Theorem 6.25]. Since E is reflexive and separable, we have the separability of E^{**}, thus E^* is separable as well [159, Theorem 4.6.8], hence $\bar{B}(0, R)$ is weakly metrizable [7, Theorem 6.31]. Let $\mathcal{T}$ denote the relative topology on $\bar{B}(0, R)$ induced by the weak topology of E (which we denote by $\sigma(E, E^*)$). We have verified that $(\bar{B}(0, R), \mathcal{T})$ is a compact metrizable space.

Geometric quantiles fit the M-estimation framework developed in [132]. In Huber's notation we let $\Theta = \mathfrak{X} = \bar{B}(0, R)$, $\rho(x, \theta) = \|\theta - x\|$, $a(x) = \|x\|$, $b(\theta) = \|\theta\| + 1$, $h(x) = -(1 + \|\ell\|_*)$. We equip Θ with the topology $\mathcal{T}$, so that we have a compact metrizable space, and this matches the topological setting considered by Huber.

Another technical detail that warrants verification is Assumption (A-2) in [132]. For each x, it requires lower semicontinuity of the function $\theta \mapsto \rho(x, \theta) = \|\theta - x\| - \ell(\theta)$ defined on $(\Theta, \mathcal{T})$. This follows from the lower semicontinuity of the norm as a function on the topological space $(E, \sigma(E, E^*))$ [7, Lemma 6.22].

4.9.4 Proofs for Section 4.4.2

Lemma 4.90. *ϕ has a well-separated minimizer if and only if ϕ is well-posed.*

Proof of Lemma 4.90. $\implies$ Let $\alpha_\star \in \arg\min \phi$ be well-separated. By the strict inequality in Definition 4.49 there cannot be another minimizer of ϕ. We consider $(\alpha_n)_{n\geq 1}$ a minimizing sequence and $\epsilon > 0$. Let $\eta = \inf_{\substack{\alpha \in E \\ \|\alpha - \alpha_\star\| \geq \epsilon}} \phi(\alpha) - \phi(\alpha_\star)$, which is positive by definition. Since $(\alpha_n)_{n\geq 1}$ is minimizing, for n large enough we obtain $\phi(\alpha_n) < \phi(\alpha_\star) + \eta$, i.e.,

$$\phi(\alpha_n) < \inf_{\substack{\alpha \in E \\ \|\alpha - \alpha_\star\| \geq \epsilon}} \phi(\alpha)$$

hence $\|\alpha_n - \alpha_\star\| < \epsilon$.

$\impliedby$ Let $\alpha_\star$ denote the minimizer of ϕ. We show that it is well-separated. Assume for the sake of contradiction that there is some $\epsilon_0 > 0$ such that $\phi(\alpha_\star) \geq \inf_{\substack{\alpha \in E \\ \|\alpha - \alpha_\star\| \geq \epsilon_0}} \phi(\alpha)$. Since $\alpha_\star$ is a minimizer of ϕ, this infimum is actually equal to $\phi(\alpha_\star)$. By an elementary property of infima there is some sequence $(\alpha_n)_{n\geq 1}$ such that $\forall n \geq 1$, $\|\alpha_n - \alpha_\star\| \geq \epsilon_0$ and $\phi(\alpha_n) \xrightarrow[n\to\infty]{} \phi(\alpha_\star)$. Hence $(\alpha_n)_{n\geq 1}$ is a minimizing sequence that does not converge in the norm topology to $\alpha_\star$. □

Proof of Proposition 4.51. 1. The assumptions of the Proposition ensure that ϕ has a unique minimizer $\alpha_\star$. By Lemma 4.90 it suffices to consider a minimizing sequence $(\alpha_n)_{n\geq 1}$ and prove that $\|\alpha_n - \alpha_\star\| \xrightarrow[n\to\infty]{} 0$.

We show first that $(\alpha_n)_{n\geq 1}$ converges in the weak topology of E to $\alpha_\star$. For this purpose we make use of Lemma 4.87: let $(\alpha_{n_k})_{k\geq 1}$ be an arbitrary subsequence. Since ϕ is coercive, $(\alpha_{n_k})_{k\geq 1}$ is bounded. E is reflexive, so the same weak compactness argument as in the proof of Theorem 4.42 yields a subsequence $(\alpha_{n_{k_j}})_{j\geq 1}$ that converges in the weak topology of E to some $\alpha_0 \in E$. It remains to prove that $\alpha_0 = \alpha_\star$. Fix some $x \in E$ and observe that the sequence $(\alpha_{n_{k_j}} - x)_{j\geq 1}$ converges weakly to $\alpha_0 - x$. Since the norm is weakly lower semicontinuous (see [7, Lemma 6.22]),

$$\|\alpha_0 - x\| - \|x\| \leq \liminf_j (\|\alpha_{n_{k_j}} - x\| - \|x\|).$$

Integrating w.r.t. x yields

$$\begin{aligned}
\phi(\alpha_0) &\leq \int_E \liminf_j (\|\alpha_{n_{k_j}} - x\| - \|x\|) d\mu(x) - \ell(\alpha_0) \\
&\overset{(i)}{\leq} \liminf_j \int_E (\|\alpha_{n_{k_j}} - x\| - \|x\|) d\mu(x) - \ell(\alpha_0) \\
&\overset{(ii)}{=} \liminf_j \left[\phi_0(\alpha_{n_{k_j}})\right] - \lim_j \ell(\alpha_{n_{k_j}}) \overset{(iii)}{=} \liminf_j \phi(\alpha_{n_{k_j}}) \overset{(iv)}{=} \phi(\alpha_\star).
\end{aligned}$$

Inequality (i) stems from Fatou's lemma for functions with an integrable lower bound, in equality (ii) we exploit the weak convergence of $\alpha_{n_{k_j}}$, equality (iii) is justified by a standard property of lim inf, and (iv) holds because $(\alpha_n)_{n\geq 1}$ is a minimizing sequence. The freshly derived inequality $\phi(\alpha_0) \leq \phi(\alpha_\star)$ combined with $\mathsf{Quant}(\mu) = \{\alpha_\star\}$ implies $\alpha_0 = \alpha_\star$ and we can conclude that

$$\alpha_n \xrightarrow[n\to\infty]{\text{weakly}} \alpha_\star. \tag{4.30}$$

We show next that the sequence of norms $(\|\alpha_n\|)_{n\geq 1}$ converges to $\|\alpha_\star\|$. Since it is a bounded sequence of real numbers, it suffices to show that it has a unique subsequential limit: we consider a subsequence $(\|\alpha_{n_k}\|)_{k\geq 1}$ that converges to some $R \geq 0$ and we prove that $R = \|\alpha_\star\|$. By (4.30), for each $x \in E$ the sequence $(\alpha_{n_k} - x)_{k\geq 1}$ converges weakly to $\alpha_\star - x$ and by weak lower semicontinuity of the norm,

$$\|\alpha_\star - x\| - \|x\| \leq \liminf_k (\|\alpha_{n_k} - x\| - \|x\|). \tag{4.31}$$

Integrating w.r.t. x we obtain as above

$$\begin{aligned}
\phi(\alpha_\star) &\leq \int_E \liminf_k (\|\alpha_{n_k} - x\| - \|x\|) d\mu(x) - \ell(\alpha_\star) \qquad (4.32)\\
&\leq \liminf_k \int_E (\|\alpha_{n_k} - x\| - \|x\|) d\mu(x) - \ell(\alpha_\star) \\
&= \liminf_k \phi(\alpha_{n_k}) \\
&= \phi(\alpha_\star).
\end{aligned}$$

The inequality (4.32) is therefore an equality, i.e., the function

$$x \mapsto \liminf_k (\|\alpha_{n_k} - x\| - \|x\|) - (\|\alpha_\star - x\| - \|x\|)$$

is nonnegative by (4.31) and has integral 0 by (4.32). Consequently, it is 0 μ-almost everywhere and there exists some $x_0 \in E$ such that $\liminf_k(\|\alpha_{n_k} - x_0\| - \|x_0\|) = \|\alpha_\star - x_0\| - \|x_0\|$, i.e., $\liminf_k \|\alpha_{n_k} - x_0\| = \|\alpha_\star - x_0\|$. We can then find some further subsequence $(\alpha_{n_{k_j}})_{j\geq 1}$ such that

$$\|\alpha_{n_{k_j}} - x_0\| \xrightarrow[j\to\infty]{} \|\alpha_\star - x_0\|.$$

But $(\alpha_{n_{k_j}} - x_0)_{j\geq 1}$ converges weakly to $\alpha_\star - x_0$, so by the Radon–Riesz property it converges in the norm topology to $\alpha_\star - x_0$, hence so does $(\alpha_{n_{k_j}})_{j\geq 1}$ to $\alpha_\star$, and in particular $\|\alpha_{n_{k_j}}\| \to_j \|\alpha_\star\|$, thus $R = \|\alpha_\star\|$ from which we obtain the convergence

$$\|\alpha_n\| \xrightarrow[n\to\infty]{} \|\alpha_\star\|. \tag{4.33}$$

By (4.30), (4.33) and the Radon–Riesz property, the first item of the Proposition is proved.

2. Let $(\alpha_n)_{n\geq 1}$ be a minimizing sequence such that $\alpha_n \in L$ for each $n \geq 1$. By Assumption (A8) and Proposition 4.19, we have $\mathsf{Med}(\mu) = \{\alpha_\star\}$ and $\alpha_\star \in L$. As seen in the proof of Proposition 4.19, we can assume w.l.o.g. that L goes through the origin: $L = \mathbb{R}v$ with $\|v\| = 1$. Using the same notations, we introduce the pushforward measure ν on $\mathbb{R}$ so that $\mathsf{Med}(\nu) = \{f(\alpha_\star)\}$. Since $\forall \alpha \in L$, $\phi_\mu(\alpha) = \phi_\nu(f(\alpha))$, we obtain that $(f(\alpha_n))_{n\geq 1}$ is a minimizing sequence for ϕ_ν. By the first item of Proposition 4.51, ϕ_ν is well-posed hence $(f(\alpha_n))$ converges to $f(\alpha_\star)$. This implies the convergence of (α_n) to $\alpha_\star$ in the norm topology. □

Proof of Theorem 4.54. Let $\Omega_0 = \{\omega : \lim_n \epsilon_n^\omega = 0\}$, let Ω_1 be the set having inner probability 1 in the second item of Proposition 4.40 and let Ω_2 be the almost-sure event from Proposition 4.36. Fix some $\omega \in \Omega_0 \cap \Omega_1 \cap \Omega_2$ and consider $(\widehat{\alpha}_n^\omega)_{n\geq 1}$ a sequence of ϵ_n^ω-empirical medians. In view of Proposition 4.51 it suffices to prove that $(\widehat{\alpha}_n^\omega)_{n\geq 1}$ is a minimizing sequence, i.e., $\phi(\widehat{\alpha}_n^\omega) \xrightarrow[n\to\infty]{} \phi(\alpha_\star)$.

Since $\omega \in \Omega_1$ there is some $\rho > 0$ such that $\alpha_\star$ and all the $\widehat{\alpha}_n^\omega$ lie in the closed ball $\bar{B}(0,\rho)$. Note that

$$\begin{aligned} 0 \leq \phi(\widehat{\alpha}_n^\omega) - \phi(\alpha_\star) &= \phi(\widehat{\alpha}_n^\omega) - \widehat{\phi}_n^\omega(\widehat{\alpha}_n^\omega) + \widehat{\phi}_n^\omega(\widehat{\alpha}_n^\omega) - \phi(\alpha_\star) \\ &\leq \sup_{\alpha\in\bar{B}(0,\rho)} \left(|\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)|\right) + \widehat{\phi}_n^\omega(\widehat{\alpha}_n^\omega) - \phi(\alpha_\star). \end{aligned} \tag{4.34}$$

Additionally we have the upper bound

$$\begin{aligned} \widehat{\phi}_n^\omega(\widehat{\alpha}_n^\omega) - \phi(\alpha_\star) &= (\widehat{\phi}_n^\omega(\widehat{\alpha}_n^\omega) - \inf(\widehat{\phi}_n^\omega)) + (\inf(\widehat{\phi}_n^\omega) - \widehat{\phi}_n^\omega(\alpha_\star)) + (\widehat{\phi}_n^\omega(\alpha_\star) - \phi(\alpha_\star)) \\ &\leq \epsilon_n^\omega + \sup_{\alpha\in\bar{B}(0,\rho)} |\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)|. \end{aligned} \tag{4.35}$$

Plugging this in (4.34) yields

$$0 \leq \phi(\widehat{\alpha}_n^\omega) - \phi(\alpha_\star) \leq \epsilon_n^\omega + 2 \sup_{\alpha\in\bar{B}(0,\rho)} |\widehat{\phi}_n^\omega(\alpha) - \phi(\alpha)|.$$

Since $\omega \in \Omega_0 \cap \Omega_2$ the right hand side of the last display converges to 0, hence $(\widehat{\alpha}_n^\omega)_{n\geq 1}$ is a minimizing sequence. We have thus obtained the inclusion

$$\Omega_0 \cap \Omega_1 \cap \Omega_2 \subset \{\omega : \|\widehat{\alpha}_n^\omega - \alpha_\star\| \to_n 0\}$$

and we conclude using Lemma 4.85. □

Proof of Theorem 4.55. Let $(\widehat{\alpha}_n)_{n\geq 1}$ be a sequence of ϵ_n-empirical quantiles. Fix some $\epsilon > 0$ for the remainder of the proof and let

$$\eta = \inf_{\substack{\alpha\in E \\ \|\alpha-\alpha_\star\|\geq\epsilon}} \phi(\alpha) - \phi(\alpha_\star),$$

so that $\eta > 0$ by Proposition 4.51 and Lemma 4.90. By definition of η

$$\forall \alpha \in E, \|\alpha - \alpha_\star\| \geq \epsilon \implies \phi(\alpha) - \phi(\alpha_\star) \geq \eta,$$

and using the same algebraic manipulations and upper bounds as in (4.34) and (4.35), we obtain the inclusions of sets valid for each $n \geq 1$:

$$\begin{aligned} \{\|\widehat{\alpha}_n - \alpha_\star\| \geq \epsilon\} &\subset \{\phi(\widehat{\alpha}_n) - \phi(\alpha_\star) \geq \eta\} \\ &\subset \{\phi(\widehat{\alpha}_n) - \widehat{\phi}_n(\widehat{\alpha}_n) + \epsilon_n + \widehat{\phi}_n(\alpha_\star) - \phi(\alpha_\star) \geq \eta\} \\ &\subset \{\phi(\widehat{\alpha}_n) - \widehat{\phi}_n(\widehat{\alpha}_n) \geq \eta/4\} \cup \{\widehat{\phi}_n(\alpha_\star) - \phi(\alpha_\star) \geq \eta/4\} \cup \{\epsilon_n \geq \eta/2\}. \end{aligned} \tag{4.36}$$

To finish the proof, it suffices by Lemma 4.85 to show that each of the three sets in (4.36) has outer probability converging to 0. Let R be as in Item 3. of Proposition 4.40. By Proposition 4.36 and its proof the (measurable) random variables $\sup_{\alpha\in\bar{B}(0,R)} |\widehat{\phi}_n(\alpha) - \phi(\alpha)|$ converge $\mathbb{P}$-almost surely to 0, hence in probability as well. Combining this with the inclusion

$$\{\phi(\widehat{\alpha}_n)-\widehat{\phi}_n(\widehat{\alpha}_n) \geq \eta/4\}\cap\{\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) \subset \bar{B}(0,R)\} \subset \Big\{ \sup_{\alpha\in\bar{B}(0,R)} |\widehat{\phi}_n(\alpha)-\phi(\alpha)| \geq \eta/4\Big\}$$

yields the convergence

$$\mathbb{P}^*\big(\{\phi(\widehat{\alpha}_n) - \widehat{\phi}_n(\widehat{\alpha}_n) \geq \eta/4\} \cap \{\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) \subset \bar{B}(0,R)\}\big) \xrightarrow[n\to\infty]{} 0.$$

Since $\mathbb{P}_*(\{\epsilon_n\text{-}\mathsf{Quant}(\widehat{\mu}_n) \subset \bar{B}(0,R)\}) \to_n 1$ the fourth item of Lemma 4.85 yields the wanted convergence

$$\mathbb{P}^*(\phi(\widehat{\alpha}_n) - \widehat{\phi}_n(\widehat{\alpha}_n) \geq \eta/4) \xrightarrow[n\to\infty]{} 0.$$

In a similar fashion we show

$$\mathbb{P}^*(\widehat{\phi}_n(\alpha_\star) - \phi(\alpha_\star) \geq \eta/4) \xrightarrow[n\to\infty]{} 0.$$

The convergence of $(\epsilon_n)_n$ in outer probability to 0 finishes the proof. □

Proof of Proposition 4.57. Let $\Omega_0, \Omega_1, \Omega_2$ be as in the proof of Theorem 4.54. Let L denote an affine line such that $\mu(L) = 1$ and define the event $\Omega_3 = \bigcap_{n\geq 1}\{X_n \in L\}$, so that $\mathbb{P}(\Omega_3) = 1$. We fix some $\omega \in \Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3$ and we consider $(\widehat{\alpha}_n^\omega)_{n\geq 1}$ a sequence of 0-empirical medians. Since $\omega \in \Omega_3$ we have $\widehat{\mu}_n^\omega(L) = 1$ for each $n \geq 1$, hence by Proposition 4.19 the empirical median $\widehat{\alpha}_n^\omega$ lies on L. We can then apply the second item of Proposition 4.51 and finish the proof as for Theorem 4.54.

The adaptation of Theorem 4.55 is similar and therefore omitted. □

Proof of Corollary 4.58. Uniformly convex Banach spaces are reflexive [186, Theorem 5.2.15], strictly convex [55, Proposition 5.2.6], and they enjoy the Radon–Riesz property [186, Theorem 5.2.18]. Uniform convexity of each space in the list was established in Corollary 4.18, it suffices to check separability.

Every finite-dimensional space is separable. Separability conditions for L^p spaces are taken from Corollary 4.13. $W^{k,p}(\Omega)$ is separable [1, Theorem 3.5]. $S_p(H)$ is separable because H is separable [68, Theorem 18.14 (c)]. □

4.10 Proofs for Section 4.5

4.10.1 Proofs for Section 4.5.1

Proof of Lemma 4.62. Given $\lambda \in \mathbb{R}_{\geq 0}$, if we replace (α, h) with $(\lambda\alpha, \lambda h)$ then both sides of the first inequality are multiplied by λ and the second inequality is left unchanged. We can thus assume w.l.o.g. that $\|\alpha\| = 1$.

If α and h are linearly dependent, i.e., $h = \lambda\alpha$ for some real λ then the inequalities rewrite as

$$\big||1+\lambda| - 1 - \lambda\big| \leq \frac{1}{2}(\lambda^2 \wedge |\lambda|^3) \quad \text{and} \quad \left|\frac{1+\lambda}{|1+\lambda|} - 1\right| \leq 2(|\lambda| \wedge \lambda^2) \tag{4.37}$$

where $\lambda \neq -1$ in the rightmost one. The validity of (4.37) is easily checked by elementary calculus.

We can therefore assume that α and h are linearly independent and we can find some $\beta \in E$ such that $\{\alpha, \beta\}$ forms an orthonormal basis of $\operatorname{span}(\{\alpha, h\})$. If we write $h = a\alpha + b\beta$ for some $(a, b) \in \mathbb{R}^2$, the inequalities of Lemma 4.62 rewrite as

$$\left|\big((1+a)^2 + b^2\big)^{1/2} - 1 - a - \frac{b^2}{2}\right| \leq \frac{1}{2}\min\big(a^2+b^2, (a^2+b^2)^{3/2}\big),$$

$$\left[\left(\frac{1+a}{((1+a)^2+b^2)^{1/2}} - 1\right)^2 + \left(\frac{1}{((1+a)^2+b^2)^{1/2}} - 1\right)^2 b^2\right]^{1/2} \leq 2\min\big((a^2+b^2)^{1/2}, a^2+b^2\big).$$

These last two inequalities can be proved using polar coordinates; this is rather tedious and therefore omitted. □

Proof of Proposition 4.64. For $x \in E$ we let φ_x denote the function $\alpha \mapsto \|\alpha - x\| - \|x\|$. The subdifferential of the norm of the Hilbert space E is given by

$$\partial N(\alpha) = \begin{cases} \{\alpha/\|\alpha\|\} & \text{if } \alpha \neq 0, \\ \{\beta : \|\beta\| \leq 1\} & \text{if } \alpha = 0. \end{cases} \tag{4.38}$$

Let $\alpha, h \in E$ be fixed. By (4.38) the vector $\mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}$ is in the subdifferential of the convex function φ_x at α, hence the following inequalities hold:

$$\varphi_x(\alpha+h) \geq \varphi_x(\alpha) + \langle \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}, h\rangle,$$
$$\varphi_x(\alpha) \geq \varphi_x(\alpha+h) + \langle \mathbb{1}_{x\neq\alpha+h}\frac{\alpha+h-x}{\|\alpha+h-x\|}, -h\rangle \text{ for every } h \in E.$$

Summing yields

$$0 \leq \varphi_x(\alpha+h) - \varphi_x(\alpha) - \langle \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}, h\rangle \leq \langle \mathbb{1}_{x\neq\alpha+h}\frac{\alpha+h-x}{\|\alpha+h-x\|} - \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}, h\rangle.$$

To transform the last line into one involving the function ϕ_0 it suffices to integrate with respect to x, e.g., $\int_E \varphi_x(\alpha+h)d\mu(x) = \phi_0(\alpha+h)$. The gradient of ϕ_0 appears if we can justify the equality

$$\int_E \langle \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}, h\rangle d\mu(x) = \langle \int_E \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|} d\mu(x), h\rangle \ , \tag{4.39}$$

where an E-valued function is integrated in the right-hand side. To make sense of such an integral we employ the theory of Bochner integration [75, Section II.2]. We let $f : E \to E$ be the function $x \mapsto \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}$: f is Borel measurable, with separable range since E is assumed separable. By Pettis's measurability theorem, f is μ-measurable (see [75, Section II.1]). Additionally, $\|f\|$ is integrable in the usual sense, thus f is Bochner integrable. With T the bounded operator $T : u \mapsto \langle u, h\rangle$, a standard property of Bochner integration yields

$$\int_E (T \circ f)(x) d\mu(x) = T\Big(\int_E f(x) d\mu(x)\Big),$$

which is exactly Equation (4.39). Replacing x with $X(\omega)$ and integrating the functions

$$\omega \mapsto \mathbb{1}_{X(\omega)\neq\alpha}\frac{\alpha-X(\omega)}{\|\alpha-X(\omega)\|}, \quad \omega \mapsto \mathbb{1}_{X(\omega)\neq\alpha+h}\frac{\alpha+h-X(\omega)}{\|\alpha+h-X(\omega)\|}$$

in the Bochner sense (which is licit by the same arguments as above), we obtain

$$0 \leq \phi_0(\alpha+h) - \phi_0(\alpha) - \langle \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{\alpha-X}{\|\alpha-X\|}\Big], h\rangle \leq \langle \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha+h}\frac{\alpha+h-X}{\|\alpha+h-X\|}\Big] - \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{\alpha-X}{\|\alpha-X\|}\Big], h\rangle \tag{4.40}$$

where the expectations denote Bochner integrals. We are now ready to prove each item of Proposition 4.64.

1. We assume that α is not an atom of μ, i.e., $\mu(\{\alpha\}) = \mathbb{P}(X = \alpha) = 0$. To establish Fréchet differentiability of ϕ_0 at α, it suffices by (4.40) to show that

$$\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha+h}\frac{\alpha+h-X}{\|\alpha+h-X\|}\Big] - \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{\alpha-X}{\|\alpha-X\|}\Big] \xrightarrow[\|h\|\to 0]{} 0. \tag{4.41}$$

We consider any sequence $(h_n)_{n\geq 1}$ such that $\|h_n\| \to 0$ and we note that the following convergence holds in the norm topology of E, for each x in $E\setminus\{\alpha\}$ (hence for μ-almost every x by the initial assumption):

$$\mathbb{1}_{x\neq\alpha+h_n}\frac{\alpha+h_n-x}{\|\alpha+h_n-x\|} \xrightarrow[n\to\infty]{} \mathbb{1}_{x\neq\alpha}\frac{\alpha-x}{\|\alpha-x\|}.$$

By the dominated convergence theorem for Bochner integrals we obtain (4.41) and we can conclude that ϕ_0 is differentiable at α with gradient

$$\nabla\phi_0(\alpha) = \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{\alpha-X}{\|\alpha-X\|}\Big],$$

hence so is ϕ with gradient $\nabla\phi(\alpha) = \nabla\phi_0(\alpha) - \ell$.

Conversely, we assume that ϕ is Fréchet differentiable at $\alpha_0 \in E$. In that case, ϕ_0 is also differentiable at α_0. If $\mu(\{\alpha_0\}) = 1$, then μ is the Dirac measure δ_{α_0} and ϕ_0 is simply the function $\alpha \mapsto \|\alpha-\alpha_0\| - \|\alpha_0\|$, which is not differentiable at α_0, hence we must have $\mu(\{\alpha_0\}) < 1$. Assume for the sake of contradiction that $\mu(\{\alpha_0\}) > 0$ and define the probability measure $\nu = (1-\mu(\{\alpha_0\}))^{-1}(\mu - \mu(\{\alpha_0\})\delta_{\alpha_0})$, as well as the corresponding objective function

$$\phi_{0,\nu} : \alpha \mapsto \frac{1}{1-\mu(\{\alpha_0\})}\phi_0(\alpha) - \frac{\mu(\{\alpha_0\})}{1-\mu(\{\alpha_0\})}(\|\alpha-\alpha_0\| - \|\alpha_0\|).$$

By construction $\nu(\{\alpha_0\}) = 0$ hence $\phi_{0,\nu}$ is differentiable at α_0. Since additionally ϕ_0 is Fréchet differentiable at α_0 and $\mu(\{\alpha_0\}) > 0$, we obtain by subtracting and scaling that the function $\alpha \mapsto \|\alpha-\alpha_0\|$ is Fréchet differentiable at α_0, which is absurd. Therefore $\mu(\{\alpha_0\}) = 0$.

2. We assume that $\mathbb{E}[\|X-\alpha\|^{-1}] < \infty$. and we define the function

$$\begin{aligned} g\colon E &\to B(E)\\ x &\mapsto \mathbb{1}_{x\neq\alpha}\nabla^2 N(\alpha-x) = \mathbb{1}_{x\neq\alpha}\frac{1}{\|\alpha-x\|}\Big(\mathrm{Id} - \frac{(\alpha-x)\otimes(\alpha-x)}{\|\alpha-x\|^2}\Big)\end{aligned}$$

where $B(E)$ is the Banach space of bounded operators on E equipped with the operator norm $\|\cdot\|_{op}$. We check next that g is indeed Bochner integrable by making use of the decomposition $g = g_1 + g_2$ where

$$g_1 : x \mapsto \mathbb{1}_{x\neq\alpha}\frac{1}{\|\alpha-x\|}\,\mathrm{Id} \quad\text{ and }\quad g_2 : x \mapsto \mathbb{1}_{x\neq\alpha}\frac{(\alpha-x)\otimes(\alpha-x)}{\|\alpha-x\|^3}.$$

g_1 is Borel measurable and range(g_1) is a subset of the line spanned by Id, hence range(g_1) is a separable subset of $B(E)$. By Pettis's measurability theorem, g_1 is μ-measurable. Moreover $\|g_1\|_{op}$ is integrable in the usual sense, hence g_1 is Bochner integrable. The function $h : E \to B(E), z \mapsto z\otimes z$ is continuous by the straightforward estimate

$$\|h(z) - h(z_0)\|_{op} \leq (\|z\| + \|z_0\|)\|z-z_0\| \text{ for every } (z, z_0) \in E^2,$$

and the Borel measurability of g_2 easily follows. Since $g_2(x)$ has rank at most 1 for each x, the function g_2 takes values in $S_2(E)$, the Hilbert space of Hilbert–Schmidt operators on E. Since E is separable, $S_2(E)$ is separable w.r.t. the Hilbert–Schmidt norm $\|\cdot\|_2$ [68, Theorem 18.14 (c)]. This norm verifies [189, Corollary 16.9]

$$\|A\|_{op} \leq \|A\|_2 \quad \text{for any } A \in S_2(E), \tag{4.42}$$

hence $S_2(E)$ is a separable subset of $B(E)$, and so is range(g_2). We have $\|h(z)\|_{op} = \|z\|^2$ thus $\|g_2(x)\|_{op} = \mathbb{1}_{x\neq\alpha}\frac{1}{\|\alpha-x\|}$ and $\|g_2\|_{op}$ is Lebesgue integrable, hence g_2 is Bochner integrable and so is g. Replacing x with $X(\omega)$ and repeating the same arguments we find that

$$\omega \mapsto \mathbb{1}_{X(\omega)\neq\alpha}\nabla^2 N(\alpha - X(\omega))$$

is Bochner integrable and it follows that H is well-defined and $H \in B(E)$. Since Bochner integrals and bounded operators commute, we have for every $(h_1, h_2) \in E^2$ that

$$\begin{aligned}\langle Hh_1, h_2\rangle &= \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha - X\|}\Big(\langle h_1, h_2\rangle - \frac{\langle h_1, \alpha - X\rangle\langle h_2, \alpha - X\rangle}{\|\alpha - X\|^2}\Big)\Big] \\ &= \langle h_1, Hh_2\rangle,\end{aligned}$$

and

$$\langle Hh_1, h_1\rangle = \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha - X\|}\Big(\|h_1\|^2 - \frac{\langle h_1, \alpha - X\rangle^2}{\|\alpha - X\|^2}\Big)\Big].$$

By Cauchy–Schwarz inequality, $\mathbb{1}_{X\neq\alpha}(\|h_1\|^2 - \frac{\langle h_1,\alpha-X\rangle^2}{\|\alpha-X\|^2}) \geq 0$, hence $\langle Hh_1, h_1\rangle \geq 0$. Regarding the Taylor expansion,

$$\begin{aligned}&\phi(\alpha + h) - \phi(\alpha) - \langle\nabla\phi(\alpha), h\rangle - \frac{1}{2}\langle Hh, h\rangle \\ &= \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\big(\|\alpha + h - X\| - \|\alpha - X\| - \langle\nabla N(\alpha - X), h\rangle - \frac{1}{2}\langle\nabla^2 N(\alpha - X)h, h\rangle\big)\Big] + \mathbb{E}[\mathbb{1}_{X=\alpha}\|h\|] \\ &\leq \mathbb{E}\Big[\frac{\|h\|^2}{\|\alpha - X\|} \wedge \frac{\|h\|^3}{\|\alpha - X\|^2}\Big] \\ &= \|h\|^2\Big(\mathbb{E}\Big[\mathbb{1}_{\|X-\alpha\|\leq\|h\|}\frac{1}{\|X - \alpha\|}\Big] + \mathbb{E}\Big[\mathbb{1}_{\|X-\alpha\|>\|h\|}\frac{\|h\|}{\|X - \alpha\|^2}\Big]\Big),\end{aligned}$$

where the inequality stems from Lemma 4.62 and $\mathbb{P}(X = \alpha) = 0$. To finish the proof it suffices to show that each expectation in the last line converges to 0 as h goes to 0. This follows from the dominated convergence theorem; for the second expectation, we have the domination

$$\begin{aligned}\mathbb{1}_{\|X-\alpha\|>\|h\|}\frac{\|h\|}{\|X - \alpha\|^2} = \frac{1}{\|h\|}\mathbb{1}_{\|X-\alpha\|>\|h\|}\frac{\|h\|^2}{\|X - \alpha\|^2} &\leq \frac{1}{\|h\|}\mathbb{1}_{\|X-\alpha\|>\|h\|}\frac{\|h\|}{\|X - \alpha\|} \\ &\leq \frac{1}{\|X - \alpha\|}.\end{aligned}$$

3. We assume additionally that μ is in $\mathcal{M}_\sim$. The real number $\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha-X\|}\Big]$ is nonzero and the operator H rewrites as a nonzero multiple of

$$\operatorname{Id} - \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha-X\|}\Big]^{-1}\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{(\alpha-X)\otimes(\alpha-X)}{\|\alpha-X\|^3}\Big].$$

By Neumann series [189, Lemma 17.2], invertibility of H follows if we prove the inequality

$$\Big\|\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha-X\|}\Big]^{-1}\mathbb{E}\Big[\mathbb{1}_{X\neq\alpha}\frac{(\alpha-X)\otimes(\alpha-X)}{\|\alpha-X\|^3}\Big]\Big\|_{op} < 1,$$

or equivalently

$$\|A\|_{op} < \mathbb{E}\Big[\mathbb{1}_{Y\neq 0}\frac{1}{\|Y\|}\Big], \tag{4.43}$$

where we let $Y = \alpha - X$ and $A = \mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{Y\otimes Y}{\|Y\|^3}\Big]$. A straightforward computation shows that the operator A is self-adjoint and nonnegative, thus by [189, Lemma 11.13] its operator norm rewrites as

$$\|A\|_{op} = \sup_{\|h\|=1}\langle Ah, h\rangle = \sup_{\|h\|=1}\mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{\langle Y,h\rangle^2}{\|Y\|^3}\Big],$$

and there exists $(h_n)_{n\geq1}$ a sequence of unit vectors such that

$$\mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{\langle Y,h_n\rangle^2}{\|Y\|^3}\Big] \xrightarrow[n\to\infty]{} \|A\|_{op}.$$

Reflexivity of E and the Eberlein–Šmulian theorem [7, Theorems 6.25 and 6.34] imply the existence of a subsequence $(h_{n_k})_{k\geq1}$ that converges in the weak topology of E to some $\tilde{h}\in E$. Since the norm is weakly lower semicontinuous, we have further $\|\tilde{h}\|\leq 1$. Applying the dominated convergence theorem along the subsequence, we obtain the additional convergence

$$\mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{\langle Y,h_{n_k}\rangle^2}{\|Y\|^3}\Big] \xrightarrow[k\to\infty]{} \mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{\langle Y,\tilde{h}\rangle^2}{\|Y\|^3}\Big].$$

Identifying the limits, applying Cauchy–Schwarz inequality and the bound $\|\tilde{h}\|\leq1$, we have

$$\|A\|_{op} = \mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{\langle Y,\tilde{h}\rangle^2}{\|Y\|^3}\Big] \overset{(i)}{\leq} \mathbb{E}\Big[\mathbb{1}_{Y\neq0}\frac{1}{\|Y\|}\Big].$$

Assume for the sake of contradiction that equality occurs in (i). By Cauchy–Schwarz the random variable

$$\mathbb{1}_{Y\neq0}\frac{1}{\|Y\|}\Big(1-\langle\frac{Y}{\|Y\|},\tilde{h}\rangle^2\Big)$$

is nonnegative, and since it has expectation zero, we must have $\langle\mathbb{1}_{Y\neq0}\frac{Y}{\|Y\|},\tilde{h}\rangle^2 = 1$ $\mathbb{P}$-almost surely. This implies equality in Cauchy–Schwarz, hence $\mathbb{1}_{Y\neq0}\frac{Y}{\|Y\|}$ and $\tilde{h}$ are proportional $\mathbb{P}$-almost surely, thus $\mathbb{P}(X \in \alpha+\mathbb{R}\tilde{h}) = \mathbb{P}(Y \in \mathbb{R}\tilde{h}) = 1$, which contradicts the assumption $\mu \in \mathcal{M}_\sim$. Inequality (i) is therefore strict and we have proved (4.43).

Since E is Banach, the bounded inverse theorem [159, Theorem 4.12-2] implies that the inverse operator H^{-1} is bounded. That H^{-1} is self-adjoint and nonnegative follows easily from these two properties being true for H.

Lastly, we let $[\cdot,\cdot]$ denote the bilinear form $[h_1,h_2] = \langle Hh_1, h_2\rangle$ which is symmetric and nonnegative (an equivalent terminology is positive semidefinite) by the second item of Proposition 4.64. As a consequence, $[\cdot,\cdot]$ satisfies the following Schwarz inequality [159, Problem 14 p.195]:

$$[h_1,h_2]^2 \leq [h_1,h_1][h_2,h_2].$$

The next chain of inequalities holds:

$$\begin{aligned}\|h\|^2 = [H^{-1}h, h] &\leq [H^{-1}h, H^{-1}h]^{1/2} \quad [h,h]^{1/2}\\ &= \langle h, H^{-1}h\rangle^{1/2} \quad \langle Hh, h\rangle^{1/2}\\ &\leq \|h\|\|H^{-1}\|_{op}^{1/2} \quad \langle Hh,h\rangle^{1/2},\end{aligned}$$

where the first (resp., second) inequality stems from the Schwarz (resp., Cauchy–Schwarz) inequality. Assuming h is nonzero, we obtain by rearranging that

$$\langle Hh, h\rangle \geq \frac{1}{\|H^{-1}\|_{op}}\|h\|^2,$$

hence

$$\inf_{\|h\|=1}\langle Hh,h\rangle \geq \frac{1}{\|H^{-1}\|_{op}} > 0.$$

4. We assume α has a neighborhood U that does not contain any atom. For h so small that $\alpha + h$ lies in U we have

$$\begin{aligned}&\|\nabla\phi(\alpha+h) - \nabla\phi(\alpha) - \mathbb{E}[\mathbb{1}_{X\neq\alpha}\nabla^2 N(\alpha - X)h]\|\\ &\overset{(i)}{=} \Big\|\mathbb{E}\Big[\mathbb{1}_{X\notin\{\alpha,\alpha+h\}}\big(\nabla N(\alpha+h-X) - \nabla N(\alpha - X) - \nabla^2 N(\alpha-X)h\big)\Big]\Big\|\\ &\leq \mathbb{E}\Big[\Big\|\mathbb{1}_{X\notin\{\alpha,\alpha+h\}}\big(\nabla N(\alpha+h-X) - \nabla N(\alpha - X) - \nabla^2 N(\alpha-X)h\big)\Big\|\Big]\\ &\overset{(ii)}{\leq} 2\mathbb{E}\Big[\frac{\|h\|}{\|\alpha - X\|} \wedge \frac{\|h\|^2}{\|\alpha-X\|^2}\Big]\\ &= 2\|h\|\Big(\mathbb{E}\Big[\mathbb{1}_{\|X-\alpha\|\leq\|h\|}\frac{1}{\|X-\alpha\|}\Big] + \mathbb{E}\Big[\mathbb{1}_{\|X-\alpha\|>\|h\|}\frac{\|h\|}{\|X-\alpha\|^2}\Big]\Big).\end{aligned} \tag{4.44}$$

Since α and $\alpha+h$ are not atoms of μ, the random variable $\mathbb{1}_{X\in\{\alpha,\alpha+h\}}$ is $\mathbb{P}$-almost surely zero, thus in (i) we omit the expectation involving this indicator. Inequality (ii) follows from Lemma 4.62. As shown in item 2. above, each expectation in (4.44) goes to 0 as h goes to 0. We thus obtain $\|\nabla\phi(\alpha+h) - \nabla\phi(\alpha) - \mathbb{E}[\mathbb{1}_{X\neq\alpha}\nabla^2 N(\alpha-X)h]\| = o(\|h\|)$ and ϕ is twice differentiable at α.

To identify the Hessian operator we need the equality

$$\mathbb{E}[\mathbb{1}_{X\neq\alpha}\nabla^2 N(\alpha - X)h] = \mathbb{E}[\mathbb{1}_{X\neq\alpha}\nabla^2 N(\alpha-X)]h. \tag{4.45}$$

Equality (4.45) is then justified by considering the evaluation operator $T : B(E) \to E, A \mapsto Ah$ and we can finally conclude that

$$\begin{aligned}\nabla^2\phi(\alpha) &= \mathbb{E}[\mathbb{1}_{X\neq\alpha}\nabla^2 N(\alpha - X)] \\ &= \mathbb{E}\left[\mathbb{1}_{X\neq\alpha}\frac{1}{\|\alpha - X\|}\left(\mathrm{Id} - \frac{(\alpha - X)\otimes(\alpha - X)}{\|\alpha - X\|^2}\right)\right].\end{aligned}$$

□

4.10.2 Proofs for Section 4.5.1

Proof of Proposition 4.70. 1. For any $\beta \in E$, $\widehat{\Psi}_n$ is twice differentiable at β with Hessian $\nabla^2\phi(\alpha_\star)/n$, which is nonnegative by Proposition 4.64, hence $\widehat{\Psi}_n$ is convex. The vector $\widehat{\beta}_n$ is the unique zero of the gradient $\nabla\widehat{\Psi}_n(\beta)$, hence the unique minimizer of $\widehat{\Psi}_n$.

2. $\alpha_\star$ is a minimizer of ϕ, which is differentiable at $\alpha_\star$ by the moment assumption and Proposition 4.64. Consequently, $\nabla\phi(\alpha_\star) = 0$ or equivalently the random element $Y = \mathbb{1}_{X\neq\alpha_\star}\frac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell$ is centered, hence so is $\nabla^2\phi(\alpha_\star)^{-1}Y$. Since E is separable and

$$\widehat{\beta}_n = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla^2\phi(\alpha_\star)^{-1}\left(\mathbb{1}_{X_i\neq\alpha_\star}\frac{\alpha_\star - X_i}{\|\alpha_\star - X_i\|} - \ell\right),$$

the central limit theorem for Hilbert spaces [168, Section 10.1] yields convergence in distribution to a centered Gaussian with covariance bilinear form Σ such that for every $(u, v) \in E$,

$$\begin{aligned}\Sigma(u,v) &= \mathbb{E}\big[\langle u, \nabla^2\phi(\alpha_\star)^{-1}Y\rangle\langle v, \nabla^2\phi(\alpha_\star)^{-1}Y\rangle\big] \\ &= \mathbb{E}\big[\langle \nabla^2\phi(\alpha_\star)^{-1}u, Y\rangle\langle \nabla^2\phi(\alpha_\star)^{-1}v, Y\rangle\big] \\ &= \mathbb{E}\big[\langle \nabla^2\phi(\alpha_\star)^{-1}u, (Y\otimes Y)\nabla^2\phi(\alpha_\star)^{-1}v\rangle\big] \\ &= \langle \nabla^2\phi(\alpha_\star)^{-1}u, \mathbb{E}[Y\otimes Y]\nabla^2\phi(\alpha_\star)^{-1}v\rangle \\ &= \langle u, \Big[\nabla^2\phi(\alpha_\star)^{-1}\mathbb{E}\big[\mathbb{1}_{X\neq\alpha_\star}(\tfrac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell)\otimes(\tfrac{\alpha_\star - X}{\|\alpha_\star - X\|} - \ell)\big]\nabla^2\phi(\alpha_\star)^{-1}\Big]v\rangle.\end{aligned}$$

This identifies the covariance operator.

By [39, Theorem 2.3.6], $(\widehat{\beta}_n)$ is uniformly tight and since compact sets are bounded, $\widehat{\beta}_n = O_{\mathbb{P}}(1)$.

3. By definition of κ, the function $\beta \mapsto n\widehat{\Psi}_n(\beta) - \kappa/2\|\beta\|^2$ is convex, hence $\widehat{\Psi}_n$ is κ/n-strongly convex and the inequality follows. □

We state a technical lemma of independent interest that will be useful in the proof of Proposition 4.71.

Lemma 4.91. *Let $Y_1, Y_2, \dots$ be i.i.d. nonnegative random variables such that $\mathbb{E}[Y_1] < \infty$.*

1. $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i\geq\sqrt{n}}Y_i$ *converges $\mathbb{P}$-almost surely to 0.*

2. Let $R > 0$. The supremum

$$\sup_{\rho \geq \frac{\sqrt{n}}{R}} \frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{Y_i < \rho} Y_i^2$$

converges $\mathbb{P}$-almost surely to 0.

Proof of Lemma 4.91. 1. For each integer $N \geq 1$, the strong law of large numbers yields the convergence $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \geq \sqrt{N}} Y_i \to_n \mathbb{E}[\mathbb{1}_{Y_1 \geq \sqrt{N}} Y_1]$ on an event Ω_N with $\mathbb{P}(\Omega_N) = 1$. Given $N \geq 1$ and $n \geq N$ we have

$$0 \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \geq \sqrt{n}} Y_i \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \geq \sqrt{N}} Y_i.$$

On the almost-sure event $\bigcap_{N \geq 1} \Omega_N$ we obtain

$$\forall N \geq 1, \quad \limsup_n \Big(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \geq \sqrt{n}} Y_i \Big) \leq \mathbb{E}[\mathbb{1}_{Y_1 \geq \sqrt{N}} Y_1].$$

By the dominated convergence theorem, $\mathbb{E}[\mathbb{1}_{Y_1 \geq \sqrt{N}} Y_1] \to 0$ as $N \to \infty$, hence the claim.

2. The following proof is adapted from [213]. The convergence $\frac{1}{n}\sum_{i=1}^n Y_i \to \mathbb{E}[Y_1]$ happens on an event Ω_0 and for each integer $M \geq 1$ we have $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \geq M} Y_i \to_n \mathbb{E}[\mathbb{1}_{Y_1 \geq M} Y_1]$ on an event Ω_M with $\mathbb{P}(\Omega_M) = 1$. In the rest of the proof we consider $\omega \in \bigcap_{M \geq 0} \Omega_M$ and we prove (the dependence on ω is omitted):

$$\sup_{\rho \geq \frac{\sqrt{n}}{R}} \frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{Y_i < \rho} Y_i^2 \xrightarrow[n \to \infty]{} 0.$$

Let $\epsilon > 0$ and fix some $M \geq 1$ that satisfies $\mathbb{E}[\mathbb{1}_{Y_1 \geq M} Y_1] < \epsilon/2$. Note that

$$\sup_{\rho \geq \frac{\sqrt{n}}{R}} \frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{Y_i < \rho} Y_i^2 \leq \sup_{\rho \geq \frac{\sqrt{n}}{R}} \Big(\frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{M \leq Y_i < \rho} Y_i^2 \Big) + \sup_{\rho \geq \frac{\sqrt{n}}{R}} \Big(\frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{Y_i < M} Y_i^2 \Big). \tag{4.46}$$

We bound each supremum in the right-hand side of (4.46) separately. Given $\rho \geq \frac{\sqrt{n}}{R}$,

$$\frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{M \leq Y_i < \rho} Y_i^2 \leq \frac{1}{\rho^2} \sum_{i=1}^{n} \mathbb{1}_{M \leq Y_i < \rho} Y_i \leq R^2 \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \geq M} Y_i. \tag{4.47}$$

By the hypothesis on ω, there exists $N \geq 1$ such that

$$n \geq N \implies \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \geq M} Y_i - \mathbb{E}[\mathbb{1}_{Y_1 \geq M} Y_1] \leq \epsilon/2. \tag{4.48}$$

Combining (4.47) and (4.48), we have

$$n \geq N \implies \sup_{\rho \geq \frac{\sqrt{n}}{R}} \Big(\frac{1}{\rho^3} \sum_{i=1}^{n} \mathbb{1}_{M \leq Y_i < \rho} Y_i^2 \Big) < R^2 \epsilon.$$

Regarding the remaining supremum in the right-hand side of (4.46), it is estimated as follows:

$$\sup_{\rho \geq \frac{\sqrt{n}}{R}} \Big(\frac{1}{\rho^3} \sum_{i=1}^n \mathbb{1}_{Y_i < M} Y_i^2 \Big) \leq \frac{R^3 M}{n^{3/2}} \sum_{i=1}^n Y_i = \frac{R^3 M}{\sqrt{n}} \frac{1}{n} \sum_{i=1}^n Y_i,$$

which is easily $< \epsilon$ for n large enough. □

Proof of Proposition 4.71. The suprema we consider below are measurable. We make use of the decomposition

$$\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta) = \Delta_1(\beta) + \Delta_2(\beta),$$

where

$$\Delta_1(\beta) = \widehat{\phi}_n\Big(\alpha_\star + \tfrac{\beta}{\sqrt{n}}\Big) - \widehat{\phi}_n(\alpha_\star) - \langle \nabla \widehat{\phi}_n(\alpha_\star), \tfrac{\beta}{\sqrt{n}} \rangle - \tfrac{1}{2} \langle \nabla^2 \widehat{\phi}_n(\alpha_\star) \tfrac{\beta}{\sqrt{n}}, \tfrac{\beta}{\sqrt{n}} \rangle,$$
$$\Delta_2(\beta) = \tfrac{1}{2} \langle \big(\nabla^2 \widehat{\phi}_n(\alpha_\star) - \nabla^2 \phi(\alpha_\star)\big) \tfrac{\beta}{\sqrt{n}}, \tfrac{\beta}{\sqrt{n}} \rangle.$$

1. We assume $\mathbb{E}[\|X - \alpha_\star\|^{-1}] < \infty$. The first item of Lemma 4.62 yields the bound:

$$\begin{aligned} n|\Delta_1(\beta)| &= \Big| \sum_{i=1}^n \|\alpha_\star - X_i + \tfrac{\beta}{\sqrt{n}}\| - \|\alpha_\star - X_i\| \\ &\qquad - \langle \mathbb{1}_{X_i \neq \alpha_\star} \nabla N(\alpha_\star - X_i), \tfrac{\beta}{\sqrt{n}} \rangle - \tfrac{1}{2} \langle \mathbb{1}_{X_i \neq \alpha_\star} \nabla^2 N(\alpha_\star - X_i) \tfrac{\beta}{\sqrt{n}}, \tfrac{\beta}{\sqrt{n}} \rangle \Big| \\ &\leq \frac{1}{2} \sum_{i=1}^n \mathbb{1}_{X_i \neq \alpha_\star} \Big(\frac{\|\beta/\sqrt{n}\|^2}{\|X_i - \alpha_\star\|} \wedge \frac{\|\beta/\sqrt{n}\|^3}{\|X_i - \alpha_\star\|^2} \Big) + \sum_{i=1}^n \mathbb{1}_{X_i = \alpha_\star} \Big\| \frac{\beta}{\sqrt{n}} \Big\| \\ &= S_n(\beta) + T_n(\beta) + \sum_{i=1}^n \mathbb{1}_{X_i = \alpha_\star} \Big\| \frac{\beta}{\sqrt{n}} \Big\|, \end{aligned} \tag{4.49}$$

where

$$S_n(\beta) = \frac{1}{2} \sum_{i=1}^n \mathbb{1}_{X_i \neq \alpha_\star} \mathbb{1}_{\|X_i - \alpha_\star\| \leq \| \frac{\beta}{\sqrt{n}} \|} \frac{\|\beta/\sqrt{n}\|^2}{\|X_i - \alpha_\star\|},$$
$$T_n(\beta) = \frac{1}{2} \sum_{i=1}^n \mathbb{1}_{X_i \neq \alpha_\star} \mathbb{1}_{\|X_i - \alpha_\star\| > \| \frac{\beta}{\sqrt{n}} \|} \frac{\|\beta/\sqrt{n}\|^3}{\|X_i - \alpha_\star\|^2}.$$

Moreover, if $\|\beta\| \leq R$, we have $S_n(\beta) \leq \frac{R^2}{2n} \sum_{i=1}^n \mathbb{1}_{X_i \neq \alpha_\star} \mathbb{1}_{\|X_i - \alpha_\star\| \leq \frac{R}{\sqrt{n}}} \frac{1}{\|X_i - \alpha_\star\|}$. The first item of Lemma 4.91 applied with $Y_i = \mathbb{1}_{X_i \neq \alpha_\star} \frac{R}{\|X_i - \alpha_\star\|}$ yields

$$\sup_{\|\beta\| \leq R} S_n(\beta) \to_n 0 \quad \text{a.s.}$$

Next, we observe that

$$\begin{aligned} \sup_{\|\beta\| \leq R} T_n(\beta) &= \sup_{r \in (0, R]} \frac{1}{2} \sum_{i=1}^n \mathbb{1}_{X_i \neq \alpha_\star} \mathbb{1}_{\|X_i - \alpha_\star\| > \frac{r}{\sqrt{n}}} \frac{(r/\sqrt{n})^3}{\|X_i - \alpha_\star\|^2} \\ &= \sup_{\rho \geq \frac{\sqrt{n}}{R}} \sum_{i=1}^n \mathbb{1}_{X_i \neq \alpha_\star} \mathbb{1}_{\|X_i - \alpha_\star\|^{-1} < \rho} \frac{\|X_i - \alpha_\star\|^{-2}}{\rho^3}, \end{aligned}$$

and the second item of Lemma 4.91 gives

$$\sup_{\|\beta\|\leq R} T_n(\beta) \to_n 0 \quad \text{a.s.}$$

Since $\alpha_\star$ is not an atom of μ, the sum in (4.49) is zero $\mathbb{P}$-almost surely, hence

$$n \sup_{\|\beta\|\leq R} |\Delta_1(\beta)| \to_n 0 \quad \text{a.s.}$$

To deal with Δ_2 we note that

$$n \sup_{\|\beta\|\leq R} |\Delta_2(\beta)| \leq \frac{R^2}{2} \|\nabla^2 \widehat{\phi}_n(\alpha_\star) - \nabla^2 \phi(\alpha_\star)\|_{op}$$

and we leverage the decomposition

$$\nabla^2 \widehat{\phi}_n(\alpha_\star) - \nabla^2 \phi(\alpha_\star) = A - B \tag{4.50}$$

where

$$A = \Big(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\neq\alpha_\star} \frac{1}{\|X_i - \alpha_\star\|} - \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}\frac{1}{\|X-\alpha_\star\|}\Big]\Big)\,\mathrm{Id},$$
$$B = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\neq\alpha_\star}\frac{(\alpha_\star - X_i)\otimes(\alpha_\star - X_i)}{\|X_i-\alpha_\star\|^3} - \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}\frac{(\alpha_\star - X)\otimes(\alpha_\star - X)}{\|X-\alpha_\star\|^3}\Big].$$

Since $\|A\|_{op} = \Big|\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{X_i\neq\alpha_\star}\frac{1}{\|X_i-\alpha_\star\|} - \mathbb{E}\Big[\mathbb{1}_{X\neq\alpha_\star}\frac{1}{\|X-\alpha_\star\|}\Big]\Big|$, the strong law of large numbers yields $\|A\|_{op} \to_n 0$ $\mathbb{P}$-almost surely. Since the operator $z\otimes z$ has rank at most one, it is a Hilbert–Schmidt operator. Consequently, B takes values in the space $S_2(E)$, which we equip with the norm $\|\cdot\|_2$. The function $f: E \to S_2(E), z \mapsto z\otimes z$ is continuous by the estimate

$$\begin{aligned}\|f(z)-f(z_0)\|_2^2 &= \|f(z)\|_2^2 + \|f(z_0)\|_2^2 - 2\langle f(z), f(z_0)\rangle_2\\ &= \|z\|^4 + \|z_0\|^4 - 2\langle z, z_0\rangle^2,\end{aligned}$$

hence B is measurable between the σ-algebras $\mathcal{F}$ and $\mathcal{B}(S_2(E))$. The space $S_2(E)$ is separable [68, Theorem 18.14 (c)], thus by Mourier's strong law of large numbers for Banach spaces [168, Corollary 7.10] we have the convergence $\|B\|_2 \to_n 0$ $\mathbb{P}$-almost surely. By inequality (4.42), $\|B\|_{op} \to_n 0$ $\mathbb{P}$-almost surely. Combining the convergence on A and B, we have

$$n \sup_{\|\beta\|\leq R} |\Delta_2(\beta)| \to_n 0 \quad \text{a.s.},$$

and this finishes the proof.

2. We assume $\mathbb{E}[\|X-\alpha_\star\|^{-2}] < \infty$ and we follow a similar path. Instead of (4.49), we bound the minimum directly by $\frac{\|\beta/\sqrt{n}\|^3}{\|X_i-\alpha_\star\|^2}$ and we use the central limit theorem to obtain $n\sup_{\|\beta\|\leq R}|\Delta_1(\beta)| = O_{\mathbb{P}}(n^{-1/2})$. Regarding Δ_2, we exploit the same decomposition (4.50). Since we assume a finite second moment for $\|X-\alpha_\star\|^{-1}$, we can leverage the

central limit theorem which yields $\|A\|_{op} = O_{\mathbb{P}}(n^{-1/2})$. By the central limit theorem for Hilbert spaces [168, Section 10.1], $\|B\|_2 = O_{\mathbb{P}}(n^{-1/2})$ thus $\|B\|_{op} = O_{\mathbb{P}}(n^{-1/2})$.

Regarding the difference $\nabla\widehat{\psi}_n - \nabla\widehat{\Psi}_n$, we use the decomposition

$$\nabla\widehat{\psi}_n(\beta) - \nabla\widehat{\Psi}_n(\beta) = D_1(\beta) + D_2(\beta),$$

where

$$D_1(\beta) = \tfrac{1}{\sqrt{n}}\nabla\widehat{\phi}_n(\alpha_\star + \tfrac{\beta}{\sqrt{n}}) - \tfrac{1}{\sqrt{n}}\nabla\widehat{\phi}_n(\alpha_\star) - \nabla^2\widehat{\phi}_n(\alpha_\star)\tfrac{\beta}{n},$$
$$D_2(\beta) = \big(\nabla^2\widehat{\phi}_n(\alpha_\star) - \nabla^2\phi(\alpha_\star)\big)\tfrac{\beta}{n}.$$

For β such that $0 < \|\beta\| \le R$, the second item of Lemma 4.62 gives

$$\begin{aligned} n^{3/2}|D_1(\beta)| &= \Big|\sum_{i=1}^n \mathbb{1}_{X_i\notin\{\alpha_\star,\alpha_\star+\frac{\beta}{\sqrt{n}}\}}\Big(\nabla N(\alpha_\star - X_i + \tfrac{\beta}{\sqrt{n}}) - \nabla N(\alpha_\star - X_i) - \nabla^2 N(\alpha_\star - X_i)\tfrac{\beta}{\sqrt{n}}\Big) \\ &\quad + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star}\nabla N(\tfrac{\beta}{\sqrt{n}}) + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star+\frac{\beta}{\sqrt{n}}}\nabla N(\tfrac{\beta}{\sqrt{n}})\Big| \\ &\le 2\sum_{i=1}^n \mathbb{1}_{X_i\neq\alpha_\star}\frac{\|\beta/\sqrt{n}\|^2}{\|X_i-\alpha_\star\|^2} + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star} + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star+\frac{\beta}{\sqrt{n}}} \\ &\le \frac{2R^2}{n}\sum_{i=1}^n \mathbb{1}_{X_i\neq\alpha_\star}\frac{1}{\|X_i-\alpha_\star\|^2} + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star} + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star+\frac{\beta}{\sqrt{n}}}, \end{aligned}$$

hence

$$\sup_{\|\beta\|\le R} n^{3/2}|D_1(\beta)| \le \frac{2R^2}{n}\sum_{i=1}^n \frac{\mathbb{1}_{X_i\neq\alpha_\star}}{\|X_i-\alpha_\star\|^2} + \sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star} + \sup_{\|\beta\|\le R}\sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star+\frac{\beta}{\sqrt{n}}}. \tag{4.51}$$

In the right-hand side of (4.51), the first summand is $O_{\mathbb{P}}(1)$ by the strong law of large numbers. Since $\alpha_\star$ is not an atom of μ, the second summand is zero $\mathbb{P}$-almost surely, hence $O_{\mathbb{P}}(1)$. The last supremum requires more work and we write $S_n = \sup_{\|\beta\|\le R}\sum_{i=1}^n \mathbb{1}_{X_i=\alpha_\star+\frac{\beta}{\sqrt{n}}}$ for convenience. Given a fixed $k \in \{1,\dots,n\}$, if $S_n \ge k$ there is some β with $\|\beta\| \le R$ such that at least k of the X_i are equal to $\alpha_\star + \frac{\beta}{\sqrt{n}}$. Therefore, each of these verifies $\|X_i - \alpha_\star\|^{-1} \ge \frac{\sqrt{n}}{R}$. By this observation we have the bound

$$\begin{aligned} \mathbb{P}(S_n \ge k) &\le \mathbb{P}\Big(\bigcup_{\substack{I\subset\{1,\dots,n\}\\|I|=k}}\bigcap_{i\in I}\{\|X_i-\alpha_\star\|^{-1} \ge \tfrac{\sqrt{n}}{R}\}\Big) \\ &\overset{(i)}{\le} \binom{n}{k}\big[\mathbb{P}(\|X_1-\alpha_\star\|^{-1} \ge \tfrac{\sqrt{n}}{R})\big]^k = \binom{n}{k}p_n^k\,, \end{aligned}$$

where $p_n = \mathbb{P}(\|X_1-\alpha_\star\|^{-1} \ge \frac{\sqrt{n}}{R})$. To obtain (i) we used the union bound and independence of the X_i. Since $\|X_i - \alpha_\star\|^{-1}$ has finite second moment, we have the estimate $p_n = o(1/n)$, thus $\mathbb{E}[S_n] = \sum_{k=1}^n \mathbb{P}(S_n \ge k) \le (1+p_n)^n = (1+o(n^{-1}))^n = 1+o(1)$. Consequently, $\mathbb{E}[S_n]$ is bounded and $S_n = O_{\mathbb{P}}(1)$. Inequality (4.51) then yields $\sup_{\|\beta\|\le R} n^{3/2}|D_1(\beta)| = O_{\mathbb{P}}(1)$. We obtain $\sup_{\|\beta\|\le R} n^{3/2}|D_2(\beta)| = O_{\mathbb{P}}(1)$ by identical bounds on $\|A\|_{op}$ and $\|B\|_{op}$ as above. □

Proof of Theorem 4.73. 1. Let $\varepsilon, \eta > 0$. By Proposition 4.70, $\widehat{\beta}_n = O_{\mathbb{P}}(1)$ so there exists $M_1 > 0$ such that $\mathbb{P}(\|\widehat{\beta}_n\| > M_1) < \eta/3$ for every $n \geq 1$. By Proposition 4.71 and the convergence assumption on ϵ_n, there is some $N \geq 1$ such that for every $n \geq N$,

$$\mathbb{P}\Big(\sup_{\|\beta\| \leq M_1+\varepsilon} |\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| > \frac{\kappa\varepsilon^2}{8n}\Big) < \eta/3 \quad \text{and} \quad \mathbb{P}^*\Big(\epsilon_n > \frac{\kappa\varepsilon^2}{8n}\Big) < \eta/3.$$

We let $\Omega_n = \{\|\widehat{\beta}_n\| \leq M_1\} \cap \{\sup_{\|\beta\| \leq M_1+\varepsilon} |\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| \leq \frac{\kappa\varepsilon^2}{8n}\} \cap \{\epsilon_n \leq \frac{\kappa\varepsilon^2}{8n}\}$ and S denotes the sphere centered at $\widehat{\beta}_n$ with radius ε. For $n \geq N$, for $\omega \in \Omega_n$ (ω is implicit in what follows) and $\beta \in S$ we have the lower bound

$$\begin{aligned}\widehat{\psi}_n(\beta) \overset{(i)}{\geq} \widehat{\Psi}_n(\beta) - \frac{\kappa\varepsilon^2}{8n} \overset{(ii)}{\geq} \widehat{\Psi}_n(\beta) + \frac{\kappa\varepsilon^2}{2n} - \frac{\kappa\varepsilon^2}{8n} &\geq \widehat{\psi}_n(\widehat{\beta}_n) + \frac{\kappa\varepsilon^2}{2n} - \frac{\kappa\varepsilon^2}{4n} \\ &= \widehat{\psi}_n(\widehat{\beta}_n) + \frac{\kappa\varepsilon^2}{4n} \\ &> \widehat{\psi}_n(\widehat{\beta}_n) + \epsilon_n, \end{aligned} \tag{4.52}$$

where (i) follows from $\|\beta\| \leq M_1 + \varepsilon$ and (ii) from the third item of Proposition 4.70. Inequality (4.52) implies that the ϵ_n-arg min of $\widehat{\psi}_n$ is a subset of the closed ball centered at $\widehat{\beta}_n$ with radius ε. Otherwise, there exists β an ϵ_n-minimizer of $\widehat{\psi}_n$ such that $\|\beta - \widehat{\beta}_n\| > \varepsilon$ and we can find $\lambda \in [0,1]$ satisfying $(1-\lambda)\beta + \lambda\widehat{\beta}_n \in S$. Thus by convexity of $\widehat{\psi}_n$,

$$\begin{aligned}\widehat{\psi}_n((1-\lambda)\beta + \lambda\widehat{\beta}_n) \leq (1-\lambda)\widehat{\psi}_n(\beta) + \lambda\widehat{\psi}_n(\widehat{\beta}_n) &\leq (1-\lambda)(\epsilon_n + \inf\widehat{\psi}_n) + \lambda\widehat{\psi}_n(\widehat{\beta}_n) \\ &\leq \epsilon_n + \widehat{\psi}_n(\widehat{\beta}_n),\end{aligned}$$

which contradicts (4.52).

Since $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$ is an ϵ_n-minimizer of $\widehat{\psi}_n$, we obtain

$$\mathbb{P}^*\big(\|\sqrt{n}(\widehat{\alpha}_n - \alpha_\star) - \widehat{\beta}_n\| > \varepsilon\big) \leq \mathbb{P}^*(\Omega_n^c) < \eta \quad \text{for every } n \geq N,$$

hence the claim.

2. Let $\varepsilon > 0$. By Proposition 4.70, $\widehat{\beta}_n = O_{\mathbb{P}}(1)$ so there exists $M_1 > 0$ such that $\mathbb{P}(\|\widehat{\beta}_n\| > M_1) < \epsilon/3$ for every $n \geq 1$. By Proposition 4.71, there is some $M_2 > 0$ such that $\forall n \geq 1, \mathbb{P}\Big(\sup_{\|\beta\| \leq M_1+1} |\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| > \frac{M_2}{2n^{3/2}}\Big) < \varepsilon/3$. We define the radius $r_n = 2(M_2/\kappa)^{1/2} n^{-1/4}$. For n larger than some N we have the bounds

$$\mathbb{P}^*\Big(\epsilon_n > \frac{M_2}{2n^{3/2}}\Big) < \varepsilon/3 \quad \text{and} \quad r_n \leq 1.$$

We let $\Omega_n = \{\|\widehat{\beta}_n\| \leq M_1\} \cap \{\sup_{\|\beta\| \leq M_1+1} |\widehat{\psi}_n(\beta) - \widehat{\Psi}_n(\beta)| \leq \frac{M_2}{2n^{3/2}}\} \cap \{\epsilon_n \leq \frac{M_2}{2n^{3/2}}\}$ and S denotes the sphere centered at $\widehat{\beta}_n$ with radius r_n.

By arguments similar to those developed in the previous item, we obtain

$$\mathbb{P}^*\big(\|\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)\| > r_n\big) \leq \mathbb{P}^*(\Omega_n^c) < \varepsilon \quad \text{for every } n \geq N,$$

hence the claim.

3. Let $\varepsilon > 0$. There exists $M_1 > 0$ such that $\mathbb{P}(\|\widehat{\beta}_n\| > M_1) < \epsilon/3$ for every $n \geq 1$. By Proposition 4.71, there is some $M_2 > 0$ such that

$$\forall n \geq 1, \mathbb{P}\Big(\sup_{\|\beta\| \leq M_1+1} |\nabla\widehat{\psi}_n(\beta) - \nabla\widehat{\Psi}_n(\beta)| > \tfrac{M_2}{n^{3/2}} \Big) < \varepsilon/3.$$

We let $r_n = (2M_2/\kappa)n^{-1/2}$ and $s_n = (M_2/\kappa)n^{-1/2}$. There is some $N \geq 1$ such that

$$n \geq N \implies \mathbb{P}^*\Big(\epsilon_n > \frac{M_2^2}{2\kappa n^2}\Big) < \varepsilon/3 \quad \text{and} \quad r_n \leq 1.$$

We put $\Omega_n = \{\|\widehat{\beta}_n\| > M_1\} \cap \{\sup_{\|\beta\| \leq M_1+1} |\nabla\widehat{\psi}_n(\beta) - \nabla\widehat{\Psi}_n(\beta)| \leq \frac{M_2}{n^{3/2}}\} \cap \{\epsilon_n \leq \frac{M_2^2}{2\kappa n^2}\}$. We fix $n \geq N$, $\omega \in \Omega_n$ (ω is implicit in what follows), a unit vector $a \in E$ and we define the convex function

$$\begin{aligned} \widehat{\gamma}_n \colon \mathbb{R} &\to \mathbb{R} \\ t &\mapsto \widehat{\psi}_n(\widehat{\beta}_n + ta). \end{aligned}$$

The quantity $g_n = \langle \nabla\widehat{\psi}_n(\widehat{\beta}_n + r_n a), a\rangle$ verifies the estimate

$$g_n \geq \langle \nabla\widehat{\Psi}_n(\widehat{\beta}_n + r_n a), a\rangle - \frac{M_2}{n^{3/2}} = \frac{r_n}{n}\langle \nabla^2\phi(\alpha_\star)a, a\rangle - \frac{M_2}{n^{3/2}} \geq \frac{M_2}{n^{3/2}}$$

and it is in the subdifferential of $\widehat{\gamma}_n$ at r_n, thus

$$\begin{aligned} \forall t \geq r_n + s_n, \quad \widehat{\gamma}_n(t) &\geq \widehat{\gamma}_n(r_n) + g_n(t - r_n) \\ &\geq \inf(\widehat{\gamma}_n) + \epsilon_n + g_n s_n - \epsilon_n \\ &\geq \inf(\widehat{\gamma}_n) + \epsilon_n + \frac{M_2^2}{\kappa n^2} - \epsilon_n \\ &> \inf(\widehat{\gamma}_n) + \epsilon_n. \end{aligned}$$

Since the inequality holds uniformly on the unit vector a, ϵ_n-minimizers of $\widehat{\psi}_n$ lie in the open ball centered at $\widehat{\beta}_n$ with radius $r_n + s_n = (3M_2/\kappa)n^{-1/2}$ and we conclude as before. □

Proof of Theorem 4.77. We use the theory developed by Van der Vaart and Wellner [261, Chapter 1.3] to make sense of convergence in distribution for nonmeasurable maps. By Proposition 4.70, the Borel measurable random element $\widehat{\beta}_n$ converges in distribution to γ, the Gaussian measure with mean 0 and covariance operator Σ. The first item of Theorem 4.73 combined with [261, Lemma 1.10.2] and Slutsky's theorem [261, p.32] yields convergence in distribution of $\sqrt{n}(\widehat{\alpha}_n - \alpha_\star)$ to γ. □

Chapter 5

Convex Fréchet ℓ-means in a metric tree

5.1 Introduction

5.1.1 Context

Statisticians commonly model data as an i.i.d. sample from an unknown probability measure μ. There is much interest in the central tendency of μ, i.e., in defining a location parameter that is representative of the whole population, and then in estimating this location parameter. When the ambient space is $\mathbb{R}^d$, a prominent measure of central tendency is the mean $\int_{\mathbb{R}^d} x \, \mathrm{d}\mu(x)$ (provided μ has one moment) and an estimator is the sample mean $\frac{1}{n}\sum_{i=1}^n X_i$, where $X_1, X_2, \ldots$ are i.i.d. random vectors with distribution μ. Fréchet [95] extended the notion of mean to the general setting of metric spaces by leveraging an optimization problem. Given a metric space (E, d) and a Borel probability measure μ on E, one says that μ has k finite moments, $k \geq 1$, if $\int_E d(\alpha, x)^k \, \mathrm{d}\mu(x)$ is finite, for some (and hence, every) $\alpha \in E$. If μ has two finite moments, a Fréchet mean (or barycenter) of μ is a minimizer of the objective function

$$\begin{aligned} E &\to \mathbb{R} \\ \alpha &\mapsto \int_E d(\alpha, x)^2 \, \mathrm{d}\mu(x). \end{aligned} \tag{5.1}$$

In fact, the definition of a Fréchet mean of μ only requires a finite first moment, since the objective function can be replaced with $\alpha \mapsto \int_E (d(\alpha, x)^2 - d(\alpha_0, x)^2) \, \mathrm{d}\mu(x)$ for some arbitrary $\alpha_0 \in E$, and the definition will not be affected by the choice of α_0. In many settings the Fréchet mean $\alpha_\star$ exists and is unique [243, 3, 4, 202, 273].

A natural estimator of $\alpha_\star$ is the sample Fréchet mean $\widehat{\alpha}_n$ obtained by minimizing the sample objective $\alpha \mapsto \frac{1}{n}\sum_{i=1}^n d(\alpha, X_i)^2$. Laws of large numbers for $(\widehat{\alpha}_n)_{n\geq 1}$ hold under a variety of assumptions on the space E [286, 33, 243] and central limit theorems have been developed when E is a Riemannian manifold [34, 31, 32, 87]. Non-Euclideanity of the space allows for new asymptotic phenomena such as stickiness [126, 136] and smeariness [125, 86, 85]. The non-asymptotic properties of the estimator have attracted much attention recently [230, 5, 167, 276, 49, 88].

Except for laws of large numbers [137] and Riemannian central limit theorems [48], these statistical results and their proofs are specific to the Fréchet mean, i.e., they are tied to the objective function (5.1). Still, other measures of central tendency are of interest. In the simplest setting of the real line $\mathbb{R}$, a major shortcoming of the sample mean is its lack of robustness to outliers, hence the need for alternatives such as the median. The population Fréchet median can be defined by replacing the squared distance in (5.1) with $d(\alpha, x)$. In order to cover a variety of location parameters, we study more general objectives of the form $\alpha \mapsto \int_E \ell(d(\alpha, x))\, \mathrm{d}\mu(x)$ where $\ell : [0, \infty) \to [0, \infty)$ is a convex nondecreasing function. We refer to minimizers of such an objective as Fréchet ℓ-means.

In exchange for generality in the objective, we constrain the ambient space to be a metric tree T, i.e., an undirected connected acyclic graph with weighted edges, where weights are understood as edge lengths and the distance between two points is the length of the (unique) shortest path between them. Metric trees arise in real-life applications, as they are an ideal model for road and communication networks. Tree-shaped networks appear naturally when modeling rivers or sparsely populated areas. A distribution system organized around a unique hub may be described as a star-like network, thus as a tree. In all these settings, the demand for service can occur at random locations across the network and these locations are distributed according to μ. Minimizing $\alpha \mapsto \int_T d(\alpha, x)\, \mathrm{d}\mu(x)$ is then akin to locating a new facility on the network with least average travel time to the demand. This median problem was initially studied in the special case where μ is discrete and supported on the vertices of the network. It gained traction among the operations research community in the 1960s, with an emphasis on the development of efficient algorithms (see, e.g., the surveys [111, 248, 249, 112]). More general objective functions were considered in [235, 44] and the case of non-discrete μ was studied in [200, 38].

A metric tree is a particular instance of a Hadamard space [45], hence results for Fréchet means in general Hadamard spaces (e.g., [243, 15, 48, 49, 88]) apply also to metric trees. There is little statistical literature on Fréchet ℓ-means in the specific setting of metric trees. Basrak [18] focuses on the Fréchet mean in a binary metric tree, and he establishes a central limit theorem for the inductive mean (a different estimator from the sample Fréchet mean). Risser et al. [96, 98] seek to compute Fréchet means on metric graphs, while Hotz et al. [126] develop laws of large numbers and central limit theorems for the Fréchet mean when the ambient space is an open book, i.e., a finite collection of copies of a Euclidean halfspace, glued with each other along their boundary. A special case of an open book is the m-spider, which can be viewed as a metric tree with one central vertex and infinitely long edges.

5.1.2 Contributions and outline

The goal of this work is to investigate the statistical properties of Fréchet ℓ-means in a metric tree T. We describe below how the chapter is organized and we give a brief summary of our contributions.

- In Section 5.2 we introduce the precise terminology and setting for our study. By leveraging the geodesic convexity of the objective function (Proposition 5.7) and the geometry of the tree (Definition 5.8), we develop a notion of directional derivatives (Definition 5.10) and we are able to locate (Proposition 5.14) and characterize (Proposition 5.15) Fréchet ℓ-means according to the signs of these directional derivatives.

- In Section 5.3 we turn to estimation using a sample analog. We observe that the topic of consistency is settled (Lemma 5.23) and we extend the notion of stickiness introduced by Hotz et al. [126] to the metric tree. An arbitrary point $c \in T$ is either sticky, partly sticky or nonsticky according to the signs of directional derivatives at c (Definition 5.24). We show that empirical stickiness is a non-asymptotic phenomenon that happens with exponential probability (Theorem 5.27). As an immediate consequence, we obtain a sticky law of large numbers (Corollary 5.28). Finally, we provide an equivalent definition of stickiness that is stated in terms of robustness to small pertubations of the population distribution (Proposition 5.29).

- In Section 5.4 we focus on Fréchet medians, i.e., when $\ell(z) = z$. We provide more specific statements on the location (Proposition 5.32) and uniqueness (Proposition 5.34) of medians. In the partly sticky case, we establish central limit theorems (Theorems 5.43 and 5.51) and non-asymptotic concentration bounds (Theorems 5.46 and 5.53).

5.2 Fréchet ℓ-means in a metric tree

5.2.1 Terminology and setting

Let us make precise what we mean by a metric tree and introduce further useful terminology.

Definition 5.1. 1. Let T denote an undirected, connected, acyclic graph with weighted edges (in the usual graph-theoretic sense). The weight of an edge is always assumed to be positive and it is understood as the length of this edge, i.e., as the distance between the corresponding adjacent vertices. We assume additionally that T has finitely many vertices. We implicitly consider a planar and isometric embedding of T in $\mathbb{R}^2$; T is then equipped with the shortest path metric d: the distance between two points of T (not necessarily vertices) is the length of the shortest path between them. Then, (T, d) is a metric space, which is referred to as a *metric tree*. In the sequel, we denote by D its diameter and by $\mathcal{B}(T)$ its Borel σ-algebra.

2. A vertex $v \in T$ is a *leaf* if it has exactly one adjacent vertex.

3. Let $m \geq 2$. T is an *m-spider* if the underlying graph-theoric tree has exactly m leaves and there is a single vertex adjacent to all of them.

Now, let us define an analog of the Lebesgue measure on (T, d).

Definition 5.2. Let $\mathcal{E}$ be the set of all edges of T. Each edge $e \in \mathcal{E}$ can be identified with a segment S_e in $\mathbb{R}$ of the same length, hence, it inherits its own Lebesgue measure, denoted by λ_e. Now, for any $A \in \mathcal{B}(T)$, set $\lambda(A) = \sum_{e \in \mathcal{E}} \lambda_e(A \cap e)$, where $A \cap e$ is identified isometrically with a subset of S_e.

Next, we introduce some relevant concepts from metric geometry. Given $x, y \in T$, a constant speed geodesic from x to y is a map γ from some interval $[a, b] \subset \mathbb{R}$ to E such that $\gamma(a) = x$, $\gamma(b) = y$ and $d(\gamma(t_1), \gamma(t_2)) = v|t_1 - t_2|$ for some $v \in [0, \infty)$ and every $t_1, t_2 \in [a, b]$. If $x \neq y$, $v = \frac{d(x,y)}{b-a}$ is called the speed of the geodesic γ. For the sake of legibility, we will often write γ_t in lieu of $\gamma(t)$. The space (T, d) is uniquely geodesic, meaning that between any two points $x, y \in T$, there always exists a geodesic from x to y and that is it unique up to reparametrization. Its image is denoted by $[x, y]$ and it is referred to as the geodesic segment joining x and y. We also define open and half-open geodesic intervals (x, y), $[x, y)$, $(x, y]$: For instance, $[x, y) = \gamma([a, b))$ for some $a \leq b$ and a geodesic γ from x to y defined on $[a, b]$.

A well-known geometric property of metric trees is that they are CAT(0); see, e.g., [45, Example 1.15(5) p.167]. By our assumptions, (T, d) is also compact and complete, hence it is a compact Hadamard space. In Hadamard spaces it is possible to develop a theory of convex analysis, convex optimization and probability that generalizes to nonlinear settings the classical results known in Hilbert spaces [15]. Here, we recall the definition of geodesic convexity in (T, d).

Definition 5.3. Let (T, d) be a metric tree as above.

1. A subset $G \subset T$ is called geodesically convex (convex, for short) if and only if for all $x, y \in G$ $[x, y] \subset G$.

2. A function $f : T \to \mathbb{R}$ is called geodesically convex (convex, for short) if and only if for every geodesic $\gamma : [0, 1] \to T$ and $t \in (0, 1)$ we have the inequality $f(\gamma_t) \leq (1-t)f(x) + tf(y)$. We call f geodesically strictly convex (strictly convex, for short), if and only if the previous inequality is strict, so long as $\gamma_0 \neq \gamma_1$.

Definition 5.4. Let (T, d) be a metric tree as above, μ be a probability measure on $(T, \mathcal{B}(T))$ and $\ell : [0, \infty) \to [0, \infty)$ be a convex and nondecreasing function, which we call the *loss function*. We define the *objective function* ϕ

$$\begin{aligned} \phi \colon T &\to \mathbb{R} \\ \alpha &\mapsto \int_T \ell\left(d(\alpha, x)\right) \mathrm{d}\mu(x). \end{aligned} \tag{5.2}$$

Minimizers of ϕ are called *Fréchet ℓ-means of μ*, and we denote by $M(\mu)$ the set of all minimizers.

Example 5.5. Examples of loss functions ℓ include:

1. $\ell : z \mapsto z^p$ where $p \in [1, \infty)$. In this setting, the minimizers of ϕ are called *Fréchet p-means* of μ. In the case $p = 1$ they are referred to as *Fréchet medians*, and when $p = 2$ as *barycenters* or just *Fréchet means*. The corresponding set of minimizers will be denoted specifically by $M_p(\mu)$.

2. $\ell : z \mapsto z^2 \mathbb{1}_{|z|\leq c} + (2c|z| - c^2)\mathbb{1}_{|z|>c}$ where $c \geq 0$. It is known as the Huber loss [133].

3. $\ell : z \mapsto 2c^2 \left((1 + \frac{z^2}{c^2})^{1/2} - 1 \right)$ where $c > 0$. It is known as the pseudo-Huber loss, which is a smooth approximation of the standard Huber loss.

The following lemma exhibits basic regularity properties of the loss.

Lemma 5.6. *Let $\ell : [0, \infty) \to [0, \infty)$ be a convex and nondecreasing function.*

1. *The left-derivative $\ell'_- : (0, \infty) \to [0, \infty)$ and right-derivative $\ell'_+ : [0, \infty) \to [0, \infty)$ of ℓ exist and are nondecreasing.*

2. *ℓ is continuous and locally Lipschitz.*

3. *For every $z \in [0, \infty)$, $\ell(z) = \ell(0) + \int_0^z \ell'_-(t)\,\mathrm{d}t = \ell(0) + \int_0^z \ell'_+(t)\,\mathrm{d}t$.*

In the next proposition, we show that the objective ϕ is well-defined and we provide other foundational properties of ϕ and $M(\mu)$.

Proposition 5.7. 1. *ϕ is well-defined, continuous and convex.*

2. *$M(\mu)$ is a nonempty, closed and convex subset of T.*

3. *ℓ is strictly convex if and only if ℓ'_+ is increasing. In that case, ϕ is strictly convex and $M(\mu)$ is a singleton.*

5.2.2 Convex calculus in a metric tree

The following section is dedicated to locating and characterizing the minimizers of ϕ.

Given a real-valued function f defined on a vector space E, the variations of f with respect to a reference point $\alpha \in E$ and in a direction $v \in E$ are naturally assessed by restricting f to the half-line $\{\alpha + tv : t \geq 0\} \subset E$ and defining the difference quotient $q : t \mapsto \frac{f(\alpha+tv)-f(\alpha)}{t}$ where $t \in (0, \infty)$. If additionally f is convex, then q is nondecreasing and bounded below; its right-sided limit is the directional derivative of f at α in the direction v [118, p.238].

In general the metric space T has no linear structure, and a point $v \in T$ does not carry by itself a notion of direction. However the restriction to the half-line in the difference quotient defined above can be replaced with the restriction to the geodesic segment $[\alpha, v]$, thus we consider the *metric difference quotient*

$$\begin{aligned} Q\colon (0,1] &\to \mathbb{R} \\ t &\mapsto \frac{\phi(\gamma_t) - \phi(\alpha)}{d(\gamma_t, \alpha)}, \end{aligned} \tag{5.3}$$

where $\gamma : [0, 1] \to [\alpha, v]$ denotes the geodesic from α to v. Since $d(\gamma_t, \alpha) = td(v, \alpha)$ and $t \mapsto \phi(\gamma_t)$ is convex, Q is nondecreasing and has a right-sided limit at 0 (which we will see is finite). Before we provide the value of this limit, we need the following definition.

Definition 5.8. Given α and v two distinct elements of T, we let $w_1, \dots, w_m$ denote the leaves of T and we define the subset

$$T_{\alpha\to v} = (\alpha, v] \cup \bigcup_{i\in\{1,\dots,m\}:\, \alpha\notin[v,w_i]} [v, w_i].$$

Alternatively, the metric space $T \setminus \{\alpha\}$ has two path-components and $T_{\alpha\to v}$ is the path-component that contains v. It is also the largest convex subset of T that contains v but not α.

Figure 5.1 illustrates this definition in two situations: either α is in the interior of an edge, or α is a vertex of T. We stress that α does not belong to $T_{\alpha\to v}$.

(a) α is in the interior of an edge (b) α is a vertex

Figure 5.1: Illustration for Definition 5.8 in two cases. $T_{\alpha\to v}$ is drawn in red and $T \setminus T_{\alpha\to v}$ is drawn in green.

As shown in the next proposition, the expression for the limit of the metric difference quotient (5.3) involves the left- and right- derivative of the univariate convex function ℓ, which already played a role in Lemma 5.6.

Proposition 5.9. *Let α and v be two distinct points in T. The following convergence holds:*

$$\frac{\phi(\alpha') - \phi(\alpha)}{d(\alpha', \alpha)} \xrightarrow[\substack{\alpha'\to\alpha \\ \alpha'\in(\alpha,v]}]{} \int_{T\setminus T_{\alpha\to v}} \ell'_+(d(\alpha,x))\,\mathrm{d}\mu(x) - \int_{T_{\alpha\to v}} \ell'_-(d(\alpha,x))\,\mathrm{d}\mu(x). \tag{5.4}$$

Consequently, the metric difference quotient $Q(t)$ converges to this finite limit as $t \to 0^+$.

Definition 5.10. We refer to the limiting value in (5.4) as *the directional derivative of ϕ at α towards v* and we denote it by $\phi'_v(\alpha)$.

Remark 5.11. If w is in $T_{\alpha\to v}$ and $w \neq v$, we note that $T_{\alpha\to w} = T_{\alpha\to v}$, thus $\phi'_w(\alpha) = \phi'_v(\alpha)$. The equality between derivatives is expected: if α' is in $(\alpha, w]$ and sufficiently close to α, then α' is in $(\alpha, v]$.

Example 5.12. 1. For Fréchet p-means, $p \geq 1$,

$$\phi'_v(\alpha) = p\left(\int_{T\setminus T_{\alpha\to v}} d(\alpha,x)^{p-1}\,\mathrm{d}\mu(x) - \int_{T_{\alpha\to v}} d(\alpha,x)^{p-1}\,\mathrm{d}\mu(x)\right).$$

2. In particular, for Fréchet medians ($p = 1$ above),

$$\begin{aligned}\phi'_v(\alpha) &= \mu(T \setminus T_{\alpha\to v}) - \mu(T_{\alpha\to v}) \\ &= 1 - 2\mu(T_{\alpha\to v}) \\ &= 2\mu(T \setminus T_{\alpha\to v}) - 1.\end{aligned}$$

   Remarkably, the directional derivative does not involve the metric d; it is expressed solely in terms of μ.

Remark 5.13. Assessing the directional derivative of ϕ along an edge is not a new idea: [235, 44] perform the computation for the sample Fréchet p-mean, [243] does so for the population Fréchet mean on a m-spider, and [185] for the sample Fréchet mean.

Next, we leverage the geometry of T and the geodesic convexity of ϕ to show a connection between the sign of the directional derivative and the location of the minimizers of ϕ.

Proposition 5.14. *Let α_0 and v two distinct points in T.*

1. *If $\phi'_v(\alpha_0) < 0$, then $M(\mu) \subset T_{\alpha_0\to v}$.*
2. *If $\phi'_v(\alpha_0) > 0$, then $M(\mu) \subset T \setminus T_{\alpha_0\to v}$.*
3. *If $\phi'_v(\alpha_0) = 0$, then $\alpha_0 \in M(\mu)$.*

As a consequence, we obtain the following first-order optimality conditions.

Proposition 5.15. *Let $\alpha \in T$.*

1. *The following are equivalent:*
   (a) *$\alpha \in M(\mu)$.*
   (b) *For every $v \in T \setminus \{\alpha\}$, $\phi'_v(\alpha) \geq 0$.*
   (c) *For every neighboring vertex v of α, $\phi'_v(\alpha) \geq 0$.*
2. *If for every $v \in T \setminus \{\alpha\}$, $\phi'_v(\alpha) > 0$, then α is the unique minimizer of ϕ, i.e., $M(\mu) = \{\alpha\}$.*
3. *Assume that α lies in the interior of an edge $[v, w]$, that $\mu(\{\alpha\}) = 0$ or $\ell'_+(0) = 0$, and that ℓ is differentiable over $(0, \infty)$. Then*

$$\begin{aligned}\alpha \in M(\mu) &\iff \phi'_v(\alpha) = \phi'_w(\alpha) = 0 \\ &\iff \int_{T_{\alpha\to v}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x) = \int_{T_{\alpha\to w}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x).\end{aligned}$$

Note that thanks to Remark 5.11, in Part 2 of Proposition 5.15, it is sufficient to only consider neighboring vertices of α.

Remark 5.16. Any v with $\phi'_v(\alpha) < 0$ is called a *descent direction at α*. By Proposition 5.15 we obtain the following alternative, which is well-known in convex optimization over vector spaces: either there exists a descent direction at α, or $\alpha \in M(\mu)$.

Example 5.17. Assume that α lies in the interior of an edge $[v, w]$.

1. For Fréchet p-means with $p > 1$, item 3. of Proposition 5.15 yields
$$\alpha \in M_p(\mu) \iff \int_{T_{\alpha \to v}} d(\alpha, x)^{p-1} \, \mathrm{d}\mu(x) = \int_{T_{\alpha \to w}} d(\alpha, x)^{p-1} \, \mathrm{d}\mu(x).$$

2. For Fréchet medians,
$$\alpha \in M_1(\mu) \iff \mu(T_{\alpha \to v} \cup \{\alpha\}) \geq \frac{1}{2} \text{ and } \mu(T_{\alpha \to w} \cup \{\alpha\}) \geq \frac{1}{2}.$$
This last optimality condition is reminiscent of the classical characterization of a median on $\mathbb{R}$ as any $m \in \mathbb{R}$ that verifies both $\mu((-\infty, m]) \geq 1/2$ and $\mu([m, \infty)) \geq 1/2$.

By Proposition 5.7, $M(\mu)$ is a nonempty convex subset of T. Under a mild additional assumption on ℓ we obtain the following more precise statement on the geometry of $M(\mu)$.

Proposition 5.18. *If ℓ is increasing, then $M(\mu)$ is a geodesic segment.*

Example 5.19. When $p > 1$ the loss defining the Fréchet p-mean is strictly convex, hence by Proposition 5.7 $M(\mu)$ is a singleton and thus a geodesic segment. For Fréchet medians ($p = 1$), the loss is not strictly convex, however it is increasing and $M(\mu)$ is a geodesic segment. In particular, it cannot contain a 3-spider.

Lastly, the following localization property involves the support of μ (i.e., the smallest closed subset of T that has μ-probability 1 [204, Theorem 2.1]) and it will prove useful later.

Proposition 5.20. *Assume ℓ is increasing and* $\mathrm{supp}(\mu) \subset G$ *where G is a closed convex subset of T. Then $M(\mu) \subset G$.*

5.3 Estimation of Fréchet ℓ-means and statistical results

5.3.1 Estimation setting

Definition 5.21. Let $(X_n)_{n \geq 1}$ be a sequence of i.i.d. T-valued random elements defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, each with distribution μ. For each $n \geq 1$, we define the *empirical measure* $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ and the *empirical objective function* $\widehat{\phi}_n : \alpha \mapsto \frac{1}{n} \sum_{i=1}^{n} \ell(d(\alpha, X_i))$. Minimizers of $\widehat{\phi}_n$ (i.e., elements of $M(\widehat{\mu}_n)$) are called *empirical Fréchet ℓ-means.*

To avoid notational overburden, the directional derivative of $\widehat{\phi}_n$ will be written as $\widehat{\phi}'_v(\alpha)$; the integer n is clear from the context and is therefore omitted.

We assume throughout that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete. By [195, Proposition 5.3.9, Proposition 5.3.13], $M(\widehat{\mu}_n)$ is a measurable closed subset of T, i.e., for all compact subsets $K \subset T$, the subset $\{\omega \in \Omega : M(\widehat{\mu}_n^{\omega}) \cap K = \emptyset\}$ is measurable. Hence, for all open subsets $O \subset T$, $\{M(\widehat{\mu}_n) \cap O = \emptyset\}$ is also measurable. Indeed, write O as the union of a sequence $(K_p)_{p\geq 1}$ of compact subsets of T (such a sequence exists because T is locally compact and Hausdorff). Then, $\{M(\widehat{\mu}_n) \cap O = \emptyset\} = \bigcap_{p\geq 1}\{M(\widehat{\mu}_n) \cap K_p = \emptyset\}$, which is measurable. In particular, for all closed subsets $F \subset T$, $\{M(\widehat{\mu}_n) \subset F\}$ is measurable, which will be useful later (e.g., in Theorem 5.27).

Replacing μ with $\widehat{\mu}_n$ in Proposition 5.7, $\widehat{\phi}_n$ is continuous, $M(\widehat{\mu}_n)$ is non-empty and there exists a measurable selection of $M(\widehat{\mu}_n)$, i.e., a minimizer $\widehat{\alpha}_n$ of $\widehat{\phi}_n$, which is a random variable [7, Theorem 18.19]. This is most useful when $M(\widehat{\mu}_n)$ is not a singleton, which may happen for instance if ℓ is not strictly convex.

5.3.2 A law of large numbers for sets

In terms of sets, estimation is successful if the stochastic set $M(\widehat{\mu}_n)$ gets closer in some sense to the true set $M(\mu)$ as $n \to \infty$. In the works [286, 247, 33, 137, 231, 89] two modes of convergence are considered for the sequence $(M(\widehat{\mu}_n))_{n\geq 1}$.

Definition 5.22.

1. $(M(\widehat{\mu}_n))_{n\geq 1}$ is *strongly consistent in outer limit* [231] (alternatively, *in Kuratowski upper limit* [89] or *in the sense of Ziezold* [286, 137]) if
$$\mathbb{P}\Big(\bigcap_{n\geq 1} \overline{\bigcup_{p\geq n} M(\widehat{\mu}_p)} \subset M(\mu)\Big) = 1.$$
2. $(M(\widehat{\mu}_n))_{n\geq 1}$ is *strongly consistent in one-sided Hausdorff distance* [231, 89] (alternatively, *in the sense of Bhattacharya–Patrangenaru* [33]) if
$$\mathbb{P}\Big(\sup_{\alpha\in M(\widehat{\mu}_n)} \inf_{\beta\in M(\mu)} d(\alpha,\beta) \xrightarrow[n\to\infty]{} 0\Big) = 1.$$

In [126, 231, 89] each of these statements is regarded as a set-valued strong law of large numbers. A caveat about Definition 5.22 is that these notions of closeness between $M(\widehat{\mu}_n)$ and $M(\mu)$ are only one-sided: there might exist some $\alpha_\star \in M(\mu)$ such that the distance of $\alpha_\star$ to $M(\widehat{\mu}_n)$ remains bounded away from 0 with positive probability. However, in the case of a unique Fréchet ℓ-mean (i.e., $M(\mu) = \{\alpha_\star\}$), Definition 5.22 yields strong consistency in the usual sense: for any sequence of measurable selections $(\widehat{\alpha}_n)_{n\geq 1}$, $d(\widehat{\alpha}_n, \alpha_\star) \to 0$ almost surely.

Since the metric space that we consider is compact, the two modes of consistency introduced above are equivalent. In [137, 231], strong consistency is obtained for a wide variety of metric spaces and functions ℓ. As an application of these results to our setting, we obtain the following strong law of large numbers.

Lemma 5.23 (Strong law of large numbers)**.** *$(M(\widehat{\mu}_n))_{n\geq 1}$ is strongly consistent in either of the senses of Definition 5.22.*

5.3.3 Stickiness

In order to simplify the exposition, throughout this subsection we add the requirement that the loss ℓ be increasing, instead of just nondecreasing as in Definition 5.4.

We leverage the geometry of the metric tree T to describe in more detail how $(M(\widehat{\mu}_n))_{n\geq 1}$ converges to $M(\mu)$. To this end, we adapt the concept of stickiness that was introduced in [126] and further explored in [136, 32, 162].

Definition 5.24. Let $c \in T$ with neighboring vertices $v_1, \dots, v_m$. Depending on which of the following disjoint and exhaustive conditions is satisfied, we say that c is:

- *sticky* if $c \in M(\mu)$ and for every $i \in \{1, \dots, m\}$, $\phi'_{v_i}(c) > 0$,
- *partly sticky* if $c \in M(\mu)$ and there exists some $i \in \{1, \dots, m\}$ such that $\phi'_{v_i}(c) = 0$,
- *nonsticky* if c is not in $M(\mu)$.

Remark 5.25. By Proposition 5.15 c is nonsticky if and only if there exists some (unique) $i \in \{1, \dots, m\}$ such that $\phi'_{v_i}(c) < 0$.

Remark 5.26. Originally Hotz et al. [126] defined stickiness in the setting of the Fréchet mean ($\ell : z \mapsto z^2$) on an open book. For us, the compact metric tree that most resembles an open book is the m-spider with center c and leaves $v_1, \dots, v_m$ (if each branch of the m-spider was unbounded, the space would be an open book with spine $\{c\}$). In [126, Definition 2.10], stickiness is defined according to the sign of the quantity

$$m_i = \int_{(c,v_i]} d(c,x)\,\mathrm{d}\mu(x) - \sum_{j\neq i} \int_{(c,v_j]} d(c,x)\,\mathrm{d}\mu(x),$$

which in our notation is exactly $-\frac{1}{2}\phi'_{v_i}(c)$.

When the directional derivative $\phi'_{v_i}(c)$ is nonzero, Proposition 5.14 helps to locate the minimizers of ϕ. As n grows, it is expected that the empirical counterpart $\widehat{\phi}'_{v_i}(c)$ becomes nonzero and has the same sign as $\phi'_{v_i}(c)$ with high probability. It is then possible to obtain identical localization constraints on the empirical Fréchet ℓ-means. The following theorem makes this intuition precise.

Theorem 5.27 (Nonasymptotic empirical stickiness)**.** *Let $c \in T$ and $n \geq 1$ be fixed.*

1. *If c is sticky, then $M(\mu) = \{c\}$ and*

$$\mathbb{P}(M(\widehat{\mu}_n) = \{c\}) \geq 1 - \sum_{i=1}^{m} \exp\left(-\frac{\phi'_{v_i}(c)^2}{2\ell'_+(D)^2} n\right).$$

2. *If c is partly sticky, we let $I = \big\{i = 1, \dots, m : \phi'_{v_i}(c) = 0\big\}$. Then, I has either one or two elements. Moreover, $\{c\} \subset M(\mu) \subset \{c\} \cup \bigcup_{i\in I} T_{c\to v_i}$.*

   (a) If $|I| = 1$, then, writing $I = \{i^*\}$,

$$\mathbb{P}\big(M(\widehat{\mu}_n) \subset \{c\} \cup T_{c\to v_{i^*}}\big) \geq 1 - \sum_{i\neq i^*} \exp\left(-\frac{\phi'_{v_i}(c)^2}{2\ell'_+(D)^2} n\right).$$

(b) *If* $|I| = 2$, *then* $\mu\left(\{c\} \cup \bigcup_{i\in I} T_{c\to v_i}\right) = 1$ *and*

$$\mathbb{P}\Big(M(\widehat{\mu}_n) \subset \{c\} \cup \bigcup_{i\in I} T_{c\to v_i}\Big) = 1.$$

3. *If* c *is nonsticky with* $\phi'_{v_i}(c) < 0$ *for some* $i \in \{1, \dots, m\}$, *then* $M(\mu) \subset T_{c\to v_i}$ *and*

$$\mathbb{P}(M(\widehat{\mu}_n) \subset T_{c\to v_i}) \geq 1 - \exp\left(-\frac{\phi'_{v_i}(c)^2}{2\ell'_+(D)^2} n\right).$$

The exponential bounds of Theorem 5.27 combined with the Borel–Cantelli lemma lead to the following asymptotic result. When c is sticky, with probability 1 the empirical sets of minimizers $M(\widehat{\mu}_n)$ are eventually equal to $\{c\}$. This justifies the use of the adjective "sticky".

Corollary 5.28 (Sticky law of large numbers)**.**

1. *If* c *is sticky, then with probability* 1, $M(\widehat{\mu}_n) = \{c\}$ *for all large enough* n.
2. *If* c *is partly sticky, let* I *as in Theorem 5.27.*
   (a) *If* $|I| = 1$, *then with probability* 1, $M(\widehat{\mu}_n) \subset \{c\} \cup \bigcup_{i\in I} T_{c\to v_i}$ *for all large enough* n.
   (b) *If* $|I| = 2$, *then with probability* 1, $M(\widehat{\mu}_n) \subset \{c\} \cup \bigcup_{i\in I} T_{c\to v_i}$ *for all* $n \geq 1$.
3. *If* c *is nonsticky with* $\phi'_{v_i}(c) < 0$ *for some* $i \in \{1, \dots, m\}$, *then with probability* 1, $M(\widehat{\mu}_n) \subset T_{c\to v_i}$ *for all large enough* n.

The earlier definition of stickiness involves the signs of the directional derivatives at c. It is therefore stated in terms of the landscape of the objective function ϕ around c. The following proposition provides an equivalent formulation of stickiness: c is sticky if and only if the equality $M(\nu) = \{c\}$ holds for every measure ν that is sufficiently close to μ. The notion of stickiness thus has an interpretation in terms of robustness.

We quantify the closeness between two probability measures ν_1, ν_2 using the total variation metric defined as $\mathrm{TV}(\nu_1, \nu_2) = \sup_{B\in\mathcal{B}(T)} |\nu_1(B) - \nu_2(B)|$ and the 1-Wasserstein metric $W_1(\nu_1, \nu_2) = \sup\{\int_T f(x)\,\mathrm{d}\nu_1(x) - \int_T f(x)\,\mathrm{d}\nu_2(x) : f \text{ is 1-Lipschitz}\}$. Total variation is stronger than 1-Wasserstein, in the sense that $W_1(\nu_1, \nu_2) \leq D\,\mathrm{TV}(\nu_1, \nu_2)$, where D is the diameter of T [264, Theorem 6.15]. Note that closeness in W_1 need not imply closeness in total variation.

Proposition 5.29.

1. *c is sticky if and only if there exists* $\varepsilon > 0$ *such that for every probability measure* ν *verifying* $\mathrm{TV}(\nu, \mu) \leq \varepsilon$ *we have* $M(\nu) = \{c\}$.
2. *Under the additional assumption that* ℓ *is differentiable with Lipschitz derivative, 1. holds in* W_1 *instead of* TV.

Remark 5.30. Connections between stickiness and robustness under perturbations were already explored in the context of stratified spaces by Huckemann et al. [136, Section 7], Bhattacharya et al. [32, Proposition 2.8] and most recently by Lammers et al. [162].

Remark 5.31. Note that stickiness can happen even when the distribution μ has a density with respect to the Lebesgue measure on T. As an example, consider the m-spider T_m ($m \geq 3$) with legs of length two. This is the metric tree with one central vertex, connected with all m leaves by edges of length 2. More formally, T_m is defined as $\{1, \dots, m\} \times [0, 2]$ where one identifies all elements of the form $(k, 0), k = 1, \dots, m$ and where $d((k, x), (l, y)) = |x - y|$ if $k = l$, $x + y$ otherwise. Now, let μ be the uniform distribution on $\{(k, x) \in T_m : k = 1, \dots, m, 0 \leq x \leq 1\}$, i.e., the distribution with constant density (with respect to the Lebesgue measure on T_m) that is equal to $1/m$ on the first half of each leg, and 0 everywhere else. Then, a straighforward computation shows that for all $\alpha = (k, x) \in T_m$ with $x \leq 1$, $\phi(\alpha) = \frac{m-1}{m}(L(1 + x) - L(x)) + \frac{1}{m}(L(x) + L(1 - x))$, where $L(u) = \int_0^u \ell(t)\,\mathrm{d}t$, for $u \geq 0$. Denote the latter expression by $F(x)$. Then, for all leaves v of T_m, $\phi'_v((1, 0)) = F'_+(0) = \left(1 - \frac{2}{m}\right)\ell(1) > 0$, since $m \geq 3$. Hence, by the second part of Proposition 5.15, there is a unique Fréchet ℓ-mean $\alpha_\star = (1, 0)$ (i.e., the central vertex of T_m) and it is sticky.

To illustrate Proposition 5.29, perturb μ by defining a distribution ν that is uniform on $\{(k, x) \in T_m : k = 1, \dots, m, 0 \leq x \leq b_k\}$, where $b_1, \dots, b_m \in (0, 2]$ are fixed numbers. Set $b = b_1 + \ldots + b_m$. Now, for ν, a similar computation shows that for all $\alpha = (k, x) \in T_m$ with $x \leq b_k$, $\phi(\alpha) = \frac{1}{b}\left(\sum_{j \neq k}(L(x + b_j) - L(x)) + L(x) + L(b_k - x)\right)$, which we denote by $F_k(x)$. Recall that $\alpha_\star = (1, 0)$ is the central vertex of T_m. Then, if we let v_k be the leaf of T_m on the k-th leg, it holds that $\phi'_{v_k}(\alpha_\star) = (F_k)'_+(0) = \frac{1}{b}\left(\sum_{j \neq k}\ell(b_j) - \ell(b_k)\right)$. If all b_j's are close enough to 1, it hence still holds that $\phi'_{v_k}(\alpha_\star) > 0$ for all $k = 1, \dots, m$, yielding $M(\nu) = \{\alpha_\star\}$, by the second part of Proposition 5.15.

5.4 The special case of Fréchet medians

Now, we restrict our focus to Fréchet medians, i.e., the case where the loss is $\ell : z \mapsto z$. Among the operations research community, the median case has generated the most interest, as it is the most intuitive in applications: the practitioner looks for a new facility on the network that minimizes the average travel time to the demand.

5.4.1 Further descriptive results

In Proposition 5.18 we observed that the set of medians $M_1(\mu)$ was a geodesic segment. Besides, for a discrete measure with uniform weights $\nu = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ on the real line $\mathbb{R}$, it is well-known that $M_1(\nu)$ contains at least one of the x_i. We provide a generalization of this fact on a tree: an extremity of $M_1(\mu)$ is a vertex of T, or it is in the support of μ (i.e., the smallest closed subset of T that has μ-probability 1 [204, Theorem 2.1]).

Proposition 5.32. *Let α_1 and α_2 denote the endpoints of the geodesic segment $M_1(\mu)$, i.e., $M_1(\mu) = [\alpha_1, \alpha_2]$. The following inclusion holds: $\{\alpha_1, \alpha_2\} \subset \mathcal{V} \cup \operatorname{supp}(\mu)$, where $\mathcal{V}$ is the set of vertices of T.*

Remark 5.33. Hakimi [109] states a weaker statement: when μ is discrete and supported on $\mathcal{V}$, he proves that $M_1(\mu) \cap \mathcal{V} \neq \emptyset$.

A measure ν on $\mathbb{R}$ has at least two medians if and only if there exists $m_1 < m_2$ such that $\nu((-\infty, m_1]) = \nu([m_2, -\infty)) = \frac{1}{2}$ [223, Corollary 2.6]. The next proposition is an extension of this fact to metric trees. It will prove useful later, so as to guarantee that $M_1(\mu)$ is a singleton.

Proposition 5.34. *μ has more than one Fréchet median if and only if there exist G_1, G_2 two disjoint closed convex subsets of T such that $\mu(G_1) = \mu(G_2) = \frac{1}{2}$. Consequently, if the support of μ is a connected subset of T (e.g., if μ has a positive density with respect to the Lebesgue measure on T), then μ has a unique Fréchet median.*

In the next subsections, a convex subset G is known that contains $M_1(\mu)$. It is fruitful to consider the metric projection on G (for the definition and basic properties of the metric projection on a closed convex subset of a Hadamard space see, e.g., [15, Theorem 2.1.12]) and transform μ into a measure supported on G, hence the following definition.

Definition 5.35. Let G be a closed convex subset of T, and let $\pi : T \to T$ denote the metric projection on G. We denote by $\pi\#\mu$ the pushforward measure of μ by π, and we write $\phi_{\pi\#\mu}$ for the objective function corresponding to $\pi\#\mu$.

Remark 5.36. Although the image of π is G, we define π as a map with codomain T so that $\pi\#\mu$ remains naturally a Borel measure on T.

The following technical proposition gathers statements on $\pi\#\mu$ that will prove useful in the next subsections.

Proposition 5.37.

1. *The set $\pi(T \setminus G)$ is finite. We write $\pi(T \setminus G) = \{v_1, \ldots, v_m\}$ and we define the sets $T_i = \pi^{-1}(\{v_i\})$.*
2. *$\pi\#\mu$ is a Borel measure on T. It rewrites explicitly as $\pi\#\mu = \mu_{|\mathring{G}} + \sum_{i=1}^m \mu(T_i)\delta_{v_i}$.*
3. *$M_1(\pi\#\mu)$ is a subset of G.*
4. *ϕ and $\phi_{\pi\#\mu}$ differ by an additive constant over G. More precisely,*
$$\forall \alpha \in G, \ \phi(\alpha) = \phi_{\pi\#\mu}(\alpha) + \sum_{i=1}^m \int_{T_i} d(v_i, x)\, \mathrm{d}\mu(x).$$
5. *The following inclusion holds: $M_1(\mu) \cap G \subset M_1(\pi\#\mu)$.*
6. *Assume that $M_1(\mu) \subset G$. Then $M_1(\mu) = M_1(\pi\#\mu)$.*

5.4.2 Further statistical results

We are now ready to return to the statistical side. In what follows we assume that $M_1(\mu) = \{\alpha_\star\}$, i.e., there is a unique Fréchet median $\alpha_\star$. A sufficient conditions for uniqueness was given in Proposition 5.34. We are thus in the classical setting of parameter estimation. However the empirical set $M_1(\widehat{\mu}_n)$ may not be a singleton; we consider therefore an arbitrary sequence of measurable selections $(\widehat{\alpha}_n)_{n\geq 1}$, as explained in Section 5.3.1.

Recall, from Definition 5.24, that $\alpha_\star$ must be either sticky or partly sticky, since it is in $M(\mu)$. When $\alpha_\star$ is sticky, Corollary 5.28 asserts that $(\widehat{\alpha}_n)_{n\geq 1}$ converges almost surely to $\alpha_\star$ at an arbitrarily fast rate. From a statistical standpoint the sticky case is thus fully elucidated, and in the rest of this section, we assume that $\alpha_\star$ is partly sticky.

Let $v_1, \dots, v_m$ denote the neighboring vertices of $\alpha_\star$. In Theorem 5.27 it was seen, for the partly sticky case, that there are at most two indices $i \in \{1, \dots, m\}$ such that $\phi'_{v_i}(\alpha_\star) = 0$, i.e., $\mu(T_{\alpha_\star \to v}) = 1/2$. Therefore, we study the properties of $\widehat{\alpha}_n$ in two distinct cases, which we denominate as follows.

Definition 5.38. We say that $\alpha_\star$ is *one-sidedly partly sticky* if there is a unique i such that $\phi'_{v_i}(\alpha_\star) = 0$. Otherwise, we say that $\alpha_\star$ is *two-sidedly partly sticky.*

The two-sided partly sticky case

Assume without loss of generality that $\phi'_{v_1}(\alpha_\star) = \phi'_{v_2}(\alpha_\star) = 0$ and for all $i \geq 3$, $\phi'_{v_i}(\alpha_\star) > 0$ (recall that $v_1, \dots, v_m$ are the vertices adjacent to $\alpha_\star$). In other words, $\mu(T_{\alpha_\star \to v_1}) = \mu(T_{\alpha_\star \to v_2}) = 1/2$ thus $\mu(T_{\alpha_\star \to v_i}) = 0$ for all $i \geq 3$ and all the mass of μ is supported on $T_{\alpha_\star \to v_1} \cup T_{\alpha_\star \to v_2}$ (in particular, $\alpha_\star$ is not an atom of μ). Note that $\alpha_\star$ may be in the interior of an edge, in which case $m = 2$.

By the law of large numbers in Lemma 5.23, it is known that $\mathbb{P}(d(\widehat{\alpha}_n, \alpha_\star) \to 0) = 1$. By Theorem 5.27, we know additionally that $\widehat{\alpha}_n \in \{\alpha_\star\} \cup T_{\alpha_\star \to v_1} \cup T_{\alpha_\star \to v_2}$ almost surely. As a consequence, $\widehat{\alpha}_n$ is eventually in the geodesic segment $[v_1, v_2]$ with probability 1. The closed convex subset on which we will project the measure μ and the data is therefore $G = [v_1, v_2]$. By assumption it contains the true median $\alpha_\star$.

G is naturally isometric to the compact interval $[-d(\alpha_\star, v_1), d(\alpha_\star, v_2)] \subset \mathbb{R}$, where $\alpha_\star$ is sent on 0. By pushing forward again, this time with target space $\mathbb{R}$, we replace the problem with the analysis of sample medians on the real line. This motivates the next definition and the lemma that follows.

Definition 5.39. Let $\gamma : [-d(\alpha_\star, v_1), d(\alpha_\star, v_2)] \to G = [v_1, v_2]$ denote the unit-speed geodesic from v_1 to v_2, and let I denote the inverse of γ. We define a new sequence of i.i.d. real valued random variables $Y_1, Y_2, \dots$ as $Y_i = I(\pi(X_i)), i \geq 1$. Moreover, for each $n \geq 1$, we set $\widehat{m}_n = I(\pi(\widehat{\alpha}_n))$ and the event $\Omega_n = \{\widehat{\alpha}_n \in [v_1, v_2]\}$. We denote by ν the distribution of the Y_i's, i.e., the pushforward measure $(I \circ \pi)\#\mu$, and by Y a random variable with distribution ν.

Remark 5.40. For convenience, we also use the notation $M_1(\cdot)$ to denote the set of medians of a measure on $\mathbb{R}$ (which is not a metric tree by our definition).

Lemma 5.41.
1. *ν is a Borel measure on $\mathbb{R}$ supported on the segment $[-d(\alpha_\star, v_1), d(\alpha_\star, v_2)]$, $Y_1, Y_2, \dots$ are i.i.d. with distribution ν and $M_1(\nu) = \{0\}$.*
2. *On the event Ω_n, $\widehat{m}_n \in M_1\left(\frac{1}{n}\sum_{k=1}^{n} \delta_{Y_k}\right)$ and $d(\widehat{\alpha}_n, \alpha_\star) = |\widehat{m}_n - 0| = |\widehat{m}_n|$.*
3. *For $i \in \{1, 2\}$, $0 < \mu((\alpha_\star, v_i)) \leq 1/2$ and*

$$\mathbb{P}(\Omega_n) \geq 1 - \left(1 - 4\mu\left((\alpha_\star, v_1)\right)^2\right)^{n/2} - \left(1 - 4\mu\left((\alpha_\star, v_2)\right)^2\right)^{n/2}.$$

Before we can state a central limit theorem, we define the following function.

Definition 5.42. The *two-sided branch mass function* Δ is

$$\begin{aligned}\Delta\colon [-d(\alpha_\star, v_1), d(\alpha_\star, v_2)] &\to [0,\infty)\\ t &\mapsto \mu\left((\alpha_\star, \gamma_t)\right).\end{aligned}$$

This function plays an important role: The next result shows that its rate of decay as $t \to 0$ drives the rate of convergence of $\widehat{\alpha}_n$ and the asymptotic distribution of a properly rescaled version of $\widehat{m}_n$.

Theorem 5.43 (Two-sided sticky central limit theorem)**.** *Assume that Δ has the following asymptotic expansion as $t \to 0$:*

$$\Delta(t) = K|t|^a + o(|t|^a), \tag{5.5}$$

for some constants $a > 0$ and $K > 0$. Let Z denote a random variable with the standard normal distribution.

1. *$n^{1/(2a)}\widehat{m}_n$ converges in distribution to the random variable* $\operatorname{sgn}(Z)\left(\frac{|Z|}{2K}\right)^{1/a}$.
2. *$n^{1/(2a)}d(\widehat{\alpha}_n, \alpha_\star)$ converges in distribution to the random variable* $\left(\frac{|Z|}{2K}\right)^{1/a}$.

Corollary 5.44. *Assume that (5.5) holds with $a = 1$ and positive K. Then $\sqrt{n}\widehat{m}_n$ is asymptotically normal with asymptotic variance $\frac{1}{4K^2}$.*

Remark 5.45.
1. For instance, if μ has a density f with respect to the Lebesgue measure on T (see Definition 5.2) and f is positive and continuous, then (5.5) holds with $a = 1$ and $K = f(\alpha_\star)$, which is reminiscent of the standard real case.
2. The assumption in Corollary 5.44 is equivalently formulated as $t \mapsto \operatorname{sgn}(t)\Delta(t)$ being differentiable at 0 with positive derivative K.

The function Δ also plays a key role in the concentration bound stated next.

Theorem 5.46. *Let $n \geq 1$ be fixed. For t such that $0 < t \leq \min(d(\alpha_\star, v_1), d(\alpha_\star, v_2))$, the quantities $\Delta(t)$ and $\Delta(-t)$ are both in $(0, \frac{1}{2}]$, and the following concentration bound holds:*

$$\mathbb{P}(d(\widehat{\alpha}_n, \alpha_\star) \geq t) \leq \left(1 - 4\Delta^2(t)\right)^{n/2} + \left(1 - 4\Delta^2(-t)\right)^{n/2}. \tag{5.6}$$

More generally, for every $t > 0$:

$$\begin{aligned}\mathbb{P}(d(\widehat{\alpha}_n, \alpha_\star) \geq t) \leq \mathbb{1}_{t \leq d(\alpha_\star, v_1)}\left(1 - 4\Delta^2(t)\right)^{n/2} + \mathbb{1}_{t \leq d(\alpha_\star, v_2)}\left(1 - 4\Delta^2(-t)\right)^{n/2}\\ + (\mathbb{1}_{t > d(\alpha_\star, v_1)} + \mathbb{1}_{t > d(\alpha_\star, v_2)})\mathbb{P}(\Omega_n^c).\end{aligned}$$

The one-sided partly sticky case

We turn to the partly sticky case with $\phi'_{v_1}(\alpha_\star) = 0$ and for all $i \geq 2$, $\phi'_{v_i}(\alpha_\star) > 0$. In other words, $\mu(T_{\alpha_\star \to v_1}) = 1/2$ and $\mu(T_{\alpha_\star \to v_i}) < 1/2$, for all $i \geq 2$. In the sequel, we denote by $\varepsilon = \min_{2 \leq i \leq m} (1/2 - \mu(T_{\alpha_\star \to v_i})) > 0$.

Note that in the two-sided partly sticky case, it held that

$$\widehat{\alpha}_n \in \{\alpha_\star\} \cup \bigcup_{i \in I} T_{\alpha_\star \to v_i} \tag{5.7}$$

almost surely, where $I = \{1, 2\}$. Here, $I = \{1\}$ and the next result shows that (5.7) no longer holds almost surely, but with exponentially large probability.

Proposition 5.47. *If $n \geq 4$, then it holds that $\widehat{\alpha}_n \in \{\alpha_\star\} \cup T_{\alpha_\star \to v_1}$ with probability at least $1 - 2e^{-n\varepsilon^2}$.*

Next, we proceed similarly to the two-sided partly sticky case. Here, the closed convex subset on which we project is $G = [\alpha_\star, v_1]$.

Definition 5.48. Let $\gamma_1 : [0, d(\alpha_\star, v_1)] \to [\alpha_\star, v_1]$ denote the unit-speed geodesic from $\alpha_\star$ to v_1, and let I denote the inverse of γ_1. For each $n \geq 1$ we define $Y_n = I(\pi(X_n))$, $\widehat{m}_n = I(\pi(\widehat{\alpha}_n))$ and the event $\Omega_n = \left\{\widehat{\alpha}_n \in [\alpha_\star, v_1]\right\}$. We denote by ν the pushforward measure $(I \circ \pi)\#\mu$, and by Y a random variable with distribution ν.

Lemma 5.49.
1. *ν is a Borel measure on $\mathbb{R}$ supported on the compact interval $[0, d(\alpha_\star, v_1)]$, the Y_n are i.i.d. with distribution ν and $M_1(\nu) = \{0\}$.*
2. *On the event Ω_n, $\widehat{m}_n \in M_1\left(\frac{1}{n}\sum_{k=1}^n \delta_{Y_k}\right)$ and $d(\widehat{\alpha}_n, \alpha_\star) = |\widehat{m}_n - 0| = \widehat{m}_n$.*
3. *The following inequalities hold: $0 < \mu((\alpha_\star, v_1)) \leq 1/2$, $0 < \phi'_{v_i}(\alpha_\star) \leq 1$ for every $i \geq 2$, and*

$$\mathbb{P}(\Omega_n) \geq 1 - \left(1 - 4\mu\left((\alpha_\star, v_1)\right)^2\right)^{n/2} - \sum_{i=2}^{m} \Big(4\mu\left(T_{\alpha_\star \to v_i}\right)\left(1 - \mu\left(T_{\alpha_\star \to v_i}\right)\right)\Big)^{n/2}.$$

Similarly as in the proof of Proposition 5.47, the sum on the right hand side of the inequality above can be bounded from above by $2e^{-n\varepsilon^2}$ (as soon as $n \geq 4$), which does not depend on m.

Definition 5.50. For each $i \in \{1, \ldots, m\}$ we define $\gamma_i : [0, d(\alpha_\star, v_i)] \to [\alpha_\star, v_i]$ the unit-speed geodesic from $\alpha_\star$ to v_i and the i-th *branch mass function*

$$\begin{aligned} \delta_i \colon [0, d(\alpha_\star, v_i)] &\to [0, \infty) \\ t &\mapsto \mu\left((\alpha_\star, \gamma_{i,t})\right). \end{aligned}$$

Theorem 5.51 (One-sided sticky central limit theorem)**.** *Assume that δ_1 has the following expansion as $t \to 0^+$:*

$$\delta_1(t) = Kt^a + o(t^a),$$

for some constants $a > 0$ and $K > 0$. Let Z denote a random variable with the standard normal distribution.

1. $n^{1/(2a)}\widehat{m}_n$ *converges in distribution to the random variable* $\frac{1}{2K}\max(0,Z)^{1/a}$.

2. $n^{1/(2a)}d(\widehat{\alpha}_n,\alpha_\star)$ *converges in distribution to the random variable* $\frac{1}{2K}\max(0,Z)^{1/a}$.

Remark 5.52. The rate of convergence $n^{1/(2a)}$ is the same as in the two-sided partly sticky case. In contrast however, the fluctuations are one-sided along the edge $[\alpha_\star, v_1]$.

Theorem 5.53. *Let $n \geq 1$ be fixed. For all positive numbers t such that $0 < t \leq \min_{1\leq i\leq m} d(\alpha_\star, v_i)$, we have $\delta_1(t) \in (0, \frac{1}{2}]$, $2\delta_i(t) + \phi'_{v_i}(\alpha_\star) \in (0,1]$ for each $i \geq 2$, and the following concentration bound holds:*

$$\mathbb{P}(d(\widehat{\alpha}_n,\alpha_\star) \geq t) \leq \left(1-4\delta_1^2(t)\right)^{n/2} + \sum_{i=2}^{m}\left(1-\left(2\delta_i(t)+\phi'_{v_i}(\alpha_\star)\right)^2\right)^{n/2}. \tag{5.8}$$

More generally, for every $t > 0$:

$$\mathbb{P}(d(\widehat{\alpha}_n,\alpha_\star) \geq t) \leq \mathbb{1}_{t\leq d(\alpha_\star,v_1)}\left(1-4\delta_1^2(t)\right)^{n/2} + \sum_{i=2}^{m}\mathbb{1}_{t\leq d(\alpha_\star,v_i)}\left(1-\left(2\delta_i(t)+\phi'_{v_i}(\alpha_\star)\right)^2\right)^{n/2}$$
$$+\sum_{i=1}^{m}\mathbb{1}_{t>d(\alpha_\star,v_i)}\mathbb{P}(\Omega_n^c).$$

Remark 5.54. Since $\phi'_{v_1}(\alpha_\star) = 0$, half of the total mass from μ is on $T_{\alpha_\star\to v_1}$ (recall that $\phi'_{v_i}(\alpha_\star) = 1-2\mu(T_{\alpha_\star\to v_i})$), while the other half is shared among the other $m-1$ branches departing from $\alpha_\star$. If the branch in direction v_i with $i \geq 2$ has very low mass, i.e., if $\mu(T_{\alpha_\star\to v_i})$ is small, then $\phi'_{v_i}(\alpha_\star)$ is close to 1 and the contribution of the term $\left(1-\left(2\delta_i(t)+\phi'_{v_i}(\alpha_\star)\right)^2\right)^{n/2}$ in (5.8) is exponentially small.

5.5 Conclusion

In this work, we considered location estimation on a metric tree, that is assumed to be bounded, with finitely many vertices. It seems that all our results can be easily extended to the case of an unbounded metric tree. However, an unbounded tree would only add technicalities that would rather put shade on the intrinsic phenomena that we were aiming at exhibiting here (namely, stickiness).

Directions for future research include establishing limit distribution results and concentration inequalities for Fréchet ℓ-means other than the median, as well as extending our understanding of the stickiness phenomenon beyond metric trees.

5.6 Proofs

5.6.1 Proofs for Section 5.2

Lemma 5.55. *Let $f : [0,\infty) \to [0,\infty)$ be nondecreasing and $a \geq 0$. The function $\ell : z \mapsto a + \int_0^z f(t)\,\mathrm{d}t$ is convex.*

Proof of Lemma 5.55. The following inequality between slopes holds: for any $0 \le z_1 < z_2 < z_3$,

$$\frac{\ell(z_2)-\ell(z_1)}{z_2-z_1} = \frac{1}{z_2-z_1}\int_{z_1}^{z_2} f(t)\,\mathrm{d}t \le f(z_2) \le \frac{1}{z_3-z_2}\int_{z_2}^{z_3} f(t)\,\mathrm{d}t = \frac{\ell(z_3)-\ell(z_2)}{z_3-z_2}, \tag{5.9}$$

thus ℓ is convex [203, Proposition 6.2.1]. □

Proof of Lemma 5.6. 1. Since ℓ is convex, it has finite left- and right-derivative at each $z > 0$, with ℓ'_- and ℓ'_+ being nondecreasing [203, Proposition 6.2.7]. Since ℓ is nondecreasing, ℓ'_- and ℓ'_+ are nonnegative, and the function $z \mapsto (\ell(z)-\ell(0))/z$ is bounded below by 0, hence $\ell'_+(0)$ is finite.

2. For a fixed $Z > 0$ and $0 \le z_1 < z_2 \le Z$, we have the estimate

$$0 \le \frac{\ell(z_2)-\ell(z_1)}{z_2-z_1} \le \ell'_-(z_2) \le \ell'_-(Z),$$

thus ℓ is locally Lipschitz.

3. By the last point, ℓ is absolutely continuous on compact intervals. Let $Z > 0$ be fixed. By the fundamental theorem of calculus [227, Theorem 7.18], ℓ is differentiable a.e. on $[0, Z]$ and for every $z \in [0, Z]$, $\ell(z) - \ell(0) = \int_0^z f_Z(t)\,\mathrm{d}t$, where f_Z denotes a derivative of ℓ. Since f_Z and ℓ'_+ coincide a.e. on $[0, Z]$, we obtain $\ell(z) - \ell(0) = \int_0^z \ell'_+(t)\,\mathrm{d}t$. We proceed similarly with ℓ'_-. □

Proof of Proposition 5.7. 1. Since T is bounded and ℓ is nondecreasing, $\ell(d(\alpha, x)) \le \ell(D)$, hence ϕ is well-defined. That ϕ is continuous follows from continuity of ℓ (seen in Lemma 5.6) and the dominated convergence theorem. Since T is Hadamard, by [203, Example 8.4.7 (i)] the map $\alpha \mapsto d(\alpha, x)$ is convex for each $x \in T$, and by the convexity and monotonicity of ℓ we obtain convexity of $\alpha \mapsto \ell(d(\alpha, x))$. That ϕ is convex follows by integration.

2. T is compact and ϕ is continuous, hence $M(\mu)$ is nonempty. By [15, Example 2.1.3], $M(\mu)$ is closed and convex.

3. If ℓ'_+ is increasing, the inequality (5.9) between slopes is strict, hence ℓ is strictly convex. If ℓ'_+ is not increasing, there exists an open interval $I \subset [0, \infty)$ where ℓ'_+ is equal to some constant C. By [203, Proposition 6.2.7], ℓ'_- is also equal to C, hence ℓ is differentiable over I with derivative C, thus ℓ is affine over I and ℓ is not strictly convex.

We suppose now that ℓ is strictly convex. Assume for the sake of contradiction that ϕ is not strictly convex: there exists a geodesic $\gamma : [0,1] \to T$ and $t \in (0,1)$ such that $\gamma_0 \neq \gamma_1$ and $\phi(\gamma_t) = (1-t)\phi(\gamma_0) + t\phi(\gamma_1)$, i.e.,

$$0 = \int_T \left((1-t)\ell(d(\gamma_0, x)) + t\ell(d(\gamma_1, x)) - \ell(d(\gamma_t, x))\right)\mathrm{d}\mu(x).$$

The function $x \mapsto (1-t)\ell(d(\gamma_0, x)) + t\ell(d(\gamma_1, x)) - \ell(d(\gamma_t, x))$ is thus nonnegative and has integral 0. Consequently, there exists $x_\star \in T$ such that

$$\ell(d(\gamma_t, x_\star)) = (1-t)\ell(d(\gamma_0, x_\star)) + t\ell(d(\gamma_1, x_\star)). \tag{5.10}$$

Since ℓ is strictly convex and nondecreasing, the function $\alpha \mapsto \ell(d(\alpha, x_\star))$ is strictly convex as well; this contradicts (5.10). By [203, Proposition 8.4.5] $M(\mu)$ is a singleton. □

Proof of Proposition 5.9. Decomposing the distance $d(\alpha', x)$ with respect to the location of x in the tree, we obtain the equality

$$\begin{aligned}\phi(\alpha') - \phi(\alpha) = &\int_{T\setminus T_{\alpha\to v}} \Big(\ell\left(d(\alpha,x)+d(\alpha,\alpha')\right) - \ell(d(\alpha,x))\Big)\,\mathrm{d}\mu(x) \\ &+ \int_{T_{\alpha'\to v}} \Big(\ell\left(d(\alpha,x)-d(\alpha,\alpha')\right) - \ell(d(\alpha,x))\Big)\,\mathrm{d}\mu(x) \\ &+ \int_{(\alpha,\alpha']} \Big(\ell(d(\alpha',x)) - \ell(d(\alpha,x))\Big)\,\mathrm{d}\mu(x).\end{aligned} \tag{5.11}$$

To obtain the limit (5.4), we consider a sequence $(\alpha'_n)_{n\geq 1}$ of points in $(\alpha, v]$ that converges to α, and we apply the dominated convergence theorem to each integral in Equation (5.11). The domination follows from the following estimate: for $\alpha' \neq \alpha$ and $x \in T$, by the convexity of ℓ:

$$\frac{\ell(d(\alpha',x)) - \ell(d(\alpha,x))}{d(\alpha',\alpha)} \leq \ell'_+(d(\alpha',x))\frac{d(\alpha',x)-d(\alpha,x)}{d(\alpha',\alpha)} \leq \ell'_+(D),$$

hence by symmetry

$$\frac{|\ell(d(\alpha',x)) - \ell(d(\alpha,x))|}{d(\alpha',\alpha)} \leq \ell'_+(D).$$

□

Proof of Proposition 5.14. 1. Since $\phi'_v(\alpha_0) < 0$, there exists a one-sided neighborhood N of α_0 such that $N \subset (\alpha_0, v]$ and $\alpha' \in N \implies \phi(\alpha') < \phi(\alpha_0)$. For the sake of contradiction assume the existence of $\alpha_\star \in M(\mu) \cap (T \setminus T_{\alpha_0\to v})$. Fix some $\alpha' \in N$ and let $\gamma : [0,1] \to [\alpha', \alpha_\star]$ be the geodesic from α' to $\alpha_\star$. For some $t \in (0,1)$, $\gamma(t) = \alpha_0$, thus

$$\phi(\alpha_0) = \phi(\gamma_t) \leq (1-t)\phi(\alpha') + t\phi(\alpha_\star) < (1-t)\phi(\alpha_0) + t\phi(\alpha_\star),$$

hence $\phi(\alpha_0) < \phi(\alpha_\star)$, a contradiction.

2. There exists a one-sided neighborhood N of α_0 such that $N \subset (\alpha_0, v]$ and $\alpha' \in N \implies \phi(\alpha') > \phi(\alpha_0)$. For the sake of contradiction assume the existence of $\alpha_\star \in M(\mu) \cap T_{\alpha_0\to v}$. Let $\gamma : [0,1] \to [\alpha_0, \alpha_\star]$ be the geodesic from α_0 to $\alpha_\star$. For small enough positive t, $\gamma(t)$ is in N, thus

$$\phi(\alpha_0) < \phi(\gamma_t) \leq (1-t)\phi(\alpha_0) + t\phi(\alpha_\star),$$

hence $\phi(\alpha_0) < \phi(\alpha_\star)$, a contradiction.

3. We let $\alpha \in T \setminus \{\alpha_0\}$ be arbitrary and we show that $\phi(\alpha_0) \leq \phi(\alpha)$.

Consider first the case where $\alpha \in T_{\alpha_0\to v}$. Letting $\gamma : [0,1] \to [\alpha_0, \alpha]$ be the geodesic from α_0 to α, $\phi(\gamma_t) \leq (1-t)\phi(\alpha_0) + t\phi(\alpha)$, thus for each $t > 0$

$$\frac{\phi(\gamma_t) - \phi(\alpha_0)}{t} \leq \phi(\alpha) - \phi(\alpha_0). \tag{5.12}$$

Since for small enough t we have $\gamma_t \in (\alpha_0, v]$, passing to the limit yields $\phi'_v(\alpha_0) \leq \phi(\alpha) - \phi(\alpha_0)$, hence $\phi(\alpha_0) \leq \phi(\alpha)$.

If $T \setminus (T_{\alpha_0 \to v} \cup \{\alpha_0\})$ is empty, the proof is over. Otherwise we pick w in this set and we consider the case where $\alpha \in T \setminus (T_{\alpha_0 \to v} \cup \{\alpha_0\})$. With the geodesic γ from α_0 to α, Equation (5.12) still holds and taking the limit yields $\phi'_w(\alpha_0) \leq \phi(\alpha) - \phi(\alpha_0)$. For α', α'' such that $\alpha' \in (\alpha_0, v]$ and $\alpha'' \in (\alpha_0, w]$, letting ψ denote the geodesic from α'' to α' and $t = d(\alpha_0, \alpha'')/d(\alpha', \alpha'')$, the convexity inequality $\phi(\psi_t) \leq (1-t)\phi(\alpha'') + t\phi(\alpha')$ rewrites as

$$\frac{\phi(\alpha_0) - \phi(\alpha')}{d(\alpha_0, \alpha')} \leq \frac{\phi(\alpha'') - \phi(\alpha_0)}{d(\alpha'', \alpha_0)}. \tag{5.13}$$

Taking limits, we obtain $0 = -\phi'_v(\alpha_0) \leq \phi'_w(\alpha_0)$, hence $0 \leq \phi(\alpha) - \phi(\alpha_0)$. □

Proof of Proposition 5.15. 1. If $\alpha \in M(\mu)$, condition (b) follows from nonnegativity of the numerator in (5.4). Suppose that every neighboring vertex v satisfies $\phi'_v(\alpha) \geq 0$. If all the $\phi'_v(\alpha)$ are positive, then by combining the inclusions of Proposition 5.14 we have $M(\mu) = \{\alpha\}$. Otherwise $\phi'_v(\alpha) = 0$ for some v and Proposition 5.14 yields $\alpha \in M(\mu)$.

2. Assume that for all $v \in T \setminus \{\alpha\}$, $\phi'_v(\alpha) > 0$. Then, by Proposition 5.14, $M(\mu) \subset T \setminus T_{\alpha \to v}$, for all $v \neq \alpha$, i.e., $M(\mu) \subset \{\alpha\}$. It also holds that $\alpha \in M(\mu)$ by the first part of this proposition, yielding the result.

3. For convenience, let us abuse notation and write that ℓ is also differentiable at 0 with $\ell'(0) = \ell'_+(0)$. We define $S(\alpha) = \int_T \ell'(d(\alpha, x))\,\mathrm{d}\mu(x)$. By the assumption, α has two neighboring vertices: v and w. By 1.(c) and the differentiability of ℓ, $\alpha \in M(\mu)$ if and only if

$$\int_{T_{\alpha \to v}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x) \leq \frac{S(\alpha)}{2} \text{ and } \int_{T_{\alpha \to w}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x) \leq \frac{S(\alpha)}{2}.$$

Note additionally that

$$S(\alpha) = \int_{T_{\alpha \to v}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x) + \int_{T_{\alpha \to w}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x) + \int_{\{\alpha\}} \ell'(d(\alpha, x))\,\mathrm{d}\mu(x).$$

Since $\mu(\{\alpha\})\ell'(0) = \mu(\{\alpha\})\ell'_+(0) = 0$, the rightmost integral is 0, and the claim follows. □

Proof of Proposition 5.18. Assume for the sake of contradiction that $M(\mu)$ is not a geodesic segment. Then it contains a 3-spider G with center c (a vertex of T) and outer vertices v_1, v_2, v_3 (which may not be vertices of T). For $i \in \{1, 2, 3\}$, since $c \in M(\mu)$ we must have $\phi'_{v_i}(c) \geq 0$. Furthermore, if $\phi'_{v_i}(c) > 0$ then by Proposition 5.14 we would have $M(\mu) \subset T \setminus T_{c \to v_i}$ which contradicts $v_i \in M(\mu)$, thus $\phi'_{v_i}(c) = 0$. Since ℓ'_+ is nonnegative and $\ell'_+(z) \geq \ell'_-(z)$ holds for each $z > 0$ we obtain the bound:

$$\begin{aligned}
0 = \phi'_{v_i}(c) &= \int_{\{c\}} \ell'_+(d(c, x))\,\mathrm{d}\mu(x) + \int_{T \setminus (T_{c \to v_i} \cup \{c\})} \ell'_+(d(c, x))\,\mathrm{d}\mu(x) - \int_{T_{c \to v_i}} \ell'_-(d(c, x))\,\mathrm{d}\mu(x) \\
&\geq \int_{T \setminus \{c\}} \ell'_-(d(c, x))\,\mathrm{d}\mu(x) - 2\int_{T_{c \to v_i}} \ell'_-(d(c, x))\,\mathrm{d}\mu(x),
\end{aligned}$$

hence $\int_{T_{c\to v_i}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) \geq \frac{1}{2}\int_{T\setminus\{c\}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x)$. Summing these inequalities, we find

$$\int_{T\setminus\{c\}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) \geq \sum_{i=1}^{3}\int_{T_{c\to v_i}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) \geq \frac{3}{2}\int_{T\setminus\{c\}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x), \tag{5.14}$$

which yields $\int_{T\setminus\{c\}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) = 0$ and $\mathbb{1}_{T\setminus\{c\}}(x)\ell'_-(d(c,x)) = 0$ for μ-almost every x. Since ℓ is increasing, $\ell_-(z) > 0$ holds for each $z > 0$, thus $\mu = \delta_c$ and $\phi'_{v_1}(c) = 1$. This is a contradiction, hence $M(\mu)$ is a geodesic segment. □

Proof of Proposition 5.20. Since T is a Hadamard space and G is closed and convex, the metric projection on G is well-defined [15, Theorem 2.1.12], and we denote it by π. Fix $\alpha \in T \setminus G$ and let $x \in G$. By the Pythagorean inequality [15, Theorem 2.1.12 (ii)] and the strict monotonicity of ℓ,

$$\ell(d(\alpha, x)) > \ell(d(\pi(\alpha), x)). \tag{5.15}$$

Since $\mathrm{supp}(\mu) \subset G$, we have further $\phi(\alpha) = \int_G \ell(d(\alpha,x))\,\mathrm{d}\mu(x) > \int_G \ell(d(\pi(\alpha),x))\,\mathrm{d}\mu(x) = \phi(\pi(\alpha))$, where the strict inequality is a consequence of (5.15). As a result, $\alpha \in T \setminus M_1(\mu)$. □

5.6.2 Proofs for Section 5.3

Proof of Lemma 5.23. [137, Theorem A.3] or [231, Theorem 3.2]. □

Proof of Theorem 5.27. 1. Let $i \in \{1,\dots,m\}$ be fixed. For each $k \in \{1,\dots,n\}$ we define the random variables

$$Y_k = \mathbb{1}_{T\setminus T_{c\to v_i}}(X_k)\ell'_+(d(c,X_k)) - \mathbb{1}_{T_{c\to v_i}}(X_k)\ell'_-(d(c,X_k)),$$

so that $\widehat{\phi}'_{v_i}(c) = \frac{1}{n}\sum_{k=1}^n Y_k$, the Y_k are i.i.d. and $|Y_k| \leq \ell'_+(d(c,X_k)) \leq \ell'_+(D)$. Since c is sticky, $\phi'_{v_i}(c) > 0$ and by Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}(\widehat{\phi}'_{v_i}(c) \leq 0) &= \mathbb{P}\big(-\widehat{\phi}'_{v_i}(c) - (-\phi'_{v_i}(c)) \geq \phi'_{v_i}(c)\big)\\ &\leq \exp\left(-\frac{\phi'_{v_i}(c)^2}{2\ell'_+(D)^2}n\right).\end{aligned}$$

We note that $\ell'_+(D) > 0$ since ℓ is increasing. Proposition 5.14 yields the equality $M(\mu) = \{c\}$ as well as the inclusion $\bigcap_{i=1}^m \{\widehat{\phi}'_{v_i}(c) > 0\} \subset \{M(\widehat{\mu}_n) = \{c\}\}$. By a union bound,

$$\mathbb{P}(M(\widehat{\mu}_n) = \{c\}) \geq 1 - \sum_{i=1}^{m}\exp\left(-\frac{\phi'_{v_i}(c)^2}{2\ell'_+(D)^2}n\right).$$

2. For the partly sticky case, the implication $i \notin I \implies \phi'_{v_i}(c) > 0$ and the equality

$$\bigcap_{i\notin I}(T \setminus T_{c\to v_i}) = \{c\} \cup \bigcup_{i\in I} T_{c\to v_i}$$

justify the inclusion $M(\mu) \subset \{c\} \cup \bigcup_{i\in I} T_{c\to v_i}$. The bound on the cardinality of I follows from the argument that led to (5.14).

When $|I| = 1$, the proof of the exponential bound is similar to the sticky case. Now, assume that $I = \{1, 2\}$. The argument that led to (5.14) yields

$$\int_{T_{c\to v_1}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) = \int_{T_{c\to v_2}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) = \frac{1}{2}\int_{T\setminus\{c\}} \ell'_-(d(c,x))\,\mathrm{d}\mu(x),$$

thus

$$\int_{T\setminus(\{c\}\cup T_{c\to v_1}\cup T_{c\to v_2})} \ell'_-(d(c,x))\,\mathrm{d}\mu(x) = 0,$$

and since ℓ is increasing, this implies

$$\mu(T \setminus (\{c\} \cup T_{c\to v_1} \cup T_{c\to v_2})) = 0.$$

Consequently, $\operatorname{supp}(\mu) \subset \{c\} \cup T_{c\to v_1} \cup T_{c\to v_2}$, thus the event $\bigcap_{k=1}^{n} \{X_k \in (\{c\} \cup T_{c\to v_1} \cup T_{c\to v_2})\}$ has probability 1 and $\mathbb{P}(\operatorname{supp}(\widehat{\mu}_n) \subset \{c\} \cup T_{c\to v_1} \cup T_{c\to v_2}) = 1$. Since ℓ was assumed to be increasing in this section, by Proposition 5.20 we obtain

$$\mathbb{P}(M(\widehat{\mu}_n) \subset \{c\} \cup T_{c\to v_1} \cup T_{c\to v_2}) = 1.$$

3. The proof in the nonsticky case is also similar to the sticky case. □

Proof of Corollary 5.28. We only deal with the sticky case, as the others are similar. Since the series $\sum_{n\geq 1}\sum_{i=1}^{m} \exp\left(-\frac{\phi'_{v_i}(c)^2}{2\ell'_+(D)^2} n\right)$ has a finite sum, the Borel–Cantelli lemma yields

$$\mathbb{P}\big(\exists N \geq 1, \forall n \geq N, M(\widehat{\mu}_n) = \{c\}\big) = 1.$$

□

Proof of Proposition 5.29. In this proof we will say for convenience that c *is μ-sticky* if c is sticky when the measure under study is μ, and we say similarly that c *is μ-partly sticky.*

1. Suppose c is μ-sticky and define $\varepsilon = \min_j \phi'_{v_j}(c)/\left(4\ell'_+(D)\right)$. Let ν be such that $\mathrm{TV}(\nu, \mu) \leq \varepsilon$, let $\widetilde{\phi}$ denote the objective function associated to ν and fix some $i \in \{1, \dots, m\}$. Recall that

$$\mathrm{TV}(\nu_1, \nu_2) = \frac{1}{2}\sup\Big\{\int_T f(x)\,\mathrm{d}\nu_1(x) - \int_T f(x)\,\mathrm{d}\nu_2(x) \mid f : T \to [-1, 1] \text{ is measurable}\Big\}$$

(see, e.g., [236, Lemma 1 p.432]). Since the function

$$\varphi : x \mapsto \mathbb{1}_{T\setminus T_{c\to v_i}}(x)\ell'_+(d(c,x)) - \mathbb{1}_{T_{c\to v_i}}(x)\ell'_-(d(c,x)) \tag{5.16}$$

is bounded by $\ell'_+(D)$ and we have the estimate

$$|\widetilde{\phi}'_{v_i}(c) - \phi'_{v_i}(c)| = \left|\int_T \varphi(x)\,\mathrm{d}\nu(x) - \int_T \varphi(x)\,\mathrm{d}\mu(x)\right| \leq 2\ell'_+(D)\,\mathrm{TV}(\nu, \mu) \leq \frac{\phi'_{v_i}(c)}{2},$$

which implies $\widetilde{\phi}'_{v_i}(c) > 0$, thus c is ν-sticky and $M(\nu) = \{c\}$.

Conversely for some $\varepsilon > 0$, we suppose that $\forall \nu, \mathrm{TV}(\nu, \mu) \leq \varepsilon \implies M(\nu) = \{c\}$ and we assume for the sake of contradiction that c is not μ-sticky. Since $c \in M(\mu)$, c is μ-partly sticky and there exists i with $\phi'_{v_i}(c) = 0$, i.e., $\int_T \varphi(x)\,\mathrm{d}\mu(x) = 0$ with φ defined in (5.16). Since ℓ is increasing, $\ell'_-(d(c, v_i)) > 0$ hence $\varphi(v_i) < 0$. Next, define the mixture measure $\nu = (1-\varepsilon)\mu + \varepsilon\delta_{v_i}$ so that

$$\widetilde{\phi}'_{v_i}(c) = (1-\varepsilon)\phi'_{v_i}(c) + \varepsilon\varphi(v_i) = \varepsilon\varphi(v_i) < 0 \tag{5.17}$$

and $\mathrm{TV}(\nu, \mu) \leq \varepsilon$. By our initial assumption, the closeness of the measures implies $M(\nu) = \{c\}$, which contradicts the inequality (5.17).

3. If ℓ is differentiable and ℓ' is M-Lipschitz, then φ is $2M$-Lipschitz. The previous arguments then readily adapt with the 1-Wasserstein distance.

□

5.6.3 Proofs for Section 5.4

Proof of Proposition 5.32. Assume for the sake of contradiction that α_1 is neither a vertex nor a point in the support. Then α_1 lies in the interior of an edge $[v, w]$, where

$$[v, \alpha_1) \cap M_1(\mu) = \emptyset. \tag{5.18}$$

Since $T \setminus \mathrm{supp}(\mu)$ is open, so is the intersection $(T \setminus \mathrm{supp}(\mu)) \cap (v, w)$, which contains α_1. Consequently, we can pick some $\alpha \in (T \setminus \mathrm{supp}(\mu)) \cap (v, \alpha_1)$ that verifies $\mu([\alpha, \alpha_1]) = 0$. This last equality implies $\mu(T_{\alpha\to v}) = \mu(T_{\alpha_1\to v})$ and $\mu(T_{\alpha\to w}) = \mu(T_{\alpha_1\to w})$, thus $\phi'_v(\alpha) = \phi'_v(\alpha_1)$ and $\phi'_w(\alpha) = \phi'_w(\alpha_1)$. Therefore α verifies the same optimality conditions as α_1 and we must have $\alpha \in M_1(\mu)$. However, by construction $\alpha \in (v, \alpha_1)$, which contradicts (5.18). We proceed identically with α_2. □

Proof of Proposition 5.34. Assume $\alpha_1 \neq \alpha_2$ are elements of $M_1(\mu)$. It is easily seen that $T \setminus T_{\alpha_1\to\alpha_2}$ and $T \setminus T_{\alpha_2\to\alpha_1}$ are disjoint, closed and convex subsets of T. A necessary optimality condition for α_1 is $\phi'_{\alpha_2}(\alpha_1) \geq 0$, which rewrites as $\mu(T \setminus T_{\alpha_1\to\alpha_2}) \geq \frac{1}{2}$. Symmetrically, $\mu(T \setminus T_{\alpha_2\to\alpha_1}) \geq \frac{1}{2}$ hence $\mu(T \setminus T_{\alpha_1\to\alpha_2}) = \mu(T \setminus T_{\alpha_2\to\alpha_1}) = \frac{1}{2}$.

Conversely, assume the existence of such G_1, G_2. Since T is compact the distance between subsets $d(G_1, G_2)$ is positive and attained for some $\alpha_1 \in G_1$, $\alpha_2 \in G_2$, i.e., $d(G_1, G_2) = d(\alpha_1, \alpha_2) > 0$. Since $G_1 \subset T \setminus T_{\alpha_1\to\alpha_2}$ and $G_2 \subset T \setminus T_{\alpha_2\to\alpha_1}$, we obtain $\mu(T \setminus T_{\alpha_1\to\alpha_2}) = \mu(T \setminus T_{\alpha_2\to\alpha_1}) = \frac{1}{2}$, thus $\phi'_{\alpha_2}(\alpha_1) = \phi'_{\alpha_1}(\alpha_2) = 0$, hence $\{\alpha_1, \alpha_2\} \subset M_1(\mu)$. □

Proof of Proposition 5.37. 1. G equipped with the induced metric is a metric tree. By the finiteness assumption on T, the tree G also has finitely many vertices. For $x \in T \setminus G$, $\pi(x)$ is clearly among the vertices of G, hence $\pi(T \setminus G)$ is finite.

2. By [15, Theorem 2.1.12] π is 1-Lipschitz, hence continuous and $\pi\#\mu$ is a Borel measure on T. Given a Borel subset B of T, note that $\pi^{-1}(B)$ rewrites as the disjoint union $\pi^{-1}(B \cap \mathring{G}) \cup \bigcup_{i=1}^m \pi^{-1}(B \cap \{v_i\})$, with $\pi^{-1}(B \cap \{v_i\}) = T_i$ if $v_i \in B$ and $\pi^{-1}(B \cap \{v_i\}) = \emptyset$ otherwise.

3. Let $\alpha \in T \setminus G$ and assume w.l.o.g. that $\alpha \in T_1$. For any $y \in G$ we have the decomposition $d(\alpha, y) = d(\alpha, v_1) + d(v_1, y)$, thus

$$\phi_{\pi\#\mu}(\alpha) = \int_T (d(\alpha, v_1) + d(v_1, \pi(x)))\,\mathrm{d}\mu(x) = d(\alpha, v_1) + \phi_{\pi\#\mu}(v_1) > \phi_{\pi\#\mu}(v_1) = \phi_{\pi\#\mu}(\pi(\alpha)).$$

As a consequence, any minimizer of $\phi_{\pi\#\mu}$ lies in G.

4. Fix $\alpha \in G$. We leverage the explicit form of $\pi\#\mu$ and we decompose the distance $d(\alpha, x)$ for $x \in T_i$:

$$\begin{aligned}
\phi_{\pi\#\mu}(\alpha) &= \int_{\mathring{G}} d(\alpha, x)\,\mathrm{d}\mu(x) + \sum_{i=1}^m \mu(T_i) d(\alpha, v_i) \\
&= \int_{\mathring{G}} d(\alpha, x)\,\mathrm{d}\mu(x) + \sum_{i=1}^m \int_{T_i} (d(\alpha, x) - d(x, v_i))\,\mathrm{d}\mu(x) \\
&= \phi(\alpha) - \sum_{i=1}^m \int_{T_i} d(v_i, x)\,\mathrm{d}\mu(x).
\end{aligned}$$

5. Let $\alpha \in M_1(\mu) \cap G$. By 4., α is in $\arg\min_{\alpha \in G} \phi(\alpha)$ and this set is equal to $M_1(\pi\#\mu)$ by 3.

6. By points 3. and 4., $M_1(\mu) = \arg\min_{\alpha \in G} \phi(\alpha) = \arg\min_{\alpha \in G} \phi_{\pi\#\mu}(\alpha) = M_1(\pi\#\mu)$. □

Proof of Lemma 5.41. 1. By Proposition 5.37, $M_1(\pi\#\mu) = M_1(\mu) = \{\alpha_\star\}$, thus $M_1(\nu) = \{0\}$.

2. On the event Ω_n, $\widehat{\alpha}_n \in M_1(\widehat{\mu}_n) \cap [v_1, v_2]$ hence $\widehat{\alpha}_n \in M_1(\pi\#\widehat{\mu}_n)$, which rewrites as $\widehat{\alpha}_n \in M_1\left(\frac{1}{n}\sum_{k=1}^n \delta_{\pi(X_k)}\right)$. On Ω_n we have therefore $\widehat{m}_n \in M_1\left(\frac{1}{n}\sum_{k=1}^n \delta_{Y_k}\right)$ and $d(\widehat{\alpha}_n, \alpha_\star) = |\widehat{m}_n - 0| = |\widehat{m}_n|$.

3. By Theorem 5.27 we have

$$\mathbb{P}(\widehat{\alpha}_n \in \{\alpha\} \cup T_{\alpha_\star \to v_1} \cup T_{\alpha_\star \to v_2}) = 1.$$

The equalities $\phi'_{v_1}(\alpha_\star) = \phi'_{v_2}(\alpha_\star) = 0$ rewrite as $\mu(T_{\alpha_\star \to v_1}) = \mu(T_{\alpha_\star \to v_2}) = \frac{1}{2}$. Moreover,

$$0 > \phi'_{\alpha_\star}(v_1) = 1 - 2\mu(T_{v_1 \to \alpha_\star}) = 1 - 2\left(\mu([\alpha_\star, v_1)) + \mu(T_{\alpha_\star \to v_2})\right) = -2\mu((\alpha_\star, v_1)).$$

Since $\mathbb{P}(\widehat{\phi}'_{\alpha_\star}(v_1) \leq 0) = \mathbb{P}(\widehat{\mu}_n(T_{v_1 \to \alpha_\star}) \geq \frac{1}{2})$ and $n\widehat{\mu}_n(T_{v_1 \to \alpha_\star})$ is a sum of n i.i.d. Bernoulli random variables, each with parameter $\mu(T_{v_1 \to \alpha_\star})$, the Chernoff bound [66, Theorem 1 and Example 3] provides

$$\mathbb{P}(\widehat{\phi}'_{\alpha_\star}(v_1) \leq 0) \leq \left(2\sqrt{\mu(T_{v_1 \to \alpha_\star})(1 - \mu(T_{v_1 \to \alpha_\star}))}\right)^n = \left(1 - 4\mu((\alpha_\star, v_1))^2\right)^{n/2} \tag{5.19}$$

We proceed similarly with v_2. Note that $\mathbb{P}(\Omega_n) \geq \mathbb{P}(M_1(\widehat{\mu}_n) \subset [v_1, v_2])$ and perform a union bound to obtain the claim. □

Proof of Theorem 5.43. 1. For each $n \geq 1$ let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ denote the order statistics of the sample $Y_1, \ldots, Y_n$. It is well-known that the set of real medians $M_1\left(\frac{1}{n}\sum_{k=1}^n \delta_{Y_k}\right)$ is the singleton $\{Y_{\left(\lfloor\frac{n}{2}\rfloor+1\right)}\}$ when n is odd and the interval $[Y_{\left(\lfloor\frac{n}{2}\rfloor\right)}, Y_{\left(\lfloor\frac{n}{2}\rfloor+1\right)}]$ when n is even.

We follow a similar path as the proof of [233, Theorem 5.10] for real quantiles. Fix $t > 0$ and let us determine the limit of $\mathbb{P}(n^{1/(2a)}\widehat{m}_n < t)$. We start with the upper bound

$$\mathbb{P}(\{n^{1/(2a)}\widehat{m}_n < t\} \cap \Omega_n) \leq \mathbb{P}\Big(\{Y_{\left(\lfloor \frac{n}{2} \rfloor\right)} < \frac{t}{n^{1/(2a)}}\} \cap \Omega_n\Big) \leq \mathbb{P}\Big(\sum_{k=1}^{n} \mathbb{1}_{Y_k < \frac{t}{n^{1/(2a)}}} \geq \Big\lfloor \frac{n}{2} \Big\rfloor\Big).$$

Letting $B_n = \sum_{k=1}^n \mathbb{1}_{Y_k < \frac{t}{n^{1/(2a)}}}$, $C_n = \frac{B_n - \mathbb{E}[B_n]}{\mathbb{V}[B_n]}$ and $p_n = \mathbb{P}(Y < \frac{t}{n^{1/(2a)}})$ we obtain

$$\mathbb{P}(\{n^{1/(2a)}\widehat{m}_n < t\} \cap \Omega_n) \leq \mathbb{P}\Big(C_n \geq \frac{\lfloor \frac{n}{2} \rfloor - np_n}{\sqrt{np_n(1-p_n)}}\Big) = F_{-C_n}\left(\frac{np_n - \lfloor \frac{n}{2} \rfloor}{\sqrt{np_n(1-p_n)}}\right). \tag{5.20}$$

Note that $\lim_n p_n = P(Y \leq 0) = \frac{1}{2}$ and as n goes to infinity,

$$p_n - \frac{1}{2} = \mathbb{P}\big(Y \in (0, t/n^{1/(2a)})\big) = \Delta(t/n^{1/(2a)}) = Kt^a n^{-1/2} + o(n^{-1/2}),$$

therefore

$$\frac{np_n - \lfloor \frac{n}{2} \rfloor}{\sqrt{np_n(1-p_n)}} \xrightarrow[n\to\infty]{} 2Kt^a.$$

By the Lyapunov central limit theorem [233, Example 1.33], C_n converges in distribution to a standard normal, hence so does $-C_n$. By Pólya's theorem [233, Proposition 1.16], $\sup_{x\in\mathbb{R}} |F_{-C_n}(x) - \Phi(x)| \xrightarrow[n\to\infty]{} 0$ (where Φ denotes the cdf of the standard normal distribution) and the RHS of (5.20) converges to $\Phi(2Kt^a)$. Moreover $\mathbb{P}(\Omega_n) \to 0$, hence

$$\limsup_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n < t) = \limsup_n \mathbb{P}(\{n^{1/(2a)}\widehat{m}_n < t\} \cap \Omega_n) \leq \Phi(2Kt^a). \tag{5.21}$$

Now, we turn to the lower bound

$$\mathbb{P}(\{n^{1/(2a)}\widehat{m}_n < t\} \cap \Omega_n) \geq 1 - \mathbb{P}\Big(\{Y_{\left(\lfloor \frac{n}{2} \rfloor + 1\right)} \geq \frac{t}{n^{1/(2a)}}\} \cap \Omega_n\Big) \geq 1 - \mathbb{P}\Big(\sum_{k=1}^{n} \mathbb{1}_{Y_k \geq \frac{t}{n^{1/(2a)}}} \geq \frac{n}{2}\Big)$$

and by the exact same techniques we find that the RHS converges to $\Phi(2Kt^a)$, thus

$$\liminf_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n < t) = \liminf_n \mathbb{P}(\{n^{1/(2a)}\widehat{m}_n < t\} \cap \Omega_n) \geq \Phi(2Kt^a).$$

Combining with (5.21) we obtain

$$\forall t > 0,\ \mathbb{P}(n^{1/(2a)}\widehat{m}_n < t) \xrightarrow[n\to\infty]{} \Phi(2Kt^a). \tag{5.22}$$

Next, fix $u > 0$ and an integer $k \geq 1$. Observe that

$$\limsup_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n \leq u) \leq \limsup_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n < u + \tfrac{1}{k}) \overset{(5.22)}{=} \Phi(2K(u + \tfrac{1}{k})^a).$$

Letting $k \to \infty$ yields $\limsup_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n \leq u) \leq \Phi(2Ku^a)$. Furthermore

$$\liminf_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n \leq u) \geq \liminf_n \mathbb{P}(n^{1/(2a)}\widehat{m}_n < u) \overset{(5.22)}{=} \Phi(2Ku^a),$$

thus

$$\forall u > 0,\ \mathbb{P}(n^{1/(2a)}\widehat{m}_n \le u) \xrightarrow[n\to\infty]{} \Phi(2Ku^a). \tag{5.23}$$

Finally, fix $t \le 0$ and note that

$$\mathbb{P}(\{n^{1/(2a)}\widehat{m}_n \le t\} \ \cap\ \Omega_n) = \mathbb{P}\Big(\sum_{k=1}^{n} \mathbb{1}_{Y_k \le \frac{t}{n^{1/(2a)}}} \ge \Big\lfloor \frac{n}{2} \Big\rfloor\Big).$$

Letting $B_n = \sum_{k=1}^n \mathbb{1}_{Y_k \le \frac{t}{n^{1/(2a)}}}$, $C_n = \frac{B_n - \mathbb{E}[B_n]}{\mathbb{V}[B_n]}$ and $p_n = \mathbb{P}(Y \le \frac{t}{n^{1/(2a)}})$ we have

$$\mathbb{P}(\{n^{1/(2a)}\widehat{m}_n \le t\} \ \cap\ \Omega_n) \le \mathbb{P}\Big(C_n \ge \frac{\lfloor \frac{n}{2} \rfloor - np_n}{\sqrt{np_n(1-p_n)}}\Big) = F_{-C_n}\left(\frac{np_n - \lfloor \frac{n}{2} \rfloor}{\sqrt{np_n(1-p_n)}}\right).$$

Note that $\lim_n p_n = P(Y \le 0) = \frac{1}{2}$ and as n goes to infinity,

$$\frac{1}{2} - p_n = \mathbb{P}\big(Y \in (t/n^{1/(2a)}, 0)\big) = \Delta(t/n^{1/(2a)}) = K|t|^a n^{-1/2} + o(n^{-1/2}),$$

from which we derive the convergence

$$\frac{np_n - \lfloor \frac{n}{2} \rfloor}{\sqrt{np_n(1-p_n)}} \xrightarrow[n\to\infty]{} -2K|t|^a.$$

The rest of the proof is similar to what was done for $t \ge 0$ and we find

$$\forall t \le 0,\ \mathbb{P}(n^{1/(2a)}\widehat{m}_n \le t) \xrightarrow[n\to\infty]{} \Phi(-2K|t|^a). \tag{5.24}$$

Combining (5.23) and (5.24), $n^{1/(2a)}\widehat{m}_n$ converges in distribution to a random variable with cdf $t \mapsto \Phi(2K\mathrm{sgn}(t)|t|^a)$, hence to the random variable $\mathrm{sgn}(Z)\left(\frac{|Z|}{2K}\right)^{1/a}$.

2. On the event Ω_n, we have the equality $d(\widehat{\alpha}_n, \alpha_\star) = |\widehat{m}_n|$. The convergence in distribution of $n^{1/(2a)}\widehat{m}_n$ and the estimate $\mathbb{P}(\Omega_n) \to 0$ are enough to obtain the claim. □

Proof of Corollary 5.44. This is a direct consequence of Theorem 5.43. □

Proof of Theorem 5.46. By Theorem 5.27 we have $\mathbb{P}(\widehat{\alpha}_n \in \{\alpha\} \cup T_{\alpha_\star\to v_1} \cup T_{\alpha_\star \to v_2}) = 1$. Next, note that

$$\begin{aligned}\mathbb{P}(d(\widehat{\alpha}_n, \alpha_\star) \ge t,\ \widehat{\alpha}_n \in \{\alpha\} \cup T_{\alpha_\star\to v_1}) &\le \mathbb{1}_{t \le d(\alpha_\star, v_1)}\mathbb{P}(\widehat{\alpha}_n \notin T_{\gamma_t \to \alpha_\star}) + \mathbb{1}_{t > d(\alpha_\star, v_1)}\mathbb{P}(\Omega_n^c) \\ &\le \mathbb{1}_{t \le d(\alpha_\star, v_1)}\mathbb{P}\big(\widehat{\phi}'_{\alpha_\star}(\gamma_t) \ge 0\big) + \mathbb{1}_{t > d(\alpha_\star, v_1)}\mathbb{P}(\Omega_n^c).\end{aligned}$$

Furthermore, $\phi'_{\alpha_\star}(\gamma_t) < 0$ and

$$\phi'_{\alpha_\star}(\gamma_t) = 1 - 2\mu(T_{\gamma_t \to \alpha_\star}) = 1 - 2\mu\left(T_{\alpha_\star \to v_2} \cup [\alpha_\star, \gamma_t)\right) = 1 - 2(\tfrac{1}{2} + \Delta(t)) = -2\Delta(t).$$

Proceeding similarly as in (5.19), we obtain

$$\mathbb{P}(d(\widehat{\alpha}_n, \alpha_\star) \ge t,\ \widehat{\alpha}_n \in \{\alpha\} \cup T_{\alpha_\star\to v_1}) \le \mathbb{1}_{t \le d(\alpha_\star, v_1)}\left(1 - 4\Delta^2(t)\right)^{n/2} + \mathbb{1}_{t > d(\alpha_\star, v_1)}\mathbb{P}(\Omega_n^c).$$

A similar bound holds for v_2 and this finishes the proof. □

Proof of Proposition 5.47. First, write that $\mathbb{P}(\widehat{\alpha}_n \notin \{\alpha_\star\} \cup T_{\alpha_\star \to v_1}) = \sum_{j=2}^m \mathbb{P}(\widehat{\alpha}_n \in T_{\alpha_\star \to v_j})$. For all $j = 1, \ldots, m$, let $N_j = \#\{k = 1, \ldots, n : X_k \in T_{\alpha_\star \to v_j}\}$ and $p_j = \mu(T_{\alpha_\star \to v_j})$. Then, for all $j = 2, \ldots, m$, Chernoff's bound [66, Theorem 1 and Example 3] yields

$$\begin{aligned}
\mathbb{P}(\widehat{\alpha}_n \in T_{\alpha_\star \to v_j}) &= \mathbb{P}(N_j \geq n/2) \\
&= \mathbb{P}(N_j/n - p_j \geq 1/2 - p_j) \\
&\leq (4p_j(1-p_j))^{n/2} .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\widehat{\alpha}_n \notin \{\alpha_\star\} \cup T_{\alpha_\star \to v_1}) &\leq \sum_{j=2}^m (4p_j(1-p_j))^{n/2} \\
&= \sum_{j=2}^m 4p_j(1-p_j)\,(4p_j(1-p_j))^{n/2-1} \\
&\leq 4\left(1-4\varepsilon^2\right)^{n/2-1} \sum_{j=2}^m p_j \\
&\leq 2\left(1-4\varepsilon^2\right)^{n/2-1} \\
&\leq 2e^{-n\varepsilon^2},
\end{aligned}$$

where, in the second to last inequality, we used the fact that $p_2 + \ldots + p_m \leq 1 - p_1 = 1/2$, and in the last inequality, the fact that $n/2 - 1 \geq n/4$, since $n \geq 4$.

□

Proof of Lemma 5.49. Similar to the proof of Lemma 5.41 □

Proof of Theorem 5.51. Similar to the proof of Theorem 5.43. □

Proof of Theorem 5.53. Similar to the proof of Theorem 5.46. □

Bibliography

[1] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140. Elsevier/Academic Press, Amsterdam, second edition, 2003. ISBN 0-12-044143-8.

[2] Jorge G. Adrover and Víctor J. Yohai. Simultaneous redescending M-estimates for regression and scale. *Comm. Statist. Theory Methods*, 29(2):243–262, 2000. ISSN 0361-0926. doi: 10.1080/03610920008832482. URL https://doi.org/10.1080/03610920008832482.

[3] Bijan Afsari. Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.*, 139(2):655–673, 2011. ISSN 0002-9939. doi: 10.1090/S0002-9939-2010-10541-5. URL https://doi.org/10.1090/S0002-9939-2010-10541-5.

[4] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011. ISSN 0036-1410. doi: 10.1137/100805741. URL https://doi.org/10.1137/100805741.

[5] A. Ahidar-Coutrix, T. Le Gouic, and Q. Paris. Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics. *Probab. Theory Related Fields*, 177(1-2):323–368, 2020. ISSN 0178-8051. doi: 10.1007/s00440-019-00950-0. URL https://doi.org/10.1007/s00440-019-00950-0.

[6] Fernando Albiac and Nigel J. Kalton. *Topics in Banach Space Theory*. Springer, [Cham], second edition, 2016. ISBN 978-3-319-31555-3; 978-3-319-31557-7. doi: 10.1007/978-3-319-31557-7. URL https://doi.org/10.1007/978-3-319-31557-7.

[7] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, third edition, 2006. ISBN 978-3-540-32696-0; 3-540-32696-0. doi: 10.1007/3-540-29587-9. URL https://doi.org/10.1007/3-540-29587-9.

[8] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, Cambridge, MA, 1985. ISBN 978-0-674-00560-0.

[9] Aloisio Araujo and Evarist Giné. *The central limit theorem for real and Banach valued random variables*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York-Chichester-Brisbane, 1980. ISBN 0-471-05304-X.

[10] Miguel A. Arcones. A remark on approximate M-estimators. *Statist. Probab. Lett.*, 38(4):311–321, 1998. ISSN 0167-7152. doi: 10.1016/S0167-7152(98)00038-8. URL https://doi.org/10.1016/S0167-7152(98)00038-8.

[11] Miguel A. Arcones and David M. Mason. A general approach to Bahadur-Kiefer representations for M-estimators. *Math. Methods Statist.*, 6(3):267–292, 1997. ISSN 1066-5307.

[12] Marc Arnaudon, Frédéric Barbaresco, and Le Yang. Medians and means in Riemannian geometry: existence, uniqueness and computation. In *Matrix information geometry*, pages 169–197. Springer, Heidelberg, 2013. ISBN 978-3-642-30231-2. doi: 10.1007/978-3-642-30232-9_8. URL https://doi.org/10.1007/978-3-642-30232-9_8.

[13] Edgar Asplund. Fréchet differentiability of convex functions. *Acta Math.*, 121:31–47, 1968. ISSN 0001-5962. doi: 10.1007/BF02391908. URL https://doi.org/10.1007/BF02391908.

[14] H. Attouch. *Variational Convergence for Functions and Operators*. Applicable Mathematics Series. Pitman (Advanced Publishing Program), Boston, MA, 1984. ISBN 0-273-08583-2.

[15] Miroslav Bačák. *Convex Analysis and Optimization in Hadamard Spaces*, volume 22 of *De Gruyter Series in Nonlinear Analysis and Applications*. De Gruyter, Berlin, 2014. ISBN 978-3-11-036103-2; 978-3-11-036162-9. doi: 10.1515/9783110361629. URL https://doi.org/10.1515/9783110361629.

[16] R. R. Bahadur. A note on quantiles in large samples. *Ann. Math. Statist.*, 37:577–580, 1966. ISSN 0003-4851. doi: 10.1214/aoms/1177699450. URL https://doi.org/10.1214/aoms/1177699450.

[17] Viorel Barbu and Teodor Precupanu. *Convexity and Optimization in Banach Spaces*. Springer, Dordrecht, fourth edition, 2012. ISBN 978-94-007-2246-0; 978-94-007-2247-7. doi: 10.1007/978-94-007-2247-7. URL https://doi.org/10.1007/978-94-007-2247-7.

[18] Bojan Basrak. Limit theorems for the inductive mean on metric trees. *J. Appl. Probab.*, 47(4):1136–1149, 2010. ISSN 0021-9002,1475-6072. doi: 10.1239/jap/1294170525. URL https://doi.org/10.1239/jap/1294170525.

[19] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[20] Gerald Beer. *Topologies on Closed and Closed Convex Sets*, volume 268 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1993. ISBN 0-7923-2531-1. doi: 10.1007/978-94-015-8149-3. URL https://doi.org/10.1007/978-94-015-8149-3.

[21] Gerald Beer and Jonathan M. Borwein. Mosco convergence and reflexivity. *Proc. Amer. Math. Soc.*, 109(2):427–436, 1990. ISSN 0002-9939. doi: 10.2307/2048005. URL https://doi.org/10.2307/2048005.

[22] Pierre C Bellec. Optimal bounds for aggregation of affine estimators. *The Annals of Statistics*, 46(1):30–59, 2018.

[23] Pierre C Bellec and Gabriel Romon. Chi-square and normal inference in high-dimensional multi-task regression. *arXiv preprint arXiv:2107.07828v1*, 2021.

[24] Pierre C Bellec and Alexandre B Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In Vladimir Panov, editor, *Modern Problems of Stochastic Analysis and Statistics, Selected Contributions In Honor of Valentin Konakov*. Springer, 2017. URL https://arxiv.org/pdf/1609.06675.pdf.

[25] Pierre C Bellec and Cun-Hui Zhang. Second order stein: Sure for sure and other applications in high-dimensional inference. *arXiv:1804.01230*, 2018. URL https://arxiv.org/pdf/1811.04121.pdf.

[26] Pierre C. Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022. ISSN 1350-7265,1573-9759. doi: 10.3150/21-BEJ1348. URL https://doi.org/10.3150/21-BEJ1348.

[27] Pierre C. Bellec and Cun-Hui Zhang. Debiasing convex regularized estimators and interval estimation in linear models. *Ann. Statist.*, 51(2):391–436, 2023. ISSN 0090-5364,2168-8966. doi: 10.1214/22-aos2243. URL https://doi.org/10.1214/22-aos2243.

[28] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[29] J. O. Berger and G. Salinetti. Approximations of bayes decision problems: the epigraphical approach. *Annals of Operations Research*, 56(1):1–13, Dec 1995. ISSN 1572-9338. doi: 10.1007/BF02031697. URL https://doi.org/10.1007/BF02031697.

[30] Quentin Bertrand, Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Handling correlated and repeated measurements with the smoothed multivariate square-root lasso. In *Advances in Neural Information Processing Systems*, pages 3961–3972, 2019.

[31] Abhishek Bhattacharya and Rabi Bhattacharya. *Nonparametric inference on manifolds*, volume 2 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge, 2012. ISBN 978-1-107-01958-4. doi: 10.1017/CBO9781139094764. URL https://doi.org/10.1017/CBO9781139094764. With applications to shape spaces.

[32] Rabi Bhattacharya and Lizhen Lin. Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proc. Amer. Math. Soc.*, 145 (1):413–428, 2017. ISSN 0002-9939,1088-6826. doi: 10.1090/proc/13216. URL https://doi.org/10.1090/proc/13216.

[33] Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.*, 31(1):1–29, 2003. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1046294456. URL https://doi.org/10.1214/aos/1046294456.

[34] Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.*, 33(3):1225–1259, 2005. ISSN 0090-5364,2168-8966. doi: 10.1214/009053605000000093. URL https://doi.org/10.1214/009053605000000093.

[35] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 08 2009. doi: 10.1214/08-AOS620. URL http://dx.doi.org/10.1214/08-AOS620.

[36] Herman J. Bierens. *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Themes in Modern Econometrics. Cambridge University Press, Cambridge, 2004. ISBN 978-0-521-54224-1.

[37] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27(4):733–767, 2001. ISSN 0196-8858. doi: 10.1006/aama.2001.0759. URL https://doi.org/10.1006/aama.2001.0759.

[38] Rafael Blanquero and Emilio Carrizosa. Solving the median problem with continuous demand on a network. *Comput. Optim. Appl.*, 56(3):723–734, 2013. ISSN 0926-6003,1573-2894. doi: 10.1007/s10589-013-9574-3. URL https://doi.org/10.1007/s10589-013-9574-3.

[39] Vladimir I. Bogachev. *Weak Convergence of Measures*, volume 234 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2018. ISBN 978-1-4704-4738-0. doi: 10.1090/surv/234. URL https://doi.org/10.1090/surv/234.

[40] Jonathan M. Borwein and Jon D. Vanderwerff. *Convex Functions: Constructions, Characterizations and Counterexamples*, volume 109 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-85005-6. doi: 10.1017/CBO9781139087322. URL https://doi.org/10.1017/CBO9781139087322.

[41] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[42] Jelena Bradic, Jianqing Fan, and Yinchu Zhu. Testability of high-dimensional linear models with non-sparse structures. *arXiv preprint arXiv:1802.09117*, 2018.

[43] Andrea Braides. *Γ-convergence for Beginners*, volume 22 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2002. ISBN 0-19-850784-4. doi: 10.1093/acprof:oso/9780198507840.001.0001. URL https://doi.org/10.1093/acprof:oso/9780198507840.001.0001.

[44] Margaret L. Brandeau and Samuel S. Chiu. Parametric facility location on a tree network with an L_p-norm cost function. *Transportation Sci.*, 22(1):59–69, 1988. ISSN 0041-1655. doi: 10.1287/trsc.22.1.59. URL https://doi.org/10.1287/trsc.22.1.59.

[45] Martin R. Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999. ISBN 3-540-64324-9. doi: 10.1007/978-3-662-12494-9. URL https://doi.org/10.1007/978-3-662-12494-9.

[46] B. M. Brown. Statistical uses of the spatial median. *J. Roy. Statist. Soc. Ser. B*, 45(1):25–30, 1983. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1983)45:1<25:SUOTSM>2.0.CO;2-P&origin=MSN.

[47] L. D. Brown and R. Purves. Measurable selections of extrema. *Ann. Statist.*, 1:902–912, 1973. ISSN 0090-5364. doi: 10.1214/aos/1176342510. URL https://doi.org/10.1214/aos/1176342510.

[48] Victor-Emmanuel Brunel. Geodesically convex M-estimation in metric spaces. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2188–2210. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/brunel23a.html.

[49] Victor-Emmanuel Brunel and Jordan Serres. Concentration of empirical barycenters in metric spaces. *arXiv preprint arXiv:2303.01144v1*, 2023.

[50] Damian Brzyski, Alexej Gossmann, Weijie Su, and Małgorzata Bogdan. Group slope–adaptive selection of groups of predictors. *Journal of the American Statistical Association*, 114(525):419–433, 2019.

[51] Peter Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, 2013.

[52] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001. ISBN 0-8218-2129-6. doi: 10.1090/gsm/033. URL https://doi.org/10.1090/gsm/033.

[53] Benoît Cadre. Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, 35(4):509–521, 2001. ISSN 0233-1888. doi: 10.1080/02331880108802751. URL https://doi.org/10.1080/02331880108802751.

[54] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, 45(2):615–646, 2017.

[55] Tianxi Cai, Tony Cai, and Zijian Guo. Individualized treatment selection: An optimal hypothesis testing approach in high-dimensional models. *arXiv preprint arXiv:1904.12891*, 2019.

[56] Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[57] Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43, 2013. ISSN 1350-7265. doi: 10.3150/11-BEJ390. URL https://doi.org/10.3150/11-BEJ390.

[58] Hervé Cardot, Peggy Cénac, and Antoine Godichon-Baggioni. Online estimation of the geometric median in Hilbert spaces: nonasymptotic confidence balls. *Ann. Statist.*, 45(2):591–614, 2017. ISSN 0090-5364. doi: 10.1214/16-AOS1460. URL https://doi.org/10.1214/16-AOS1460.

[59] Henri Cartan. *Differential Calculus*. Hermann, Paris; Houghton Mifflin Co., Boston, Mass., 1971. Exercises by C. Buttin, F. Rideau and J. L. Verley, Translated from the French.

[60] Rui Caseiro, Pedro Martins, João F. Henriques, and Jorge Batista. A nonparametric riemannian framework on tensor field with application to foreground segmentation. *Pattern Recognition*, 45(11):3997–4017, 2012. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2012.04.011. URL https://www.sciencedirect.com/science/article/pii/S0031320312001689.

[61] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *arXiv preprint arXiv:1903.10603*, 2019.

[62] Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

[63] Anirvan Chakraborty and Probal Chaudhuri. The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Ann. Statist.*, 42(3):1203–1231, 2014. ISSN 0090-5364. doi: 10.1214/14-AOS1226. URL https://doi.org/10.1214/14-AOS1226.

[64] Anirvan Chakraborty and Probal Chaudhuri. The deepest point for distributions in infinite dimensional spaces. *Stat. Methodol.*, 20:27–39, 2014. ISSN 1572-3127. doi: 10.1016/j.stamet.2013.04.004. URL https://doi.org/10.1016/j.stamet.2013.04.004.

[65] Probal Chaudhuri. On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91(434):862–872, 1996. ISSN 0162-1459. doi: 10.2307/2291681. URL https://doi.org/10.2307/2291681.

[66] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952. ISSN 0003-4851. doi: 10.1214/aoms/1177729330. URL https://doi.org/10.1214/aoms/1177729330.

[67] Jérôme-Alexis Chevalier, Alexandre Gramfort, Joseph Salmon, and Bertrand Thirion. Statistical control for spatio-temporal MEG/EEG source imaging with desparsified multi-task Lasso. In *Thirty-fourth Conference on Neural Information Processing Systems*, 2020. URL https://papers.nips.cc/paper/2020/hash/1359aa933b48b754a2f54adb688bfa77-Abstract.html.

[68] John B. Conway. *A Course in Operator Theory*, volume 21 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2000. ISBN 0-8218-2065-6. doi: 10.1090/gsm/021. URL https://doi.org/10.1090/gsm/021.

[69] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000. doi: 10.1017/CBO9780511801389.

[70] Gianni Dal Maso. *An Introduction to* Γ*-Convergence*, volume 8 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser Boston, Inc., Boston, MA, 1993. ISBN 0-8176-3679-X. doi: 10.1007/978-1-4612-0327-8. URL https://doi.org/10.1007/978-1-4612-0327-8.

[71] Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

[72] Andreas Defant and Klaus Floret. *Tensor Norms and Operator Ideals*, volume 176 of *North-Holland Mathematics Studies*. North-Holland Publishing Co., Amsterdam, 1993. ISBN 0-444-89091-2.

[73] Holger Dette and Kevin Kokot. Detecting relevant differences in the covariance operators of functional time series: a sup-norm approach. *Ann. Inst. Statist. Math.*, 74(2):195–231, 2022. ISSN 0020-3157,1572-9052. doi: 10.1007/s10463-021-00795-2. URL https://doi.org/10.1007/s10463-021-00795-2.

[74] Holger Dette, Kevin Kokot, and Alexander Aue. Functional data analysis in the Banach space of continuous functions. *Ann. Statist.*, 48(2):1168–1192, 2020.

ISSN 0090-5364,2168-8966. doi: 10.1214/19-AOS1842. URL https://doi.org/10.1214/19-AOS1842.

[75] J. Diestel and J. J. Uhl, Jr. *Vector Measures*. Mathematical Surveys, No. 15. American Mathematical Society, Providence, R.I., 1977. With a foreword by B. J. Pettis.

[76] Michael X. Dong and Roger J.-B. Wets. Estimating density functions: a constrained maximum likelihood approach. *J. Nonparametr. Statist.*, 12(4):549–595, 2000. ISSN 1048-5252. doi: 10.1080/10485250008832822. URL https://doi.org/10.1080/10485250008832822.

[77] A. L. Dontchev and T. Zolezzi. *Well-Posed Optimization Problems*, volume 1543 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1993. ISBN 3-540-56737-2. doi: 10.1007/BFb0084195. URL https://doi.org/10.1007/BFb0084195.

[78] A. Ya. Dorogovtsev. On asymptotic normality of the least square estimators of an infinite-dimensional parameter. *Ukrainian Math. J.*, 45(1):48–58, 1993. ISSN 0041-5995,1573-9376.

[79] A. Ya. Dorogovtsev, N. Zerek, and A. G. Kukush. Asymptotic properties of estimates of nonlinear regression in Hilbert space. *Teor. Veroyatnost. i Mat. Statist.*, (35):36–43, 122, 1986. ISSN 0131-6982.

[80] A. Ya. Dorogovtsev, N. Zerek, and A. G. Kukush. Weak convergence to the normal distribution of an estimator for an infinite-dimensional parameter. *Teor. Veroyatnost. i Mat. Statist.*, (37):39–46, 135, 1987. ISSN 0131-6982.

[81] R. M. Dudley. A course on empirical processes. In *École d'été de probabilités de Saint-Flour, XII—1982*, volume 1097 of *Lecture Notes in Math.*, pages 1–142. Springer, Berlin, 1984. doi: 10.1007/BFb0099432. URL https://doi.org/10.1007/BFb0099432.

[82] R. M. Dudley. *Uniform central limit theorems*, volume 142 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, New York, second edition, 2014. ISBN 978-0-521-73841-5; 978-0-521-49884-5.

[83] Richard M. Dudley. Consistency of M-estimators and one-sided bracketing. In *High dimensional probability (Oberwolfach, 1996)*, volume 43 of *Progr. Probab.*, pages 33–58. Birkhäuser, Basel, 1998. doi: 10.1007/978-3-0348-8829-5. URL https://doi.org/10.1007/978-3-0348-8829-5.

[84] Jitka Dupačová and Roger Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.*, 16(4):1517–1549, 1988. ISSN 0090-5364. doi: 10.1214/aos/1176351052. URL https://doi.org/10.1214/aos/1176351052.

[85] Benjamin Eltzner. Geometrical smeariness—a new phenomenon of Fréchet means. *Bernoulli*, 28(1):239–254, 2022. ISSN 1350-7265,1573-9759. doi: 10.3150/21-bej1340. URL https://doi.org/10.3150/21-bej1340.

[86] Benjamin Eltzner and Stephan F. Huckemann. A smeary central limit theorem for manifolds with application to high-dimensional spheres. *Ann. Statist.*, 47 (6):3360–3381, 2019. ISSN 0090-5364. doi: 10.1214/18-AOS1781. URL https://doi.org/10.1214/18-AOS1781.

[87] Benjamin Eltzner, Fernando Galaz-García, Stephan Huckemann, and Wilderich Tuschmann. Stability of the cut locus and a central limit theorem for Fréchet means of Riemannian manifolds. *Proc. Amer. Math. Soc.*, 149(9):3947–3963, 2021. ISSN 0002-9939,1088-6826. doi: 10.1090/proc/15429. URL https://doi.org/10.1090/proc/15429.

[88] Paul Escande. On the concentration of the minimizers of empirical risks. *arXiv preprint arXiv:2304.00809v3*, 2023.

[89] Steven N. Evans and Adam Q. Jaffe. Limit theorems for fréchet mean sets. *Bernoulli*, 2023. Forthcoming, https://www.e-publications.org/ims/submission/BEJ/user/submissionFile/49299?confirm=a1a771f4 (version: 2023-09-08).

[90] Marián Fabian, Petr Habala, Petr Hájek, Vicente Montesinos Santalucía, Jan Pelant, and Václav Zizler. *Functional Analysis and Infinite-Dimensional Geometry*, volume 8 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer-Verlag, New York, 2001. ISBN 0-387-95219-5. doi: 10.1007/978-1-4757-3480-5. URL https://doi.org/10.1007/978-1-4757-3480-5.

[91] Marián Fabian, Petr Habala, Petr Hájek, Vicente Montesinos, and Václav Zizler. *Banach Space Theory*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. ISBN 978-1-4419-7514-0. doi: 10.1007/978-1-4419-7515-7. URL https://doi.org/10.1007/978-1-4419-7515-7.

[92] Aasa Feragen, Megan Owen, Jens Petersen, Mathilde M. W. Wille, Laura H. Thomsen, Asger Dirksen, and Marleen de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In James C. Gee, Sarang Joshi, Kilian M. Pohl, William M. Wells, and Lilla Zöllei, editors, *Information Processing in Medical Imaging*, pages 74–85, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-38868-2.

[93] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer, New York, 2006. ISBN 0-387-30369-3; 978-0387-30369-7. Theory and practice.

[94] P.T. Fletcher, Conglin Lu, S.M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004. doi: 10.1109/TMI.2004.831793.

[95] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré*, 10:215–310, 1947. ISSN 0365-320X.

[96] Sébastien Gadat, Ioana Gavra, and Laurent Risser. How to calculate the barycenter of a weighted graph. *Math. Oper. Res.*, 43(4):1085–1118, 2018. ISSN 0364-765X. doi: 10.1287/moor.2017.0896. URL https://doi.org/10.1287/moor.2017.0896.

[97] Leszek Gasiński and Nikolaos S. Papageorgiou. *Exercises in analysis. Part 1.* Problem Books in Mathematics. Springer, Cham, 2014. ISBN 978-3-319-06175-7; 978-3-319-06176-4.

[98] Ioana Gavra and Laurent Risser. Online barycenter estimation of large weighted graphs. *arXiv preprint arXiv:1803.11137v1*, 2018.

[99] Gery Geenens. Curse of dimensionality and related issues in nonparametric functional regression. *Stat. Surv.*, 5:30–43, 2011. ISSN 1935-7516. doi: 10.1214/09-SS049. URL https://doi.org/10.1214/09-SS049.

[100] Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008. ISSN 0006-3444. doi: 10.1093/biomet/asn031. URL https://doi.org/10.1093/biomet/asn031.

[101] Daniel Gervini. Technical supplement: Robust functional estimation using the median and spherical principal components. https://sites.uwm.edu/gervini/files/2016/04/SpMed_Supp-2hv7tgq.pdf, 2008.

[102] Charles J. Geyer. On the asymptotics of constrained M-estimation. *Ann. Statist.*, 22(4):1993–2010, 1994. ISSN 0090-5364. doi: 10.1214/aos/1176325768. URL https://doi.org/10.1214/aos/1176325768.

[103] C. Gini and L. Galvani. Di talune estensioni dei concetti di media ai caratteri qualitativi. Metron 8, Nr. 1, 2 (1929), 3-209 (1929)., 1929.

[104] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2021.

[105] Antoine Godichon-Baggioni. Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms: L^p and almost sure rates of convergence. *J. Multivariate Anal.*, 146:209–222, 2016. ISSN 0047-259X. doi: 10.1016/j.jmva.2015.09.013. URL https://doi.org/10.1016/j.jmva.2015.09.013.

[106] Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

[107] Zijian Guo, Claude Renaux, Peter Bühlmann, and T Tony Cai. Group inference in high dimensions with applications to hierarchical testing. *arXiv preprint arXiv:1909.01503*, 2019.

[108] Shelby J. Haberman. Concavity and estimation. *Ann. Statist.*, 17(4):1631–1661, 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347385. URL https://doi.org/10.1214/aos/1176347385.

[109] S. L. Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Oper. Res.*, 12:450–459, 1964. ISSN 0030-364X. doi: 10.1287/opre.12.3.450.

[110] J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417, 12 1948. ISSN 0006-3444. doi: 10.1093/biomet/35.3-4.414. URL https://doi.org/10.1093/biomet/35.3-4.414.

[111] Gabriel Y. Handler and Pitu B. Mirchandani. *Location on networks theory and algorithms*. MIT Press Series in Signal Processing, Optimization, and Control. MIT Press, Cambridge, Mass.-London, 1979. ISBN 0-262-08090-7.

[112] Pierre Hansen, Martine Labbé, Dominique Peeters, and Jacques-François Thisse. Single facility location on networks. In *Surveys in combinatorial optimization (Rio de Janeiro, 1985)*, volume 132 of *North-Holland Math. Stud.*, pages 113–145. North-Holland, Amsterdam, 1987. ISBN 0-444-70136-2. doi: 10.1016/S0304-0208(08)73234-1. URL https://doi.org/10.1016/S0304-0208(08)73234-1.

[113] Wolfgang Karl Härdle and Léopold Simar. *Applied Multivariate Statistical Analysis*. Springer International Publishing, Cham, 2019. ISBN 978-3-030-26006-4.

[114] P. Harmand, D. Werner, and W. Werner. *M-Ideals in Banach Spaces and Banach Algebras*, volume 1547 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1993. ISBN 3-540-56814-X. doi: 10.1007/BFb0084355. URL https://doi.org/10.1007/BFb0084355.

[115] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[116] Christian Hess. Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. *Ann. Statist.*, 24(3):1298–1315, 1996. ISSN 0090-5364. doi: 10.1214/aos/1032526970. URL https://doi.org/10.1214/aos/1032526970.

[117] Christian Hess and Raffaello Seri. Generic consistency for approximate stochastic programming and statistical problems. *SIAM J. Optim.*, 29(1):290–317, 2019.

ISSN 1052-6234. doi: 10.1137/17M1156769. URL https://doi.org/10.1137/17M1156769.

[118] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms. I*, volume 305 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. ISBN 3-540-56850-6. Fundamentals.

[119] Nils Lid Hjort and David Pollard. Asymptotics for minimisers of convex processes. Statistical Research Report, Univ. Oslo, 1993.

[120] J. Hoffmann-Jørgensen. *Probability with a View toward Statistics. Vol. II*. Chapman & Hall Probability Series. Chapman & Hall, New York, 1994. ISBN 0-412-05231-8. doi: 10.1007/978-1-4899-3019-4. URL https://doi.org/10.1007/978-1-4899-3019-4.

[121] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 2008. ISSN 0090-5364. doi: 10.1214/009053607000000677. URL https://doi.org/10.1214/009053607000000677.

[122] Roger A. Horn and Charles R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1991. ISBN 0-521-30587-X. doi: 10.1017/CBO9780511840371. URL https://doi.org/10.1017/CBO9780511840371.

[123] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. doi: 10.1017/9781139020411.

[124] Lajos Horváth and Piotr Kokoszka. *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer, New York, 2012. ISBN 978-1-4614-3654-6. doi: 10.1007/978-1-4614-3655-3. URL https://doi.org/10.1007/978-1-4614-3655-3.

[125] T. Hotz and S. Huckemann. Intrinsic means on the circle: uniqueness, locus and asymptotics. *Ann. Inst. Statist. Math.*, 67(1):177–193, 2015. ISSN 0020-3157,1572-9052. doi: 10.1007/s10463-013-0444-7. URL https://doi.org/10.1007/s10463-013-0444-7.

[126] Thomas Hotz, Stephan Huckemann, Huiling Le, J. S. Marron, Jonathan C. Mattingly, Ezra Miller, James Nolen, Megan Owen, Vic Patrangenaru, and Sean Skwerer. Sticky central limit theorems on open books. *Ann. Appl. Probab.*, 23(6):2238–2258, 2013. ISSN 1050-5164,2168-8737. doi: 10.1214/12-AAP899. URL https://doi.org/10.1214/12-AAP899.

[127] Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2015. ISBN 978-0-470-01691-6. doi: 10.1002/9781118762547. URL https://doi.org/10.1002/9781118762547.

[128] J.G. (https://math.stackexchange.com/users/56861/j g). Limiting value of the variance of the χ_n distribution (square root of χ_n^2 distribution). Mathematics Stack Exchange. URL https://math.stackexchange.com/q/3376610. URL:https://math.stackexchange.com/q/3376610 (version: 2019-10-01).

[129] Shouchuan Hu and Nikolas S. Papageorgiou. *Handbook of Multivalued Analysis. Vol. I*, volume 419 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1997. ISBN 0-7923-4682-3. doi: 10.1007/978-1-4615-6359-4. URL https://doi.org/10.1007/978-1-4615-6359-4. Theory.

[130] Junzhou Huang and Tong Zhang. The benefit of group sparsity. *Ann. Statist.*, 38 (4):1978–2004, 2010. ISSN 0090-5364,2168-8966. doi: 10.1214/09-AOS778. URL https://doi.org/10.1214/09-AOS778.

[131] Peter J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964. ISSN 0003-4851. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

[132] Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif., 1967.

[133] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6. doi: 10.1002/9780470434697. URL https://doi.org/10.1002/9780470434697.

[134] Stephan Huckemann and Herbert Ziezold. Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab.*, 38(2):299–319, 2006. ISSN 0001-8678,1475-6064. doi: 10.1239/aap/1151337073. URL https://doi.org/10.1239/aap/1151337073.

[135] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica*, 20(1):1–58, 2010. ISSN 1017-0405,1996-8507.

[136] Stephan Huckemann, Jonathan C. Mattingly, Ezra Miller, and James Nolen. Sticky central limit theorems at isolated hyperbolic planar singularities. *Electron. J. Probab.*, 20:no. 78, 34, 2015. ISSN 1083-6489. doi: 10.1214/EJP.v20-3887. URL https://doi.org/10.1214/EJP.v20-3887.

[137] Stephan F. Huckemann. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *Ann. Statist.*, 39(2):1098–1124, 2011. ISSN 0090-5364,2168-8966. doi: 10.1214/10-AOS862. URL https://doi.org/10.1214/10-AOS862.

[138] Stephan F. Huckemann and Benjamin Eltzner. Data analysis on nonstandard spaces. *Wiley Interdiscip. Rev. Comput. Stat.*, 13(3):Paper No. e1526, 19, 2021. ISSN 1939-5108,1939-0068. doi: 10.1002/wics.1526. URL https://doi.org/10.1002/wics.1526.

[139] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[140] Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[141] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622, 2018.

[142] Robert I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40:633–643, 1969. ISSN 0003-4851. doi: 10.1214/aoms/1177697731. URL https://doi.org/10.1214/aoms/1177697731.

[143] jlewk (https://math.stackexchange.com/users/484640/jlewk). Is it true that $A \geq B$ implies $B^\dagger \geq A^\dagger$ for singular positive semi-definite matrices? Mathematics Stack Exchange, . URL https://math.stackexchange.com/q/3682798. URL:https://math.stackexchange.com/q/3682798 (version: 2020-05-20).

[144] jlewk (https://math.stackexchange.com/users/484640/jlewk). Lipschitz continuity of $\sqrt{A}$. Mathematics Stack Exchange, . URL https://math.stackexchange.com/q/3968118. URL:https://math.stackexchange.com/q/3968118 (version: 2020-12-31).

[145] Jürgen Jost. *Nonpositive Curvature: Geometric and Analytic Aspects*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1997. ISBN 3-7643-5736-3. doi: 10.1007/978-3-0348-8918-6. URL https://doi.org/10.1007/978-3-0348-8918-6.

[146] Peter Kall. On approximations and stability in stochastic programming. In *Parametric optimization and related topics (Plaue, 1985)*, volume 35 of *Math. Res.*, pages 387–407. Akademie-Verlag, Berlin, 1987.

[147] H. Karcher. Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.*, 30(5):509–541, 1977. ISSN 0010-3640. doi: 10.1002/cpa.3160300502. URL https://doi.org/10.1002/cpa.3160300502.

[148] J. H. B. Kemperman. The median of a finite measure on a banach space. In Yadolah Dodge, editor, *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam, 1987.

[149] D. G. Kendall, D. Barden, T. K. Carne, and H. Le. *Shape and shape theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1999. ISBN 0-471-96823-4. doi: 10.1002/9780470317006. URL https://doi.org/10.1002/9780470317006.

[150] J. Kiefer. On Bahadur's representation of sample quantiles. *Ann. Math. Statist.*, 38:1323–1342, 1967. ISSN 0003-4851. doi: 10.1214/aoms/1177698690. URL https://doi.org/10.1214/aoms/1177698690.

[151] JeanKyung Kim and David Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1): 191–219, 1990. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1176347498. URL https://doi.org/10.1214/aos/1176347498.

[152] Alan J. King and Roger J.-B. Wets. Epi-consistency of convex stochastic programs. *Stochastics Stochastics Rep.*, 34(1-2):83–92, 1991. ISSN 1045-1129. doi: 10.1080/17442509108833676. URL https://doi.org/10.1080/17442509108833676.

[153] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 2000. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1015957397. URL https://doi.org/10.1214/aos/1015957397.

[154] Roger Koenker. *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 2005. ISBN 978-0-521-60827-5; 0-521-60827-9. doi: 10.1017/CBO9780511754098. URL https://doi.org/10.1017/CBO9780511754098.

[155] V. Koltchinskii. Bahadur-Kiefer approximation for spatial quantiles. In *Probability in Banach spaces, 9 (Sandjberg, 1993)*, volume 35 of *Progr. Probab.*, pages 401–415. Birkhäuser Boston, Boston, MA, 1994.

[156] V. Koltchinskii. Spatial quantiles and their Bahadur-Kiefer representations. In *Asymptotic statistics (Prague, 1993)*, Contrib. Statist., pages 361–367. Physica, Heidelberg, 1994.

[157] V. Koltchinskii. *M*-estimation and spatial quantiles. In *Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994)*, volume 109 of *Lect. Notes Stat.*, pages 235–250. Springer, New York, 1996. doi: 10.1007/978-1-4612-2380-1_16. URL https://doi.org/10.1007/978-1-4612-2380-1_16.

[158] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017. ISSN 1350-7265,1573-9759. doi: 10.3150/15-BEJ730. URL https://doi.org/10.3150/15-BEJ730.

[159] Erwin Kreyszig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, New York-London-Sydney, 1978. ISBN 0-471-50731-8.

[160] A. G. Kukush. Asymptotic normality of the estimator of an infinite-dimensional parameter in the model with a smooth regression function. *Math. Methods Statist.*, 5(3):343–356, 1996. ISSN 1066-5307,1934-8045.

[161] Hui Hsiung Kuo. *Gaussian Measures in Banach Spaces*. Lecture Notes in Mathematics, Vol. 463. Springer-Verlag, Berlin-New York, 1975.

[162] Lars Lammers, Do Tran Van, Tom M. W. Nye, and Stephan F. Huckemann. Types of Stickiness in BHV Phylogenetic Tree Spaces and Their Degree. In *Geometric science of information. Part I*, volume 14071 of *Lecture Notes in Comput. Sci.*, pages 357–365. Springer, Cham, 2023. ISBN 978-3-031-38270-3; 978-3-031-38271-0. doi: 10.1007/978-3-031-38271-0_35. URL https://doi.org/10.1007/978-3-031-38271-0_35.

[163] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 10 2000. doi: 10.1214/aos/1015957395. URL http://dx.doi.org/10.1214/aos/1015957395.

[164] Huiling Le. Locating Fréchet means with application to shape spaces. *Adv. in Appl. Probab.*, 33(2):324–338, 2001. ISSN 0001-8678,1475-6064. doi: 10.1239/aap/999188316. URL https://doi.org/10.1239/aap/999188316.

[165] Denis Le Bihan, Jean-François Mangin, Cyril Poupon, Chris A. Clark, Sabina Pappata, Nicolas Molko, and Hughes Chabriat. Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, 13(4):534–546, 2001. doi: https://doi.org/10.1002/jmri.1076. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.1076.

[166] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields*, 168(3-4):901–917, 2017. ISSN 0178-8051,1432-2064. doi: 10.1007/s00440-016-0727-z. URL https://doi.org/10.1007/s00440-016-0727-z.

[167] Thibaut Le Gouic, Quentin Paris, Philippe Rigollet, and Austin J. Stromme. Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. *J. Eur. Math. Soc. (JEMS)*, 25(6):2229–2250, 2023. ISSN 1435-9855,1435-9863. doi: 10.4171/jems/1234. URL https://doi.org/10.4171/jems/1234.

[168] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, Berlin, 1991. ISBN 3-540-52013-9. doi: 10.1007/978-3-642-20212-4. URL https://doi.org/10.1007/978-3-642-20212-4.

[169] Stephen John Leese. *Set-valued functions and selectors*. PhD thesis, Keele University, July 1974. URL https://eprints.keele.ac.uk/6034/.

[170] Carlos Antonio León and Jean-Claude Massé. A counterexample on the existence of the L_1-median. *Statist. Probab. Lett.*, 13(2):117–120, 1992. ISSN 0167-7152. doi: 10.1016/0167-7152(92)90085-J. URL https://doi.org/10.1016/0167-7152(92)90085-J.

[171] Christophe Ley and Thomas Verdebout. *Modern directional statistics*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Boca Raton, FL, 2017. ISBN 978-1-4987-0664-3.

[172] Christopher Liaw, Abbas Mehrabian, Yaniv Plan, and Roman Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric aspects of functional analysis*, pages 277–299. Springer, 2017.

[173] Rong Rong Lin, Hai Zhang Zhang, and Jun Zhang. On reproducing kernel banach spaces: Generic definitions and unified framework of constructions. *Acta Mathematica Sinica, English Series*, 38(8):1459–1483, Aug 2022. ISSN 1439-7617. doi: 10.1007/s10114-022-1397-7. URL https://doi.org/10.1007/s10114-022-1397-7.

[174] Karim Lounici, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

[175] Roberto Lucchetti. *Convexity and Well-Posed Problems*, volume 22 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer, New York, 2006. ISBN 978-0387-28719-5; 0-387-28719-1. doi: 10.1007/0-387-31082-7. URL https://doi.org/10.1007/0-387-31082-7.

[176] Roberto Lucchetti and R. J.-B. Wets. Convergence of minima of integral functionals, with applications to optimal control and stochastic optimization. *Statist. Decisions*, 11(1):69–84, 1993. ISSN 0721-2631.

[177] Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc. (JEMS)*, 22(3):925–965, 2020. ISSN 1435-9855,1435-9863. doi: 10.4171/jems/937. URL https://doi.org/10.4171/jems/937.

[178] Françoise Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. *C. R. Acad. Sci. Paris Sér. I Math.*, 303(7):289–292, 1986. ISSN 0249-6291.

[179] Jacopo Mandozzi and Peter Bühlmann. Hierarchical testing in the high-dimensional setting with correlated variables. *Journal of the American Statistical Association*, 111(513):331–343, 2016.

[180] Jacopo Mandozzi and Peter Bühlmann. A sequential rejection testing method for high-dimensional regression with correlated variables. *The international journal of biostatistics*, 12(1):79–95, 2016.

[181] Kanti V. Mardia and Peter E. Jupp. *Directional statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000. ISBN 0-471-95333-4. Revised reprint of *Statistics of directional data* by Mardia [MR0336854 (49 #1627)].

[182] André Mas. A sufficient condition for the CLT in the space of nuclear operators—application to covariance of random functions. *Statist. Probab. Lett.*, 76(14): 1503–1509, 2006. ISSN 0167-7152,1879-2103. doi: 10.1016/j.spl.2006.03.010. URL https://doi.org/10.1016/j.spl.2006.03.010.

[183] Mathurin Massias, Quentin Bertrand, Alexandre Gramfort, and Joseph Salmon. Support recovery and sup-norm convergence rates for sparse pivotal estimation. *arXiv preprint arXiv:2001.05401*, 2020.

[184] Charles A. McCarthy. c_p. *Israel J. Math.*, 5:249–271, 1967. ISSN 0021-2172. doi: 10.1007/BF02771613. URL https://doi.org/10.1007/BF02771613.

[185] Andrew McCormack and Peter Hoff. The Stein effect for Fréchet means. *Ann. Statist.*, 50(6):3647–3676, 2022. ISSN 0090-5364,2168-8966. doi: 10.1214/22-aos2245. URL https://doi.org/10.1214/22-aos2245.

[186] Robert E. Megginson. *An Introduction to Banach Space Theory*, volume 183 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998. ISBN 0-387-98431-3. doi: 10.1007/978-1-4612-0603-3. URL https://doi.org/10.1007/978-1-4612-0603-3.

[187] Nicolai Meinshausen. Group bound: confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(5):923–945, 2015.

[188] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[189] Reinhold Meise and Dietmar Vogt. *Introduction to Functional Analysis*, volume 2 of *Oxford Graduate Texts in Mathematics*. The Clarendon Press, Oxford University Press, New York., 1997. ISBN 0-19-851485-9. Translated from the German by M. S. Ramanujan and revised by the authors.

[190] metamorphy (https://math.stackexchange.com/users/543769/metamorphy). Inequality involving the gamma function: $\int_0^a x^{a-1}e^{-x}dx > \frac{1}{2}\gamma(a)$. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/3590470. URL:https://math.stackexchange.com/q/3590470 (version: 2020-03-22).

[191] P. Milasevic and G. R. Ducharme. Uniqueness of the spatial median. *Ann. Statist.*, 15(3):1332–1333, 1987. ISSN 0090-5364. doi: 10.1214/aos/1176350511. URL https://doi.org/10.1214/aos/1176350511.

[192] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015. ISSN 1350-7265. doi: 10.3150/14-BEJ645. URL https://doi.org/10.3150/14-BEJ645.

[193] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

[194] Ritwik Mitra and Cun-Hui Zhang. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics*, 10(2): 1829–1873, 2016.

[195] Ilya Molchanov. *Theory of Random Sets*, volume 87 of *Probability Theory and Stochastic Modelling*. Springer-Verlag, London, 2017. ISBN 978-1-4471-7347-2; 978-1-4471-7349-6. Second edition of [MR2132405].

[196] Umberto Mosco. Convergence of convex sets and of solutions of variational inequalities. *Advances in Math.*, 3:510–585, 1969. ISSN 0001-8708. doi: 10.1016/0001-8708(69)90009-7. URL https://doi.org/10.1016/0001-8708(69)90009-7.

[197] Karl Mosler and Pavlo Mozharovskyi. Choosing among notions of multivariate depth statistics. *Statist. Sci.*, 37(3):348–368, 2022. ISSN 0883-4237,2168-8745. doi: 10.1214/21-sts827. URL https://doi.org/10.1214/21-sts827.

[198] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. ISSN 1935-8237. doi: 10.1561/2200000060. URL http://dx.doi.org/10.1561/2200000060.

[199] Wojciech Niemiro. Asymptotics for M-estimators defined by convex minimization. *Ann. Statist.*, 20(3):1514–1533, 1992. ISSN 0090-5364. doi: 10.1214/aos/1176348782. URL https://doi.org/10.1214/aos/1176348782.

[200] Paul T. Nkansah and H. T. David. Network median problems with continuously distributed demand. *Transportation Sci.*, 20(3):213–219, 1986. ISSN 0041-1655. doi: 10.1287/trsc.20.3.213. URL https://doi.org/10.1287/trsc.20.3.213.

[201] Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

[202] Shin-ichi Ohta. Barycenters in Alexandrov spaces of curvature bounded below. *Adv. Geom.*, 12(4):571–587, 2012. ISSN 1615-715X. doi: 10.1515/advgeom-2011-058. URL https://doi.org/10.1515/advgeom-2011-058.

[203] Athanase Papadopoulos. *Metric Spaces, Convexity and Nonpositive Curvature*, volume 6 of *IRMA Lectures in Mathematics and Theoretical Physics*. European Mathematical Society (EMS), Zürich, second edition, 2014. ISBN 978-3-03719-132-3. doi: 10.4171/132. URL https://doi.org/10.4171/132.

[204] K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Probability and Mathematical Statistics, No. 3. Academic Press, Inc., New York-London, 1967.

[205] Riccardo Passeggeri and Nancy Reid. On quantiles, continuity and robustness. *arXiv preprint arXiv:2206.06998v4*, 2022.

[206] Victor Patrangenaru and Leif Ellingson. *Nonparametric Statistics on Manifolds and their Applications to Object Data Analysis*. CRC Press, Boca Raton, FL, 2016. ISBN 978-1-4398-2050-6. With a foreword by Victor Pambuccian.

[207] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[208] Aleksander Pełčzyński and MichałWojciechowski. Spaces of functions with bounded variation and Sobolev spaces without local unconditional structure. *J. Reine Angew. Math.*, 558:109–157, 2003. ISSN 0075-4102. doi: 10.1515/crll.2003.036. URL https://doi.org/10.1515/crll.2003.036.

[209] Xavier Pennec, Stefan Sommer, and Tom Fletcher, editors. *Riemannian Geometric Statistics in Medical Image Analysis*. Elsevier/Academic Press, London, 2020. ISBN 978-0-12-814725-2.

[210] Michael D. Perlman. On the strong consistency of approximate maximum likelihood estimators. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics*, pages 263–281, 1972.

[211] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073.

[212] J. Pfanzagl. On the measurability and consistency of minimum contrast estimates. *Metrika*, 14(1):249–272, Dec 1969. ISSN 1435-926X. doi: 10.1007/BF02613654. URL https://doi.org/10.1007/BF02613654.

[213] Iosif Pinelis. Convergence in probability of a supremum. MathOverflow, 2022. URL https://mathoverflow.net/q/429734. URL: https://mathoverflow.net/q/429734 (version: 2022-09-04).

[214] David Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984. ISBN 0-387-90990-7. doi: 10.1007/978-1-4612-5254-2. URL https://doi.org/10.1007/978-1-4612-5254-2.

[215] David Pollard. New ways to prove central limit theorems. *Econometric Theory*, 1(3):295–313, 1985.

[216] J. O. Ramsay and B. W. Silverman. *Applied Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 0-387-95414-7. doi: 10.1007/b98886. URL https://doi.org/10.1007/b98886. Methods and case studies.

[217] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005. ISBN 978-0387-40080-8; 0-387-40080-X.

[218] M. M. Rao and Z. D. Ren. *Theory of Orlicz Spaces*, volume 146 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, 1991. ISBN 0-8247-8478-2.

[219] R. Ranga Rao. Relations between Weak and Uniform Convergence of Measures with Applications. *The Annals of Mathematical Statistics*, 33(2):659–680, 1962. doi: 10.1214/aoms/1177704588. URL https://doi.org/10.1214/aoms/1177704588.

[220] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[221] Yogesh Rathi, Allen Tannenbaum, and Oleg Michailovich. Segmenting images on the tensor manifold. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383010.

[222] Stephen M. Robinson and Roger J.-B. Wets. Stability in two-stage stochastic programming. *SIAM J. Control Optim.*, 25(6):1409–1416, 1987. ISSN 0363-0129. doi: 10.1137/0325077. URL https://doi.org/10.1137/0325077.

[223] Gabriel Romon. Statistical properties of approximate geometric quantiles in infinite-dimensional Banach spaces. *arXiv preprint arXiv:2211.00035v3*, 2022.

[224] Gabriel Romon and Victor-Emmanuel Brunel. Convex generalized Fréchet means in a metric tree. *arXiv preprint arXiv:2310.17435v2*, 2023.

[225] Frank Alexander Ross. Editor's note on the center of population and point of minimum travel. *Journal of the American Statistical Association*, 25(172):447–452, 1930. doi: 10.1080/01621459.1930.10502217. URL https://doi.org/10.1080/01621459.1930.10502217.

[226] Johannes O. Royset and Roger J-B Wets. Variational analysis of constrained M-estimators. *The Annals of Statistics*, 48(5):2759–2790, 2020. doi: 10.1214/19-AOS1905. URL https://doi.org/10.1214/19-AOS1905.

[227] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Book Co., New York, third edition, 1987. ISBN 0-07-054234-1.

[228] Manfred Schäl. A selection theorem for optimization problems. *Arch. Math. (Basel)*, 25:219–224, 1974. ISSN 0003-889X. doi: 10.1007/BF01238668. URL https://doi.org/10.1007/BF01238668.

[229] René L. Schilling. *Measures, Integrals and Martingales*. Cambridge University Press, Cambridge, second edition, 2017. ISBN 978-1-316-62024-3.

[230] Christof Schötz. Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electron. J. Stat.*, 13(2):4280–4345, 2019. doi: 10.1214/19-EJS1618. URL https://doi.org/10.1214/19-EJS1618.

[231] Christof Schötz. Strong laws of large numbers for generalizations of Fréchet mean sets. *Statistics*, 56(1):34–52, 2022. ISSN 0233-1888. doi: 10.1080/02331888.2022.2032063. URL https://doi.org/10.1080/02331888.2022.2032063.

[232] Robert Serfling. Quantile functions for multivariate analysis: approaches and applications. volume 56, pages 214–232. 2002. doi: 10.1111/1467-9574.00195. URL https://doi.org/10.1111/1467-9574.00195. Special issue: Frontier research in theoretical statistics, 2000 (Eindhoven).

[233] Jun Shao. *Mathematical Statistics*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2003. ISBN 0-387-95382-5. doi: 10.1007/b97553. URL https://doi.org/10.1007/b97553.

[234] Alexander Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.*, 17(2):841–858, 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347146. URL https://doi.org/10.1214/aos/1176347146.

[235] D. R. Shier and P. M. Dearing. Optimal locations for a class of nonlinear, single-facility location problems on a network. *Oper. Res.*, 31(2):292–303, 1983. ISSN 0030-364X,1526-5463. doi: 10.1287/opre.31.2.292. URL https://doi.org/10.1287/opre.31.2.292.

[236] Albert N. Shiryaev. *Probability. 1*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, third edition, 2016. ISBN 978-0-387-72205-4; 978-0-387-72206-1.

[237] Albert N. Shiryaev. *Probability. 2*, volume 95 of *Graduate Texts in Mathematics*. Springer, New York, third edition, 2019. ISBN 978-0-387-72207-8; 978-0-387-72208-5. doi: 10.1007/978-0-387-72208-5. URL https://doi.org/10.1007/978-0-387-72208-5.

[238] Beatriz Sinova, Gil González-Rodríguez, and Stefan Van Aelst. M-estimators of location for functional data. *Bernoulli*, 24(3):2328–2357, 2018. ISSN 1350-7265. doi: 10.3150/17-BEJ929. URL https://doi.org/10.3150/17-BEJ929.

[239] Christopher G. Small. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277, 1990. ISSN 03067734, 17515823. doi: 10.2307/1403809. URL https://doi.org/10.2307/1403809.

[240] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

[241] Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[242] Benjamin Stucky and Sara van de Geer. Asymptotic confidence regions for high-dimensional structured sparsity. *IEEE Transactions on Signal Processing*, 66(8): 2178–2190, 2018.

[243] Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, volume 338 of *Contemp. Math.*, pages 357–390. Amer. Math. Soc., Providence, RI, 2003. doi: 10.1090/conm/338/06080. URL https://doi.org/10.1090/conm/338/06080.

[244] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99 (4):879–898, 2012.

[245] Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

[246] Kondagunta Sundaresan and Srinivasa Swaminathan. *Geometry and Nonlinear Analysis in Banach Spaces*, volume 1131 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1985. ISBN 3-540-15237-7. doi: 10.1007/BFb0075323. URL https://doi.org/10.1007/BFb0075323.

[247] Harald Sverdrup-Thygeson. Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces. *Ann. Statist.*, 9(1):141–145, 1981. ISSN 0090-5364,2168-8966. URL http://links.jstor.org/sici?sici=0090-5364(198101)9:1<141:SLOLNF>2.0.CO;2-Y&origin=MSN.

[248] Barbaros C. Tansel, Richard L. Francis, and Timothy J. Lowe. Location on networks: a survey. I. The p-center and p-median problems. *Management Sci.*, 29(4):482–497, 1983. ISSN 0025-1909. doi: 10.1287/mnsc.29.4.482. URL https://doi.org/10.1287/mnsc.29.4.482.

[249] Barbaros C. Tansel, Richard L. Francis, and Timothy J. Lowe. Location on networks: a survey. II. Exploiting tree network structure. *Management Sci.*, 29 (4):498–511, 1983. ISSN 0025-1909. doi: 10.1287/mnsc.29.4.498. URL https://doi.org/10.1287/mnsc.29.4.498.

[250] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[251] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. ISSN 0035-9246. URL http://links.jstor.org/sici?sici=0035-9246(1996)58:1<267:RSASVT>2.0.CO;2-G&origin=MSN.

[252] N. N. Vakhania, V. I. Tarieladze, and S. A. Chobanyan. *Probability Distributions on Banach Spaces*, volume 14 of *Mathematics and its Applications (Soviet Series)*. D. Reidel Publishing Co., Dordrecht, 1987. ISBN 90-277-2496-2. doi: 10.1007/978-94-009-3873-1. URL https://doi.org/10.1007/978-94-009-3873-1.

[253] M. Valadier. La multiapplication médianes conditionnelles. *Travaux Sém. Anal. Convexe*, 12(2):exp. no. 21, 15, 1982.

[254] M. Valadier. La multi-application médianes conditionnelles. *Z. Wahrsch. Verw. Gebiete*, 67(3):279–282, 1984. ISSN 0044-3719. doi: 10.1007/BF00535005. URL https://doi.org/10.1007/BF00535005.

[255] Sara van de Geer and Benjamin Stucky. χ 2-confidence sets in high-dimensional regression. In *Statistical analysis for high-dimensional data*, pages 279–306. Springer, 2016.

[256] Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[257] A. W. van der Vaart. Infinite-dimensional M-estimation. In *Probability theory and mathematical statistics (Vilnius, 1993)*, pages 715–733. TEV, Vilnius, 1994. ISBN 90-6764-178-2.

[258] A. W. van der Vaart. Efficiency of infinite-dimensional M-estimators. *Statist. Neerlandica*, 49(1):9–30, 1995. ISSN 0039-0402,1467-9574. doi: 10.1111/j.1467-9574.1995.tb01452.x. URL https://doi.org/10.1111/j.1467-9574.1995.tb01452.x.

[259] A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9; 0-521-78450-6. doi: 10.1017/CBO9780511802256. URL https://doi.org/10.1017/CBO9780511802256.

[260] Aad van der Vaart. Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, volume 1781 of *Lecture Notes in Math.*, pages 331–457. Springer, Berlin, 2002. ISBN 3-540-43736-3.

[261] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. doi: 10.1007/978-1-4757-2545-2. URL https://doi.org/10.1007/978-1-4757-2545-2.

[262] V. S. Varadarajan. On the convergence of sample probability distributions. *Sankhyā*, 19:23–26, 1958. ISSN 0036-4452.

[263] Roman Vershynin. *High-Dimensional Probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. doi: 10.1017/9781108231596. URL

https://doi.org/10.1017/9781108231596. An introduction with applications in data science, With a foreword by Sara van de Geer.

[264] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

[265] V. F. Šolohovič. On stability of extremal problems in Banach space. *Mathematics of the USSR-Sbornik*, 14(3):417–427, apr 1971. doi: 10.1070/sm1971v014n03abeh002625. URL https://doi.org/10.1070/sm1971v014n03abeh002625.

[266] Daniel H. Wagner. Survey of measurable selection theorems. *SIAM J. Control Optim.*, 15(5):859–903, 1977. ISSN 0363-0129. doi: 10.1137/0315056. URL https://doi.org/10.1137/0315056.

[267] Daniel H. Wagner. Survey of measurable selection theorems: an update. In *Measure theory, Oberwolfach 1979 (Proc. Conf., Oberwolfach, 1979)*, volume 794 of *Lecture Notes in Math.*, pages 176–219. Springer, Berlin-New York, 1980. doi: 10.1007/BFb0088224. URL https://doi.org/10.1007/BFb0088224.

[268] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[269] Martin J. Wainwright. *High-Dimensional Statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. ISBN 978-1-108-49802-9. doi: 10.1017/9781108627771. URL https://doi.org/10.1017/9781108627771. A non-asymptotic viewpoint.

[270] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016. doi: 10.1146/annurev-statistics-041715-033624. URL https://doi.org/10.1146/annurev-statistics-041715-033624.

[271] Jinde Wang. Asymptotics of least-squares estimators for constrained nonlinear regression. *Ann. Statist.*, 24(3):1316–1326, 1996. ISSN 0090-5364. doi: 10.1214/aos/1032526971. URL https://doi.org/10.1214/aos/1032526971.

[272] Alfred Weber. *Über den Standort der Industrien*. Tübingen, J.C.B. Mohr (Paul Siebeck)., 1909. English translation by Freidrich, C. J. (1929), *Alfred Weber's Theory of Location of Industries*, University of Chicago Press.

[273] Takumi Yokota. Convex functions and barycenter on CAT(1)-spaces of small radii. *J. Math. Soc. Japan*, 68(3):1297–1323, 2016. ISSN 0025-5645,1881-1167. doi: 10.2969/jmsj/06831297. URL https://doi.org/10.2969/jmsj/06831297.

[274] Takumi Yokota. Convex functions and p-barycenter on CAT(1)-spaces of small radii. *Tsukuba J. Math.*, 41(1):43–80, 2017. ISSN 0387-4982,2423-821X. doi: 10.21099/tkbjm/1506353559. URL https://doi.org/10.21099/tkbjm/1506353559.

[275] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006. ISSN 1369-7412,1467-9868. doi: 10.1111/j.1467-9868.2005.00532.x. URL https://doi.org/10.1111/j.1467-9868.2005.00532.x.

[276] Ho Yun and Byeong U. Park. Exponential concentration for geometric-median-of-means in non-positive curvature spaces. *Bernoulli*, 29(4):2927–2960, 2023. ISSN 1350-7265,1573-9759. doi: 10.3150/22-bej1569. URL https://doi.org/10.3150/22-bej1569.

[277] Mihail Zervos. On the epiconvergence of stochastic optimization problems. *Math. Oper. Res.*, 24(2):495–508, 1999. ISSN 0364-765X. doi: 10.1287/moor.24.2.495. URL https://doi.org/10.1287/moor.24.2.495.

[278] Cun-Hui Zhang. Statistical inference for high-dimensional data. *Mathematisches Forschungsinstitut Oberwolfach: Very High Dimensional Semiparametric Models, Report*, (48):28–31, 2011.

[279] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 08 2008. doi: 10.1214/07-AOS520. URL https://doi.org/10.1214/07-AOS520.

[280] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[281] Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel Banach spaces for machine learning. *J. Mach. Learn. Res.*, 10:2741–2775, 2009. ISSN 1532-4435. doi: 10.1109/IJCNN.2009.5179093. URL https://doi.org/10.1109/IJCNN.2009.5179093.

[282] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

[283] Qing Zhou and Seunghyun Min. Uncertainty quantification under group sparsity. *Biometrika*, 104(3):613–632, 2017.

[284] Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113 (524):1583–1600, 2018.

[285] Yinchu Zhu and Jelena Bradic. Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364, 2018.

[286] Herbert Ziezold. On expected figures and a strong law of large numbers for random elements in quasi-metric spaces. In *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the Eighth European Meeting of Statisticians (Tech. Univ. Prague, Prague, 1974), Vol. A*, pages 591–602. Reidel, Dordrecht-Boston, Mass., 1977. ISBN 90-277-0852-5.

[287] C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific Publishing Co., Inc., River Edge, NJ, 2002. ISBN 981-238-067-1. doi: 10.1142/9789812777096. URL https://doi.org/10.1142/9789812777096.

[288] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000. ISSN 0090-5364,2168-8966. doi: 10.1214/aos/1016218226. URL https://doi.org/10.1214/aos/1016218226.

ECOLE DOCTORALE DE MATHEMATIQUES HADAMARD

Titre : Contributions à la statistique de grande dimension, de dimension infinie et dans les espaces métriques

Mots clés : Régression multivariée, Quantile géométrique, Moyenne de Fréchet, Statistique en grande dimension, Statistique en dimension infinie, Statistique dans les espaces métriques

Résumé : Trois problèmes sont abordés dans cette thèse: l'inférence en régression multi-tâche de grande dimension, les quantiles géométriques dans les espaces normés de dimension infinie, et les moyennes de Fréchet généralisées dans les arbres métriques. Premièrement, nous considérons un modèle de régression multi-tâche avec une hypothèse de sparsité sur les lignes de la matrice paramètre. L'estimation est faite en haute dimension avec l'estimateur Lasso multi-tâche. Afin de corriger le biais induit par la pénalité, nous introduisons un nouvel objet dépendant uniquement des données que nous appelons matrice d'interaction. Cet outil nous permet d'établir des résultats asymptotiques avec des lois limites normales ou χ^2. Il en découle des intervalles de confiance et des ellipsoïdes de confiance, qui sont valides dans des régimes de sparsité qui ne sont pas couverts par la littérature existante. Deuxièmement, nous étudions le quantile géométrique, qui généralise le quantile classique au cadre des espaces normés. Nous commençons par fournir de nouveaux résultats sur l'existence et l'unicité des quantiles géométriques. L'estimation est effectuée avec un M-estimateur approché et nous examinons ses propriétés asymptotiques en dimension infinie. Quand le quantile théorique n'est pas unique, nous utilisons la théorie de la convergence variationnelle pour obtenir des résultats asymptotiques sur les sous-suites dans la topologie faible. Quand le quantile théorique est unique, nous montrons que l'estimateur est consistant pour la topologie de la norme dans une large classe d'espaces de Banach, en particulier dans les espaces séparables et uniformément convexes. Dans les Hilbert séparables nous démontrons des représentations de Bahadur–Kiefer de l'estimateur, dont découle immédiatement la normalité asymptotique à la vitesse paramétrique. Finalement, nous considérons des mesures de tendance centrale pour des données vivant sur un réseau, qui est modélisé par un arbre métrique. Les paramètres de localisation que nous étudions sont appelés moyennes de Fréchet généralisées: elles sont obtenues en remplaçant le carré dans la définition de la moyenne de Fréchet par une fonction de perte convexe et croissante. Nous élaborons une notion de dérivée directionnelle dans l'arbre, ce qui nous aide à localiser et caractériser les minimiseurs. Nous examinons les propriétés statistiques du M-estimateur correspondant: nous étendons le concept de moyenne collante au contexte des arbres métriques, puis nous obtenons un théorème collant non-asymptotique et une loi des grands nombres collante. Pour la médiane de Fréchet, nous établissons des bornes de concentration non-asymptotiques et des théorèmes central limite collants.

Title : Contributions to high-dimensional, infinite-dimensional and nonlinear statistics

Keywords : Multi-task regression, Geometric quantile, Fréchet mean, High-dimensional statistics, Infinite-dimensional statistics, Nonlinear statistics

Abstract : Three topics are explored in this thesis: inference in high-dimensional multi-task regression, geometric quantiles in infinite-dimensional Banach spaces and generalized Fréchet means in metric trees. First, we consider a multi-task regression model with a sparsity assumption on the rows of the unknown parameter matrix. Estimation is performed in the high-dimensional regime using the multi-task Lasso estimator. To correct for the bias induced by the penalty, we introduce a new data-driven object that we call the interaction matrix. This tool lets us develop normal and chi-square asymptotic distribution results, from which we obtain confidence intervals and confidence ellipsoids in sparsity regimes that are not covered by the existing literature. Second, we study the geometric quantile, which generalizes the classical univariate quantile to normed spaces. We begin by providing new results on the existence and uniqueness of geometric quantiles. Estimation is then conducted with an approximate M-estimator and we investigate its large-sample properties in infinite dimension. When the population quantile is not uniquely defined, we leverage the theory of variational convergence to obtain asymptotic statements on subsequences in the weak topology. When there is a unique population quantile, we show that the estimator is consistent in the norm topology for a wide range of Banach spaces including every separable uniformly convex space. In separable Hilbert spaces, we establish novel Bahadur–Kiefer representations of the estimator, from which asymptotic normality at the parametric rate follows. Lastly, we consider measures of central tendency for data that lives on a network, which is modeled by a metric tree. The location parameters that we study are called generalized Fréchet means: they obtained by relaxing the square in the definition of the Fréchet mean to an arbitrary convex nondecreasing loss. We develop a notion of directional derivative in the tree, which helps us locate and characterize the minimizers. We examine the statistical properties of the corresponding M-estimator: we extend the notion of stickiness to the setting of metrics trees, and we state a non-asymptotic sticky theorem, as well as a sticky law of large numbers. For the Fréchet median, we develop non-asymptotic concentration bounds and sticky central limit theorems.

Institut Polytechnique de Paris
91120 Palaiseau, France