Estimation beyond Missing (Completely) at Random

Tianyi Ma[*], Kabir A. Verchand[*,°], Thomas B. Berrett[†], Tengyao Wang[‡] and Richard J. Samworth[*]

[*]Statistical Laboratory, University of Cambridge
[°]Industrial and Systems Engineering, Georgia Institute of Technology
[†]Department of Statistics, University of Warwick
[‡]Department of Statistics, London School of Economics

October 15, 2024

Abstract

We study the effects of missingness on the estimation of population parameters. Moving beyond restrictive missing completely at random (MCAR) assumptions, we first formulate a missing data analogue of Huber's arbitrary ϵ-contamination model. For mean estimation with respect to squared Euclidean error loss, we show that the minimax quantiles decompose as a sum of the corresponding minimax quantiles under a heterogeneous, MCAR assumption, and a robust error term, depending on ϵ, that reflects the additional error incurred by departure from MCAR.

We next introduce natural classes of *realisable ϵ-contamination models*, where an MCAR version of a base distribution P is contaminated by an arbitrary missing not at random (MNAR) version of P. These classes are rich enough to capture various notions of biased sampling and sensitivity conditions, yet we show that they enjoy improved minimax performance relative to our earlier arbitrary contamination classes for both parametric and nonparametric classes of base distributions. For instance, with a univariate Gaussian base distribution, consistent mean estimation over realisable ϵ-contamination classes is possible even when ϵ and the proportion of missingness converge (slowly) to 1. Finally, we extend our results to the setting of departures from missing at random (MAR) in normal linear regression with a realisable missing response.

1 Introduction

A major theme of modern statistical research concerns problems where we wish to make inference about (some aspect of) a target population, but do not have access to an independent sample of size n from this distribution. Departures from this idealised scenario may take many different forms: spatial, temporal or some other form of dependence may be present (Cressie, 2015; Brockwell and Davis, 1991), or (some of) our data may be drawn from a source distribution that is different from, but related to, our target population, as in

transfer learning (Cai and Wei, 2021; Reeve, Cannings and Samworth, 2021). In a similar vein, the field of robust statistics aims to draw reliable inference when some of our data may be contaminated (Huber, 1964).

One of the most common ways in which observed data may fail to represent a sample from a target population is when components may be missing or unobserved. Even the relatively benign setting where data are missing completely at random (MCAR)—that is, when the data generating and missingness mechanisms are independent—presents substantial challenges for practitioners and theoreticians alike. A significant, ongoing research effort has therefore sought to introduce appropriate methodology under the MCAR hypothesis in several contemporary statistical problems, including sparse linear regression (Loh and Wainwright, 2012; Belloni, Rosenbaum and Tsybakov, 2017), classification (Cai and Zhang, 2019; Sell, Berrett and Cannings, 2024), sparse or high-dimensional principal component analysis (Elsener and van de Geer, 2019; Zhu, Wang and Samworth, 2022; Yan, Chen and Fan, 2024), covariance and precision matrix estimation (Lounici, 2014; Loh and Tan, 2018) and high-dimensional changepoint estimation (Xie, Huang and Willett, 2012; Follain, Wang and Samworth, 2022).

Despite this progress, it is frequently argued that MCAR should be regarded very much as the exception rather than the rule in applications. For instance, supporters of one political party may be less likely than other voters to respond to survey requests (Kennedy et al., 2018), while in education, efforts to model the value added by teachers may be hindered by large numbers of students with incomplete records and the tendency for those students to be lower achieving (McCaffrey and Lockwood, 2011). Likewise, in epidemiology, individuals with depression may be less likely to participate in a survey than those without depression (Prince, 2012), while metabolomic data are typically subject to a high proportion of non-MCAR missingness due to a metabolite-specific missingness mechanism in which more abundant analytes are more likely to be observed (Do et al., 2018; McKennan, Ober and Nicolae, 2020).

The most well-studied alternative to MCAR is the missing at random (MAR) hypothesis (Little and Rubin, 2014; Seaman et al., 2013; Farewell, Daniel and Seaman, 2022). The main virtue of this assumption is that, in well-specified, identifiable parametric models, likelihood-based methods may retain parametric rates of convergence to population estimands. On the other hand, it also has several drawbacks: first, it may well still be too restrictive as an appropriate missingness model for practical data sets (e.g. in the examples of the previous paragraph). Second, its links to likelihood-based methods and simple missingness patterns limit its applicability; third, even in simple parametric models, population parameters may be unidentifiable under MAR (see Section 2.2.1); and finally, it fails to measure proximity to the MCAR class in an appropriate, continuous fashion, and may therefore be unable to capture the essence of a given statistical challenge.

Our goal in this paper is to commence a line of work that seeks to understand the extent to which (non-MCAR) missingness affects our ability to estimate population parameters. We primarily focus here on the most basic statistical problem of mean estimation, though we extend our results to regression settings where the response variable may be missing in Section 5. In order to address the fundamental difficulty of the challenge, we introduce Huber-style models that interpolate between MCAR and larger classes that allow much more general dependence relationships between the data generating and missingness mechanisms. We measure performance of estimators via their squared Euclidean error, but since this loss

function is unbounded and we have a positive probability under our models of observing no data, the minimax risk is infinite (so uninformative for the purposes of comparing estimators). Instead, we work with the recently-developed minimax quantile framework; see Section 2.4.

We begin in Section 2 by introducing a formal framework for studying missing data via extended measurable spaces, which allow for missing components. Our main statistical models are what we refer to as *arbitrary ϵ-contamination* and *realisable ϵ-contamination* models. In the former, we perturb a distribution P that is subject to MCAR missingness by an additional mixture component (having corresponding mixture proportion ϵ) that may be an arbitrary distribution on our extended measurable space; in particular, this latter mixture component may be viewed as an MNAR version of an arbitrary distribution P'. On the other hand, in our realisable classes, although we again allow mixture perturbations of a base distribution P subject to MCAR missingness, we now require the contamination component to be an MNAR version of P itself. Although we are not aware of previous studies of these realisable classes, we believe that in many practical settings, it is appropriate to regard our data (whether observed or not) as arising from a particular base distribution, and the missingness mechanism only playing a role thereafter (even if it is potentially dependent on the data). Such a setting would result in our observed data as being from a realisable model, and these classes therefore form a natural way to restrict the vast array of different possible dependence relationships between data generating and missingness mechanisms. As we establish in this work, they also offer the potential for improved performance guarantees relative to those available for arbitrary contamination models.

In Section 3, we study the minimax quantiles of our squared Euclidean error loss function over arbitrary ϵ-contamination models. Theorem 3 provides an upper bound via an iterative imputation version of the robust descent algorithm of Depersin and Lecué (2022b). This bound decomposes as a sum of an MCAR term and a term quantifying the effect of contamination from MCAR. A corresponding lower bound on the minimax quantile given in Theorem 4 reveals that, at least when the covariance matrix of our base distribution is diagonal, the upper bound in Theorem 3 is optimal up to multiplicative constants in terms of its behaviour under departures from MCAR, and is optimal up to logarithmic factors in the dimension and quantile level in the MCAR term.

We turn our attention in Section 4.1 to realisable contamination of a Gaussian base distribution. Focusing for now on the univariate case for simplicity of exposition, we introduce a minimum Kolmogorov distance estimator, and show in Theorems 6 and 7 that it achieves the minimax optimal rate for both the MCAR and MCAR departure terms, except for a possible logarithmic dependence in an intermediate effective contamination level regime that vanishes with the effective sample size. This latter result also reveals the surprising fact that consistent mean estimation is possible in this model even in settings where the proportion of missingness and the proportion of MNAR contamination converge (slowly) to 1. Section 4.2 concerns more general realisable models, where our base distribution is only required to satisfy moment or ψ_r-Orlicz norm conditions with $r \geq 1$. Theorems 10 and 11 provide upper and lower bounds on the minimax quantiles that match up to universal constants under both conditions. For both our Gaussian and our nonparametric classes of base distributions, we also discuss multivariate extensions of these results.

Table 1 presents a selection of our findings for univariate mean estimation problems. These illustrate the benefits in terms of improved worst-case performance of working with realisable, as opposed to arbitrary, contamination. It is interesting to see, for instance, that

when the effective contamination level κ is small, the minimax quantile rate for a Gaussian distribution under arbitrary contamination agrees with the corresponding rate for a general distribution with finite variance under realisable contamination. It is also notable that, while under arbitrary contamination the minimax quantiles are infinite as soon as $\epsilon \geq q/(1+q)$, where q denotes the MCAR observation proportion, under realisable contamination this threshold converges to 1 with the sample size.

	Arbitrary contamination		**Realisable contamination**	
Base distribution	Minimax rate	ϵ condition	Minimax rate	ϵ condition
Gaussian	$\mathscr{M}_0 + \sigma^2\kappa^2$	$\epsilon < \dfrac{q}{1+q}$	$\mathscr{M}_0 + \dfrac{\sigma^2 \log^2(1+\kappa)}{\log\{nq(1-\epsilon)\}}$	$\epsilon < 1 - o_n(1)$
Sub-Gaussian	$\mathscr{M}_0 + \sigma^2\kappa^2 \log\left(\frac{1}{\kappa}\right)$	$\epsilon < \dfrac{q}{1+q}$	$\mathscr{M}_0 + \sigma^2\big(\kappa^2 \wedge \log(1+\kappa)\big)$	$\epsilon < 1 - o_n(1)$
Finite variance	$\mathscr{M}_0 + \sigma^2\kappa$	$\epsilon < \dfrac{q}{1+q}$	$\mathscr{M}_0 + \sigma^2(\kappa^2 \wedge \kappa)$	$\epsilon < 1 - o_n(1)$

Table 1: A comparison of minimax rates under arbitrary and realisable ϵ-contamination for different univariate base distribution classes. Here, $\mathscr{M}_0 := \frac{\sigma^2 \log(1/\delta)}{nq(1-\epsilon)}$ denotes the MCAR minimax $(1-\delta)$th quantile rate for estimating the mean of a base distribution P having variance (or squared sub-Gaussian norm) σ^2 based on $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathsf{MCAR}_{(q(1-\epsilon),P)}$ (see (1) below), and $\kappa := \frac{\epsilon}{q(1-\epsilon)}$ denotes the effective contamination level. In the Gaussian realisable rate, we have ignored a potential logarithmic multiplicative factor in a regime where the overall rate remains polynomial in the effective sample size $nq(1-\epsilon)$. The results for arbitrary contamination are provided in Section C.3 and Theorem 4, while the results for realisable contamination are given in Section 4.

Extensions of our results to normal linear regression models with realisable missing responses are discussed in Section 5. Here, we show that even when the contamination proportion ϵ is allowed to grow slowly to 1, consistent estimation of the vector of regression coefficients remains achievable under a mild regularity assumption on the design.

All of our proofs are deferred to the Appendix.

1.1 Related work

The ϵ-contamination models that form the bedrock of our framework for the analysis of the effects of missing data are inspired by related models in the robust statistics literature (Huber, 1964). Recently, there has been a concentration of research effort attempting to argue that statistical procedures achieve the optimal dependence on ϵ in different statistical problems, thereby providing evidence of their robustness. For instance, for fully observed data, Chen, Gao and Ren (2018) demonstrate the optimality in this sense of the Tukey median (Tukey, 1975) for mean estimation, as well as a matrix depth estimator of a covariance matrix. Since the Tukey median is computationally intractable, various alternatives have been considered in the both the statistics and theoretical computer science literature (see, e.g., Diakonikolas and Kane, 2023, and references therein). Other problems studied within this framework include linear regression (Bakshi and Prasad, 2021; Pensia, Jog and Loh,

2024+), nonparametric regression (Gao, 2020) and robust clustering (Liu and Moitra, 2023; Jana, Fan and Kulkarni, 2024).

Our realisable contamination models are related to several previous attempts to study restricted forms of missing not at random and biased sampling. For instance, Vardi (1985) introduced a biased sampling model and, under the assumption that the sampling mechanism is known, studied nonparametric estimation of the distribution function. Under this oracle model, Gill, Vardi and Wellner (1988) studied classical asymptotics and efficiency guarantees for the nonparametric maximum likelihood estimator; see also Bickel and Ritov (1991) for similar guarantees in a linear regression setting. Later, Aronow and Lee (2013) and Sahoo, Lei and Wager (2022) introduced likelihood ratio constraints to perform estimation in situations where the sampling mechanism may be unknown. In the causal inference literature, efforts to restrict unobserved confounding have led to the introduction of similar restrictions known as sensitivity conditions (Rosenbaum, 1987; Zhao, Small and Bhattacharya, 2019). As we show in our discussion following Proposition 2, our realisable contamination classes can be understood as generalisations of these notions. In a different direction, and with a view towards computational efficiency, Daskalakis et al. (2018) considered estimating population parameters in a biased sampling model induced by truncation to a known set; Kontonis, Tzamos and Zampetakis (2019) and Diakonikolas et al. (2024) studied the computational and statistical consequences of the absence of knowledge of this truncation set. Distributions obtained by truncation are missing not at random and as such can be captured when $\epsilon = 1$ by our realisable ϵ-contamination classes.

1.2 Notation

For $d \in \mathbb{N}$, we let $[d] := \{1, \dots, d\}$ and write $2^{[d]}$ for the power set of $[d]$. For $a, b \in \mathbb{R}$, we let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. We also define $\log_+(x) := \log(x) \vee 1$ for $x > 0$. If I is an arbitrary index set, then for functions $f, g : I \to \mathbb{R}$, we write $f \gtrsim g$ if there exists a universal constant $c > 0$ such that $f(i) \geq cg(i)$ for all $i \in I$, and write $f \lesssim g$ if there exists a universal constant $C > 0$ such that $f(i) \leq Cg(i)$ for all $i \in I$.

For $S \subseteq [d]$, we define $\mathbf{1}_S \in \{0,1\}^d$ by $(\mathbf{1}_S)_j := \mathbb{1}_{\{j \in S\}}$; for $j \in [d]$, we write $e_j \in \mathbb{R}^d$ for the jth standard basis vector. We denote the unit Euclidean sphere in $\mathbb{R}^d$ by $\mathbb{S}^{d-1}$. The sets $\mathcal{S}^{d\times d}$, $\mathcal{S}^{d\times d}_+$ and $\mathcal{S}^{d\times d}_{++}$ denote the set of symmetric, symmetric positive semidefinite and symmetric positive definite matrices in $\mathbb{R}^{d\times d}$ respectively. For $A \in \mathbb{R}^{d\times d}$, we write $\|A\|_{\mathrm{op}}$ for its operator (spectral) norm and $\|A\|_\infty$ for its maximum absolute entry. Further, for $A \in \mathcal{S}^{d\times d}_+$, we let $\mathbf{r}(A) := \mathrm{tr}(A)/\|A\|_{\mathrm{op}}$ denote the effective rank of A, with the convention that $0/0 := 0$. Given $(a_1, \dots, a_d)^\top \in \mathbb{R}^d$, let $\mathrm{diag}(a_1, \dots, a_d) \in \mathbb{R}^{d\times d}$ denote the diagonal matrix with entries $a_1, \dots a_d$, and let $I_d := \mathrm{diag}(1, \dots, 1)$ denote the identity matrix in $\mathbb{R}^{d\times d}$.

For a topological space $(\mathcal{X}, \tau)$, we let $\mathcal{B}(\mathcal{X})$ denote the Borel σ-algebra of $\mathcal{X}$, and let $\mathcal{P}(\mathcal{X})$ denote the set of all probability measures on $\big(\mathcal{X}, \mathcal{B}(\mathcal{X})\big)$. For two measures μ_1, μ_2 on $\big(\mathcal{X}, \mathcal{B}(\mathcal{X})\big)$, we write $\mu_1 \ll \mu_2$ if μ_1 is absolutely continuous with respect to μ_2. We write $\lambda \in \mathcal{P}(\mathbb{R})$ for Lebesgue measure on $\mathbb{R}$. Given a collection $\mathcal{Q}$ of distributions, we define $\mathcal{Q}^{\otimes n} := \{Q^{\otimes n} : Q \in \mathcal{Q}\}$. For a random variable X taking values in $\mathcal{X}$, we let $\mathsf{Law}(X) \in \mathcal{P}(\mathcal{X})$ denote the distribution of X, and $\mathrm{supp}(X) \subseteq \mathcal{X}$ denote the support of X, i.e. the intersection of all closed sets $C \subseteq \mathcal{X}$ with $\mathbb{P}(X \in C) = 1$.

For $\theta \in \mathbb{R}$ and $\sigma \in [0, \infty)$, we let $\Phi_{(\theta,\sigma)}(\cdot)$ and $\phi_{(\theta,\sigma)}(\cdot)$ denote the distribution and density functions of the $\mathsf{N}(\theta, \sigma^2)$ distribution respectively, with the shorthand that $\Phi := \Phi_{(0,1)}$ and

$\phi := \phi_{(0,1)}$.

2 Statistical setting

2.1 The extended space $\mathcal{X}_\star$ and classical models of missing data

In this section, we introduce spaces that are convenient for models of missing data. Let $d \in \mathbb{N}$ and, for $j \in [d]$, let $(\mathcal{X}_j, \tau_j)$ denote a topological space equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X}_j)$. We use the symbol $\star$ to denote a missing element[1] and define, for each $j \in [d]$, the *extended space* $\mathcal{X}_{j,\star} := \mathcal{X}_j \cup \{\star\}$, equipped with the topology $\tau_{j,\star} := \tau_j \cup \{A \cup \{\star\} : A \in \tau_j\}$ and corresponding Borel σ-algebra $\mathcal{B}(\mathcal{X}_{j,\star}) = \mathcal{B}(\mathcal{X}_j) \cup \{A \cup \{\star\} : A \in \mathcal{B}(\mathcal{X}_j)\}$. Given a measure μ_j on $\big(\mathcal{X}_j, \mathcal{B}(\mathcal{X}_j)\big)$, we define the *extended measure* $\mu_{j,\star}$ on $\big(\mathcal{X}_{j,\star}, \mathcal{B}(\mathcal{X}_{j,\star})\big)$ by

$$\mu_{j,\star}(A) := \mu_j(A) \quad \text{and} \quad \mu_{j,\star}(A \cup \{\star\}) := \mu_j(A) + 1$$

for $A \in \mathcal{B}(\mathcal{X}_j)$. It is also convenient to define the product spaces $\mathcal{X} := \prod_{j=1}^d \mathcal{X}_j$ and $\mathcal{X}_\star := \prod_{j=1}^d \mathcal{X}_{j,\star}$, equipped with their product σ-algebras $\mathcal{B}(\mathcal{X}) := \otimes_{j\in[d]} \mathcal{B}(\mathcal{X}_j)$ and $\mathcal{B}(\mathcal{X}_\star) := \otimes_{j\in[d]} \mathcal{B}(\mathcal{X}_{j,\star})$ respectively.

We will often reason about missing data via *revelation vectors* $\omega \in \{0,1\}^d$, which together with an element $x \in \mathcal{X}$ induce an element of the extended space $\mathcal{X}_\star$ through the binary operator $\circledast : \mathcal{X} \times \{0,1\}^d \to \mathcal{X}_\star$, where the jth component of $x \circledast \omega$ is defined by

$$(x \circledast \omega)_j := \begin{cases} x_j & \text{if } \omega_j = 1 \\ \star & \text{if } \omega_j = 0, \end{cases}$$

for $j \in [d]$. The following example gives a concrete illustration of the abstract notation.

Example 1. Let $X \sim \mathsf{N}(0,1)$ and let Ω be a binary random variable satisfying $\mathbb{P}(\Omega = 1 \,|\, X = x) = g(x)$ for some Borel measurable function $g : \mathbb{R} \to [0,1]$. Then the $\mathbb{R}_\star$-valued random variable $X \circledast \Omega$ admits a density $f_\star : \mathbb{R}_\star \to [0,\infty)$ with respect to the extended Lebesgue measure $\lambda_\star$, where

$$f_\star(z) := \begin{cases} g(z)\phi(z) & \text{if } z \in \mathbb{R} \\ 1 - \int_{\mathbb{R}} g(x)\phi(x)\,\mathrm{d}x & \text{if } z = \star. \end{cases}$$ ♢

An advantage of working with the extended measurable space $\mathcal{X}_\star$ is that it allows us to give succinct definitions of three classical models of missingness: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). For each definition, we will let $X \sim P \in \mathcal{P}(\mathcal{X})$ and $\pi \in \mathcal{P}(2^{[d]})$. To define the *MCAR distribution*, let Ω be a random vector in $\{0,1\}^d$, independent of X, such that $\mathbb{P}(\Omega = \mathbf{1}_S) = \pi(S)$ for $S \subseteq [d]$, and define[2]

$$\mathsf{MCAR}_{(\pi,P)} := \mathsf{Law}(X \circledast \Omega) \in \mathcal{P}(\mathcal{X}_\star). \tag{1}$$

[1] When $\mathcal{X}_j = \mathbb{R}$, we adopt the conventions that $\star \cdot 0 := 0 =: 0 \cdot \star$, that $x \cdot \star := \star =: \star \cdot x$ for $x \in \mathcal{X}_\star \setminus \{0\}$ and that $\star + x = \star$ for $x \in \mathcal{X}_\star$.

[2] When $d = 1$, we may identify π with $q := \pi(\{1\})$, and write $\mathsf{MCAR}_{(q,P)}$ in place of $\mathsf{MCAR}_{(\pi,P)}$.

Next, the *family of MAR distributions* is the subset of $\mathcal{P}(\mathcal{X}_\star)$ given by[3]

$$\mathsf{MAR}_{(\pi,P)} \coloneqq \big\{\mathsf{Law}(X \circledast \Omega') : \; X \sim P,\, \mathbb{P}(\Omega' = \mathbf{1}_S) = \pi(S)\; \forall S \subseteq [d], \\ \mathbb{P}(\Omega' = \omega \,|\, X = x) = \mathbb{P}(\Omega' = \omega \,|\, X \circledast \omega = x \circledast \omega)\; \forall \omega \in \{0,1\}^d, P\text{-a.e. } x \in \mathcal{X}\big\}. \tag{2}$$

The missing at random definition captures the intuitive idea that 'missingness depends only on the observed variables' and is tailored toward likelihood-based methods for which it implies the so-called *ignorability* of the missingness mechanism (Seaman et al., 2013, Section 5). Finally, we define the corresponding *family of MNAR distributions* $\mathsf{MNAR}_{(\pi,P)} \subseteq \mathcal{P}(\mathcal{X}_\star)$ as

$$\mathsf{MNAR}_{(\pi,P)} \coloneqq \big\{\mathsf{Law}(X \circledast \Omega') : X \sim P, \mathbb{P}(\Omega' = \mathbf{1}_S) = \pi(S)\; \forall S \subseteq [d]\big\} \subseteq \mathcal{P}(\mathcal{X}_\star), \tag{3}$$

According to these definitions, $\mathsf{MCAR}_{(\pi,P)} \in \mathsf{MAR}_{(\pi,P)} \subseteq \mathsf{MNAR}_{(\pi,P)}$. When the distribution π is not fixed, we let

$$\mathsf{MAR}_P \coloneqq \bigcup_{\rho \in \mathcal{P}(2^{[d]})} \mathsf{MAR}_{(\rho,P)} \quad \text{and} \quad \mathsf{MNAR}_P \coloneqq \bigcup_{\rho \in \mathcal{P}(2^{[d]})} \mathsf{MNAR}_{(\rho,P)}. \tag{4}$$

2.2 Models of departures from M(C)AR

2.2.1 Identifiability issues under the missing at random assumption

The MAR assumption (2) is arguably the most widely adopted form of departure from the restrictive MCAR assumption in statistical practice. However, in Example 2 below, we show that even in simple parametric scenarios, this can lead to identifiability issues that preclude consistent estimation; in particular, even though the the MAR assumption holds, it is nonetheless impossible to identify the population mean.

Example 2. For $\theta = (\theta_1, \theta_2)^\top \in [0,1]^2$, define $P_\theta \in \mathcal{P}\big(\{0,1\}^3\big)$ such that for $X \coloneqq (X_1, X_2, X_3)^\top \sim P_\theta$, we have $X_1 \sim \mathsf{Ber}(\theta_1)$, $X_2 \sim \mathsf{Ber}(\theta_2)$ independently, and $X_3 = X_1 + X_2 \,(\mathrm{mod}\, 2)$. We next specify a missingness mechanism via

$$\Omega \mid X = \begin{cases} (0,0,1) & \text{if } X_3 = 1, \\ (0,1,1) & \text{with probability } 1/2 \text{ if } X_3 = 0, \\ (1,0,1) & \text{with probability } 1/2 \text{ if } X_3 = 0, \end{cases}$$

and note that if $X \sim P_\theta$, then $R_\theta \coloneqq \mathsf{Law}(X \circledast \Omega) \in \mathsf{MAR}_{P_\theta}$. We have

$$R_\theta\big(\{(\star,\star,1)\}\big) = \theta_1(1-\theta_2) + (1-\theta_1)\theta_2, \quad R_\theta\big(\{(\star,1,0)\}\big) = R_\theta\big(\{(1,\star,0)\}\big) = \frac{\theta_1\theta_2}{2}$$

and

$$R_\theta\big(\{(\star,0,0)\}\big) = R_\theta\big(\{(0,\star,0)\}\big) = \frac{(1-\theta_1)(1-\theta_2)}{2}.$$

Thus, $R_{(\theta_1,\theta_2)} = R_{(\theta_2,\theta_1)}$ for all $(\theta_1,\theta_2)^\top \in [0,1]^2$, so that it is impossible to identify the parameter. This symmetry additionally implies that the population log-likelihood may not admit a unique global maximiser. Indeed, letting $\theta^* \in [0,1]^2$ denote the true parameter, the population log-likelihood is given by $\mathcal{L}(\theta) \coloneqq \log \mathbb{E}_{\theta^*}\big\{R_\theta(\{X \circledast \omega\})\big\}$, which is symmetric in the components of θ. ♢

[3] A formal definition of the conditional probabilities in (2) can be provided through the notion of disintegrations, whose existence is assumed here (and is guaranteed when $\mathcal{X}_j$ is a Polish space for each $j \in [d]$); see Section G.

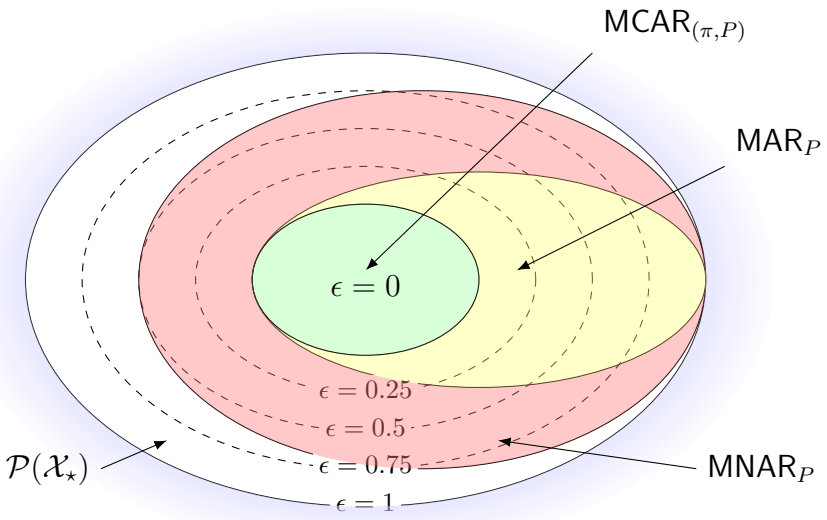


Figure 1: An illustration of the arbitrary ϵ-contamination model $\mathcal{P}^{\mathrm{arb}}(P, \epsilon, \pi)$, which interpolates between $\mathsf{MCAR}_{(\pi,P)}$ and $\mathcal{P}(\mathcal{X}_\star)$.

2.2.2 Huber-style models of departure from MCAR

Given the failure of the MAR assumption to ensure the tractability of the mean estimation problem, and in light of dual representation of the incompatibility index given by Berrett and Samworth (2023, Theorem 2), it is natural to model departures from MCAR via a nonparametric, Huber-style contamination model. In particular, given $P \in \mathcal{P}(\mathcal{X})$, $\epsilon \in [0, 1]$ and $\pi \in \mathcal{P}(2^{[d]})$, we define the *arbitrary ϵ-contamination model*

$$\mathcal{P}^{\mathrm{arb}}(P, \epsilon, \pi) := \Big\{(1-\epsilon)\mathsf{MCAR}_{(\pi,P)} + \epsilon Q : Q \in \mathcal{P}(\mathcal{X}_\star)\Big\}. \tag{5}$$

This family comprises mixture distributions in which one of the mixture components can be an arbitrary distribution on $\mathcal{X}_\star$. One way to think about such distributions is via the following algorithm for drawing an observation $Z \sim (1-\epsilon)\mathsf{MCAR}_{(\pi,P)} + \epsilon Q$, where $Q \in \mathcal{P}(\mathcal{X}_\star)$. We first generate $W \sim \mathsf{Ber}(\epsilon)$; if $W = 0$, we then draw Ω and X independently, according to $\mathbb{P}(\Omega = \mathbf{1}_S \,|\, W = 0) = \pi(S)$ for $S \subseteq [d]$ and $X \,|\, \{W = 0\} \sim P$, and finally set $Z := X \circledast \Omega$. On the other hand, if $W = 1$, then we draw $Z \,|\, \{W = 1\} \sim Q$. The arbitrary ϵ-contamination model allows us to interpolate in a continuous way between $\mathcal{P}^{\mathrm{arb}}(P, 0, \pi) = \mathsf{MCAR}_{(\pi,P)}$ and $\mathcal{P}^{\mathrm{arb}}(P, 1, \pi) = \mathcal{P}(\mathcal{X}_\star)$; see Figure 1.

An attraction of the arbitrary contamination model is its generality. Nevertheless, in many practical settings where one considers data arising from a particular distribution that are then subjected to some form of missingness, it may be preferable to seek classes to interpolate between $\mathsf{MCAR}_{(\pi,P)}$ and MNAR_P. To this end, a key definition in our framework is that of the *realisable ϵ-contamination model*

$$\mathcal{R}(P, \epsilon, \pi) := \Big\{(1-\epsilon)\mathsf{MCAR}_{(\pi,P)} + \epsilon\, Q : Q \in \mathsf{MNAR}_P\Big\}; \tag{6}$$

see Figure 2. In this model, the contamination mixture component is restricted to being a partially-observed version of $X \sim P$, where nevertheless the observation pattern may both be different from that in the uncontaminated component, and dependent on X. Thus,

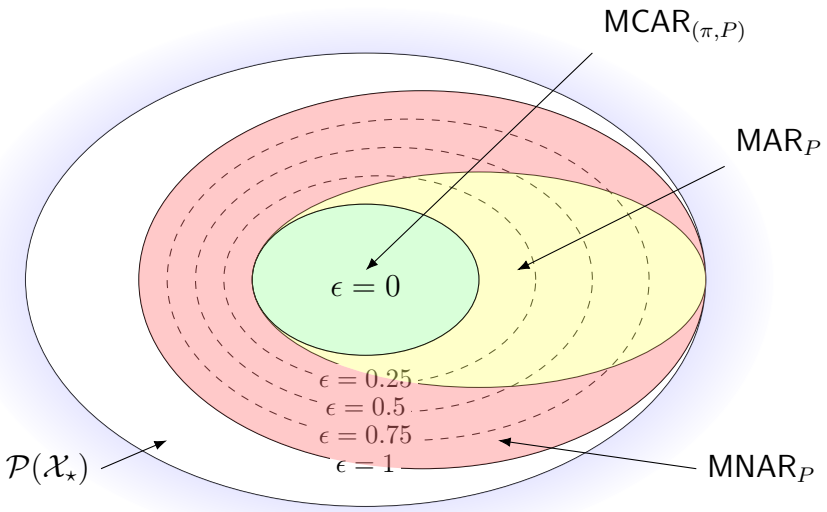


Figure 2: An illustration of the realisable ϵ-contamination model $\mathcal{R}(P, \epsilon, \pi)$, which interpolates between $\mathsf{MCAR}_{(\pi,P)}$ and MNAR_P.

the realisable contamination model $\mathcal{R}(P, \epsilon, \pi)$ represents a (still nonparametric) subclass of $\mathcal{P}^{\mathrm{arb}}(P, \epsilon, \pi)$, with the potential to yield improved rates of mean estimation. On the other hand, noting that $\mathcal{R}(P, 1, \pi) = \mathsf{MNAR}_P$, in Example 2, the distribution R_θ belongs to MNAR_P but not to $\mathcal{R}(P, \epsilon, \pi)$ for any $\epsilon \in [0, 1)$, so the inability to estimate its mean consistently does not contradict our minimax upper bounds established in Theorem 10.

2.2.3 Regression with missing response

A practical application of our framework for mean estimation with missingness is to regression problems where the response variable may be missing. Here, we consider a d-dimensional (random) covariate vector and a real-valued response variable Y. Given a disintegration $(P_{Y|x})_{x\in\mathbb{R}^d}$ of the joint distribution of (X, Y) into conditional distributions on $\mathbb{R}$, and $q \in (0, 1]$, we define the collection of *missing at random (MAR) response distributions* as

$$\mathsf{MAR}^{\mathrm{Res}}_{(q,P_{Y|X})} := \Big\{ \mathsf{Law}(Y \circledast \Omega \,|\, X) : \; Y \mid X \sim P_{Y|X}, \mathrm{supp}(\Omega) = \{0,1\}, \\ \Omega \perp\!\!\!\perp Y \,|\, X \text{ and } \mathbb{P}(\Omega = 1 \,|\, X) \geq q \Big\}, \quad (7)$$

and the corresponding collection of *missing not at random (MNAR) response distributions* as

$$\mathsf{MNAR}^{\mathrm{Res}}_{(P_{Y|X})} := \Big\{ \mathsf{Law}(Y \circledast \Omega \,|\, X) : Y \mid X \sim P_{Y|X} \text{ and } \mathrm{supp}(\Omega) = \{0,1\} \Big\}. \quad (8)$$

We note two crucial differences between the classes defined in (7) and (8) above. First, in the missing at random setting, we require that the missingness mechanism Ω is conditionally independent of the response Y given the covariate vector X, whereas in the latter setting, the mechanism may depend arbitrarily on the response as well as the covariate. Second, under the missing at random setting, we require a lower bound on the probability of observing a

particular response given its corresponding covariate vector X. This latter condition implies a one-sided version of a so-called *strict overlap condition* (see, e.g., Hirano, Imbens and Ridder, 2003, Assumption 4(ii)). By contrast, we impose no such assumption in the missing not at random setting. Given $\epsilon \in [0,1]$, we define our *realisable ϵ-contamination model with a missing response* as

$$\mathcal{R}^{\mathrm{Res}}(P_{Y|X}, \epsilon, q) := \Big\{(1-\epsilon)\mathsf{MAR}^{\mathrm{Res}}_{(q,P_{Y|X})} + \epsilon Q : \; Q \in \mathsf{MNAR}^{\mathrm{Res}}_{(P_{Y|X})}\Big\}. \tag{9}$$

2.3 Characterisation of realisability

As we have argued previously, the realisable ϵ-contamination model is often very natural in settings where our data are observed subject to missingness. It is therefore of great interest to characterise distributions $R \in \mathcal{R}(P, \epsilon, \pi)$, and this is achieved in Theorem 1 below through integrals of bounded, continuous functions with respect to R. Given a topological space $(\mathcal{Z}, \tau_{\mathcal{Z}})$, it is convenient to write $C_{\mathrm{b}}(\mathcal{Z})$ for the set of bounded, continuous functions on $\mathcal{Z}$. For $f \in C_{\mathrm{b}}(\mathcal{X}_\star)$, we also define $f_{\max} : \mathcal{X} \to \mathbb{R}$ by $f_{\max}(x) := \max_{\omega \in \{0,1\}^d} f(x \circledast \omega)$.

Theorem 1. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be locally compact Hausdorff spaces*[4] *and let $\mathcal{X} := \prod_{j=1}^d \mathcal{X}_j$. Assume that every open set in $\mathcal{X}$ is σ-compact. Fix $P \in \mathcal{P}(\mathcal{X})$, $\epsilon \in (0,1]$, $\pi \in \mathcal{P}(2^{[d]})$. Let $R \in \mathcal{P}(\mathcal{X}_\star)$, and define a signed measure on $\mathcal{X}_\star$ by $Q := \epsilon^{-1}\{R - (1-\epsilon)\mathsf{MCAR}_{(\pi,P)}\}$. Then $R \in \mathcal{R}(P, \epsilon, \pi)$ if and only if $Q \in \mathcal{P}(\mathcal{X}_\star)$ and*

$$P(f_{\max}) \geq Q(f) \tag{10}$$

for all $f \in C_{\mathrm{b}}(\mathcal{X}_\star)$.

An important special case of Theorem 1, and indeed the main content of its proof, concerns the setting where $\epsilon = 1$. Here, the result states that a distribution Q belongs to MNAR_P if and only if (10) holds, and is a consequence of a generalised version of Farkas's lemma (Farkas, 1902), due to Craven and Koliha (1977). An explanation of the relevance of this seemingly-unrelated lemma, which amounts to a proof of the theorem in the case where $\mathcal{X}$ is finite, is provided before the proof of the full result in Section B.1.

In the univariate case with $\mathcal{X} = \mathbb{R}$, Proposition 2 below provides a more explicit characterisation of realisability. It is also convenient here to write $\mathcal{R}(P, \epsilon, q)$ in place of $\mathcal{R}(P, \epsilon, \pi)$ when $q := \pi(\{1\})$.

Proposition 2. *Let $P \in \mathcal{P}(\mathbb{R})$ and assume that P has density p with respect to a Borel measure μ. Let $\epsilon \in [0,1]$, $\pi \in \mathcal{P}\big(\{\emptyset, \{1\}\}\big)$, and define $q := \pi(\{1\})$. Then $R \in \mathcal{R}(P, \epsilon, q)$ if and only if $R \ll \mu_\star$ and there exists a Borel measurable function $m : \mathbb{R} \to [0,1]$ such that*

$$\frac{\mathrm{d}R}{\mathrm{d}\mu_\star}(z) = \begin{cases} q(1-\epsilon) \cdot p(z) + \epsilon \cdot m(z)p(z) & \text{if } z \in \mathbb{R} \\ 1 - q(1-\epsilon) - \epsilon \int_{\mathbb{R}} m(x)p(x)\,\mathrm{d}\mu(x) & \text{if } z = \star. \end{cases} \tag{11}$$

In a similar fashion to the quantity M in the discussion following Theorem 1, the function $m : \mathbb{R} \to [0,1]$ admits an interpretation as a missingness mechanism for the MNAR component. More generally, Proposition 2 reveals that univariate realisability is characterised

[4]For the convenience of the reader, definitions of these terms from topology are provided in Section B.1.

via rejection sampling. To see this in the extreme case when $\epsilon = 1$, consider a distribution $R \in \mathcal{P}(\mathbb{R}_\star)$ such that $R \ll \mu_\star$, and for $Z \sim R$, let g denote the conditional density with respect to μ of Z given that $\{Z \neq \star\}$. By Proposition 2, $R \in \mathsf{MNAR}_P$ if and only if $g(x)/p(x) \leq 1/R\big(\{\star\}\big)$ for μ-almost all $x \in \mathbb{R}$. Thus any $R \in \mathsf{MNAR}_P$ can be obtained via rejection sampling from P.

Let us now compare our realisable class in (6) with related notions in the (primarily causal inference) literature, in the univariate case. Let $Z \coloneqq X \circledast \Omega$, where $X \sim P \ll \mu$ and where Ω is a random variable taking values in $\{0, 1\}$ that need not be independent of X. Define $h : \mathbb{R} \to [0, 1]$ by

$$h(x) \coloneqq \mathbb{P}(Z \neq \star \,|\, X = x).$$

Proposition 2 yields that $\mathsf{Law}(Z) \in \mathcal{R}(P, \epsilon, q)$ if and only if $q(1-\epsilon) \leq h(x) \leq q(1-\epsilon)+\epsilon$ for μ-almost all $x \in \mathbb{R}$. The notion of Γ-*biased sampling* of Sahoo, Lei and Wager (2022) (see also Aronow and Lee (2013)) can be stated as the condition on h and $\overline{q} \coloneqq \mathbb{P}(Z \neq \star)$ that $\Gamma^{-1} \leq h(x)/\overline{q} \leq \Gamma$ for some $\Gamma \geq 1$ and μ-almost all $x \in \mathbb{R}$. In a similar spirit, the *marginal sensitivity condition* of Zhao, Small and Bhattacharya (2019, Definition 1) asks that there exists $\Lambda \geq 1$ such that

$$\frac{1}{\Lambda} \leq \frac{h(x)}{1-h(x)} \cdot \frac{1-\overline{q}}{\overline{q}} \leq \Lambda$$

for μ-almost all $x \in \mathbb{R}$, while the classical *sensitivity condition* of Rosenbaum (1987) reads as

$$\frac{1}{\Lambda} \leq \frac{h(x_1)}{1-h(x_1)} \cdot \frac{1-h(x_2)}{h(x_2)} \leq \Lambda$$

for $(\mu \otimes \mu)$-almost all $(x_1, x_2) \in \mathbb{R}^2$. These classes all belong to $\mathcal{R}(P, \epsilon, q)$ for some $\epsilon \in [0, 1)$ and $q \in (0, 1]$. Here, for simplicity of exposition, we have presented versions of these conditions without covariates. Nevertheless the comparison remains valid when covariates are included; see Section 5.

2.4 Minimax quantile framework

In a traditional minimax analysis, the randomness in the loss function evaluated at our data is handled via a reduction to its expectation, namely the minimax risk. As mentioned in the introduction, this minimax risk is infinite in the problems that we consider, so does not provide a meaningful way of comparing different statistical procedures. We therefore adopt the minimax quantile framework of Ma, Verchand and Samworth (2024), which also offers the benefit of retaining all of the distributional information, e.g. regarding tail behaviour, in the loss function.

To introduce this paradigm in generality, we let (Θ, d) be a non-empty pseudo-metric space and for $\theta \in \Theta$, let $\mathcal{P}_\theta$ denote a family of probability measures on a measurable space $(\mathcal{Z}, \mathcal{C})$. Further, let $g : [0, \infty) \to [0, \infty)$ denote an increasing function and define the loss $L : \Theta \times \Theta \to [0, \infty)$ by $L(\theta, \theta') \coloneqq g\big(d(\theta, \theta')\big)$. Write $\widehat{\Theta}$ for the set of estimators of θ, i.e. the set of measurable functions from $\mathcal{Z}$ to Θ. For $\widehat{\theta} \in \widehat{\Theta}$, $P_\theta \in \mathcal{P}_\theta$ and a quantile level $\delta \in (0, 1]$, we write

$$\mathrm{Quantile}\big(1-\delta; P_\theta, L(\widehat{\theta}, \theta)\big) \coloneqq \inf\Big\{r \in [0, \infty) : \; P_\theta\big\{L(\widehat{\theta}, \theta) \leq r\big\} \geq 1-\delta\Big\},$$

and consider the *minimax* $(1-\delta)$*th quantile*, defined as

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, L) := \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta \in \Theta} \sup_{P_\theta \in \mathcal{P}_\theta} \text{Quantile}\big(1-\delta; P_\theta, L(\widehat{\theta}, \theta)\big), \tag{12}$$

where $\mathcal{P}_\Theta := \{\mathcal{P}_\theta : \theta \in \Theta\}$. If there exists $\widehat{\theta} \in \widehat{\Theta}$ such that with P_θ-probability at least $1-\delta$, we have $L(\widehat{\theta}, \theta) \leq \text{UB}(\delta)$ for all $\theta \in \Theta$ and $P_\theta \in \mathcal{P}_\theta$, then $\mathcal{M}(\delta, \mathcal{P}_\Theta, L) \leq \text{UB}(\delta)$. For the squared Euclidean error loss $L(\widehat{\theta}, \theta) = \|\widehat{\theta} - \theta\|_2^2$, we will slightly abuse notation by writing $\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big)$ in place of $\mathcal{M}(\delta, \mathcal{P}_\Theta, L)$.

3 Mean estimation under arbitrary contamination

Throughout this section, we take $\mathcal{X} = \mathbb{R}^d$. The set of distributions on $\mathbb{R}^d$ with mean vector $\theta \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathcal{S}_+^{d \times d}$ is denoted

$$\mathcal{P}(\theta, \Sigma) := \Big\{P \in \mathcal{P}(\mathbb{R}^d) : \ \mathbb{E}_P(X) = \theta, \ \text{Cov}_P(X) = \Sigma\Big\}. \tag{13}$$

Given $\epsilon \in [0,1]$ and $\pi \in \mathcal{P}(2^{[d]})$, it is convenient to define a specialised version of our arbitrary contamination model by

$$\mathcal{P}^{\text{arb}}(\theta, \Sigma, \epsilon, \pi) := \bigcup_{P \in \mathcal{P}(\theta, \Sigma)} \mathcal{P}^{\text{arb}}(P, \epsilon, \pi). \tag{14}$$

We now describe the robust descent algorithm with iterative imputation in Algorithm 1, and quantify its performance in Theorem 3. This algorithm depends on universal constants $A_1^{-1}, A_2, A_3 > 0$; sufficiently large values for our theoretical guarantees can be determined from Theorem 3 and Lemma 17 as well as Depersin and Lecué (2022b, Theorem 2.1), though these values are probably far from optimal. The performance of Algorithm 1 is governed by the matrix $\Sigma^{\text{IPW}} \in \mathcal{S}_+^{d \times d}$, with entries

$$\big(\Sigma^{\text{IPW}}\big)_{jk} := \frac{q_{jk}}{q_j q_k} \cdot \Sigma_{jk},$$

for $j, k \in [d]$, where $q_{jk} := \sum_{S \subseteq [d] : \{j,k\} \subseteq S} \pi(S)$ and $q_j := q_{jj}$. Further define $q_{\min} := \min_{j \in [d]} q_j$.

Theorem 3. *Let $\epsilon \in [0, 1/2)$, $n \in \mathbb{N}$ and $\delta \in (0,1)$. Let $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} P \in \mathcal{P}^{\text{arb}}\big(\theta_0, \Sigma, \epsilon, \pi\big)$ and let $\widehat{\theta}_n = \text{ITERATIVE\_ROBUST\_DESCENT}(Z_1, \ldots, Z_n; \epsilon, \delta)$ from Algorithm 1, with $A_1 = 10^{-9}$, $A_2 = 300$ and $A_3 = 180{,}000$. Taking*

$$T := 1 + \Big\lceil \log_+\big(A_1 \cdot \{\mathbf{r}(\Sigma^{\text{IPW}}) + \log(24d/\delta)\}\big) \Big\rceil,$$

as in the algorithm, we have that if both

$$q_{\min} \geq 10^{13} \cdot \left\{ \left(\epsilon + \frac{T \log(3T/\delta)}{n} \right) \vee \frac{300 T \log(6T/\delta)}{n} \right\} \vee \frac{8 \log(6d/\delta)}{n}$$

and $\delta \geq 6Te^{-n/(720{,}000T)} \vee de^{-n/(2T)}$, then there exists a universal constant $C > 0$ such that, with probability at least $1-\delta$,

$$\|\widehat{\theta}_n - \theta_0\|_2^2 \leq C \left(\frac{T \,\text{tr}(\Sigma^{\text{IPW}})}{n} + \frac{T \|\Sigma^{\text{IPW}}\|_{\text{op}} \log(6T/\delta)}{n} + \|\Sigma^{\text{IPW}}\|_{\text{op}} \epsilon \right).$$

Algorithm 1 ITERATIVE_ROBUST_DESCENT for robust mean estimation with iterative imputation

Input: Data $z_1, \ldots, z_n \in \mathbb{R}^d_\star$, contamination parameter $\epsilon \geq 0$, and tolerance parameter $\delta > 0$

Output: An estimator $\widehat{\theta}_n$ of θ_0

function ITERATIVE_ROBUST_DESCENT$(z_1, \ldots, z_n; \epsilon, \delta)$
  $T \leftarrow 1 + \lceil \log_+\big(A_1\{\mathbf{r}(\Sigma^{\mathrm{IPW}}) + \log(24d/\delta)\}\big)\rceil$, $\epsilon' \leftarrow 2\epsilon + 2T\log(3T/\delta)/n$ and $M \leftarrow \lceil (A_2 n\epsilon'/T) \vee A_3 \log(6T/\delta)\rceil$
  for $i \in [n]$ and $j \in [d]$ **do**
    $\omega_{ij} \leftarrow \mathbb{1}_{\{z_{ij} \neq \star\}}$
  end for
  Randomly partition $\big[T\lfloor n/T \rfloor\big]$ into T disjoint sets $(S^{(t)})_{t\in[T]}$ of equal cardinality
  for $j \in [d]$ **do**
    $I_j \leftarrow \{i \in S^{(1)} : \omega_{ij} = 1\}$
    $\widehat{\theta}_j^{(1)} \leftarrow$ UNIVARIATE_TRIMMED_MEAN$\big((z_{ij})_{i\in I_j}; \epsilon, \delta\big)$; see Algorithm 3
  end for
  for $t \in \{2, \ldots, T\}$ **do**
    Randomly partition $S^{(t)}$ into $M+1$ disjoint sets $(B_m^{(t)})_{m\in[M+1]}$, where the first M have cardinality $\lfloor |S^{(t)}|/M \rfloor$
    for $(m, j) \in [M] \times [d]$ **do**
      $\bar{\omega}_{mj}^{(t)} \leftarrow \mathbb{1}_{\{\sum_{i\in B_m^{(t)}} \omega_{ij} > 0\}}$
      $\bar{z}_{mj}^{(t)} \leftarrow \bar{\omega}_{mj}^{(t)} \cdot \dfrac{\sum_{i\in B_m^{(t)}} \omega_{ij} z_{ij}}{\sum_{i\in B_m^{(t)}} \omega_{ij}} + (1 - \bar{\omega}_{mj}^{(t)}) \cdot \widehat{\theta}_j^{(t-1)}$
    end for
    $\widehat{\theta}^{(t)} \leftarrow$ ROBUST_BLOCK_DESCENT$(\bar{z}_1^{(t)}, \ldots, \bar{z}_M^{(t)})$; see Algorithm 4
  end for
  return $\widehat{\theta}_n \leftarrow \widehat{\theta}^{(T)}$
end function

Since

$$T \leq 1 + \log_+\big(A_1\{d + \log(24d/\delta)\}\big) \lesssim \log d + \log_+ \log(1/\delta),$$

Theorem 3 yields that, with $\mathcal{P}_\Theta^{\mathrm{arb}} := \big\{\mathcal{P}^{\mathrm{arb}}(\theta, \Sigma, \epsilon, \pi) : \theta \in \Theta\big\}$,

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta^{\mathrm{arb}}, \|\cdot\|_2^2\big) \lesssim \big\{\log d + \log_+ \log(1/\delta)\big\} \underbrace{\left\{\frac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{n} + \frac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} \log(1/\delta)}{n}\right\}}_{\text{MCAR term}} + \underbrace{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\epsilon}_{\text{MCAR departure}}.$$

As indicated, our upper bound decomposes into a sum of two distinct components: an MCAR term and an ϵ-dependent term that captures the effect of departure from MCAR. The first of these terms further decomposes as the sum of a risk component[5] and a term that captures the dependence on the quantile level δ.

[5]Strictly speaking, this is a slight abuse of terminology, since the MCAR risk is infinite whenever $q_{\min} < 1$; however, it is the risk when $q_{\min} = 1$, and the terminology reflects the fact that the term does not depend on the quantile level δ.

In the strong contamination model, Hu and Reingold (2021, Theorem 2) provide an estimator $\widehat{\theta}^{\mathrm{HR}}$ satisfying the upper bound

$$\|\widehat{\theta}^{\mathrm{HR}} - \theta_0\|_2^2 \lesssim \frac{d\|\Sigma\|_{\mathrm{op}} \log d}{nq_{\min}} + \frac{d\|\Sigma\|_{\mathrm{op}} \log(1/\delta)}{nq_{\min}} + \frac{\|\Sigma\|_{\mathrm{op}}\epsilon}{q_{\min}}$$

with probability at least $1-\delta$, provided that $\delta \gtrsim de^{-cnq_{\min}}$ for an appropriately small $c > 0$. Since our Algorithm 1 applies the ROBUST_BLOCK_DESCENT algorithm of Depersin and Lecué (2022b) iteratively, and since that algorithm has performance guarantees under the strong contamination model, it follows that our bound in Theorem 3 also holds in the strong contamination model, and this facilitates a comparison of our conclusion with that of Hu and Reingold (2021). The improvements of our bound when $\delta \geq \exp(-e^d)$ arise from the facts that

$$\mathrm{tr}(\Sigma^{\mathrm{IPW}}) \leq d\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}, \quad \|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} \leq \frac{\|\Sigma\|_{\mathrm{op}}}{q_{\min}},$$

and

$$\log\bigl(d\log(1/\delta)\bigr)\biggl\{\frac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{n} + \frac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\log(1/\delta)}{n}\biggr\} \leq 2\biggl\{\frac{d\|\Sigma\|_{\mathrm{op}}\log d}{nq_{\min}} + \frac{d\|\Sigma\|_{\mathrm{op}}\log(1/\delta)}{nq_{\min}}\biggr\}.$$

These gains may be significant: for instance, when $\delta = e^{-d}$, we obtain that with probability at least $1-\delta$,

$$\|\widehat{\theta}_n - \theta_0\|_2^2 \lesssim \frac{d\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\log d}{n} + \|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} \cdot \epsilon.$$

By contrast, Hu and Reingold (2021) obtain that with probability at least $1-\delta$,

$$\|\widehat{\theta}^{\mathrm{HR}} - \theta_0\|_2^2 \lesssim \frac{d^2\|\Sigma\|_{\mathrm{op}}}{nq_{\min}} + \frac{\|\Sigma\|_{\mathrm{op}}}{q_{\min}} \cdot \epsilon.$$

As another example, to illustrate the effect of heterogeneous missingness across coordinates, if $d \geq 2$, $\Sigma = I_d + \mathbf{1}_{[d]}\mathbf{1}_{[d]}^\top$, $q_1 = 1/d$ and $q_j = 1$ for $j \geq 2$, then

$$\begin{aligned}\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} = \|I_d + \mathbf{1}_{[d]}\mathbf{1}_{[d]}^\top + (2d-2)e_1e_1^\top\|_{\mathrm{op}} &\leq 1 + \mathrm{tr}\bigl(\mathbf{1}_{[d]}\mathbf{1}_{[d]}^\top + (2d-2)e_1e_1^\top\bigr) \\ &= 3d - 1 \leq d(d+1) = \frac{\|\Sigma\|_{\mathrm{op}}}{q_{\min}}.\end{aligned}$$

On the other hand, due to the sample splitting in Algorithm 1, our condition on δ may be slightly stronger than that of Hu and Reingold (2021), e.g. when $q_{\min} \gtrsim 1/T$.

The optimality of our procedure can be deduced from the following minimax lower bound.

Theorem 4. *Let $\Sigma \in \mathcal{S}_{++}^{d\times d}$ be diagonal, $\pi \in \mathcal{P}(2^{[d]}), \epsilon \in [0,1]$, $\delta \in (0, 1/4]$, $\Theta := \mathbb{R}^d$ and $\mathcal{P}_\theta := \mathcal{P}^{\mathrm{arb}}(\theta, \Sigma, \epsilon, \pi)^{\otimes n}$ for $\theta \in \Theta$. Then*

$$\mathcal{M}\bigl(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\bigr)\begin{cases}\gtrsim \dfrac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{n} + \dfrac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\log(1/\delta)}{n} + \epsilon\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} & \text{if } \epsilon < \frac{q_{\min}}{1+q_{\min}} \\ = \infty & \text{if } \epsilon \geq \frac{q_{\min}}{1+q_{\min}}.\end{cases}$$

From Theorem 4, we see that up to multiplicative universal constants and when Σ is diagonal, Algorithm 1 has optimal behaviour under departures from MCAR and, up to a logarithmic factor in d and an iterated logarithmic factor in $1/\delta$, it adapts to the MCAR minimax quantile rate in settings where the MCAR term dominates. In Theorem 3, we imposed a condition of the form $q_{\min} \gtrsim \epsilon + T\log(d/\delta)/n$ for our upper bound, where the second part of this condition asks for the expected number of MCAR observations per coordinate to be at least poly-logarithmic in d and $1/\delta$. On the other hand, from Theorem 4, we see that $q_{\min} > \epsilon/(1-\epsilon)$ is necessary to ensure finite error with high probability.

4 Mean estimation under realisable contamination

In the arbitrary contamination setting of Section 3, we saw that the contamination fraction ϵ has a severe effect on our ability to estimate a population mean. The aim of this section, then, is to explore the potential benefits of restricting the form of contamination to MNAR observations from the same base distribution as our uncontaminated observations.

4.1 Gaussian realisable model

4.1.1 Univariate case

In this subsection, we consider Gaussian base distributions, and for $\theta \in \mathbb{R}$, as well as fixed $\sigma > 0$, $\epsilon \in [0,1)$ and $q \in (0,1]$, we write $\mathcal{R}(\theta) := \mathcal{R}\big(\mathsf{N}(\theta,\sigma^2),\epsilon,q\big)$ as shorthand. To gain intuition, recall the characterisation of univariate realisable distributions in Proposition 2: $R \in \mathcal{R}(\theta)$ if and only if both $R \ll \lambda_\star$ and the restriction $h : \mathbb{R} \to [0,\infty)$ of $\mathrm{d}R/\mathrm{d}\lambda_\star$ to $\mathbb{R}$ satisfies

$$h(x) \in \big[q(1-\epsilon)\phi_{(\theta,\sigma)}(x),\, \{q(1-\epsilon)+\epsilon\}\cdot\phi_{(\theta,\sigma)}(x)\big], \tag{15}$$

for λ-almost all x. In Figure 3(a), we plot a $\mathsf{N}(0,1)$-realisable h; on the other hand, in Figure 3(b), we consider the same function h and demonstrate that h is not $\mathsf{N}(1/2,1)$-realisable. This suggests that it may be possible to identify the mean by checking whether the condition in (15) is verified. Indeed, if $R \in \mathcal{R}(\theta_0)$, then for any $\theta \neq \theta_0$ and for $|x-\theta_0|$ sufficiently large, we have

$$h(x) \notin \big[q(1-\epsilon)\phi_{(\theta,\sigma)}(x),\, \{q(1-\epsilon)+\epsilon\}\cdot\phi_{(\theta,\sigma)}(x)\big].$$

Motivated by this observation, and given data $Z_1,\ldots,Z_n \in \mathbb{R}_\star$, we define $\mathcal{D} := \{i \in [n] : Z_i \neq \star\}$ and define an estimator $\widehat{\theta}^{\mathrm{AE}}_n$ as

$$\widehat{\theta}^{\mathrm{AE}}_n(Z_1,\ldots,Z_n) := \frac{1}{2}\cdot\Big(\max_{i\in\mathcal{D}} Z_i + \min_{i\in\mathcal{D}} Z_i\Big), \tag{16}$$

where we adopt the convention that $\widehat{\theta}^{\mathrm{AE}}_n := 0$ when $\mathcal{D} = \emptyset$. Thus, $\widehat{\theta}^{\mathrm{AE}}$ simply outputs the average of the extreme observed values. In the realisable model $\mathcal{R}(\theta_0)$, its performance as an estimator of θ_0 is summarised in the following theorem.

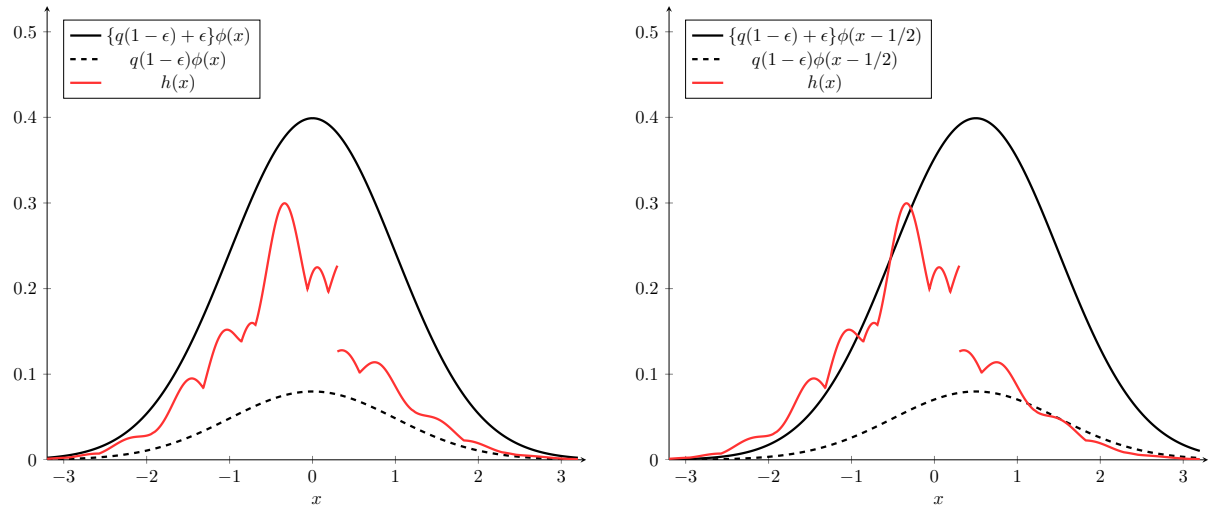


Figure 3: An example of a Gaussian-realisable distribution. Let $q = 1$ and $\epsilon = 0.8$. Panel (a) plots (i) $\{q(1-\epsilon)+\epsilon\} \cdot \phi(x)$ as a solid black curve, (ii) $q(1-\epsilon) \cdot \phi(x)$ as a dashed black curve and (iii) $\{q(1-\epsilon) + m(x)\} \cdot \phi(x)$ as a solid red curve, for some $m : \mathbb{R} \to [0,1]$. Note that the red curve is realisable by $\mathsf{N}(0,1)$. By contrast, panel (b) plots the red curve with no changes and uses $\phi(x-1/2)$ in place of $\phi(x)$ for the two black curves. In this case, the red curve is not realisable by $\mathsf{N}(1/2,1)$.

Theorem 5. *Let $\theta_0 \in \mathbb{R}$, $\epsilon \in [0,1)$, $q \in (0,1]$, $\sigma > 0$, $n \in \mathbb{N}$, $\delta \in [4e^{-nq(1-\epsilon)/8}, 1]$, and consider $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} R \in \mathcal{R}(\theta_0)$. Then with probability at least $1-\delta$,*

$$\big(\widehat{\theta}_n^{\text{AE}} - \theta_0\big)^2 \lesssim \frac{\sigma^2 \log^2(8/\delta)}{\log\big(nq(1-\epsilon)\big)} + \frac{\sigma^2 \log^2\big(1 + \frac{6\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)}.$$

We can interpret $nq(1-\epsilon)$ as the effective sample size from the MCAR component of R: on average, a proportion $1-\epsilon$ of our observations come from this MCAR component, and a proportion q of these are not missing. The condition $\delta \geq 4e^{-nq(1-\epsilon)/8}$ is therefore an effective sample size condition that asks for more MCAR observations for a higher confidence guarantee. Some condition of this form is necessary for the finiteness of the minimax quantile; see Theorem 6 below. One of the interesting features of the conclusion of Theorem 5 is that, if we consider ϵ and q as fixed, then consistent mean estimation in the Gaussian realisable model $\mathcal{R}(\theta_0)$ is possible. In fact, we can even achieve consistency when ϵ converges slowly to 1 and q converges slowly to zero, a stark contrast with the conclusions drawn in the arbitrary contamination model of Section 3. Moreover, these results are achievable via a very simple estimator that does not require knowledge of ϵ (or q or δ). On the other hand, the rate of convergence is only guaranteed to be logarithmic in the effective sample size (with the other problem parameters held fixed). Nevertheless, our high-probability minimax lower bound in Theorem 6 below shows that this is the best that one can hope for within this model, at least when ϵ is a positive constant.

Theorem 6. *Let $\epsilon \in [0,1)$, $q \in (0,1]$, $\sigma > 0$, $\delta \in (0, 1/4]$ and $n \in \mathbb{N}$. Suppose further that*

$$\log\left(1 + \frac{\epsilon}{q(1-\epsilon)}\right) \leq \log\bigl(nq(1-\epsilon)\bigr). \tag{17}$$

Then, writing $\Theta \coloneqq \mathbb{R}$, as well as $\mathcal{P}_\theta \coloneqq \bigl\{R^{\otimes n} : R \in \mathcal{R}(\theta)\bigr\}$ for $\theta \in \Theta$, we have

$$\mathcal{M}\bigl(\delta, \mathcal{P}_\Theta, |\cdot|^2\bigr) \begin{cases} \geq \dfrac{\sigma^2 \log(1/\delta)}{40nq(1-\epsilon)} + \dfrac{\sigma^2 \log^2\bigl(1 + \frac{\epsilon}{q(1-\epsilon)}\bigr)}{32\log\bigl(nq(1-\epsilon)\bigr)} & \text{if } \delta \geq \dfrac{\{1-q(1-\epsilon)\}^n}{2} \\ = \infty & \text{if } \delta < \dfrac{\{1-q(1-\epsilon)\}^n}{2}. \end{cases}$$

Condition (17) is a mild effective sample size assumption. When $q(1-\epsilon) \geq 1/2$, we have $\{1-q(1-\epsilon)\}^n \in [e^{-2nq(1-\epsilon)}, e^{-nq(1-\epsilon)}]$, so the range of δ for which we have a finite minimax $(1-\delta)$th quantile guarantee in Theorem 5 is almost optimal. Comparing the bounds in Theorem 6 with those in Theorem 5, we see that the second terms match up to a universal constant multiplicative factor. On the other hand, the first term in the lower bound in Theorem 6 may be much smaller than the corresponding term in Theorem 5, both in terms of its dependence on the effective sample size and on the quantile level.

To address the potential deficiency of the average of extremes estimator highlighted in the previous paragraph, we now introduce a minimum Kolmogorov distance estimator. Let $\widehat{R}_n \in \mathcal{P}(\mathbb{R}_\star)$ denote the empirical distribution of $Z_1, \ldots, Z_n \in \mathbb{R}_\star$, so that

$$\widehat{R}_n(B) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i \in B\}} \quad \text{for } B \in \mathcal{B}(\mathbb{R}_\star).$$

Let $\mathcal{A} \coloneqq \{(-\infty, t] : t \in \mathbb{R}\}$ denote the set of all closed lower half intervals on $\mathbb{R}$. For $R_1, R_2 \in \mathcal{P}(\mathbb{R}_\star)$ and $\mathcal{Q} \subseteq \mathcal{P}(\mathbb{R}_\star)$, define

$$d_{\mathrm{K}}(R_1, R_2) \coloneqq \sup_{A \in \mathcal{A}} \bigl|R_1(A) - R_2(A)\bigr| \quad \text{and} \quad d_{\mathrm{K}}(R_1, \mathcal{Q}) \coloneqq \inf_{Q \in \mathcal{Q}} d_{\mathrm{K}}(R_1, Q)$$

to be the Kolmogorov distance between R_1 and R_2, and the Kolmogorov distance between R_1 and the set $\mathcal{Q}$ respectively. Then, the minimum Kolmogorov distance estimator $\widehat{\theta}_n^{\mathrm{K}}$ for the Gaussian realisable class is defined as

$$\widehat{\theta}_n^{\mathrm{K}} \coloneqq \operatorname*{sargmin}_{\theta \in \mathbb{R}} d_{\mathrm{K}}\bigl(\widehat{R}_n, \mathcal{R}(\theta)\bigr),$$

where sargmin denotes the smallest element of the argmin set; this is well-defined since the function $\theta \mapsto d_{\mathrm{K}}\bigl(\widehat{R}_n, \mathcal{R}(\theta)\bigr)$ is continuous with $d_{\mathrm{K}}\bigl(\widehat{R}_n, \mathcal{R}(\theta)\bigr) \to 1$ as $|\theta| \to \infty$ and $d_{\mathrm{K}}\bigl(\widehat{R}_n, \mathcal{R}(0)\bigr) < 1$. We illustrate the Kolmogorov projection in Figure 4, and discuss its computation via a linear program in Section 4.1.3.

Theorem 7. *Let $\theta_0 \in \mathbb{R}$, $\epsilon \in [0,1)$, $q \in (0,1]$, $\sigma > 0$, $\delta \in (0,1]$, $n \geq \frac{e}{q(1-\epsilon)}$ and consider $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} R \in \mathcal{R}(\theta_0)$. Suppose that*

$$\delta \geq 4\exp\left\{-\frac{\bigl\{nq(1-\epsilon)\bigr\}^{31/36}}{6400\log\bigl(nq(1-\epsilon)\bigr)}\right\}, \tag{18}$$

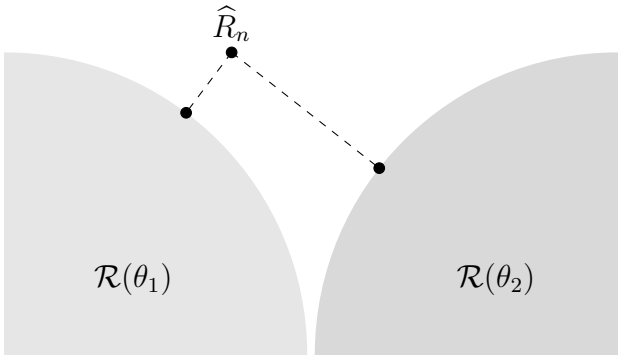


Figure 4: Illustration of the Kolmogorov projection onto two distinct realisable sets. The realisable sets are disjoint when $\theta_1 \neq \theta_2$, by Lemma 24.

and

$$\log\left(1 + \frac{4\epsilon}{q(1-\epsilon)}\right) \leq \frac{5}{216}\log\big(nq(1-\epsilon)\big). \tag{19}$$

Then with probability at least $1-\delta$,

$$(\widehat{\theta}_n^{\mathrm{K}} - \theta_0)^2 \lesssim C_{n,q,\epsilon,\delta}\left\{\frac{\sigma^2\log(4/\delta)}{nq(1-\epsilon)} + \frac{\sigma^2\log^2\big(1+\frac{4\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)}\right\}, \tag{20}$$

where

$$C_{n,q,\epsilon,\delta} := \begin{cases} 1 & \text{if } \log\big(1+\frac{4\epsilon}{q(1-\epsilon)}\big) \leq 2\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}} \\ \log\big(nq(1-\epsilon)\big) & \text{if } 2\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}} < \log\big(1+\frac{4\epsilon}{q(1-\epsilon)}\big) \leq 4\sqrt{\frac{\log(nq(1-\epsilon))\log(4/\delta)}{\{nq(1-\epsilon)\}^{31/36}}} \\ 1 & \text{if } \log\big(1+\frac{4\epsilon}{q(1-\epsilon)}\big) > 4\sqrt{\frac{\log(nq(1-\epsilon))\log(4/\delta)}{\{nq(1-\epsilon)\}^{31/36}}}. \end{cases}$$

The lower bound (18) on δ and the effective sample size condition (19) are both similar to those seen in Theorems 5 and 6. The main benefit of the minimum Kolmogorov distance estimator is that it is able to match both terms in the high-probability minimax lower bound of Theorem 6 up to a multiplicative universal constant, except in an intermediate parameter regime, where it may incur a multiplicative factor that is logarithmic in the effective sample size. Even in this middle regime, which covers the phase transition where the two terms in the bound (20) are equal, the rate remains polynomial in the effective sample size.

4.1.2 Multivariate extension

We now consider a simple multivariate extension of the Gaussian realisable model from the previous subsection, where for each observation, we either observe all coordinates simultaneously or none of them. Thus, for $P \in \mathcal{P}(\mathbb{R}^d)$, $\epsilon \in [0,1)$ and $\pi \in \mathcal{P}(\{\emptyset, [d]\})$, we define

$$\mathcal{R}_{\emptyset,[d]}(P,\epsilon,\pi) := \Big\{(1-\epsilon)\mathsf{MCAR}_{(\pi,P)} + \epsilon Q : Q \in \mathsf{MNAR}_{(\rho,P)},\ \rho \in \mathcal{P}\big(\{\emptyset,[d]\}\big)\Big\}. \tag{21}$$

For $\theta \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}_{++}^{d\times d}$, we write $\mathcal{R}_{\emptyset,[d]}(\theta) := \mathcal{R}_{\emptyset,[d]}\big(\mathsf{N}_d(\theta,\Sigma),\epsilon,\pi\big)$. Given $Z_1,\dots,Z_n \in \mathbb{R}^d_\star$ and $v \in \mathbb{S}^{d-1}$, let $\widehat{\theta}_n^{\mathrm{K}}(v)$ denote the one-dimensional minimum Kolmogorov distance estimator based on $Z_1^{(v)},\dots,Z_n^{(v)}$, where $Z_i^{(v)} := v^\top Z_i \cdot \mathbb{1}_{\{Z_i\in\mathbb{R}^d\}} + \star \cdot \mathbb{1}_{\{Z_i\notin\mathbb{R}^d\}}$ for $i \in [n]$. Let $\mathcal{N}$ denote a $(1/4)$-net in Euclidean norm of $\mathbb{S}^{d-1}$ with $|\mathcal{N}| \le 9^d$, which exists by, e.g., Vershynin (2018, Corollary 4.2.13). We define the multivariate minimum Kolmogorov distance estimator $\widehat{\theta}_n^{\mathrm{MK}}$ as

$$\widehat{\theta}_n^{\mathrm{MK}} := \operatorname*{sargmin}_{\theta\in\mathbb{R}^d} \max_{v\in\mathcal{N}} \big(v^\top\theta - \widehat{\theta}_n^{\mathrm{K}}(v)\big)^2,$$

where sargmin here denotes the smallest element of the argmin set in the lexicographic ordering.

Theorem 8. *Fix $d,n \in \mathbb{N}$, $\theta_0 \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}_{++}^{d\times d}$, $\epsilon \in [0,1)$, $\delta \in (0,1]$, $\pi \in \mathcal{P}(\{\emptyset,[d]\})$ and let $Z_1,\dots,Z_n \overset{\mathrm{iid}}{\sim} R \in \mathcal{R}_{\emptyset,[d]}(\theta_0)$. Writing $q := \pi([d])$, suppose that $nq(1-\epsilon) \ge e$,*

$$\delta \ge 4\exp\left\{ d\log 9 - \frac{\big\{nq(1-\epsilon)\big\}^{31/36}}{6400\log\big(nq(1-\epsilon)\big)}\right\},$$

and

$$\log\left(1+\frac{4\epsilon}{q(1-\epsilon)}\right) \le \frac{5}{216}\log\big(nq(1-\epsilon)\big).$$

Then with probability at least $1-\delta$,

$$\|\widehat{\theta}_n^{\mathrm{MK}} - \theta_0\|_2^2 \lesssim C_{n,q,\epsilon,\delta/(4\cdot 9^d)}\left\{\frac{\|\Sigma\|_{\mathrm{op}}\big(d+\log(4/\delta)\big)}{nq(1-\epsilon)} + \frac{\|\Sigma\|_{\mathrm{op}}\log^2\big(1+\frac{4\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)}\right\},$$

where $C_{n,q,\epsilon,\delta} > 0$ was defined in Theorem 7.

Theorem 8 reveals in particular that if we treat $\epsilon, q, \delta, \|\Sigma\|_{\mathrm{op}}$ as constants and if $\frac{d\log n}{n^{31/36}} \to 0$ as $n\to\infty$, then $\widehat{\theta}_n^{\mathrm{MK}}$ is a consistent estimator of θ_0. To facilitate comparisons with alternative estimators, we take $\Sigma = \sigma^2 I_d$ for simplicity. A naive application of the univariate minimum Kolmogorov distance estimator in each coordinate would, via Theorem 7 and a union bound, only yield a squared Euclidean error bound of order

$$C_{n,q,\epsilon,\delta/d}\left\{\frac{d\sigma^2\log(4d/\delta)}{nq(1-\epsilon)} + \frac{d\sigma^2\log^2\big(1+\frac{4\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)}\right\}$$

with probability at least $1-\delta$. Thus, in the first term, the dimension and quantile terms appear in a multiplicative as opposed to additive way, and the second term is inflated by a factor d. Similarly, if we were to apply the average of extremes estimator in each coordinate, then Theorem 5 and a union bound would give a squared Euclidean error bound of order

$$\frac{d\sigma^2\log^2(8d/\delta)}{\log\big(nq(1-\epsilon)\big)} + \frac{d\sigma^2\log^2\big(1+\frac{6\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)}.$$

In fact, a high-probability minimax lower bound is available in this setting: letting $\mathcal{P}_\theta := \{R^{\otimes n} : R \in \mathcal{R}_{\emptyset,[d]}(\mathsf{N}_d(\theta, \sigma^2 I_d), \epsilon, \pi)\}$, we have by combining Proposition 47 and Theorem 6 that

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2) \gtrsim \frac{\sigma^2(d + \log(1/\delta))}{nq(1-\epsilon)} + \frac{\sigma^2 \log^2(1 + \frac{\epsilon}{q(1-\epsilon)})}{\log(nq(1-\epsilon))}.$$

Thus, up to the logarithmic factor in the effective sample size in the middle regime of the bound in Theorem 8, the multivariate minimum Kolmogorov distance estimator is minimax rate optimal in this setting.

4.1.3 Computing the Kolmogorov distance between the empirical distribution and a realisable set

Let $Z_1, \ldots, Z_n \in \mathbb{R}_\star$, let $m := \sum_{i=1}^n \mathbb{1}_{\{Z_i \neq \star\}}$ and let $-\infty =: Z_{(0)} < Z_{(1)} \leq \cdots \leq Z_{(m)} < Z_{(m+1)} := \infty$ denote the ordered observed data. Further let $\widehat{R}_n$ be the empirical distribution of $Z_1, \ldots, Z_n$ and let $P \in \mathcal{P}(\mathbb{R})$ be such that $P \ll \lambda$. The following lemma and subsequent discussion provide an efficient way of computing the Kolmogorov distance between $\widehat{R}_n$ and the realisable set $\mathcal{R}(P, \epsilon, q)$ via linear programming.

Lemma 9. *Let $\epsilon \in [0,1)$ and $q \in [0,1]$. Writing $V_0 := 0$ and letting $\mathcal{V}$ denote the set of $(V_1, \ldots, V_{m+1})^\top \in [0,1]^{m+1}$ such that*

$$q(1-\epsilon) \cdot P\big((Z_{(i)}, Z_{(i+1)})\big) \leq V_{i+1} - V_i \leq \{q(1-\epsilon) + \epsilon\} \cdot P\big((Z_{(i)}, Z_{(i+1)})\big) \tag{22}$$

for all $i \in \{0\} \cup [m]$, we have

$$d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(P, \epsilon, q)\big) = \inf_{(V_1,\ldots,V_{m+1})^\top \in \mathcal{V}} \max_{i \in \{0\} \cup [m]} \left\{ \left| \frac{i}{n} - V_i \right| \vee \left| \frac{i}{n} - V_{i+1} \right| \right\}. \tag{23}$$

We can now rewrite the optimisation problem (23) as the following linear program:

$$\begin{aligned}
\text{minimise} \quad & t \\
\text{subject to} \quad & -t \leq \frac{i}{n} - V_i \leq t, && i \in \{0\} \cup [m] \\
& -t \leq \frac{i}{n} - V_{i+1} \leq t, && i \in \{0\} \cup [m] \\
& q(1-\epsilon) \cdot P\big((Z_{(i)}, Z_{(i+1)})\big) \leq V_{i+1} - V_i \\
& \qquad \leq \{q(1-\epsilon) + \epsilon\} \cdot P\big((Z_{(i)}, Z_{(i+1)})\big), && i \in \{0\} \cup [m].
\end{aligned}$$

This can be solved efficiently using standard software, e.g. `lpSolve` (Berkelaar et al., 2023) in `R`.

4.2 Nonparametric realisable models

4.2.1 Univariate case

We now broaden our scope from the Gaussian realisable setting of Section 4.1 and seek to determine the minimax quantiles for mean estimation, again over realisable classes, but

now with nonparametric families of base distributions, subject only to moment or tail decay conditions. To this end, for $\theta \in \mathbb{R}$, $\sigma > 0$ and $r \geq 2$, we define the class of distributions $\mathcal{P}_{L^r}(\theta, \sigma^2)$ with a finite rth moment:

$$\mathcal{P}_{L^r}(\theta, \sigma^2) := \Big\{ P \in \mathcal{P}(\mathbb{R}) : \mathbb{E}_P(X) = \theta,\, \mathbb{E}_P\big(|X - \theta|^r\big) \leq \sigma^r \Big\}. \tag{24}$$

Similarly, for $r \geq 1$, we consider tail decay conditions specified by Orlicz norms with Orlicz functions $\psi_r : t \mapsto e^{t^r} - 1$, and define

$$\mathcal{P}_{\psi_r}(\theta, \sigma^2) := \Big\{ P \in \mathcal{P}(\mathbb{R}) : \mathbb{E}_P(X) = \theta,\, \mathbb{E}_P\big\{\psi_r(|X - \theta|/\sigma)\big\} \leq 1 \Big\}. \tag{25}$$

Thus, if $X \sim P$ and we write $\|X\|_{\psi_r} := \inf\big\{t > 0 : \mathbb{E}\big(\psi_r(|X|/t)\big) \leq 1\big\}$, then $P \in \mathcal{P}_{\psi_r}(\theta, \sigma^2)$ if and only if $\mathbb{E}(X) = \theta$ and $\|X - \theta\|_{\psi_r} \leq \sigma$. We also remark that $\mathcal{P}_{\psi_1}(\theta, \sigma^2)$ and $\mathcal{P}_{\psi_2}(\theta, \sigma^2)$ correspond to classes of sub-exponential and sub-Gaussian distributions with mean θ respectively.

Upper bounds on the minimax quantiles for mean estimation over realisable classes with base distributions belonging to the classes in (24) and (25) are provided in Theorem 10 below. In the general setting of Theorem 10*(a)* where we only have a moment bound on the base distribution, the median of means estimator from Algorithm 2 is employed to obtain the logarithmic dependence on the quantile level δ, though this comes at the expense of the estimator being δ-dependent. On the other hand, in the more specialised setting of Theorem 10*(b)*, the sample mean of the observed data (which is δ-independent) suffices to obtain the logarithmic dependence on δ.

Theorem 10. *Let $\theta_0 \in \mathbb{R}$, $\epsilon \in [0, 1)$, $q \in (0, 1]$, $\sigma > 0$ and $\delta \in (0, 1]$.*

(a) Let $r \geq 2$, $P \in \mathcal{P}_{L^r}(\theta_0, \sigma^2)$ and $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} R \in \mathcal{R}(P, \epsilon, q)$. Assume further that $\delta \geq 2\exp\big(1 - nq(1-\epsilon)/8\big)$. Let $M := \lceil \log(2/\delta) \rceil$, $\mathcal{D} := \{i \in [n] : Z_i \neq \star\}$ and $\widehat{\theta}_n^{\text{MoM}} := \textsc{Median_of_Means}\big((Z_i)_{i \in \mathcal{D}}; M\big)$ from Algorithm 2. Then, with probability at least $1 - \delta$,

$$\big(\widehat{\theta}_n^{\text{MoM}} - \theta_0\big)^2 \lesssim \frac{\sigma^2 \log(2e/\delta)}{nq(1-\epsilon)} + \sigma^2 \cdot \left\{ \left(\frac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \left(\frac{\epsilon}{q(1-\epsilon)} \right)^{2/r} \right\}.$$

(b) Let $r \geq 1$, $P \in \mathcal{P}_{\psi_r}(\theta_0, \sigma^2)$, $R \in \mathcal{R}(P, \epsilon, q)$ and $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} R$. Assume further that $\delta \geq 8\exp\big(-nq(1-\epsilon)/8\big)$. Let $\mathcal{D} := \{i \in [n] : Z_i \neq \star\}$ and $\widehat{\theta}_n := |\mathcal{D}|^{-1} \sum_{i \in \mathcal{D}} Z_i$. Then, with probability at least $1 - \delta$,

$$(\widehat{\theta}_n - \theta_0)^2 \lesssim \frac{\sigma^2 \log(8/\delta)}{nq(1-\epsilon)} + \sigma^2 \cdot \left\{ \left(\frac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)} \right) \right\}.$$

In both parts of Theorem 10, the first term in the bound reflects the error incurred in estimating the mean of a distribution P belonging to either of the classes in (24) or (25) based on a sample of size $\lceil n(1-\epsilon) \rceil$ from $\mathsf{MCAR}_{(q,P)}$. The second terms arise from the contamination present in distributions $R \in \mathcal{R}(P, \epsilon, q)$. When the *effective contamination level* $\kappa := \frac{\epsilon}{q(1-\epsilon)}$ is small in the sense that $\kappa \leq 1$, in both parts of the theorem, the error

incurred from the contamination is at most of order $\sigma^2\kappa^2$. This is a substantial improvement on the corresponding term in the lower bound in Theorem 4 over arbitrary (non-realisable) contaminations of $\mathcal{P}_{L^2}(\theta_0, \sigma)$, which is of order $\sigma^2\kappa$. On the other hand, when $\kappa > 1$, the contribution to the error from the contamination term depends on the tail behaviour of P: when $P \in \mathcal{P}_{L^r}(\theta_0, \sigma^2)$, it is at most of order $\sigma^2\kappa^{2/r}$, whereas when $P \in \mathcal{P}_{\psi_r}(\theta_0, \sigma^2)$ the bound can be improved to order $\log^{2/r}(2+2\kappa)$. Again, these bounds represent a stark contrast with the lower bound in the arbitrary contamination model setting of Theorem 4, which is infinite in this regime. Finally, we remark that an attractive feature of the methods employed in Theorem 10 is that they do no require knowledge of κ.

Theorem 11 provides a complementary lower bound on the minimax quantiles in this realisable setting:

Theorem 11. *Let $\epsilon \in [0, 1)$, $q \in (0, 1]$, $\sigma > 0$ and $\delta \in (0, 1/4]$.*

(a) Let $r \geq 2$, $\Theta := \mathbb{R}$ and $\mathcal{P}_\theta := \{R^{\otimes n} : R \in \mathcal{R}(P, \epsilon, q),\ P \in \mathcal{P}_{L^r}(\theta, \sigma^2)\}$ for $\theta \in \Theta$. Then

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \begin{cases} \gtrsim \dfrac{\sigma^2 \log(1/\delta)}{nq(1-\epsilon)} + \sigma^2 \cdot \left\{ \left(\dfrac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \left(\dfrac{\epsilon}{q(1-\epsilon)} \right)^{2/r} \right\} & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \delta \geq e^{-nq(1-\epsilon)/2} \\ = \infty \qquad \text{if } \delta < \frac{(1-q(1-\epsilon))^n}{2}. \end{cases}$$

(b) Let $r \geq 1$, $\Theta := \mathbb{R}$ and $\mathcal{P}_\theta := \{R^{\otimes n} : R \in \mathcal{R}(P, \epsilon, q),\ P \in \mathcal{P}_{\psi_r}(\theta, \sigma^2)\}$ for $\theta \in \Theta$. Then

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \begin{cases} \gtrsim \dfrac{\sigma^2 \log(1/\delta)}{nq(1-\epsilon)} + \sigma^2 \cdot \left\{ \left(\dfrac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \log^{2/r}\left(2 + \dfrac{2\epsilon}{q(1-\epsilon)} \right) \right\} & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } \delta \geq e^{-nq(1-\epsilon)/2} \\ = \infty \qquad \text{if } \delta < \frac{(1-q(1-\epsilon))^n}{2}. \end{cases}$$

When $q(1-\epsilon) \leq c < 1$, we have $\{1 - q(1-\epsilon)\}^n > e^{-c'nq(1-\epsilon)}$ for some c' depending only on c, which reveals that the lower bound on δ in Theorem 10 for which can control the $(1-\delta)$th minimax quantile is essentially optimal. Thus, taken together, Theorems 10 and 11 determine the minimax quantiles of the quadratic loss function for mean estimation over our realisable classes up to universal constants. As a special case, these results demonstrate that when the effective contamination level κ is less than 1, the minimax quantile over realisable classes with base distribution $P \in \mathcal{P}_{L^2}(\theta_0, \sigma)$ scales as $\sigma^2\kappa^2$, which coincides with the minimax rate of mean estimation over Gaussian classes in the arbitrary contamination model; see Section C.3. Theorem 11*(b)* further reveals that, while careful examination of the tails enabled consistent estimation in the Gaussian realisable setting of Section 4.1, no such strategy can yield consistent estimation when the class is broadened to include all sub-Gaussian distributions with a fixed sub-Gaussian norm. More generally, the optimal rates of Theorems 10 and 11 provide a quantification of the benefits of realisable classes in terms of improved rates of mean estimation compared with the arbitrary contamination models of Section 3.

4.2.2 Multivariate extension

The aim of this subsection is to show that the univariate results of Section 4.2.1 extend to the problem of estimating a multivariate mean, under the same simplifying assumption on the set of possible observation patterns as that considered in Section 4.1.2. To this end, and by analogy with (24) and (25), for $d \in \mathbb{N}$, $\theta \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}_{++}^{d\times d}$ and $r > 0$, define

$$\mathcal{P}_{d,L^r}(\theta,\Sigma) := \big\{P \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_P(X) = \theta, \mathsf{Law}_{X\sim P}(v^\top X) \in \mathcal{P}_{L^r}(v^\top\theta, v^\top\Sigma v)\ \forall v \in \mathbb{S}^{d-1}\big\}$$

and

$$\mathcal{P}_{d,\psi_r}(\theta,\Sigma) := \big\{P \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_P(X) = \theta, \mathsf{Law}_{X\sim P}(v^\top X) \in \mathcal{P}_{\psi_r}(v^\top\theta, v^\top\Sigma v)\ \forall v \in \mathbb{S}^{d-1}\big\}.$$

Recall the definition of the realisable classes $\mathcal{R}_{\emptyset,[d]}(P,\epsilon,\pi)$ from (21).

Theorem 12. *Let $\theta_0 \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}_{++}^{d\times d}$, $\epsilon \in [0,1)$, $\delta \in (0,1]$, $\pi \in \mathcal{P}(\{\emptyset,[d]\})$ and $q := \pi([d])$.*

(a) Let $r \geq 2$, let $P \in \mathcal{P}_{d,L^r}(\theta_0,\Sigma)$ and let $Z_1,\dots,Z_n \overset{\text{iid}}{\sim} R \in \mathcal{R}_{\emptyset,[d]}(P,\epsilon,q)$. Further, let $\mathcal{D} := \{i \in [n] : Z_i \in \mathbb{R}^d\}$ and let $\widehat{\theta}_n := \textsc{Robust\_Descent}\big((Z_i)_{i\in\mathcal{D}}; 0, \delta\big)$ from Algorithm 5. There exists a universal constant $C \geq 8$ such that if $nq(1-\epsilon) \geq \mathbf{r}(\Sigma)/C$ and $\delta \geq 2\exp\big(-nq(1-\epsilon)/C\big)$, then with probability at least $1-\delta$,

$$\|\widehat{\theta}_n - \theta_0\|_2^2 \lesssim \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}}\log(2/\delta)}{nq(1-\epsilon)} + \|\Sigma\|_{\mathrm{op}}\bigg\{\bigg(\frac{\epsilon}{q(1-\epsilon)}\bigg)^2 \wedge \bigg(\frac{\epsilon}{q(1-\epsilon)}\bigg)^{2/r}\bigg\}.$$

(b) Let $r \geq 1$, $P \in \mathcal{P}_{d,\psi_r}(\theta_0,\Sigma)$, $R \in \mathcal{R}_{\emptyset,[d]}(P,\epsilon,q)$ and $Z_1,\dots,Z_n \overset{\text{iid}}{\sim} R$. Assume further that $nq(1-\epsilon) \geq \mathbf{r}(\Sigma)$ and that $\delta \geq 8\exp\big(-nq(1-\epsilon)/8\big)$. Let $\mathcal{D} := \{i \in [n] : Z_i \neq \star\}$ and $\widehat{\theta}_n := |\mathcal{D}|^{-1}\sum_{i\in\mathcal{D}} Z_i$. Then, with probability at least $1-\delta$,

$$\|\widehat{\theta}_n - \theta_0\|_2^2 \lesssim \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}}\log(2/\delta)}{nq(1-\epsilon)} + \|\Sigma\|_{\mathrm{op}}\bigg\{\bigg(\frac{\epsilon}{q(1-\epsilon)}\bigg)^2 \wedge \log^{2/r}\bigg(2 + \frac{2\epsilon}{q(1-\epsilon)}\bigg)\bigg\}.$$

Thus, even in the multivariate setting considered in Theorem 12, the realisable classes permit the same improvements in performance over fully-observed arbitrary contamination models as we saw in the univariate case in Theorem 10. Moreover, the second (contamination) terms retain this improvement in a dimension-free manner. By a small modification of the two-point construction of the proof of Theorem 11 so that the difference in means is in the direction of the leading eigenvector of Σ, together with Ma, Verchand and Samworth (2024, Proposition 10) or Depersin and Lecué (2022a, Theorem 4), we see that both bounds in Theorem 12 are minimax rate-optimal.

A difference between Theorem 12 and Theorem 10 is that in the $\mathcal{P}_{d,L^r}(\theta_0,\Sigma)$ model, we employ the Robust_Descent method of Algorithm 5 instead of the (coordinate-wise) Median_of_Means algorithm. This is to ensure that the first (MCAR) term attains the optimal rate in the multivariate setting. Nevertheless both algorithms in Theorem 12 remain adaptive to the unknown effective contamination level κ.

5 Extension: Linear regression with realisable missing response

The ideas of mean estimation with realisable missing observations developed in Section 4 extend to linear regression with a realisable missing response, as we now demonstrate. Given $\theta_0 \in \mathbb{R}^d$ and $\sigma > 0$, consider a random design normal linear model

$$Y_i = X_i^\top \theta_0 + \zeta_i,$$

where $(X_1, \zeta_1), \dots, (X_n, \zeta_n)$ are independent with $\zeta_1, \dots, \zeta_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$. In this section, we suppose that, instead of the desired responses $Y_1, \dots, Y_n$, we are only able to observe corrupted versions that are subject to realisable missingness. Thus, more precisely, we observe $Z_i := (1 - B_i) \cdot Y_i \circledast \Omega_i^{(1)} + B_i \cdot Y_i \circledast \Omega_i^{(2)}$ for $i \in [n]$, where $B_1, \dots, B_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\epsilon)$ are independent of the independent quadruples $(X_i, \zeta_i, \Omega_i^{(1)}, \Omega_i^{(2)})_{i \in [n]}$, where $\Omega_i^{(1)} | \{X_i = x\} \sim \mathsf{Ber}(q_x)$ with $\inf_{x \in \mathbb{R}^d} q_x \geq q > 0$ and where $\Omega_i^{(1)} \perp\!\!\!\perp \zeta_i | X_i$. We impose no restriction on the dependence between $\Omega_i^{(2)}$ and (X_i, ζ_i). Thus, when $\epsilon = 0$, the pairs $(X_1, Z_1), \dots, (X_n, Z_n)$ are missing at random (MAR). We summarise the conditional distribution of the observed responses by writing $Z_1 | \{X_1 = x\} \sim R_x \in \mathcal{R}^{\text{Res}}\big(\mathsf{N}(x^\top \theta_0, \sigma^2), \epsilon, q\big)$.

Our main result in this section relies on what we will call a (β, γ)-regular design assumption, for $\beta \in (0, 1/2], \gamma > 0$.

Assumption 1 ((β, γ)-regular design)**.** For all $v \in \mathbb{R}^d$, there exists a set $\mathcal{T} \subseteq [n]$ such that $|\mathcal{T}| \geq 2\beta n$ and $|X_i^\top v| > \gamma \|v\|_2$ for all $i \in \mathcal{T}$.

It is convenient to let $\mathcal{C}_{(\beta,\gamma)}$ denote the set of $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$ for which the design is (β, γ)-regular. The following lemma shows that a random design where the distribution is not concentrated on a hyperplane is (β, γ)-regular for some $\beta \in (0, 1/2]$ and $\gamma > 0$, with high probability when the sample size is sufficiently large.

Lemma 13. *Let $\delta \in (0, 1]$, $\gamma > 0$ and $X_1, \dots, X_n \overset{\text{iid}}{\sim} P \in \mathcal{P}(\mathbb{R}^d)$. Further define $\beta := \frac{1}{3} \inf_{v \in \mathbb{S}^{d-1}} \mathbb{P}\big(|X_1^\top v| > \gamma\big)$.*

(a) The infimum in the definition of β is attained, and if $P(H) < 1$ for every hyperplane $H \subseteq \mathbb{R}^d$, then $\beta > 0$ for sufficiently small $\gamma > 0$.

(b) There exists a universal constant $c > 0$ such that if $\frac{d + \log(1/\delta)}{n} \leq c\beta^2$, then, with probability at least $1 - \delta$, the $n \times d$ matrix with ith row $X_i^\top$ is a (β, γ)-regular design.

To illustrate Lemma 13*(b)* with a specific example, suppose that $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{N}_d(0, \Sigma)$ for some $\Sigma \in \mathcal{S}_{++}^{d \times d}$, and let $\lambda_{\min}(\Sigma) > 0$ denote the minimum eigenvalue of Σ. Then by Lemma 13*(b)*, there exists a universal constant $c_1 > 0$ such that if $\frac{d + \log(1/\delta)}{n} \leq c_1$, then with probability at least $1 - \delta$, the $n \times d$ matrix with ith row $X_i^\top$ is a $\big(2\Phi(-1)/3, \lambda_{\min}^{1/2}(\Sigma)\big)$-regular design.

Given $R_1, R_2 \in \mathcal{P}(\mathbb{R}_\star)$, we define their *symmetrised Kolmogorov distance* by

$$d_{\text{K}}^{\text{sym}}(R_1, R_2) := \sup_{A \in \mathcal{A}^{\text{sym}}} |R_1(A) - R_2(A)|,$$

where $\mathcal{A}^{\mathrm{sym}} := \{(-\infty, t] : t \in \mathbb{R}\} \cup \{[t, \infty) : t \in \mathbb{R}\}$. This may be larger than $d_{\mathrm{K}}(R_1, R_2)$ when $R_1(\{\star\}) \neq R_2(\{\star\})$. Now, for $\theta \in \mathbb{R}^d$, we define the empirical distribution $\widehat{R}_{n,\theta}$ by $\widehat{R}_{n,\theta}(B) := n^{-1} \sum_{i=1}^n \mathbb{1}_{\{Z_i - X_i^\top \theta \in B\}}$ for $B \in \mathcal{B}(\mathbb{R}_\star)$. We let

$$\mathcal{R}_0^{\mathrm{Lin}} := \mathcal{R}\big(\mathsf{N}(0, \sigma^2), 1 - q(1-\epsilon), 1\big),$$

and define the Kolmogorov distance estimator as

$$\widehat{\theta}_n^{\mathrm{K}} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} d_{\mathrm{K}}^{\mathrm{sym}}\big(\widehat{R}_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}\big). \tag{26}$$

The realisable set $\mathcal{R}_0^{\mathrm{Lin}}$ ensures that the selection probability $h_x := \mathbb{P}(Z_1 \neq \star \mid X_1 = x)$ satisfies the sandwich relation $q(1-\epsilon) \leq q_x(1-\epsilon) \leq h_x \leq 1$ for all $x \in \mathbb{R}^d$. Writing $\widetilde{R}_{i,\theta} := \mathsf{Law}(Z_i - x_i^\top \theta)$ for $i \in [n]$ and $\theta \in \mathbb{R}^d$, as well as $R_{n,\theta} := \frac{1}{n} \sum_{i=1}^n \widetilde{R}_{i,\theta}$, it then follows from Proposition 2 that $R_{n,\theta_0} \in \mathcal{R}_0^{\mathrm{Lin}}$.

Theorem 14. *Let $n, d \in \mathbb{N}, \epsilon \in [0,1), q \in (0,1], \delta \in (0,1]$ and $\theta_0 \in \mathbb{R}^d$. Let $(x_1, \ldots, x_n) \in \mathcal{C}_{(\beta,\gamma)}$ for some $\beta \in (0, 1/2]$ and $\gamma > 0$, and let $Z_1, \ldots, Z_n$ be independent with $Z_i | \{X_i = x_i\} \sim R_{x_i} \in \mathcal{R}^{\mathrm{Res}}\big(\mathsf{N}(x_i^\top \theta_0, \sigma^2), \epsilon, q\big)$. Then there exists a universal constant $C_1 > 0$ such that if*

$$\frac{n^{31/36}}{\log n} \geq C_1\big\{d + \log(1/\delta)\big\} \quad \textit{and} \quad \log\left(1 + \frac{4(1 - \beta q(1-\epsilon))}{\beta q(1-\epsilon)}\right) \leq \frac{\log n}{18}, \tag{27}$$

then with probability at least $1 - \delta$, conditional on $X_1 = x_1, \ldots, X_n = x_n$,

$$\big\|\widehat{\theta}_n^{\mathrm{K}} - \theta_0\big\|_2^2 \lesssim \sigma^2 \cdot \frac{\log^2\big(1 + \frac{4(1-\beta q(1-\epsilon))}{\beta q(1-\epsilon)}\big)}{\gamma^2 \log\big(nq(1-\epsilon)\big)}.$$

One of the main consequences of Theorem 14 is that if we consider $\beta, \gamma, q, \epsilon$ and δ as constants, but allow d to grow subject to the first part of (27) holding, then under the assumptions of Theorem 14, we have that $\widehat{\theta}_n^{\mathrm{K}}$ is a consistent estimator of θ_0 in squared Euclidean norm as $n \to \infty$. In fact, similar to Section 4.1, this conclusion continues to hold even if we allow ϵ to converge slowly to 1 and q to converge slowly to zero. This lies in stark contrast to the complete-case arbitrary contamination setting in which for any constant ϵ, consistent estimation is impossible (see, e.g., the discussion following Gao, 2020, Theorem 3.2). Moreover, when the parameters $\beta, \gamma, q, \epsilon$ and δ are positive constants, the optimality of the rate $1/\log n$ follows from our mean estimation lower bound (Theorem 6).

Acknowledgements: The research of TM, KAV and RJS was supported by RJS's European Research Council (ERC) Advanced Grant 101019498. The work of KAV was supported in part by National Science Foundation grant DMS-2210734. The work of TBB was supported by Engineering and Physical Sciences Research Council (EPSRC) New Investigator Award EP/W016117/1 and ERC Starting Grant 101163546. The research of TW was supported by EPSRC New Investigator Award EP/T02772X/1.

References

Aronow, P. M. and Lee, D. K. (2013) Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika*, **100**, 235–240.

Bakshi, A. and Prasad, A. (2021) Robust linear regression: Optimal rates in polynomial time. In *Symposium on Theory of Computing*, 102–115.

Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017) Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **79**, 939–956.

Berkelaar, M. et al. (2023) *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*. R package version 5.6.20.

Berrett, T. B. and Samworth, R. J. (2023) Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility. *The Annals of Statistics*, **51**, 2170–2193.

Bickel, P. J. and Ritov, J. (1991) Large sample theory of estimation in biased sampling regression models. I. *The Annals of Statistics*, **19**, 797–816.

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K. (1989) Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, **36**, 929–965.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*. Springer Science & Business Media.

Cai, T. T. and Wei, H. (2021) Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, **49**, 100–128.

Cai, T. T. and Zhang, L. (2019) High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **81**, 675–705.

Chen, M., Gao, C. and Ren, Z. (2018) Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, **46**, 1932–1960.

Craven, B. D. and Koliha, J. J. (1977) Generalizations of Farkas' theorem. *SIAM Journal on Mathematical Analysis*, **8**, 983–997.

Cressie, N. (2015) *Statistics for Spatial Data*. John Wiley & Sons.

Daskalakis, C., Gouleakis, T., Tzamos, C. and Zampetakis, M. (2018) Efficient statistics, in high dimensions, from truncated samples. In *Foundations of Computer Science*, 639–649.

Depersin, J. and Lecué, G. (2022a) Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probability Theory and Related Fields*, **183**, 997–1025.

Depersin, J. and Lecué, G. (2022b) Robust sub-Gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, **50**, 511–536.

Diakonikolas, I. and Kane, D. M. (2023) *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press.

Diakonikolas, I., Kane, D. M., Pittas, T. and Zarifis, N. (2024) Statistical query lower bounds for learning truncated Gaussians. In *Conference on Learning Theory*, 1336–1363.

Do, K. T., Wahl, S., Raffler, J., Molnos, S., Laimighofer, M., Adamski, J., Suhre, K., Strauch, K., Peters, A., Gieger, C., Langenberg, C., Stewart, I. D., Theis, F. J., Grallert, H., Kastenmüller and Krumsiek, J. (2018) Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics*, **14**, 1–18.

Dudley, R. M. (2018) *Real Analysis and Probability*. CRC Press.

Elsener, A. and van de Geer, S. (2019) Sparse spectral estimation with missing and corrupted measurements. *Stat*, **8**, e229.

Farewell, D., Daniel, R. and Seaman, S. (2022) Missing at random: a stochastic process perspective. *Biometrika*, **109**, 227–241.

Farkas, J. (1902) Theorie der einfachen Ungleichungen. *Journal für die Reine und Angewandte Mathematik*, **1902**, 1–27.

Follain, B., Wang, T. and Samworth, R. J. (2022) High-dimensional changepoint estimation with heterogeneous missingness. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **84**, 1023–1055.

Folland, G. B. (1999) *Real Analysis: Modern Techniques and their Applications*. John Wiley & Sons.

Gao, C. (2020) Robust regression via mutivariate regression depth. *Bernoulli*, **26**, 1139–1170.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988) Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, 1069–1112.

Götze, F., Sambale, H. and Sinulis, A. (2021) Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, **26**, 1–22.

Hirano, K., Imbens, G. W. and Ridder, G. (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**, 1161–1189.

Horn, R. A. (1990) The Hadamard product. In *Proceedings of Symposia in Applied Mathematics*, 87–169.

Horn, R. A. and Johnson, C. R. (2012) *Matrix Analysis*. Cambridge University Press.

Hu, L. and Reingold, O. (2021) Robust mean estimation on highly incomplete data with arbitrary outliers. In *Conference on Artificial Intelligence and Statistics*, 1558–1566.

Huber, P. J. (1964) Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 73–101.

Jana, S., Fan, J. and Kulkarni, S. (2024) A general theory for robust clustering via trimmed mean. *arXiv preprint arXiv:2401.05574*.

Kechris, A. (2012) *Classical Descriptive Set Theory*. Springer Science & Business Media.

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D. et al. (2018) An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, **82**, 1–33.

Kontonis, V., Tzamos, C. and Zampetakis, M. (2019) Efficient truncated statistics with unknown truncation. In *Foundations of Computer Science*, 1578–1595.

Lerasle, M. and Oliveira, R. I. (2011) Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.

Little, R. J. and Rubin, D. B. (2014) *Statistical Analysis with Missing Data*. John Wiley & Sons.

Liu, A. and Moitra, A. (2023) Robustly learning general mixtures of Gaussians. *Journal of the ACM*, **70**, 1–53.

Loh, P.-L. and Tan, X. L. (2018) High-dimensional robust precision matrix estimation: Cell-wise corruption under ϵ-contamination. *Electronic Journal of Statistics*, **12**, 1429–1467.

Loh, P.-L. and Wainwright, M. J. (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *The Annals of Statistics*, **40**, 1637.

Lounici, K. (2014) High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, **20**, 1029–1058.

Lugosi, G. and Mendelson, S. (2021) Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, **49**, 393–410.

Ma, T., Verchand, K. A. and Samworth, R. J. (2024) High-probability minimax lower bounds. *arXiv preprint arXiv:2406.13447*.

Massart, P. (1990) The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, **18**, 1269–1283.

McCaffrey, D. F. and Lockwood, J. R. (2011) Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, **5**, 773–797.

McDiarmid, C. (1998) Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics* (Habib, M., McDiarmid, C., Ramirez-Alfonsin, J. and Reed, B., eds.), 195–248, Springer.

McKennan, C., Ober, C. and Nicolae, D. (2020) Estimation and inference in metabolomics with non-random missing data and latent factors. *The Annals of Applied Statistics*, **14**, 789–808.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018) *Foundations of Machine Learning*. MIT Press.

Munkres, J. (2014) *Topology*. Pearson, 2nd ed.

Pensia, A., Jog, V. and Loh, P.-L. (2024+) Robust regression with covariate filtering: Heavy tails and adversarial contamination. *Journal of the American Statistical Association (to appear)*.

Polyanskiy, Y. and Wu, Y. (2024) *Information Theory: From Coding to Learning*. Cambridge University Press.

Prince, M. (2012) Epidemiology. In *Core Psychiatry* (Wright, P., Stern, J. and Phelan, M., eds.), 115–129, Elsevier Health Sciences.

Reeve, H. W. (2024) A short proof of the Dvoretzky–Kiefer–Wolfowitz–Massart inequality. *arXiv preprint arXiv:2403.16651*.

Reeve, H. W. J., Cannings, T. I. and Samworth, R. J. (2021) Adaptive transfer learning. *The Annals of Statistics*, **49**, 3618–3649.

Rosenbaum, P. R. (1987) Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, **74**, 13–26.

Sahoo, R., Lei, L. and Wager, S. (2022) Learning from a biased sample. *arXiv preprint arXiv:2209.01754*.

Schaefer, H. H. (1971) *Topological Vector Spaces*. Springer-Verlag.

Seaman, S., Galati, J., Jackson, D. and Carlin, J. (2013) What Is Meant by "Missing at Random"? *Statistical Science*, **28**, 257–268.

Sell, T., Berrett, T. B. and Cannings, T. I. (2024) Nonparametric classification with missing data. *The Annals of Statistics*, **52**, 1178–1200.

Tanguy, K. (2015) Some superconcentration inequalities for extrema of stationary Gaussian processes. *Statistics and Probability Letters*, **106**, 239–246.

Tukey, J. W. (1975) Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, vol. 2, 523–531.

van der Vaart, A. W. (1998) *Asymptotic Statistics*. Cambridge University Press, 1st ed.

Vardi, Y. (1985) Empirical distributions in selection bias models. *The Annals of Statistics*, **13**, 178–203.

Vershynin, R. (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Wainwright, M. J. (2019) *High-dimensional Statistics: A Non-asymptotic Viewpoint*, vol. 48. Cambridge University Press.

Xie, Y., Huang, J. and Willett, R. (2012) Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, **7**, 12–27.

Yan, Y., Chen, Y. and Fan, J. (2024) Inference for heteroskedastic PCA with missing data. *The Annals of Statistics*, **52**, 729–756.

Zhao, Q., Small, D. S. and Bhattacharya, B. B. (2019) Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **81**, 735–761.

Zhivotovskiy, N. (2024) Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, **29**, 1–28.

Zhu, Z., Wang, T. and Samworth, R. J. (2022) High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **84**, 2000–2031.

A Notation used in proofs

For a measurable space $(\mathcal{Z}, \mathcal{C})$ and probability measures $P, Q \in \mathcal{P}(\mathcal{Z})$, we write $P \perp Q$ if P and Q are singular. The Lebesgue decomposition theorem yields the unique decomposition $P = P_{\mathrm{ac}} + P_{\mathrm{sing}}$ where $P_{\mathrm{ac}} \ll Q$ and where $P_{\mathrm{sing}} \perp Q$. For a convex function $f : (0, \infty) \to \mathbb{R}$, we let $M_f := \lim_{x\to\infty} f(x)/x \in (-\infty, \infty]$ denote its *maximal slope*. We then define the *f-divergence* between P and Q to be

$$\mathrm{Div}_f(P, Q) := \int_{\mathcal{Z}} f\left(\frac{\mathrm{d}P_{\mathrm{ac}}}{\mathrm{d}Q}\right) \mathrm{d}Q + M_f \cdot P_{\mathrm{sing}}(\mathcal{Z}). \tag{28}$$

As important examples, if $f(x) = |x - 1|/2$, then we obtain the total variation distance $\mathrm{TV}(P, Q) := \sup_{A\in\mathcal{C}} |P(A) - Q(A)|$, while if $f(x) = x \log x$, then the resulting f-divergence is the Kullback–Leibler divergence

$$\mathrm{KL}(P, Q) := \begin{cases} \int_{\mathcal{Z}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q & \text{if } P \ll Q \\ \infty & \text{otherwise.} \end{cases}$$

Finally, if $f(x) = (x - 1)^2$, then we obtain the χ^2-divergence

$$\chi^2(P, Q) := \begin{cases} \int_{\mathcal{Z}} \left(\frac{dP}{dQ} - 1\right)^2 dQ & \text{if } P \ll Q \\ \infty & \text{otherwise.} \end{cases}$$

Recalling the spaces $\mathcal{X}_1, \ldots, \mathcal{X}_d$ from Section 2.1, for a set $S \in 2^{[d]} \setminus \{\emptyset\}$, let $\mathcal{X}_S := \prod_{j\in S} \mathcal{X}_j$, and also define $\mathcal{X}_\emptyset := \{\star\}$ and $\mathcal{X} := \prod_{j=1}^d \mathcal{X}_j$. Given $x = (x_1, \ldots, x_d) \in \mathcal{X}$ and $S \in 2^{[d]} \setminus \{\emptyset\}$, we define $x_S := (x_j)_{j\in S}$, with $x_\emptyset := \star$. For $S \subseteq [d]$, we define $\mathcal{X}_j^{(S)} := \mathcal{X}_j$ if $j \in S$ and $\mathcal{X}_j^{(S)} := \{\star\}$ if $j \notin S$, and also set $\mathcal{X}^{(S)} := \prod_{j=1}^d \mathcal{X}_j^{(S)}$. Next, we let

$$\mathcal{B}^{(S)}(\mathcal{X}_\star) := \big\{A \in \mathcal{B}(\mathcal{X}_\star) : \forall z = (z_1, \ldots, z_d) \in A,\, z_j \neq \star,\, \forall j \in S \text{ and } z_k = \star,\, \forall k \notin S\big\}.$$

Given $S \subseteq [d]$, we write $\mathcal{G}_S$ for the set of real-valued functions on $\mathcal{X}_S$, and also write $\mathcal{G}_\star$ for the set of real-valued functions on $\mathcal{X}_\star$. A function $f \in \mathcal{G}_\star$ may be identified with the sequence of functions $(f_S : S \subseteq [d])$, where $f_S \in \mathcal{G}_S$ for each S. Formally, this identification is via the bijection $\psi : \prod_{S\subseteq[d]} \mathcal{G}_S \to \mathcal{G}_\star$ given by $\psi\big((f_{S'} : S' \subseteq [d])\big)(z) \coloneqq f_S(z_S)$ for $z \in \mathcal{X}^{(S)}$ and $S \subseteq [d]$. In other words, we evaluate $f \in \mathcal{G}_\star$ at $z \in \mathcal{X}_\star$ by setting S to be the coordinates in z that are not equal to $\star$, and then computing $f_S(z_S)$.

B Proofs from Section 2

B.1 Proof of Theorem 1

We begin with a sketch of the proof in the setting where $\mathcal{X}$ is finite, both to explain the relevance of (a generalisation of) Farkas's lemma in this context, and to provide intuition for the more technical arguments that follow. Let $X \sim P \in \mathcal{P}(\mathcal{X})$ and let $Q \coloneqq \mathsf{Law}(X \circledast \Omega)$ for some random vector Ω taking values in $\{0,1\}^d$. We write $M = (M_{S,x})_{S\subseteq[d],x\in\mathcal{X}} \coloneqq \big(\mathbb{P}(\Omega = \mathbf{1}_S \,|\, X = x)\big)_{S\subseteq[d],x\in\mathcal{X}} \in [0,1]^{2^{[d]}\times\mathcal{X}}$ to summarise the missingness mechanism. Now write $\mathbb{A} \in [0,1]^{\mathcal{X}_\star\times(2^{[d]}\times\mathcal{X})}$ for the matrix with

$$\mathbb{A}_{z,(S,x)} \coloneqq P(\{x\})\mathbb{1}_{\{z_S = x_S\}} \prod_{j\in S^c} \mathbb{1}_{\{z_j=\star\}},$$

so that each column of $\mathbb{A}$ has at most one non-zero entry. Then

$$(\mathbb{A}M)_z = \sum_{S\subseteq[d]} \sum_{x\in\mathcal{X}} P(\{x\}) M_{S,x} \mathbb{1}_{\{z_S=x_S\}} \prod_{j\in S^c} \mathbb{1}_{\{z_j=\star\}} = Q(\{z\}).$$

Now, for $x \in \mathcal{X}$, write $\sigma_x \in \{0,1\}^{2^{[d]}\times\mathcal{X}}$ for the vector with $(\sigma_x)_{(S,x')} \coloneqq \mathbb{1}_{\{x=x'\}}$, so that $\sigma_x^\top M = \sum_{S\in 2^{[d]}} M_{S,x}$, and form the matrix $\mathbb{B} \coloneqq (\sigma_x^\top)_{x\in\mathcal{X}} \in \{0,1\}^{\mathcal{X}\times(2^{[d]}\times\mathcal{X})}$. We can then define $\mathcal{J} \coloneqq \{M \in [0,1]^{2^{[d]}\times\mathcal{X}} : \mathbb{B}M = \mathbf{1}_\mathcal{X}\}$ to denote the set of valid mechanisms. We deduce that $Q \in \mathsf{MNAR}_P$ if and only if there exists $M \in \mathcal{J}$ such that $\mathbb{A}M = Q$. By Farkas's lemma, this latter condition is equivalent to the statement that there does not exist $(y,w) = \big((y_z)_{z\in\mathcal{X}_\star}, (w_x)_{x\in\mathcal{X}}\big) \in \mathbb{R}^{\mathcal{X}_\star} \times \mathbb{R}^\mathcal{X}$ such that $\sum_{z\in\mathcal{X}_\star} Q(\{z\})y_z + \sum_{x\in\mathcal{X}} w_x < 0$ and $0 \le (\mathbb{A}^\top y + \mathbb{B}^\top w)_{(S,x)} = P(\{x\})y_{x\circledast\mathbf{1}_S} + w_x$ for each $S \subseteq [d]$ and $x \in \mathcal{X}$. The search for such a pair (y,w) amounts to a constrained optimisation problem, whose solution for each fixed y is to take $w_x = -P(\{x\}) \min_{S\subseteq[d]} y_{x\circledast\mathbf{1}_S}$ for $x \in \mathcal{X}$. Then

$$\sum_{z\in\mathcal{X}_\star} Q(\{z\})y_z + \sum_{x\in\mathcal{X}} w_x = \sum_{z\in\mathcal{X}_\star} Q(\{z\})y_z - \sum_{x\in\mathcal{X}} P(\{x\}) \min_{S\subseteq[d]} y_{x\circledast\mathbf{1}_S},$$

so the condition that there does not exist (y,w) for which this quantity is negative corresponds to (10) after identifying y with $-f$.

Moving now to the proof of the full theorem, we require several preliminary topological results that are stated and proved in Section F. We will also use the generalisation of Farkas's lemma below. Recall that if X is a real vector space, then the *algebraic dual* of X, denoted X^*, is the vector space of linear functions $f : X \to \mathbb{R}$. Whenever X' is a subspace of this algebraic dual, we say X' *separates points* if for every $x_1, x_2 \in X$ with $x_1 \neq x_2$, there exists

$f \in X'$ with $f(x_1) \neq f(x_2)$. The *weak topology* on X generated by X' is the coarsest topology such that $f^{-1}(U)$ is open in X for every $f \in X'$ and open set $U \subseteq \mathbb{R}$. Now let Y be another real vector space and let Y' be a subspace of its algebraic dual. A linear map $T : X \to Y$ is (X', Y')-*weakly continuous* if it is continuous when X and Y are equipped with the weak topologies generated by X' and Y' respectively. Where X' and Y' are clear from context, we will abbreviate this terminology by simply referring to T as weakly continuous.

Theorem 15 (Craven and Koliha, 1977, Theorem 2)**.** *Let X and Y be real vector spaces, and let X' and Y' be subspaces of the algebraic duals of X and Y, respectively, that separate points. Given $y \in Y$, a weakly continuous linear map $T : X \to Y$, and a convex cone $K \subseteq X$ such that $T(K)$ is weakly closed in Y, the following are equivalent:*

(a) $Tx = y$ has a solution $x \in K$;

(b) If $g \in Y'$ satisfies $g(Tx) \geq 0$ for all $x \in K$, then $g(y) \geq 0$.

For any topological space $\mathcal{Z}$, we write $C_{\mathrm{b}}(\mathcal{Z})$ for the space of bounded continuous real-valued functions on $\mathcal{Z}$. Let $\mathcal{M}(\mathcal{Z})$ denote the space of finite, signed Borel measures on $\mathcal{Z}$ and let $\mathcal{M}_+(\mathcal{Z})$ be the subspace of (non-negative) finite Borel measures. We call $\mathcal{Z}$ a *Hausdorff space* if, given any distinct $z_1, z_2 \in \mathcal{Z}$, we can find disjoint open subsets V_1, V_2 such that $z_1 \in V_1$, $z_2 \in V_2$. The space $\mathcal{Z}$ is *locally compact* if every point in $\mathcal{Z}$ has a compact neighbourhood, i.e. if for every $z \in \mathcal{Z}$, we can find an open set $U \subseteq \mathcal{Z}$ and a compact set $K \subseteq \mathcal{Z}$ such that $z \in U \subseteq K$.

The main content of the proof of Theorem 1 is Proposition 16 below. Observe that the restriction of the bijection ψ in Section A to the set $\{(f_S : S \subseteq [d]) : f_S \in C_{\mathrm{b}}(\mathcal{X}_S)\ \forall S \subseteq [d]\}$ has image $C_{\mathrm{b}}(\mathcal{X}_\star)$. This identifies $C_{\mathrm{b}}(\mathcal{X}_\star)$ with $\big(C_{\mathrm{b}}(\mathcal{X}_S) : S \in 2^{[d]}\big)$, but henceforth we will not be explicit about this identification, and will simply write $f = (f_S : S \in 2^{[d]}) \in C_{\mathrm{b}}(\mathcal{X}_\star)$. Given such an $f = (f_S : S \in 2^{[d]}) \in C_{\mathrm{b}}(\mathcal{X}_\star)$, we can express the function $f_{\max}$ from Section 2.3 as $f_{\max}(x) \coloneqq \max_{S \in 2^{[d]}} f_S(x_S)$ for $x \in \mathcal{X}$.

Proposition 16. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be locally compact Hausdorff spaces, and let $\mathcal{X} \coloneqq \prod_{j=1}^d \mathcal{X}_j$. Assume that every open set in $\mathcal{X}$ is σ-compact. If $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{X}_\star)$, then $Q \in \mathsf{MNAR}_P$ if and only if*

$$P(f_{\max}) \geq Q(f)$$

for all $f \in C_{\mathrm{b}}(\mathcal{X}_\star)$.

Proof. Recall the definition of $\phi_{\mathcal{Z}} : C_{\mathrm{b}}(\mathcal{Z}) \to \mathcal{M}(\mathcal{Z})^*$ before Lemma 31. We endow $\mathcal{M}(\mathcal{Z})$ with the weak topology generated by $\phi_{\mathcal{Z}}\big(C_{\mathrm{b}}(\mathcal{Z})\big)$, for $\mathcal{Z} \in \{\mathcal{X}, \mathcal{X}_\star, \mathcal{X} \times 2^{[d]}\}$. This ensures that $\phi_{\mathcal{Z}}(g)$ is weakly continuous for every $g \in C_{\mathrm{b}}(\mathcal{Z})$.

Let $h : \mathcal{X} \times 2^{[d]} \to \mathcal{X}_\star$ be the continuous function defined by $h(x, S) \coloneqq x \circledast \mathbf{1}_S$. Then h induces a linear map $h_* : \mathcal{M}(\mathcal{X} \times 2^{[d]}) \to \mathcal{M}(\mathcal{X}_\star)$ given by $h_*(\mu)(B) \coloneqq \mu\big(h^{-1}(B)\big)$ (see Figure 5 below). Similarly, let $j : \mathcal{X} \times 2^{[d]} \to \mathcal{X}$ be the projection map $j(x, S) \coloneqq x$, and define its induced map $j_* : \mathcal{M}(\mathcal{X} \times 2^{[d]}) \to \mathcal{M}(\mathcal{X})$. We have $\{g \circ h : g \in C_{\mathrm{b}}(\mathcal{X}_\star)\} \subseteq C_{\mathrm{b}}(\mathcal{X} \times 2^{[d]})$ and similarly $\{g \circ j : g \in C_{\mathrm{b}}(\mathcal{X})\} \subseteq C_{\mathrm{b}}(\mathcal{X} \times 2^{[d]})$, we have by Schaefer (1971, Theorem IV.2.1) that both h_* and j_* are weakly continuous. By Lemma 33, the linear map $T = (h_*, j_*) : \mathcal{M}(\mathcal{X} \times 2^{[d]}) \to \mathcal{M}(\mathcal{X}_\star) \times \mathcal{M}(\mathcal{X})$ is continuous when we endow the image space with the product topology, which by Lemma 32 is the same as the weak topology on $\mathcal{M}(\mathcal{X}_\star) \times \mathcal{M}(\mathcal{X})$ generated by $\phi_{\mathcal{X}_\star}\big(C_{\mathrm{b}}(\mathcal{X}_\star)\big) \times \phi_{\mathcal{X}}\big(C_{\mathrm{b}}(\mathcal{X})\big)$.

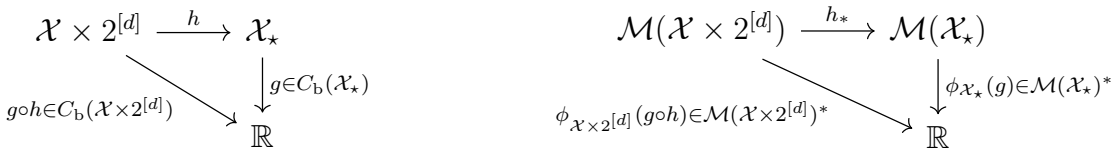


Figure 5: Schematic diagrams of various maps defined in the proof. The fact that the maps in the right panel commute follows from the fact that $h_*(\mu)(g) = \mu(g \circ h)$ for all $g \in C_{\mathrm{b}}(\mathcal{X}_\star)$.

Define K to be the convex cone $\mathcal{M}_+(\mathcal{X} \times 2^{[d]})$. We claim that $h_*(K) = \mathcal{M}_+(\mathcal{X}_\star)$. It is clear that $h_*(K) \subseteq \mathcal{M}_+(\mathcal{X}_\star)$ since for any $\mu \in K$ and any $g \in C_{\mathrm{b}}(\mathcal{X}_\star)$ such that $g \geq 0$, we have by Folland (1999, Proposition 10.1) that $h_*(\mu)(g) = \mu(g \circ h) \geq 0$. For the surjectivity, define $i : \mathcal{X}_\star \to \mathcal{X} \times 2^{[d]}$ by $i(z) := (z \odot \mathbf{1}_{\{j : z_j \neq \star\}}, \{j : z_j \neq \star\})$ and let $i_* : \mathcal{M}(\mathcal{X}_\star) \to \mathcal{M}(\mathcal{X} \times 2^{[d]})$ be its induced linear map. By the same argument as above, we have $i_*(\mathcal{M}_+(\mathcal{X}_\star)) \subseteq K$. For ν on $\mathcal{M}_+(\mathcal{X}_\star)$, we have $\nu = h_*\big(i_*(\nu)\big)$, and the surjectivity is established since $i_*(\nu) \in K$. Consequently,

$$h_*(K) = \mathcal{M}_+(\mathcal{X}_\star) = \bigcap_{g \in C_{\mathrm{b}}(\mathcal{X}_\star) : g \geq 0} \{\nu \in \mathcal{M}(\mathcal{X}_\star) : \nu(g) \geq 0\} = \bigcap_{g \in C_{\mathrm{b}}(\mathcal{X}_\star) : g \geq 0} \big(\phi_{\mathcal{X}_\star}(g)\big)^{-1}\big([0, \infty)\big)$$

is a weakly closed set. A similar argument shows that $j_*(K) = \mathcal{M}_+(\mathcal{X})$ is a weakly closed set. Thus, $T(K)$ is weakly closed set in $\mathcal{M}(\mathcal{X}_\star) \times \mathcal{M}(\mathcal{X})$, by Lemma 32.

By definition, $Q \in \mathsf{MNAR}_P$ if and only if there exists $\mu_0 \in K$ such that $T(\mu_0) = (Q, P)$. Therefore, by Lemma 31, we can apply the generalised Farkas' lemma (Lemma 15) to obtain that

$$\begin{aligned} Q \in \mathsf{MNAR}_P \iff & \bigcap_{\mu \in K} \big\{(f, g) \in C_{\mathrm{b}}(\mathcal{X}_\star) \times C_{\mathrm{b}}(\mathcal{X}) : h_*(\mu)(f) + j_*(\mu)(g) \geq 0\big\} \\ & \subseteq \big\{(f, g) \in C_{\mathrm{b}}(\mathcal{X}_\star) \times C_{\mathrm{b}}(\mathcal{X}) : Q(f) + P(g) \geq 0\big\}. \end{aligned}$$

Now, for any $(f, g) \in C_{\mathrm{b}}(\mathcal{X}_\star) \times C_{\mathrm{b}}(\mathcal{X})$ and $\mu \in K$, we have

$$h_*(\mu)(f) + j_*(\mu)(g) = \sum_{S \in 2^{[d]}} \int_{\mathcal{X}} \{(f \circ h)(x, S) + g(x)\} \, d\mu(x, S).$$

Hence, (f, g) satisfies $h_*(\mu)(f) + j_*(\mu)(g) \geq 0$ for all $\mu \in K$ if and only if $(f \circ h)(x, S) + g(x) \geq 0$ for all $x \in \mathcal{X}$ and $S \in 2^{[d]}$. Since $P(g)$ is increasing in g, it therefore suffices to check that for each $f \in C_{\mathrm{b}}(\mathcal{X}_\star)$ the function $g_f \in C_{\mathrm{b}}(\mathcal{X})$ given by $g_f(x) := -\min_{S \in 2^{[d]}} (f \circ h)(x, S) = -\min_{S \in 2^{[d]}} f_S(x_S)$ satisfies $Q(f) + P(g_f) \geq 0$. Substituting $f' := -f$, we have

$$\begin{aligned} Q \in \mathsf{MNAR}_P & \iff Q(f) + P(g_f) \geq 0 \text{ for all } f \in C_{\mathrm{b}}(\mathcal{X}_\star) \\ & \iff Q(f') \leq P(f'_{\max}) \text{ for all } f' \in C_{\mathrm{b}}(\mathcal{X}_\star) \end{aligned}$$

as desired. □

The proof of Theorem 1 now follows in a straightforward fashion.

Proof of Theorem 1. From the definition, $R \in \mathcal{R}(P, \epsilon, \pi)$ if and only if $Q \in \mathsf{MNAR}_P$, which by Proposition 16 occurs if and only if $Q \in \mathcal{P}(\mathcal{X}_\star)$ and $P(f_{\max}) \geq Q(f)$ for all $f \in C_{\mathrm{b}}(\mathcal{X}_\star)$. □

B.2 Proof of Proposition 2

Proof of Proposition 2. Suppose that $R \in \mathcal{R}(P, \epsilon, q)$ and let $A \in \mathcal{B}(\mathbb{R}_\star)$ be such that $\mu_\star(A) = 0$. Recall that if $X \sim P$, $B \sim \mathsf{Bern}(\epsilon)$, $\Omega^{(1)} \sim \mathsf{Bern}(q)$ and $\Omega^{(2)} \sim \mathsf{Bern}(q_2)$ for some $q_2 \in [0,1]$ with $B \perp\!\!\!\perp (X, \Omega^{(1)}, \Omega^{(2)})$ and $\Omega^{(1)} \perp\!\!\!\perp X$, then we can generate $Z \sim R$ via $Z := (1-B) \cdot (X \circledast \Omega^{(1)}) + B \cdot (X \circledast \Omega^{(2)})$. Then by definition of $\mu_\star$, we must have $A \in \mathcal{B}(\mathbb{R})$ and $\mu(A) = 0$. Since $P \ll \mu$, it follows that

$$0 = P(A) = \mathbb{P}(X \in A) \geq \mathbb{P}(Z \in A) = R(A).$$

This proves that $R \ll \mu_\star$. Now define $m : \mathbb{R} \to [0,1]$ by $m(x) := \mathbb{P}(\Omega^{(2)} = 1 \,|\, X = x)$. Then for any $A \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\mathbb{P}(Z \in A) &= (1-\epsilon) \cdot \mathbb{P}(X \in A,\, \Omega^{(1)} = 1) + \epsilon \cdot \mathbb{P}(X \in A,\, \Omega^{(2)} = 1) \\ &= q(1-\epsilon) \cdot \int_A p(x)\, \mathrm{d}\mu(x) + \epsilon \cdot \int_A m(x) p(x)\, \mathrm{d}\mu(x).\end{aligned}$$

Hence, $\frac{\mathrm{d}R}{\mathrm{d}\mu_\star}(x) = q(1-\epsilon) \cdot p(x) + \epsilon \cdot m(x)p(x)$ for $x \in \mathbb{R}$, and $\frac{\mathrm{d}R}{\mathrm{d}\mu_\star}(\star) = \mathbb{P}(Z = \star) = 1 - q(1-\epsilon) - \epsilon \int_{\mathbb{R}} m(x)p(x)\, \mathrm{d}\mu(x)$.

Conversely, suppose that $R \in \mathcal{P}(\mathbb{R}_\star)$ satisfies $R \ll \mu_\star$, and there exists a Borel measurable function $m : \mathbb{R} \to [0,1]$ such that $\mathrm{d}R/\mathrm{d}\mu_\star$ satisfies (11). Given $X \sim P$, define a random variable $\Omega^{(2)}$ taking values in $\{0,1\}$ such that $\mathbb{P}(\Omega^{(2)} = 1 \,|\, X = x) = m(x)$ for $x \in \mathbb{R}$. Let $B \sim \mathsf{Bern}(\epsilon)$ and $\Omega^{(1)} \sim \mathsf{Bern}(q)$ be such that $B \perp\!\!\!\perp (X, \Omega^{(1)}, \Omega^{(2)})$ and $\Omega^{(1)} \perp\!\!\!\perp X$. Then $Z := (1-B) \cdot (X \circledast \Omega^{(1)}) + B \cdot (X \circledast \Omega^{(2)}) \sim R$ and hence by construction $R \in \mathcal{R}(P, \epsilon, q)$.

This completes the proof, but we also provide an alternative proof of the converse statement using Theorem 1. Again suppose that $R \in \mathcal{P}(\mathbb{R}_\star)$ satisfies $R \ll \mu_\star$, and that $\mathrm{d}R/\mathrm{d}\mu_\star$ satisfies (11). Define $Q := \epsilon^{-1}\{R - (1-\epsilon)\mathsf{MCAR}_{(\pi,P)}\} \in \mathcal{M}(\mathbb{R}_\star)$ as in Theorem 1, and let $f = (f_{\{1\}}, f_\emptyset) \in C_{\mathrm{b}}(\mathbb{R}_\star)$. Note that by definition, $f_\emptyset \in \mathbb{R}$ is a constant and $f_{\max}(x) = f_{\{1\}}(x) \vee f_\emptyset$ for all $x \in \mathbb{R}$. Moreover, since $\mathsf{MCAR}_{(\pi,P)} \in \mathcal{R}(P, 0, \pi)$, we have by the argument in the direct part of the proof that $\mathsf{MCAR}_{(\pi,P)} \ll \mu_\star$ with $\frac{\mathrm{d}\mathsf{MCAR}_{(\pi,P)}}{\mathrm{d}\mu_\star}(x) = q \cdot p(x)$ for $x \in \mathbb{R}$ and $\frac{\mathrm{d}\mathsf{MCAR}_{(\pi,P)}}{\mathrm{d}\mu_\star}(\star) = 1 - q$, so

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu_\star}(z) = \begin{cases} m(z)p(z) & \text{if } z \in \mathbb{R} \\ 1 - \int_{\mathbb{R}} m(x)p(x)\, \mathrm{d}\mu(x) & \text{if } z = \star. \end{cases}$$

Hence $Q \in \mathcal{P}(\mathbb{R}_\star)$, and

$$\begin{aligned}P(f_{\max}) &= \int_{\mathbb{R}} \big(f_{\{1\}}(x) \vee f_\emptyset\big) p(x)\, \mathrm{d}\mu(x) \\ &\geq \int_{\mathbb{R}} \big\{m(x) f_{\{1\}}(x) + \big(1 - m(x)\big) f_\emptyset\big\} p(x)\, \mathrm{d}\mu(x) = Q(f),\end{aligned}$$

where the inequality follows from the fact that $\max(a,b)$ is at least as large as any convex combination of a and b, for $a, b \in \mathbb{R}$. We conclude that $R \in \mathcal{R}(P, \epsilon, q)$, by Theorem 1. □

C Proofs from Section 3

C.1 Proof of Theorem 3

We begin with some lemmas. Recalling the way that we can generate $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} P \in \mathcal{P}^{\text{arb}}\big(\theta_0, \Sigma, \epsilon, \pi\big)$ from Section 2.2.2, we let $\mathcal{I}_n \subseteq [n]$ denote the 'inliers', or the indices of the uncontaminated observations. Likewise, we denote by $\mathcal{O}_n \subseteq [n]$ the 'outliers', or the indices of the contaminated observations so that $\mathcal{I}_n \cup \mathcal{O}_n = [n]$.

Lemma 17. *Let $n, M \in \mathbb{N}$ be such that $n/M \geq 4$. Let $Z_1, \dots, Z_n \overset{\text{iid}}{\sim} P \in \mathcal{P}^{\text{arb}}\big(\theta_0, \Sigma, \epsilon, \pi\big)$, with corresponding observation patterns $\Omega_1, \dots, \Omega_n \in \{0,1\}^d$. Randomly select M disjoint sets $(B_m)_{m \in [M]} \subseteq [n]$ such that $|B_m| = \lfloor n/M \rfloor$, and for $\theta = (\theta_1, \dots, \theta_d)^\top \in \mathbb{R}^d$, $m \in [M]$ and $j \in [d]$, define*

$$\overline{\Omega}_{mj} \coloneqq \mathbb{1}_{\{\sum_{i \in B_m} \Omega_{ij} > 0\}} \qquad \text{and} \qquad \overline{Z}_{mj} \coloneqq \frac{\sum_{i \in B_m} \Omega_{ij} Z_{ij}}{\sum_{i \in B_m} \Omega_{ij}} \cdot \overline{\Omega}_{mj} + \theta_j \cdot (1 - \overline{\Omega}_{mj}). \tag{29}$$

Then, writing $\overline{Z}_m \coloneqq (\overline{Z}_{m1}, \dots, \overline{Z}_{md})^\top$, for all $m \in [M]$ such that $B_m \subseteq \mathcal{I}_n$, we have

$$\operatorname{tr}\Big(\mathbb{E}\big\{(\overline{Z}_m - \theta_0)(\overline{Z}_m - \theta_0)^\top\big\}\Big) \leq \frac{2}{|B_m|} \cdot \operatorname{tr}\big(\Sigma^{\text{IPW}}\big) + \frac{\|\theta - \theta_0\|_2^2}{e|B_m| q_{\min}} \tag{30a}$$

and

$$\big\|\mathbb{E}\big[(\overline{Z}_m - \theta_0)(\overline{Z}_m - \theta_0)^\top\big]\big\|_{\text{op}} \leq \frac{6}{|B_m|} \cdot \big\|\Sigma^{\text{IPW}}\big\|_{\text{op}} + \frac{\|\theta - \theta_0\|_2^2}{e|B_m| q_{\min}}. \tag{30b}$$

Proof. We compute the entries of the matrix $(\overline{Z}_m - \theta_0)(\overline{Z}_m - \theta_0)^\top$, beginning with those on the diagonal. For $j \in [d]$, let

$$A_{jj} \coloneqq \mathbb{E}\bigg(\frac{|B_m| q_j}{\sum_{i \in B_m} \Omega_{ij}} \cdot \mathbb{1}_{\{\sum_{i \in B_m} \Omega_{ij} > 0\}}\bigg) \leq 2, \tag{31}$$

where the inequality follows by the first part of Lemma 19. Further, let $E_{jj} \coloneqq (1 - q_j)^{|B_m|}$. For $i \in \mathcal{I}_n$, we can write $Z_i = X_i \circledast \Omega_i$, where $\mathbb{E}(X_i) = \theta_0$, $\operatorname{Cov}(X_i) = \Sigma$ and $X_i \perp\!\!\!\perp \Omega_i$. Hence, for any $m \in [M]$ such that $B_m \subseteq \mathcal{I}_n$ and any $j \in [d]$,

$$\begin{aligned}
\mathbb{E}\big\{(\overline{Z}_{mj} - \theta_{0,j})^2\big\} &= \mathbb{E}\big[\big\{\overline{\Omega}_{mj}(\overline{Z}_{mj} - \theta_{0,j}) + (1 - \overline{\Omega}_{mj})(\theta_j - \theta_{0,j})\big\}^2\big] \\
&= \mathbb{E}\big\{\big(\overline{\Omega}_{mj}(\overline{Z}_{mj} - \theta_{0,j})\big)^2\big\} + \mathbb{E}\big\{(1 - \overline{\Omega}_{mj})^2(\theta_j - \theta_{0,j})^2\big\} \\
&= \mathbb{E}\bigg\{\bigg(\frac{\sum_{i \in B_m} \Omega_{ij}(X_{ij} - \theta_{0,j})}{\sum_{i \in B_m} \Omega_{ij}} \cdot \mathbb{1}_{\{\sum_{i \in B_m} \Omega_{ij} > 0\}}\bigg)^2\bigg\} + \mathbb{P}(\overline{\Omega}_{mj} = 0)(\theta_j - \theta_{0,j})^2 \\
&= \mathbb{E}\bigg(\frac{\Sigma_{jj}}{\sum_{i \in B_m} \Omega_{ij}} \cdot \mathbb{1}_{\{\sum_{i \in B_m} \Omega_{ij} > 0\}}\bigg) + (1 - q_j)^{|B_m|}(\theta_j - \theta_{0,j})^2 \\
&= A_{jj} \cdot \frac{\Sigma_{jj}^{\text{IPW}}}{|B_m|} + E_{jj} \cdot (\theta_j - \theta_{0,j})^2.
\end{aligned} \tag{32}$$

Turning to the off-diagonal entries, for any $m \in [M]$ such that $B_m \subseteq \mathcal{I}_n$ and any distinct $j, k \in [d]$,

$$\begin{aligned}
&\mathbb{E}\big\{(\overline{Z}_{mj} - \theta_{0,j})(\overline{Z}_{mk} - \theta_{0,k})\big\} \\
&= \mathbb{E}\big[\big\{\overline{\Omega}_{mj}(\overline{Z}_{mj} - \theta_{0,j}) + (1 - \overline{\Omega}_{mj})(\theta_j - \theta_{0,j})\big\}\big\{\overline{\Omega}_{mk}(\overline{Z}_{mk} - \theta_{0,k}) + (1 - \overline{\Omega}_{mk})(\theta_k - \theta_{0,k})\big\}\big] \\
&= \mathbb{E}\big\{\big(\overline{\Omega}_{mj}(\overline{Z}_{mj} - \theta_{0,j})\big)\big(\overline{\Omega}_{mk}(\overline{Z}_{mk} - \theta_{0,k})\big)\big\} + \mathbb{E}\big\{(1 - \overline{\Omega}_{mj})(1 - \overline{\Omega}_{mk})(\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k})\big\},
\end{aligned}$$

where in the final step, the cross-terms vanish as $\mathbb{E}(X_i) = \theta_0$. Without loss of generality, we assume that $1 \in B_m$. For the first term, we first define

$$\begin{aligned}
A_{jk} &:= \mathbb{E}\left\{\frac{(|B_m|q_j)(|B_m|q_k)}{\big(1 + \sum_{i \in B_m \setminus \{1\}} \Omega_{ij}\big) \cdot \big(1 + \sum_{i \in B_m \setminus \{1\}} \Omega_{ik}\big)}\right\} \\
&\leq \mathbb{E}\left\{\frac{(|B_m|q_j)^2}{\big(1 + \sum_{i \in B_m \setminus \{1\}} \Omega_{ij}\big)^2}\right\}^{1/2} \mathbb{E}\left\{\frac{(|B_m|q_k)^2}{\big(1 + \sum_{i \in B_m \setminus \{1\}} \Omega_{ik}\big)^2}\right\}^{1/2} \leq \frac{2|B_m|^2}{(|B_m| - 1)^2} \leq 4, \quad (33)
\end{aligned}$$

where the first inequality follows from Cauchy–Schwarz and the second inequality follows from the second part of Lemma 19, and the final inequality uses the fact that $|B_m| \geq 4$. We then have

$$\begin{aligned}
&\mathbb{E}\big\{\overline{\Omega}_{mj}(\overline{Z}_{mj} - \theta_{0,j}) \cdot \overline{\Omega}_{mk}(\overline{Z}_{mk} - \theta_{0,k})\big\} \\
&\qquad = \Sigma_{jk} \cdot \mathbb{E}\left\{\frac{\big(\sum_{i \in B_m} \Omega_{ij}\Omega_{ik}\big)\overline{\Omega}_{mj}\overline{\Omega}_{mk}}{\big(\sum_{i \in B_m} \Omega_{ij}\big) \cdot \big(\sum_{i \in B_m} \Omega_{ik}\big)}\right\} \\
&\qquad = \Sigma_{jk} \cdot |B_m| \cdot \mathbb{E}\left\{\frac{\Omega_{1j}\Omega_{1k}}{\big(\sum_{i \in B_m} \Omega_{ij}\big) \cdot \big(\sum_{i \in B_m} \Omega_{ik}\big)}\right\} \\
&\qquad = \Sigma_{jk} \cdot |B_m| \cdot \mathbb{P}(\Omega_{1j} = \Omega_{1k} = 1) \cdot \mathbb{E}\left\{\frac{1}{\big(1 + \sum_{i \in B_m \setminus \{1\}} \Omega_{ij}\big) \cdot \big(1 + \sum_{i \in B_m \setminus \{1\}} \Omega_{ik}\big)}\right\} \\
&\qquad = A_{jk} \cdot \frac{\Sigma_{jk}q_{jk}}{|B_m|q_jq_k},
\end{aligned}$$

where the first equality follows from substituting the definition of $\overline{Z}_{mj}$ on the event $\{\overline{\Omega}_{mj} = 1\}$ (and similarly for k) and the second equality follows by symmetry. For the second term, we have

$$\begin{aligned}
\mathbb{E}\big\{(1 - \overline{\Omega}_{mj})(1 - \overline{\Omega}_{mk})(\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k})\big\} &= \mathbb{P}(\overline{\Omega}_{mj} = \overline{\Omega}_{mk} = 0) \cdot (\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}) \\
&= (1 - q_j - q_k + q_{jk})^{|B_m|} \cdot (\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}) \\
&=: E_{jk} \cdot (\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}).
\end{aligned}$$

Combining these two equalities then yields

$$\mathbb{E}\big[(\overline{Z}_{mj} - \theta_{0,j})(\overline{Z}_{mk} - \theta_{0,k})\big] = A_{jk} \cdot \frac{1}{|B_m|} \cdot \Sigma^{\mathrm{IPW}}_{jk} + E_{jk} \cdot (\theta_j - \theta_{0,j})(\theta_k - \theta_{0,k}). \quad (34)$$

Therefore, by (32) and (34),

$$\mathbb{E}\big\{(\overline{Z}_m - \theta_0)(\overline{Z}_m - \theta_0)^\top\big\} = \frac{1}{|B_m|} \cdot A \odot \Sigma^{\mathrm{IPW}} + E \odot \big\{(\theta - \theta_0)(\theta - \theta_0)^\top\big\},$$

where $A \coloneqq (A_{jk})_{j,k\in[d]}$ and $E \coloneqq (E_{jk})_{j,k\in[d]}$. The desired inequality (30a) then follows as

$$\begin{aligned}\operatorname{tr}\big(\mathbb{E}\big\{(\bar{Z}_m-\theta_0)(\bar{Z}_m-\theta_0)^\top\big\}\big) &= \frac{1}{|B_m|}\cdot\sum_{j=1}^d A_{jj}\Sigma^{\mathrm{IPW}}_{jj} + \sum_{j=1}^d E_{jj}(\theta_j-\theta_{0,j})^2 \\ &\le \frac{2}{|B_m|}\cdot\operatorname{tr}(\Sigma^{\mathrm{IPW}}) + \frac{\|\theta-\theta_0\|_2^2}{e|B_m|q_{\min}},\end{aligned}$$

where the inequality follows by (31) and Lemma 20.

For inequality (30b), we define a matrix $A' = (A'_{jk}) \in \mathbb{R}^{d\times d}$ by $A'_{jk} \coloneqq A_{jk}$ for $j \neq k$ and

$$A'_{jj} \coloneqq \mathbb{E}\left\{\frac{(|B_m|q_j)^2}{\big(1+\sum_{i\in B_m\setminus\{1\}}\Omega_{ij}\big)^2}\right\} \le 2, \tag{35}$$

where the inequality follows from the second part of Lemma 19 and the assumption that $|B_m| \ge 4$. Note that A' is a positive semi-definite matrix, as it is the expectation of a positive semi-definite matrix. Now

$$\begin{aligned}&\big\|\mathbb{E}\big\{(\bar{Z}_m-\theta_0)(\bar{Z}_m-\theta_0)^\top\big\}\big\|_{\mathrm{op}} \\ &\quad= \left\|\frac{1}{|B_m|}\cdot A\odot\Sigma^{\mathrm{IPW}} + E\odot\big\{(\theta-\theta_0)(\theta-\theta_0)^\top\big\}\right\|_{\mathrm{op}} \\ &\quad\le \frac{1}{|B_m|}\cdot\big\|A'\odot\Sigma^{\mathrm{IPW}}\big\|_{\mathrm{op}} + \frac{1}{|B_m|}\cdot\big\|(A-A')\odot\Sigma^{\mathrm{IPW}}\big\|_{\mathrm{op}} + \big\|E\odot\big\{(\theta-\theta_0)(\theta-\theta_0)^\top\big\}\big\|_{\mathrm{op}} \\ &\quad\overset{(i)}{\le} \frac{1}{|B_m|}\cdot\|A'\|_\infty\big\|\Sigma^{\mathrm{IPW}}\big\|_{\mathrm{op}} + \frac{1}{|B_m|}\cdot\big\|A-A'\big\|_\infty\big\|\Sigma^{\mathrm{IPW}}\big\|_{\mathrm{op}} + \big\|E\odot\big\{(\theta-\theta_0)(\theta-\theta_0)^\top\big\}\big\|_{\mathrm{op}} \\ &\quad\overset{(ii)}{\le} \frac{6}{|B_m|}\big\|\Sigma^{\mathrm{IPW}}\big\|_{\mathrm{op}} + \frac{\|\theta-\theta_0\|_2^2}{e|B_m|q_{\min}},\end{aligned}$$

where the first term in step (i) follows from Lemma 18 since A' is positive semidefinite, the second term in step (i) follows since $A - A'$ is diagonal, and step (ii) follows from the inequalities (31), (33) and (35), as well as Lemma 20. □

The following lemma can be deduced from Horn (1990, 3.1(e), p. 95), but for the convenience of the reader, we provide a short proof here.

Lemma 18. *Let $A, B \in \mathcal{S}^{d\times d}$ and further suppose that A is positive semi-definite. Then $\|A\odot B\|_{\mathrm{op}} \le \|A\|_\infty\|B\|_{\mathrm{op}}$.*

Proof. The proof largely follows that of Horn (1990). Since

$$\begin{pmatrix}A & A\\ A & A\end{pmatrix} \in \mathcal{S}^{2d\times 2d} \quad\text{and}\quad \begin{pmatrix}\|B\|_{\mathrm{op}}I_d & B\\ B & \|B\|_{\mathrm{op}}I_d\end{pmatrix} \in \mathcal{S}^{2d\times 2d}$$

are both positive semi-definite, by the Schur product theorem (Horn and Johnson, 2012, Theorem 7.5.3(a)), their Hadamard product

$$\begin{pmatrix}\|B\|_{\mathrm{op}}(I_d\odot A) & A\odot B\\ A\odot B & \|B\|_{\mathrm{op}}(I_d\odot A)\end{pmatrix}$$

is also positive semi-definite. Hence, for any $v \in \mathbb{R}^d$, we have

$$\begin{aligned} 0 &\le \begin{pmatrix} v^\top & -v^\top \end{pmatrix} \begin{pmatrix} \|B\|_{\mathrm{op}}(I_d \odot A) & A \odot B \\ A \odot B & \|B\|_{\mathrm{op}}(I_d \odot A) \end{pmatrix} \begin{pmatrix} v \\ -v \end{pmatrix} \\ &= 2\|B\|_{\mathrm{op}} v^\top (I \odot A) v - 2 v^\top (A \odot B) v, \end{aligned}$$

so $\|A \odot B\|_{\mathrm{op}} \le \|B\|_{\mathrm{op}} \|I \odot A\|_{\mathrm{op}} \le \|A\|_\infty \|B\|_{\mathrm{op}}$. □

Lemma 19. *Let $Y \sim \mathsf{Bin}(n, q)$ for some $n \in \mathbb{N}$ and $q \in (0, 1]$. Then*

$$\mathbb{E}\big(Y^{-1} \cdot \mathbb{1}_{\{Y>0\}}\big) \le \frac{2}{nq} \quad \text{and} \quad \mathbb{E}\big\{(Y+1)^{-2}\big\} \le \frac{2}{n^2 q^2}.$$

Proof. We have

$$\begin{aligned} \mathbb{E}\big\{(Y+1)^{-1}\big\} &= \sum_{y=0}^{n} (y+1)^{-1} \binom{n}{y} q^y (1-q)^{n-y} \\ &= \sum_{y=0}^{n} \frac{1}{q(n+1)} \binom{n+1}{y+1} q^{y+1} (1-q)^{n-y} \\ &= \frac{1}{q(n+1)} \sum_{y=1}^{n+1} \binom{n+1}{y} q^y (1-q)^{n+1-y} \le \frac{1}{nq}. \end{aligned}$$

The first inequality in the statement then follows as $y^{-1} \le 2(y+1)^{-1}$ for all $y \ge 1$. Similarly, we have

$$\begin{aligned} \mathbb{E}\big\{(Y+1)^{-1}(Y+2)^{-1}\big\} &= \sum_{y=0}^{n} (y+1)^{-1}(y+2)^{-1} \binom{n}{y} q^y (1-q)^{n-y} \\ &= \sum_{y=0}^{n} \frac{1}{q^2(n+1)(n+2)} \binom{n+2}{y+2} q^{y+2} (1-q)^{n-y} \le \frac{1}{n^2 q^2}. \end{aligned}$$

The second inequality in the statement then follows since $(y+1)^{-2} \le 2(y+1)^{-1}(y+2)^{-1}$ for all $y \ge 0$. □

Lemma 20. *Under the set up in the proof of Lemma 17, we have*

$$\|E\|_\infty \le \frac{1}{e|B_m| q_{\min}} \quad \text{and} \quad \|E \odot \big\{(\theta - \theta_0)(\theta - \theta_0)^\top\big\}\|_{\mathrm{op}} \le \frac{\|\theta - \theta_0\|_2^2}{e|B_m| q_{\min}}.$$

Proof. We will make use of the following inequality

$$(1-x)^k \le \frac{1}{ekx} \quad \text{for all } x \in (0, 1] \text{ and } k \in \mathbb{N}. \tag{36}$$

To see this, note that $k \log(1-x) \le -kx \le -\log(kx) - 1$. Hence, for each $j \in [d]$,

$$E_{jj} = (1 - q_j)^{|B_m|} \le \frac{1}{e|B_m| q_{\min}},$$

and for each $j, k \in [d]$,

$$E_{jk} = (1 - q_j - q_k + q_{jk})^{|B_m|} \leq \frac{1}{e|B_m|(q_j + q_k - q_{jk})} \leq \frac{1}{e|B_m|q_{\min}},$$

where the final inequality follows since $q_{jk} \leq q_k$, so that $q_j + q_k - q_{jk} \geq q_j \geq q_{\min}$. This establishes the first inequality.

For the second bound, we have

$$\big|\big[E \odot \big\{(\theta - \theta_0)(\theta - \theta_0)^\top\big\}\big]_{jk}\big| \leq \frac{1}{e|B_m|q_{\min}} \cdot |\theta_j - \theta_{0,j}| \cdot |\theta_k - \theta_{0,k}|.$$

Hence

$$\big\|E \odot \big\{(\theta - \theta_0)(\theta - \theta_0)^\top\big\}\big\|_{\mathrm{op}} \leq \frac{1}{e|B_m|q_{\min}} \big\||\theta - \theta_0| \cdot |\theta - \theta_0|^\top\big\|_{\mathrm{op}} = \frac{\|\theta - \theta_0\|_2^2}{e|B_m|q_{\min}},$$

where $|\theta - \theta_0|$ denotes the entrywise absolute value, and the inequality follows from the fact[6] that if $A = (A_{jk}), B = (B_{jk}) \in \mathcal{S}^{d\times d}$ are such that $|A_{jk}| \leq B_{jk}$ for all $j, k \in [d]$, then $\|A\|_{\mathrm{op}} \leq \|B\|_{\mathrm{op}}$. □

We require the following lemma, which is an analogue of Depersin and Lecué (2022b, Lemma 1). We state and prove it here as we require slight changes in several parts of the proof.

Lemma 21. *Let $Z_1, \ldots, Z_n \overset{\mathrm{iid}}{\sim} P \in \mathcal{P}^{\mathrm{arb}}\big(\theta_0, \Sigma, \epsilon, \pi\big)$. Let $n \geq 5$, $\delta \in [2e^{-(n-4)/720{,}000}, 1)$ and $\epsilon \in [0, 1/3000]$. Take $M \in \mathbb{N}$ such that*

$$\frac{n}{4} \geq M \geq 300\big(2\epsilon n + \log(2/\delta)\big) \vee 180{,}000 \log(2/\delta),$$

and for each $m \in [M]$, let $\bar{Z}_m$ be as in (29), for some $\theta \in \mathbb{R}^d$. Then, with probability at least $1 - \delta$, for all $v \in \mathbb{S}^{d-1}$, there are at least $99M/100$ blocks m such that

$$|v^\top(\bar{Z}_m - \theta_0)| \leq 4000\sqrt{\frac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{n}} + 100\sqrt{\frac{M\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}}{n}} + 30\|\theta - \theta_0\|_2\sqrt{\frac{M}{nq_{\min}}} =: r^{\mathrm{IPW}}.$$

Proof. By Lemma 38, we find that $\mathcal{E}_1 := \{|\mathcal{O}_n| \leq 2\epsilon n + \log(2/\delta)\}$ occurs with probability at least $1 - \delta/2$. Henceforth we will work on this event. By definition of M, there are at most $M/300$ blocks that are contaminated. Let $(\overline{Y}_m)_{m\in[M]}$ denote the uncontaminated block means so that $(\overline{Y}_m)_{m\in[M]}$ are independent and identically distributed, and at most $M/300$ of the vectors $(\overline{Y}_m)_{m\in[M]}$ and $(\bar{Z}_m)_{m\in[M]}$ are not equal. Let

$$\Gamma := |B_m| \cdot \mathbb{E}\big\{(\overline{Y}_m - \theta_0)(\overline{Y}_m - \theta_0)^\top\big\} \quad \text{and} \quad r' := 2800\sqrt{\frac{\mathrm{tr}(\Gamma)}{n}} + \sqrt{\frac{1200\|\Gamma\|_{\mathrm{op}}}{|B_m|}}.$$

[6] To see this, observe that $v^\top A v \leq |v|^\top |A||v| \leq |v|^\top B|v|$ for all $v \in \mathbb{R}^d$, where $|A|$ denotes the entrywise absolute value of A.

We begin by establishing that the claim of the lemma holds with r^{IPW} replaced by r'. To this end, it suffices to show that

$$\sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \mathbb{1}_{\{|v^\top(\overline{Y}_m-\theta_0)|>r'\}} \le \frac{M}{150}.$$

Define $\psi : \mathbb{R} \to [0,1]$ by

$$\psi(t) := \begin{cases} 0 & \text{if } t < 1/2 \\ 2(t-1/2) & \text{if } 1/2 \le t < 1 \\ 1 & \text{if } t \ge 1, \end{cases}$$

so that ψ is 2-Lipschitz and $\mathbb{1}_{\{t>1\}} \le \psi(t) \le \mathbb{1}_{\{t>1/2\}}$. Then

$$\begin{aligned}
\sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \mathbb{1}_{\{|v^\top(\overline{Y}_m-\theta_0)|>r'\}} &= \sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \Big\{ \mathbb{1}_{\{|v^\top(\overline{Y}_m-\theta_0)|>r'\}} - \mathbb{P}\big(|v^\top(\overline{Y}_m-\theta_0)|>r'/2\big) \\
&\qquad + \mathbb{P}\big(|v^\top(\overline{Y}_m-\theta_0)|>r'/2\big) \Big\} \\
&\le \sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \left\{ \psi\left(\frac{|v^\top(\overline{Y}_m-\theta_0)|}{r'}\right) - \mathbb{E}\psi\left(\frac{|v^\top(\overline{Y}_m-\theta_0)|}{r'}\right) \right\} \\
&\qquad + \sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \mathbb{P}\big(|v^\top(\overline{Y}_m-\theta_0)|>r'/2\big) =: A + B.
\end{aligned}$$

Towards bounding B, we apply Markov's inequality to obtain that for every $v \in \mathbb{S}^{d-1}$,

$$\mathbb{P}\big(|v^\top(\overline{Y}_m-\theta_0)|>r'/2\big) \le \frac{\mathbb{E}\big\{v^\top(\overline{Y}_m-\theta_0)(\overline{Y}_m-\theta_0)^\top v\big\}}{(r'/2)^2} \le \frac{\|\Gamma\|_{\mathrm{op}}}{|B_m|(r'/2)^2} \le \frac{1}{300},$$

where the final inequality follows from the definition of r'. Therefore, $B \le M/300$. For the first term, we have $A = (A - \mathbb{E}A) + \mathbb{E}A$. By the bounded differences inequality (Vershynin, 2018, Theorem 2.9.1) and our choice of M, we have with probability at least $1 - \delta/2$ that

$$A - \mathbb{E}A \le \sqrt{\frac{M\log(2/\delta)}{2}} \le \frac{M}{600}.$$

Now let $\varepsilon_1, \ldots, \varepsilon_M$ be independent Rademacher random variables. Then

$$\begin{aligned}
\mathbb{E}A &= \mathbb{E} \sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \left\{ \psi\left(\frac{|v^\top(\overline{Y}_m-\theta_0)|}{r'}\right) - \mathbb{E}\psi\left(\frac{|v^\top(\overline{Y}_m-\theta_0)|}{r'}\right) \right\} \\
&\overset{(i)}{\le} 2\mathbb{E} \sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \varepsilon_m \psi\left(\frac{|v^\top(\overline{Y}_m-\theta_0)|}{r'}\right) \\
&\overset{(ii)}{\le} 4\mathbb{E} \sup_{v\in\mathbb{S}^{d-1}} \sum_{m=1}^{M} \varepsilon_m \cdot \frac{v^\top(\overline{Y}_m-\theta_0)}{r'}
\end{aligned}$$

$$= \frac{4}{r'}\mathbb{E}\biggl\{\biggl\|\sum_{m=1}^{M}\varepsilon_m(\overline{Y}_m - \theta_0)\biggr\|_2\biggr\} \leq \frac{4}{r'}\mathbb{E}\biggl\{\biggl\|\sum_{m=1}^{M}\varepsilon_m(\overline{Y}_m - \theta_0)\biggr\|_2^2\biggr\}^{1/2}$$
$$= \frac{4}{r'}\sqrt{\frac{\mathrm{tr}(\Gamma)M}{|B_m|}} \leq \frac{M}{600},$$

where step (i) follows from (one-sided) symmetrisation (see, e.g., Boucheron, Lugosi and Massart, 2013, Lemma 11.4), step (ii) follows from Talagrand's contraction inequality (e.g., Vershynin, 2018, Exercise 6.7.7) since $\psi(|\cdot|)$ is 2-Lipschitz, and the final inequality follows from the definition of r'. The claim, with r^{IPW} replaced by r', follows by a union bound. The final result then follows from applying Lemma 17 to bound both $\mathrm{tr}(\Gamma)$ and $\|\Gamma\|_{\mathrm{op}}$, using the facts that $|B_m| \geq 3n/(4M)$ and $M \geq 180{,}000\log 2$, together with the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$. □

The following corollary is a variant of Depersin and Lecué (2022b, Theorem 2.1).

Corollary 22. *In the setting of Lemma 21, and with $\overline{Z}_1, \ldots, \overline{Z}_M$ defined as in Lemma 17 for some $\theta \in \mathbb{R}^d$, take*

$$\widetilde{\theta}_n \coloneqq \textsc{Robust\_Block\_Descent}(\overline{Z}_1, \ldots, \overline{Z}_M)$$

from Algorithm 4. Then with probability at least $1-\delta$,

$$\bigl\|\widetilde{\theta}_n - \theta_0\bigr\|_2 \leq 4\cdot 10^6\sqrt{\frac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{n}} + 9\cdot 10^4\sqrt{\frac{M\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}}{n}} + 3\cdot 10^4\|\theta-\theta_0\|_2\sqrt{\frac{M}{nq_{\min}}}.$$

Proof. Let $\mathcal{E}$ denote the event on which for all $A \in \mathcal{S}_+^{d\times d}$ with $\mathrm{tr}(A) = 1$, there are at least $9M/10$ blocks for which $\|A^{1/2}(\overline{Z}_m - \theta_0)\|_2 \leq 8r^{\mathrm{IPW}}$. We claim that $\mathbb{P}(\mathcal{E}) \geq 1 - e^{-M/180{,}000}$.

The proof of the claim follows that of Depersin and Lecué (2022b, Proposition 1), replacing their Lemma 1 with our Lemma 21 and their r with our r^{IPW}. Moreover, the remainder of the proof of Depersin and Lecué (2022b, Theorem 2.1) then carries over on our event $\mathcal{E}$. □

Equipped with these preliminary lemmas, we turn to the proof of Theorem 3.

Proof of Theorem 3. Let $S^{(1)}, \ldots, S^{(T)}$ be as in Algorithm 1. For $t \in [T]$, let $\epsilon^{(t)}$ denote the proportion of outliers, or contaminated observations, in the set $S^{(t)}$. Then, combining Lemma 38 and a union bound, we deduce that with probability at least $1-\delta/3$,

$$\max_{t\in[T]}\epsilon^{(t)} \leq 2\epsilon + \frac{\log(3T/\delta)}{\lfloor n/T\rfloor} \leq 2\epsilon + \frac{2T\log(3T/\delta)}{n} =: \epsilon'.$$

Henceforth we work on the event $\mathcal{E}_1 \coloneqq \bigl\{\max_{t\in[T]}\epsilon^{(t)} \leq \epsilon'\bigr\}$.

Let M be as in Algorithm 1. Then, if we take $A_2 = 1200$ and $A_3 = 180{,}000$ in Algorithm 1, it follows that for any $t \in [T]$,

$$\frac{n}{4T} \geq M \geq 300\biggl(\frac{2\epsilon^{(t)}n}{T} + \log(6T/\delta)\biggr) \vee 180{,}000\log(6T/\delta).$$

Moreover, we have both $\delta/(3T) \geq 2e^{-n/(720{,}000T)}$ and $\max_{t\in[T]} \epsilon^{(t)} \leq 1/3000$. We therefore apply Corollary 22 by first conditioning on $\widehat{\theta}^{(t)}$, with θ and n in that statement taken to be $\widehat{\theta}^{(t)}$ and $\lfloor n/T \rfloor$ respectively, and then taking a further expectation, to find that for each $t \in [T-1]$, there is an event $\mathcal{E}_2^{(t)}$ with $\mathbb{P}\big((\mathcal{E}_2^{(t)})^c\big) \leq \delta/(3T)$ such that on $\mathcal{E}_2^{(t)}$, we have

$$\begin{aligned}\|\widehat{\theta}^{(t+1)} - \theta_0\|_2 &\leq 5 \cdot 10^6 \bigg(\sqrt{\frac{T \operatorname{tr}(\Sigma^{\mathrm{IPW}})}{n}} + \sqrt{\frac{TM\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}}{n}} \bigg) + 4 \cdot 10^4 \|\widehat{\theta}^{(t)} - \theta_0\|_2 \sqrt{\frac{TM}{nq_{\min}}} \\ &=: a + b \cdot \|\widehat{\theta}^{(t)} - \theta_0\|_2.\end{aligned}$$

By the assumed lower bound on $q_{\min}$, we have $b \leq 1/2$. Moreover, since Algorithm 1 is initialised with the coordinate-wise trimmed mean and by assumption $q_{\min} \geq 8\log(6d/\delta)/n$, we combine Lugosi and Mendelson (2021, Theorem 1) with Lemma 38*(b)* and a union bound to obtain that there is an event $\mathcal{E}_3$ with $\mathbb{P}(\mathcal{E}_3^c) \leq \delta/3$ such that on $\mathcal{E}_3$, we have

$$\|\widehat{\theta}^{(1)} - \theta_0\|_2^2 \leq \frac{14000T \operatorname{tr}(\Sigma^{\mathrm{IPW}}) \log(24d/\delta)}{n} + 9216 \operatorname{tr}(\Sigma^{\mathrm{IPW}}) \cdot \epsilon'.$$

Thus, taking $A_1 = 10^{-9}$, we find that $2^{-(T-1)}\|\widehat{\theta}^{(1)} - \theta_0\|_2 \leq a$. Therefore,

$$\|\widehat{\theta}^{(T)} - \theta_0\|_2 \leq \sum_{\ell=0}^{T-2} ab^\ell + b^{T-1} \cdot \|\widehat{\theta}^{(1)} - \theta_0\|_2 \leq 2a + 2^{-(T-1)} \cdot \|\widehat{\theta}^{(1)} - \theta_0\|_2 \leq 3a.$$

Hence, $\mathbb{P}\big(\mathcal{E}_1 \cap \bigcap_{t=1}^{T-1} \mathcal{E}_2^{(t)} \cap \mathcal{E}_3\big) \geq 1 - \delta$ and on $\mathcal{E}_1 \cap \bigcap_{t=1}^{T-1} \mathcal{E}_2^{(t)} \cap \mathcal{E}_3$, we have that

$$\|\widehat{\theta}_n - \theta_0\|_2 \leq 3a = 1.5 \cdot 10^7 \cdot \bigg(\sqrt{\frac{T \operatorname{tr}(\Sigma^{\mathrm{IPW}})}{n}} + \sqrt{\frac{TM\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}}{n}} \bigg),$$

with probability at least $1-\delta$. The final bound then follows upon substituting the definition of M into the display above. □

C.2 Proof of Theorem 4

Proof of Theorem 4. First, note that when $\epsilon = 0$, by Proposition 47, we have

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \gtrsim \frac{\operatorname{tr}(\Sigma^{\mathrm{IPW}})}{n} + \frac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} \log(1/\delta)}{n}. \tag{37}$$

Now we consider the case $\epsilon \in \big(0, \frac{q_{\min}}{1+q_{\min}}\big)$. Without loss of generality, assume that $\Sigma^{\mathrm{IPW}}_{11} = \max_{j\in[d]} \Sigma^{\mathrm{IPW}}_{jj}$, and let $a := (\alpha+\alpha^2)/2$ and $b := (3\alpha+\alpha^2)/2$ for some $\alpha \in (0, 1/3]$ to be chosen later. Define random vectors $X^{(1)} = (X_1^{(1)}, \ldots, X_d^{(1)})^\top \sim P^{(1)}$ and $X^{(2)} = (X_1^{(2)}, \ldots, X_d^{(2)})^\top \sim P^{(2)}$ with independent components satisfying

$$X_1^{(1)} := \begin{cases} -\sqrt{\frac{\Sigma_{11}}{2\alpha}} & \text{with probability } \alpha \\ 0 & \text{with probability } 1-2\alpha, \\ \sqrt{\frac{\Sigma_{11}}{2\alpha}} & \text{with probability } \alpha \end{cases} \quad X_1^{(2)} := \begin{cases} -\sqrt{\frac{\Sigma_{11}}{2\alpha}} & \text{with probability } a \\ 0 & \text{with probability } 1-a-b \\ \sqrt{\frac{\Sigma_{11}}{2\alpha}} & \text{with probability } b, \end{cases}$$

and $X_j^{(1)} \overset{d}{=} X_j^{(2)} \sim \mathsf{N}(0, \Sigma_{jj})$ for $j \in \{2, \dots, d\}$. Then

$$\mathrm{Var}(X_1^{(2)}) = \frac{(a + b + 2ab - a^2 - b^2)\Sigma_{11}}{2\alpha} = \Sigma_{11}.$$

Thus $\mathrm{Cov}(X^{(1)}) = \mathrm{Cov}(X^{(2)}) = \Sigma$, so $P^{(\ell)} \in \mathcal{P}\big(\mathbb{E}(X^{(\ell)}), \Sigma\big)$ for $\ell \in \{1, 2\}$, and

$$\big\|\mathbb{E}(X^{(1)}) - \mathbb{E}(X^{(2)})\big\|_2^2 = \frac{\alpha\Sigma_{11}}{2}.$$

Moreover, by Lemma 45, we have

$$\begin{aligned}\mathrm{TV}\big(\mathsf{MCAR}_{(\pi,P^{(1)})}, \mathsf{MCAR}_{(\pi,P^{(2)})}\big) = \mathrm{ATV}(P^{(1)}, P^{(2)}, \pi) &= \sum_{S:1\in S} \pi(S) \cdot \mathrm{TV}\big(P_S^{(1)}, P_S^{(2)}\big) \\ &= q_1 \cdot \mathrm{TV}\big(P_1^{(1)}, P_1^{(2)}\big) = \frac{q_1}{2}\left(\frac{\alpha - \alpha^2}{2} + \alpha^2 + \frac{\alpha + \alpha^2}{2}\right) \le q_1\alpha.\end{aligned}$$

We then pick $\alpha = \epsilon/(3q_1) < 1/3$ since $\epsilon < q_{\min}$ so that

$$\mathrm{TV}\big(\mathsf{MCAR}_{(\pi,P^{(1)})}, \mathsf{MCAR}_{(\pi,P^{(2)})}\big) \le \epsilon \le \frac{\epsilon}{1-\epsilon}, \quad \text{and} \quad \big\|\mathbb{E}(X^{(1)}) - \mathbb{E}(X^{(2)})\big\|_2^2 = \frac{\epsilon\Sigma_{11}^{\mathrm{IPW}}}{6}.$$

Consequently, by Ma, Verchand and Samworth (2024, Theorem 4 and Lemma 25), we have

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \ge \frac{\big\|\mathbb{E}(X^{(1)}) - \mathbb{E}(X^{(2)})\big\|_2^2}{4} = \frac{\epsilon\Sigma_{11}^{\mathrm{IPW}}}{24}. \tag{38}$$

Combining (37) and (38) yields the desired result.

Next, we consider the case where $\epsilon \ge \frac{q_{\min}}{1+q_{\min}}$. Without loss of generality, assume that $q_1 = q_{\min}$. Let $\theta^{(1)} := (2t^{1/2}, 0, \dots, 0)^\top \in \mathbb{R}^d$ for some $t > 0$ and let $\theta^{(2)} := 0 \in \mathbb{R}^d$. Writing $P^{(1)} := \mathsf{N}(\theta^{(1)}, \Sigma)$ and $P^{(2)} := \mathsf{N}(\theta^{(2)}, \Sigma)$, we have by Lemma 45 that

$$\begin{aligned}\mathrm{TV}\big(\mathsf{MCAR}_{(\pi,P^{(1)})}, \mathsf{MCAR}_{(\pi,P^{(2)})}\big) = \mathrm{ATV}(P^{(1)}, P^{(2)}; \pi) &= \sum_{S:1\in S} \pi(S) \cdot \mathrm{TV}(P_S^{(1)}, P_S^{(2)}) \\ &= q_1 \mathrm{TV}\big(P_{\{1\}}^{(1)}, P_{\{1\}}^{(2)}\big) \le q_1 \le \frac{\epsilon}{1-\epsilon}.\end{aligned}$$

Hence, by Ma, Verchand and Samworth (2024, Theorem 4 and Lemma 25), we see that $\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \ge t$. Since $t > 0$ was arbitrary, the result follows. □

C.3 Univariate arbitrary contamination lower bounds

The lower bounds in Proposition 23 are presented primarily to ensure the completeness of Table 1. Corresponding upper bounds are attained by the median in the Gaussian case (Chen, Gao and Ren, 2018, Theorem 2.1), and a trimmed mean (Lugosi and Mendelson, 2021, Theorem 1 and the subsequent remark) in the sub-Gaussian case, in both cases applied to the observed data.

Proposition 23. *Let $\epsilon \in [0, 1)$, $q \in (0, 1]$, $\sigma > 0$, $\delta \in (0, 1/4]$ and $\kappa := \frac{\epsilon}{q(1-\epsilon)}$.*

(a) Let $\Theta := \mathbb{R}$ *and let* $\mathcal{P}_\theta := \{P_0^{\otimes n} : P_0 \in \mathcal{P}^{\mathrm{arb}}(\mathsf{N}(\theta, \sigma^2), \epsilon, q)\}$ *for* $\theta \in \Theta$. *Then*

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \begin{cases} \gtrsim \dfrac{\sigma^2 \log(1/\delta)}{nq(1-\epsilon)} + \sigma^2\kappa^2 & \text{if } \epsilon < \frac{q}{1+q} \\ = \infty & \text{if } \epsilon \geq \frac{q}{1+q}. \end{cases}$$

(b) Let $\Theta := \mathbb{R}$ *and let* $\mathcal{P}_\theta := \{P_0^{\otimes n} : P_0 \in \mathcal{P}^{\mathrm{arb}}(P, \epsilon, q),\ P \in \mathcal{P}_{\psi_2}(\theta, \sigma^2)\}$ *for* $\theta \in \Theta$. *Then*

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \begin{cases} \gtrsim \dfrac{\sigma^2 \log(1/\delta)}{nq(1-\epsilon)} + \sigma^2\kappa^2 \log(1/\kappa) & \text{if } \epsilon < \frac{q}{1+q} \\ = \infty & \text{if } \epsilon \geq \frac{q}{1+q}. \end{cases}$$

Proof. (a) First consider the case where $\epsilon < \frac{q}{1+q}$. Let $X_1 \sim \mathsf{N}(0, \sigma^2) =: P_1$ and $X_2 \sim \mathsf{N}(2\sigma\kappa, \sigma^2) =: P_2$. By Pinsker's inequality, $\mathrm{TV}(P_1, P_2) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P_1, P_2)} = \kappa$, so that by Lemma 45,

$$\mathrm{TV}\big(\mathsf{MCAR}_{(q,P_1)}, \mathsf{MCAR}_{(q,P_2)}\big) = q \cdot \mathrm{TV}(P_1, P_2) \leq \frac{\epsilon}{1-\epsilon}.$$

Hence, by Ma, Verchand and Samworth (2024, Lemma 25), we have

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \frac{(\mathbb{E}X_1 - \mathbb{E}X_2)^2}{4} = \sigma^2\kappa^2. \tag{39}$$

Further, by choosing the contamination distribution $Q \in \mathcal{P}(\mathbb{R}_\star)$ such that $Q\big(\{\star\}\big) = 1$ and applying Proposition 46*(a)*, we deduce that

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \gtrsim \frac{\sigma^2 \log(1/\delta)}{nq(1-\epsilon)}. \tag{40}$$

Combining (39) and (40) yields the lower bound for $\epsilon < \frac{q}{1+q}$.

Next consider the case where $\epsilon \geq \frac{q}{1+q}$. Let $a > 0$, $X_1 \sim P_1 := \mathsf{N}(0, \sigma^2)$ and $X_2 \sim P_2 := \mathsf{N}(a\sigma, \sigma^2)$. By Lemma 45,

$$\mathrm{TV}\big(\mathsf{MCAR}_{(q,P_1)}, \mathsf{MCAR}_{(q,P_2)}\big) = q \cdot \mathrm{TV}(P_1, P_2) \leq q \leq \frac{\epsilon}{1-\epsilon}.$$

Hence, by Ma, Verchand and Samworth (2024, Lemma 25), we have

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \frac{(\mathbb{E}X_1 - \mathbb{E}X_2)^2}{4} = \frac{\sigma^2 a^2}{4}.$$

Since this holds for all $a > 0$, we deduce that $\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) = \infty$ in this case.

(b) Let $c_1 > 0$ be a universal constant that will be specified later. Define $P_1 \in \mathcal{P}(\mathbb{R})$ by $P_1\big((t, \infty)\big) := e^{-t^2/(c_1\sigma)^2}$ for $t \geq 0$. Define $P_2 \in \mathcal{P}(\mathbb{R})$ by

$$P_2(\{0\}) := \kappa, \quad P_2\big((t, \infty)\big) := \begin{cases} e^{-t^2/(c_1\sigma)^2} & \text{if } 0 \leq t \leq c_1\sigma\sqrt{\log(1/\kappa)} \\ 0 & \text{if } t > c_1\sigma\sqrt{\log(1/\kappa)}. \end{cases}$$

Let $X_1 \sim P_1$ and $X_2 \sim P_2$. Since $\mathbb{P}(|X_\ell| \geq t) \leq e^{-t^2/(c_1\sigma)^2}$ for $t \geq 0$ and $\ell \in \{1,2\}$, we have by Vershynin (2018, Proposition 2.5.2) that $\|X_\ell\|_{\psi_2} \leq c_1 C_2 \sigma$ for $\ell \in \{1,2\}$, where $C_2 > 0$ is a universal constant. Thus by Vershynin (2018, Lemma 2.6.8), there exists a universal constant $C_3 > 0$ such that $\|X_\ell - \mathbb{E}X_\ell\|_{\psi_2} \leq c_1 C_2 C_3 \sigma$. Hence, taking $c_1 \coloneqq (C_2 C_3)^{-1}$, we have $P_\ell \in \mathcal{P}_{\psi_2}\big(\mathbb{E}(X_\ell), \sigma^2\big)$ for $\ell \in \{1,2\}$. Moreover, $\mathrm{TV}(P_1, P_2) = P_2(\{0\}) = \kappa$, so that by Lemma 45,

$$\mathrm{TV}\big(\mathsf{MCAR}_{(q,P_1)}, \mathsf{MCAR}_{(q,P_2)}\big) = q \cdot \mathrm{TV}(P_1, P_2) = \frac{\epsilon}{1-\epsilon}.$$

Now, integrating by parts yields

$$\begin{aligned}\mathbb{E}X_1 - \mathbb{E}X_2 &= \int_{c_1\sigma\sqrt{\log(1/\kappa)}}^{\infty} x \cdot \frac{2x}{(c_1\sigma)^2} e^{-x^2/(c_1\sigma)^2}\,\mathrm{d}x \\ &= \kappa \cdot c_1\sigma\sqrt{\log(1/\kappa)} + \int_{c_1\sigma\sqrt{\log(1/\kappa)}}^{\infty} e^{-x^2/(c_1\sigma)^2}\,\mathrm{d}x \gtrsim \sigma\kappa\sqrt{\log(1/\kappa)}.\end{aligned}$$

Hence, by Ma, Verchand and Samworth (2024, Lemma 25),

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \frac{(\mathbb{E}X_1 - \mathbb{E}X_2)^2}{4} \gtrsim \sigma^2\kappa^2 \log(1/\kappa). \tag{41}$$

By (41) and applying Proposition 46*(a)* with contamination distribution $Q \in \mathcal{P}(\mathbb{R}_\star)$ satisfying $Q(\{\star\}) = 1$ as in *(a)*, we obtain the desired lower bound for $\epsilon < \frac{q}{1+q}$. Finally, the lower bound for $\epsilon \geq \frac{q}{1+q}$ follows from part *(a)*. □

D Proofs from Section 4

D.1 Proofs from Section 4.1

D.1.1 Proof of Theorem 5

Proof of Theorem 5. Given $R \in \mathcal{R}(\theta_0)$, there exists a random vector $(X_0, \Omega_0^{(1)}, \Omega_0^{(2)}, W_0)$ taking values in $\mathbb{R} \times \{0,1\}^3$ such that $W_0 \perp\!\!\!\perp (X_0, \Omega_0^{(1)}, \Omega_0^{(2)})$, $W_0 \sim \mathsf{Ber}(\epsilon)$, $\Omega_0^{(1)} \perp\!\!\!\perp (X_0, \Omega_0^{(2)})$, $\Omega_0^{(1)} \sim \mathsf{Ber}(q)$, $X_0 \sim \mathsf{N}(\theta_0, \sigma^2)$ and

$$R = \mathsf{Law}\big((1 - W_0) \cdot X_0 \circledast \Omega_0^{(1)} + W_0 \cdot X_0 \circledast \Omega_0^{(2)}\big).$$

Note that if $Z_0 \coloneqq (1-W_0)\cdot X_0 \circledast \Omega_0^{(1)} + W_0 \cdot X_0 \circledast \Omega_0^{(2)}$, then $Z_0 \mid \{W_0 = 0\} \sim \mathsf{MCAR}_{(\mathsf{N}(\theta_0,\sigma^2),q)}$. We then generate $(X_i, \Omega_i^{(1)}, \Omega_i^{(2)}, W_i)_{i=1}^n \overset{\mathrm{iid}}{\sim} \mathsf{Law}(X_0, \Omega_0^{(1)}, \Omega_0^{(2)}, W_0)$, and set $Z_i \coloneqq (1 - W_i) \cdot X_i \circledast \Omega_i^{(1)} + W_i \cdot X_i \circledast \Omega_i^{(2)}$ for $i \in [n]$, so that $Z_1, \ldots, Z_n \overset{\mathrm{iid}}{\sim} R$.

Now define the inliers as $\mathcal{I} \coloneqq \{i \in [n] : W_i = 0\}$, the outliers as $\mathcal{O} \coloneqq \{i \in [n] : W_i = 1\}$, and the observed indices as $\mathcal{D} \coloneqq \{i \in [n] : Z_i \neq \star\}$. Equipped with this notation, we note the following pair of structural properties

$$\max_{i \in \mathcal{I}\cap\mathcal{D}} X_i \leq \max_{i\in\mathcal{D}} Z_i \leq \max_{i \in (\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}} X_i \quad \text{and} \quad \min_{i \in (\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}} X_i \leq \min_{i\in\mathcal{D}} Z_i \leq \min_{i \in \mathcal{I}\cap\mathcal{D}} X_i.$$

We deduce the sandwich relation

$$\frac{1}{2}\cdot\Big(\max_{i\in\mathcal{I}\cap\mathcal{D}} X_i + \min_{i\in(\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}} X_i\Big) \le \widehat{\theta}^{\mathrm{AE}} \le \frac{1}{2}\cdot\Big(\max_{i\in(\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}} X_i + \min_{i\in\mathcal{I}\cap\mathcal{D}} X_i\Big). \tag{42}$$

Now $X_1,\dots,X_n$ and $\mathcal{I}\cap\mathcal{D}$ are independent, and similarly $X_1,\dots,X_n$ and $\mathcal{O}$ are independent, so $(X_i)_{i\in\mathcal{I}\cap\mathcal{D}}|(\mathcal{I}\cap\mathcal{D}) \overset{\text{iid}}{\sim} \mathsf{N}(\theta_0,\sigma^2)$ and $(X_i)_{i\in(\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}}|\big((\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}\big) \overset{\text{iid}}{\sim} \mathsf{N}(\theta_0,\sigma^2)$. We let $N_1 \coloneqq |\mathcal{I}\cap\mathcal{D}|$ and $N_2 \coloneqq |(\mathcal{I}\cap\mathcal{D})\cup\mathcal{O}|$, and define

$$B_\ell \coloneqq \sigma\sqrt{2\log N_\ell} + \frac{\sigma}{2}\cdot\frac{\log\log N_\ell + \log(4\pi)}{\sqrt{2\log N_\ell}},$$

for $\ell\in\{1,2\}$. Let $\mathcal{E}_1 \coloneqq \{N_1 \ge nq(1-\epsilon)/2\}$ and $\mathcal{E}_2 \coloneqq \{|\mathcal{O}| \le 3n\epsilon\}$. By Tanguy (2015, Theorem 3), there exists a universal constant $C_1'>0$ such that for $\ell\in\{1,2\}$ and $\delta>0$,

$$\mathbb{P}\bigg(\bigg\{\Big|\max_{i\in[N_\ell]} X_i - \theta - B_\ell\Big| \ge \frac{C_1'\sigma\log(8/\delta)}{\log^{1/2} N_\ell}\bigg\}\bigcap\mathcal{E}_1 \;\bigg|\; N_\ell\bigg) \le \frac{\delta}{4}$$

and

$$\mathbb{P}\bigg(\bigg\{\Big|\min_{i\in[N_\ell]} X_i - \theta + B_\ell\Big| \ge \frac{C_1'\sigma\log(8/\delta)}{\log^{1/2} N_\ell}\bigg\}\bigcap\mathcal{E}_1 \;\bigg|\; N_\ell\bigg) \le \frac{\delta}{4}.$$

Combining these inequalities with the sandwich relation (42) yields

$$\mathbb{P}\bigg(\bigg\{\big|\widehat{\theta}_n^{\mathrm{AE}} - \theta_0\big| \ge B_2 - B_1 + \frac{2C_1'\sigma\log(8/\delta)}{\log^{1/2} N_1}\bigg\}\bigcap\mathcal{E}_1 \;\bigg|\; N_1, N_2\bigg) \le \frac{2\delta}{3}. \tag{43}$$

Using the inequality $\sqrt{a}-\sqrt{b} \le (a-b)/\sqrt{b}$ for $0<b<a$ and the fact that $x\mapsto \frac{\log\log x+\log(4\pi)}{\log^{1/2} x}$ is decreasing for $x \ge \exp\big(\frac{e^2}{4\pi}\big)$, we deduce that on $\mathcal{E}_1\cap\mathcal{E}_2$,

$$\begin{aligned} B_2 - B_1 \le \frac{2\sigma\log(N_2/N_1)}{\log^{1/2} N_1} &\le \frac{2\sigma\log\big(1+3n\epsilon/N_1\big)}{\log^{1/2} N_1} \\ &\le \frac{2\sigma\log\big(1+\frac{6\epsilon}{q(1-\epsilon)}\big)}{\log^{1/2}\big(nq(1-\epsilon)/2\big)} \le \frac{3\sigma\log\big(1+\frac{6\epsilon}{q(1-\epsilon)}\big)}{\log^{1/2}\big(nq(1-\epsilon)\big)}. \end{aligned} \tag{44}$$

Now, we first assume that $\epsilon \ge n^{-1}\log(4/\delta)$. By Lemma 38, we have $\mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2) \ge 1-\delta/2$, since by assumption, $nq(1-\epsilon) \ge 8\log(4/\delta)$. Moreover, combining the inequalities (43) and (44) yields that on $\mathcal{E}_1\cap\mathcal{E}_2$,

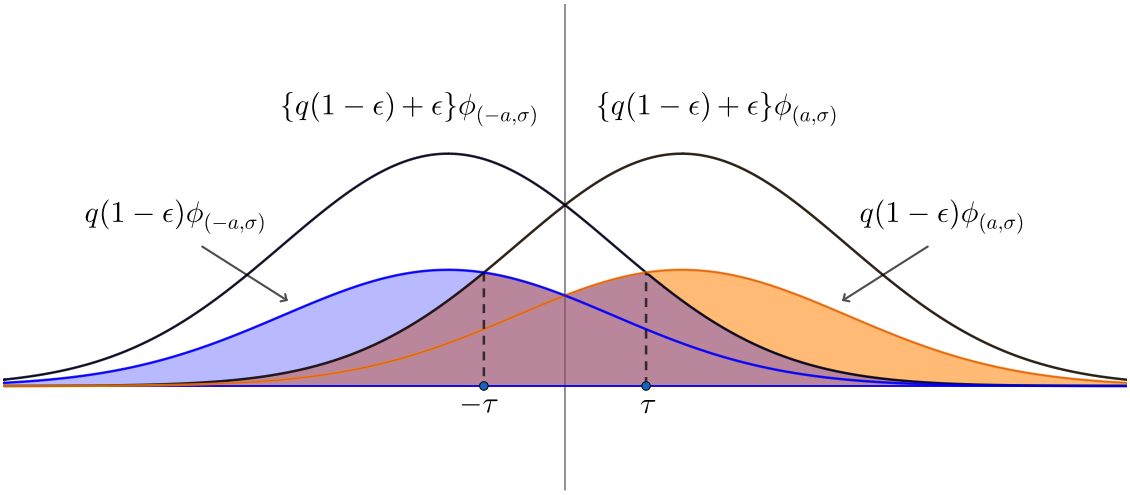
$$\big|\widehat{\theta}_n^{\mathrm{AE}} - \theta_0\big| \le \frac{3\sigma\log\big(1+\frac{6\epsilon}{q(1-\epsilon)}\big)}{\log^{1/2}\big(nq(1-\epsilon)\big)} + \frac{2C_1'\sigma\log(8/\delta)}{\log^{1/2} N_1} \le C_1\sigma\cdot\frac{\log\big(1+\frac{6\epsilon}{q(1-\epsilon)}\big)+\log(8/\delta)}{\log^{1/2}\big(nq(1-\epsilon)\big)}, \tag{45}$$

where $C_1 \coloneqq 3(1+C_1')$. Hence, (45) holds with probability at least $1-\delta$ when $\epsilon \ge n^{-1}\log(4/\delta)$. Finally, consider the case in which $\epsilon < n^{-1}\log(4/\delta)$. Then, since $\mathcal{R}\big(\mathsf{N}(\theta_0,\sigma^2),\epsilon,q\big) \subseteq \mathcal{R}\big(\mathsf{N}(\theta_0,\sigma^2), n^{-1}\log(4/\delta), q\big)$, we have by (45) that

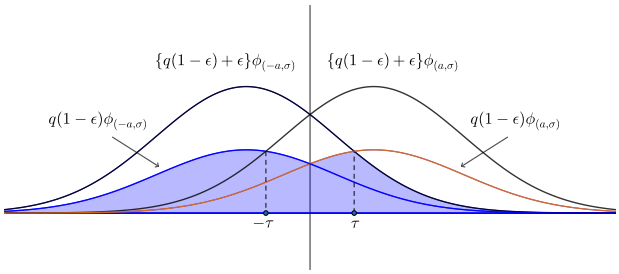
$$\begin{aligned} \big|\widehat{\theta}_n^{\mathrm{AE}} - \theta_0\big| &\le C_1\sigma\cdot\frac{\log\big(1+\frac{6\log(4/\delta)}{nq(1-\epsilon)}\big)+\log(8/\delta)}{\log^{1/2}\big(nq(1-\epsilon)\big)} \le C_1\sigma\cdot\frac{\frac{6\log(4/\delta)}{nq(1-\epsilon)}+\log(8/\delta)}{\log^{1/2}\big(nq(1-\epsilon)\big)} \\ &\le C_1\sigma\cdot\frac{2\log(8/\delta)}{\log^{1/2}\big(nq(1-\epsilon)\big)} \le 2C_1\sigma\cdot\frac{\log\big(1+\frac{6\epsilon}{q(1-\epsilon)}\big)+\log(8/\delta)}{\log^{1/2}\big(nq(1-\epsilon)\big)}, \end{aligned}$$

with probability at least $1-\delta$. □

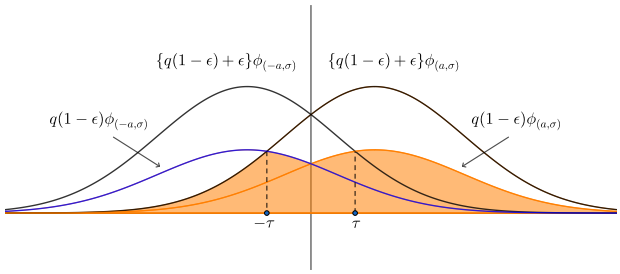
D.1.2 Proof of Theorem 6



(a) The two black curves are $\{q(1-\epsilon)+\epsilon\}\phi_{(-a,\sigma)}$ and $\{q(1-\epsilon)+\epsilon\}\phi_{(a,\sigma)}$ respectively, as labelled in the figure. The blue curve is $q(1-\epsilon)\phi_{(-a,\sigma)}$, and the orange curve is $q(1-\epsilon)\phi_{(a,\sigma)}$.



(b) The curve above the blue region illustrates the function f_1 in (47).



(c) The curve above the orange region illustrates the function f_2 in (48).

Figure 6: Construction of the lower bound in Theorem 6.

Proof of Theorem 6. Consider the construction illustrated in Figure 6. For $a > 0$ to be specified later, let

$$\tau := \frac{\sigma^2}{2a} \cdot \log\left(1 + \frac{\epsilon}{q(1-\epsilon)}\right) \tag{46}$$

denote the unique point in $\mathbb{R}$ where $\{q(1-\epsilon)+\epsilon\}\phi_{(-a,\sigma)}(\tau) = q(1-\epsilon)\phi_{(a,\sigma)}(\tau)$. Next, define the function $f_1 : \mathbb{R} \to \mathbb{R}$ as

$$f_1(x) := \begin{cases} q(1-\epsilon)\phi_{(-a,\sigma)}(x) & \text{if } x \le 0 \\ q(1-\epsilon)\phi_{(a,\sigma)}(x) & \text{if } 0 < x \le \tau \\ \big\{q(1-\epsilon)+\epsilon\big\} \cdot \phi_{(-a,\sigma)}(x) & \text{if } x > \tau. \end{cases} \tag{47}$$

Similarly, we note that $-\tau$ is the unique point satisfying $\{q(1-\epsilon)+\epsilon\}\phi_{(a,\sigma)}(-\tau) = q(1-$

$\epsilon)\phi_{(-a,\sigma)}(-\tau)$ and define the function $f_2 : \mathbb{R} \to \mathbb{R}$ as

$$f_2(x) \coloneqq \begin{cases} \{q(1-\epsilon)+\epsilon\} \cdot \phi_{(a,\sigma)}(x) & \text{if } x \le -\tau \\ q(1-\epsilon)\phi_{(-a,\sigma)}(x) & \text{if } -\tau < x \le 0 \\ q(1-\epsilon)\phi_{(a,\sigma)}(x) & \text{if } x > 0. \end{cases} \tag{48}$$

Note that $\int_{\mathbb{R}} f_\ell(x)\,\mathrm{d}x \le q(1-\epsilon)+\epsilon \le 1$ for $\ell \in \{1,2\}$, so we may construct $P_1, P_2 \in \mathcal{P}(\mathbb{R}_\star)$ with Radon–Nikodym derivatives

$$\frac{\mathrm{d}P_\ell}{\mathrm{d}\lambda_\star}(z) \coloneqq f_\ell(z)\mathbb{1}_{\{z\in\mathbb{R}\}} + \left(1 - \int_{\mathbb{R}} f_\ell(x)\,\mathrm{d}x\right)\mathbb{1}_{\{z=\star\}} \quad \text{for} \quad \ell \in \{1,2\},$$

where $\lambda_\star$ denotes the extension of the Lebesgue measure to $\mathbb{R}_\star$ as defined in Section 1.2. Then, by Proposition 2, $P_1 \in \mathcal{R}\big(\mathsf{N}(-a,\sigma^2),\epsilon,q\big)$ and $P_2 \in \mathcal{R}\big(\mathsf{N}(a,\sigma^2),\epsilon,q\big)$. Since $P_1(\{\star\}) = P_2(\{\star\})$ and $f_1(x) = f_2(x)$ for $x \in [-\tau,\tau]$, we compute

$$\begin{aligned}
\mathrm{KL}(P_1,P_2) &= \int_{-\infty}^{-\tau} q(1-\epsilon)\phi_{(-a,\sigma)}(x)\log\left(\frac{q(1-\epsilon)\phi_{(-a,\sigma)}(x)}{\{q(1-\epsilon)+\epsilon\}\phi_{(a,\sigma)}(x)}\right)\mathrm{d}x \\
&\quad + \int_{\tau}^{\infty} \{q(1-\epsilon)+\epsilon\}\phi_{(-a,\sigma)}(x)\log\left(\frac{\{q(1-\epsilon)+\epsilon\}\phi_{(-a,\sigma)}(x)}{q(1-\epsilon)\phi_{(a,\sigma)}(x)}\right)\mathrm{d}x \\
&= q(1-\epsilon)\left\{\frac{2a^2}{\sigma^2} - \log\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\right\}\{1-\Phi_{(0,\sigma)}(\tau-a)\} \\
&\quad + \{q(1-\epsilon)+\epsilon\}\left\{\frac{2a^2}{\sigma^2} + \log\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\right\}\{1-\Phi_{(0,\sigma)}(\tau+a)\} \\
&\quad + 2a\big[q(1-\epsilon)\phi_{(0,\sigma)}(\tau-a) - \{q(1-\epsilon)+\epsilon\}\phi_{(0,\sigma)}(\tau+a)\big] \\
&= \frac{2aq(1-\epsilon)}{\sigma^2}(a-\tau)\{1-\Phi_{(0,\sigma)}(\tau-a)\} + \frac{2a\{q(1-\epsilon)+\epsilon\}}{\sigma^2}(a+\tau)\{1-\Phi_{(0,\sigma)}(\tau+a)\} \\
&\quad + 2a\big[q(1-\epsilon)\phi_{(0,\sigma)}(\tau-a) - \{q(1-\epsilon)+\epsilon\}\phi_{(0,\sigma)}(\tau+a)\big].
\end{aligned}$$

Next, set

$$a \coloneqq \frac{\sigma}{4}\cdot\log\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\cdot\log^{-1/2}\big(nq(1-\epsilon)\big) > 0,$$

so that by substituting this definition into (46), we obtain

$$\tau = 2\sigma\log^{1/2}\big(nq(1-\epsilon)\big) \quad \text{and} \quad a \le \frac{\tau}{8},$$

where the inequality follows from our assumption (17). Hence, by the Mills ratio bound $1-\Phi_{(0,\sigma)}(x) \le \sigma^2\phi_{(0,\sigma)}(x)/x$ for $x > 0$, we have

$$\begin{aligned}
\mathrm{KL}(P_1,P_2) &\le 2a\{q(1-\epsilon)+\epsilon\}\phi_{(0,\sigma)}(\tau+a) \\
&\qquad + 2a\big[q(1-\epsilon)\phi_{(0,\sigma)}(\tau-a) - \{q(1-\epsilon)+\epsilon\}\phi_{(0,\sigma)}(\tau+a)\big] \\
&= 2aq(1-\epsilon)\phi_{(0,\sigma)}(\tau-a) \\
&= \frac{\sigma}{2}\cdot\log\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\cdot\log^{-1/2}\big(nq(1-\epsilon)\big)\cdot q(1-\epsilon)\cdot\phi_{(0,\sigma)}(\tau-a)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\sigma}{2} \cdot \log^{1/2}\bigl(nq(1-\epsilon)\bigr) \cdot q(1-\epsilon) \cdot \phi_{(0,\sigma)}(7\tau/8) \\
&= \frac{q(1-\epsilon)}{2\sqrt{2\pi}} \cdot \log^{1/2}\bigl(nq(1-\epsilon)\bigr) \cdot \exp\biggl\{-\frac{1}{2} \cdot \Bigl(\frac{7}{8}\Bigr)^2 \cdot 4\log\bigl(nq(1-\epsilon)\bigr)\biggr\} \\
&\leq \frac{q(1-\epsilon)}{2\sqrt{2\pi}} \cdot \log^{1/2}\bigl(nq(1-\epsilon)\bigr) \cdot \bigl\{nq(1-\epsilon)\bigr\}^{-3/2} \leq \frac{1}{5n}.
\end{aligned}
$$

Thus, $\mathrm{KL}(P_1^{\otimes n}, P_2^{\otimes n}) \leq 1/5 < \log\bigl(\frac{1}{4\delta(1-\delta)}\bigr)$ for $\delta \in (0, 1/4]$, so by Ma, Verchand and Samworth (2024, Theorem 4 and Corollary 6), we deduce that for $\delta \in (0, 1/4]$,

$$
\mathcal{M}\bigl(\delta, \mathcal{P}_\Theta, |\cdot|^2\bigr) \geq a^2 = \frac{\sigma^2 \log^2\bigl(1 + \frac{\epsilon}{q(1-\epsilon)}\bigr)}{16\log\bigl(nq(1-\epsilon)\bigr)}. \tag{49}
$$

Finally, note that $\mathsf{MCAR}_{(q(1-\epsilon),\mathsf{N}(\theta,\sigma^2))} \in \mathcal{R}(\mathsf{N}(\theta,\sigma^2), \epsilon, q)$ for all $\theta \in \mathbb{R}$, since we can choose the contamination distribution Q such that $Q(\{\star\}) = 1$. Therefore, by Proposition 46*(a)*, we have that for $\delta \in (0, 1/4]$,

$$
\mathcal{M}\bigl(\delta, \mathcal{P}_\Theta, |\cdot|^2\bigr) \begin{cases} \geq \dfrac{\sigma^2 \log(1/\delta)}{20nq(1-\epsilon)} & \text{if } \delta \geq \dfrac{\{1-q(1-\epsilon)\}^n}{2} \\ = \infty & \text{if } \delta < \dfrac{\{1-q(1-\epsilon)\}^n}{2}. \end{cases} \tag{50}
$$

Combining (49) and (50) yields the desired result. □

D.1.3 Proof of Theorem 7

In order to prove Theorem 7, we require a preliminary lemma.

Lemma 24. *Let $\theta_1, \theta_2 \in \mathbb{R}$ be distinct, and set $a := |\theta_1 - \theta_2|/2$. Then, writing $b := \frac{1}{2}\log\bigl(1 + \frac{4\epsilon}{q(1-\epsilon)}\bigr)$, there exists a continuous and strictly increasing function $f_{\mathrm{K},b} : (0,\infty) \to (0,1]$ such that*

$$
d_{\mathrm{K}}\bigl(\mathcal{R}(\theta_1), \mathcal{R}(\theta_2)\bigr) \geq f_{\mathrm{K},b}(a).
$$

Moreover,

$$
f_{\mathrm{K},b}(a) \geq q(1-\epsilon) \cdot \frac{a}{\sigma} \cdot \phi\biggl(\frac{a}{\sigma} + \frac{\sigma b}{a}\biggr) \quad \text{when } b \leq 1/2,
$$

and

$$
f_{\mathrm{K},b}(a) \geq q(1-\epsilon) \cdot \Phi\biggl(\frac{a}{\sigma} - \frac{2\sigma b}{a}\biggr) - \{q(1-\epsilon)+\epsilon\} \cdot \Phi\biggl(-\frac{a}{\sigma} - \frac{2\sigma b}{a}\biggr) \quad \text{when } b > 1/2.
$$

Proof. Since d_{K} is translation invariant, we may assume without loss of generality that $\theta_1 = -a$ and $\theta_2 = a$. By Proposition 2, if $R_\ell \in \mathcal{R}(\theta_\ell)$ for $\ell \in \{1,2\}$, then each admits a density $h_\ell : \mathbb{R}_\star \to \mathbb{R}$ with respect to the extended Lebesgue measure $\lambda_\star$ such that $h_\ell(x)/\phi_{(\theta_\ell,\sigma)}(x) \in [q(1-\epsilon),\, q(1-\epsilon)+\epsilon]$ for all $x \in \mathbb{R}$. Let $\tau := \frac{\sigma^2}{2a} \cdot \log\bigl(1 + \frac{\epsilon}{q(1-\epsilon)}\bigr) \leq \frac{\sigma^2 b}{a}$, so that $q(1-\epsilon)\phi_{(-a,\sigma)}(-\tau) = \{q(1-\epsilon)+\epsilon\}\phi_{(a,\sigma)}(-\tau)$, see Figure 6.

When $b \le 1/2$,

$$\begin{aligned}
d_{\mathrm{K}}\big(\mathcal{R}(\theta_1), \mathcal{R}(\theta_2)\big) &= \inf_{R_1 \in \mathcal{R}(\theta_1),\, R_2 \in \mathcal{R}(\theta_2)} \sup_{A \in \mathcal{A}} |R_1(A) - R_2(A)| \\
&\ge \inf_{R_1 \in \mathcal{R}(\theta_1),\, R_2 \in \mathcal{R}(\theta_2)} \big\{ R_1\big((-\infty, -\sigma^2 b/a]\big) - R_2\big((-\infty, -\sigma^2 b/a]\big) \big\} \\
&\ge q(1-\epsilon) \cdot \Phi_{(\theta_1,\sigma)}(-\sigma^2 b/a) - \{q(1-\epsilon) + \epsilon\} \cdot \Phi_{(\theta_2,\sigma)}(-\sigma^2 b/a) \\
&= q(1-\epsilon) \cdot \Phi\bigg(\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) - \{q(1-\epsilon) + \epsilon\} \cdot \Phi\bigg(-\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) =: f_{\mathrm{K},b}(a).
\end{aligned}$$

Now $f_{\mathrm{K},b}$ is continuously differentiable, with

$$\begin{aligned}
f'_{\mathrm{K},b}(a) &= q(1-\epsilon) \cdot \bigg(\frac{1}{\sigma} + \frac{\sigma b}{a^2}\bigg)\phi\bigg(\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) - \big\{q(1-\epsilon) + \epsilon\big\} \cdot \bigg(-\frac{1}{\sigma} + \frac{\sigma b}{a^2}\bigg)\phi\bigg(-\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) \\
&> \bigg(\frac{1}{\sigma} + \frac{\sigma b}{a^2}\bigg) \cdot \sigma \bigg\{ q(1-\epsilon) \cdot \phi_{(-a,\sigma)}\bigg(-\frac{\sigma^2 b}{a}\bigg) - \big\{q(1-\epsilon) + \epsilon\big\} \cdot \phi_{(a,\sigma)}\bigg(-\frac{\sigma^2 b}{a}\bigg) \bigg\} \\
&\ge \bigg(\frac{1}{\sigma} + \frac{\sigma b}{a^2}\bigg) \cdot \sigma \Big(q(1-\epsilon) \cdot \phi_{(-a,\sigma)}(-\tau) - \big\{q(1-\epsilon) + \epsilon\big\} \cdot \phi_{(a,\sigma)}(-\tau) \Big) = 0,
\end{aligned}$$

so that $f_{\mathrm{K},b}$ is strictly increasing as well. Moreover,

$$\begin{aligned}
f_{\mathrm{K},b}(a) &= q(1-\epsilon) \cdot \bigg\{ \Phi\bigg(\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) - \Phi\bigg(-\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) \bigg\} - \epsilon \cdot \Phi\bigg(-\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) \\
&\ge q(1-\epsilon) \cdot \frac{2a}{\sigma} \cdot \phi\bigg(\frac{a}{\sigma} + \frac{\sigma b}{a}\bigg) - \epsilon \cdot \Phi\bigg(-\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg), \qquad (51)
\end{aligned}$$

where the final inequality follows from the mean value theorem $\Phi\big(\frac{a}{\sigma} - \frac{\sigma b}{a}\big) - \Phi\big(-\frac{a}{\sigma} - \frac{\sigma b}{a}\big) = \frac{2a}{\sigma} \cdot \phi(x') \ge \frac{2a}{\sigma} \cdot \phi\big(\frac{a}{\sigma} + \frac{\sigma b}{a}\big)$, where $x' \in \big[-\frac{a}{\sigma} - \frac{\sigma b}{a}, \frac{a}{\sigma} - \frac{\sigma b}{a}\big]$. Next notice that

$$\begin{aligned}
\epsilon \cdot \Phi\bigg(-\frac{a}{\sigma} - \frac{\sigma b}{a}\bigg) \le \frac{\epsilon}{a/\sigma + \sigma b/a} \cdot \phi\bigg(\frac{a}{\sigma} + \frac{\sigma b}{a}\bigg) &\le \frac{\epsilon a}{\sigma b} \cdot \phi\bigg(\frac{a}{\sigma} + \frac{\sigma b}{a}\bigg) \\
&\le q(1-\epsilon) \cdot \frac{a}{\sigma} \cdot \phi\bigg(\frac{a}{\sigma} + \frac{\sigma b}{a}\bigg), \qquad (52)
\end{aligned}$$

where the first inequality follows from the Mills ratio bound $\Phi(-x) \le \phi(x)/x$ for $x > 0$, and the final inequality follows from the fact that $\log(1+x) \ge x/2$ for $x \in [0, 2]$, so that $b = \frac{1}{2}\log\big(1 + \frac{4\epsilon}{q(1-\epsilon)}\big) \ge \frac{\epsilon}{q(1-\epsilon)}$. Therefore, by (51) and (52) we deduce that

$$f_{\mathrm{K},b}(a) \ge q(1-\epsilon) \cdot \frac{a}{\sigma} \cdot \phi\bigg(\frac{a}{\sigma} + \frac{\sigma b}{a}\bigg),$$

when $b \le 1/2$.

On the other hand, when $b > 1/2$,

$$\begin{aligned}
d_{\mathrm{K}}\big(\mathcal{R}(\theta_1), \mathcal{R}(\theta_2)\big) &\ge \inf_{R_1 \in \mathcal{R}(\theta_1),\, R_2 \in \mathcal{R}(\theta_2)} \big\{ R_1\big((-\infty, -2\sigma^2 b/a]\big) - R_2\big((-\infty, -2\sigma^2 b/a]\big) \big\} \\
&\ge q(1-\epsilon) \cdot \Phi_{(\theta_1,\sigma)}(-2\sigma^2 b/a) - \{q(1-\epsilon) + \epsilon\} \cdot \Phi_{(\theta_2,\sigma)}(-2\sigma^2 b/a) \\
&= q(1-\epsilon) \cdot \Phi\bigg(\frac{a}{\sigma} - \frac{2\sigma b}{a}\bigg) - \{q(1-\epsilon) + \epsilon\} \cdot \Phi\bigg(-\frac{a}{\sigma} - \frac{2\sigma b}{a}\bigg) =: f_{\mathrm{K},b}(a).
\end{aligned}$$

Similarly to the previous case, $f_{\mathrm{K},b}$ is continuously differentiable and strictly increasing. □

Proof of Theorem 7. We first derive an upper bound on $d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(\theta_0)\big)$. Let $\mathcal{D} \coloneqq \{i \in [n] : Z_i \neq \star\}$ and $\bar{q} \coloneqq \mathbb{P}(Z_1 \neq \star)$, so that with the convention that $0/0 \coloneqq 0$,

$$\begin{aligned}\sup_{A\in\mathcal{A}} |\widehat{R}_n(A) - R(A)| &= \sup_{A\in\mathcal{A}} \left| \frac{|\mathcal{D}|}{n} \cdot \frac{1}{|\mathcal{D}|} \sum_{i\in\mathcal{D}} \mathbb{1}_{\{Z_i \in A\}} - \bar{q} \cdot \mathbb{P}(Z_1 \in A | Z_1 \neq \star) \right| \\ &\leq \frac{|\mathcal{D}|}{n} \cdot \sup_{A\in\mathcal{A}} \left| \frac{1}{|\mathcal{D}|} \sum_{i\in\mathcal{D}} \mathbb{1}_{\{Z_i \in A\}} - \mathbb{P}(Z_1 \in A | Z_1 \neq \star) \right| + \left| \frac{|\mathcal{D}|}{n} - \bar{q} \right|. \end{aligned} \tag{53}$$

Now, since $\bar{q} \geq q(1-\epsilon)$, we have by our lower bound on δ that

$$\log\Big(\frac{4}{\delta}\Big) \leq \frac{\big\{nq(1-\epsilon)\big\}^{31/36}}{6400 \log\big(nq(1-\epsilon)\big)} \leq \frac{nq(1-\epsilon)}{6400} \leq \frac{n\bar{q}}{6400}.$$

Hence, by Bernstein's inequality (Vershynin, 2018, Theorem 2.8.4), with probability at least $1 - \delta/2$,

$$\left| \frac{|\mathcal{D}|}{n} - \bar{q} \right| \leq \sqrt{\frac{4\bar{q}\log(4/\delta)}{n}} < \bar{q}. \tag{54}$$

Furthermore, by the Dvoretzky–Kiefer–Wolfowitz–Massart inequality (Massart, 1990; Reeve, 2024),

$$\sup_{A\in\mathcal{A}} \left| \frac{1}{|\mathcal{D}|} \sum_{i\in\mathcal{D}} \mathbb{1}_{\{Z_i \in A\}} - \mathbb{P}(Z_1 \in A \,|\, Z_1 \neq \star) \right| \leq \sqrt{\frac{\log(4/\delta)}{2|\mathcal{D}|}}, \tag{55}$$

with probability at least $1 - \delta/2$. Combining (53), (54) and (55) we deduce that, with probability at least $1 - \delta$,

$$\begin{aligned} d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(\theta_0)\big) \leq \sup_{A\in\mathcal{A}} |\widehat{R}_n(A) - R(A)| &\leq \sqrt{\frac{|\mathcal{D}|}{n}} \cdot \sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{4\bar{q}\log(4/\delta)}{n}} \\ &\leq \sqrt{\frac{\bar{q}\log(4/\delta)}{n}} + \sqrt{\frac{4\bar{q}\log(4/\delta)}{n}} \\ &\leq 3\sqrt{\frac{\{q(1-\epsilon)+\epsilon\}\log(4/\delta)}{n}} =: r_n. \end{aligned} \tag{56}$$

We now work on the event $\mathcal{E} \coloneqq \big\{d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(\theta_0)\big) \leq r_n\big\}$, which occurs with probability at least $1 - \delta$ by (56). If $\theta \in \mathbb{R}$ satisfies $d_{\mathrm{K}}\big(\mathcal{R}(\theta), \mathcal{R}(\theta_0)\big) > 2r_n$, then on the event $\mathcal{E}$,

$$d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(\theta)\big) \geq d_{\mathrm{K}}\big(\mathcal{R}(\theta), \mathcal{R}(\theta_0)\big) - d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(\theta_0)\big) > r_n \geq d_{\mathrm{K}}\big(\widehat{R}_n, \mathcal{R}(\theta_0)\big),$$

so $\widehat{\theta}_n^{\mathrm{K}} \neq \theta$. Therefore, with $f_{\mathrm{K},b}$ as defined in Lemma 24 and $b \coloneqq \frac{1}{2}\log\big(1 + \frac{4\epsilon}{q(1-\epsilon)}\big)$, we deduce that on $\mathcal{E}$,

$$\begin{aligned} |\widehat{\theta}_n^{\mathrm{K}} - \theta_0| &\leq \sup\big\{|\theta - \theta_0| : \theta \in \mathbb{R},\, d_{\mathrm{K}}\big(\mathcal{R}(\theta), \mathcal{R}(\theta_0)\big) \leq 2r_n\big\} \\ &\leq 2\sup\big\{a \geq 0 : f_{\mathrm{K},b}(a) \leq 2r_n\big\} = 2\inf\big\{a \geq 0 : f_{\mathrm{K},b}(a) \geq 2r_n\big\}, \end{aligned} \tag{57}$$

where the second inequality follows since by Lemma 24, $d_{\mathrm{K}}\big(\mathcal{R}(\theta),\mathcal{R}(\theta_0)\big) \geq f_{\mathrm{K},b}\big(\frac{|\theta-\theta_0|}{2}\big)$, and the final equality follows since $f_{\mathrm{K},b}$ is a strictly increasing and continuous function.

When $b \leq 1/2$, we have by (57) and Lemma 24 that on $\mathcal{E}$,

$$\begin{aligned}
|\widehat{\theta}_n^{\mathrm{K}} - \theta_0| &\leq 2\inf\big\{a \geq 0 : f_{\mathrm{K},b}(a) \geq 2r_n\big\} \\
&\leq 2\inf\left\{a \geq 0 : q(1-\epsilon)\cdot\frac{a}{\sigma}\cdot\phi\left(\frac{a}{\sigma}+\frac{\sigma b}{a}\right) \geq 6\sqrt{\frac{\{q(1-\epsilon)+\epsilon\}\log(4/\delta)}{n}}\right\} \\
&= 2\sigma\inf\left\{a \geq 0 : a\cdot\phi\left(a+\frac{b}{a}\right) \geq \sqrt{\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\cdot\frac{36\log(4/\delta)}{nq(1-\epsilon)}}\right\}. \qquad (58)
\end{aligned}$$

Now suppose further that $b \leq \sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}}$. The assumption on δ means that $b \leq \sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}} \leq 1/80$ and thus $1+\frac{4\epsilon}{q(1-\epsilon)} < 5/4$. Let $a := 20\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}}$, so that $a \leq 1/4$. Moreover, $b/a \leq 1/20$, so $a + b/a \leq 3/10$. Therefore,

$$\begin{aligned}
a\cdot\phi(a+b/a) \geq 20\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}}\cdot\phi(3/10) &\geq \sqrt{\frac{5}{4}\cdot\frac{36\log(4/\delta)}{nq(1-\epsilon)}} \\
&\geq \sqrt{\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\cdot\frac{36\log(4/\delta)}{nq(1-\epsilon)}}.
\end{aligned}$$

Hence, by (58), we have on $\mathcal{E}$ that $|\widehat{\theta}_n^{\mathrm{K}} - \theta_0| \leq 40\sigma\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}}$ when $b \leq \sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}}$.

Next, we consider the case $\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}} < b \leq 2\sqrt{\frac{\log(nq(1-\epsilon))\log(4/\delta)}{(nq(1-\epsilon))^{31/36}}}$. Then $b \leq 1/40$ and we again have $1+\frac{4\epsilon}{q(1-\epsilon)} < 5/4$. Let $a := 20b$, so that $a \leq 1/2$. Then

$$\begin{aligned}
a\cdot\phi(a+b/a) > 20\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}}\cdot\phi\left(\frac{1}{2}+\frac{1}{20}\right) &\geq \sqrt{\frac{5}{4}\cdot\frac{36\log(4/\delta)}{nq(1-\epsilon)}} \\
&\geq \sqrt{\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\cdot\frac{36\log(4/\delta)}{nq(1-\epsilon)}}.
\end{aligned}$$

Hence, by (58), we have on $\mathcal{E}$ that $|\widehat{\theta}_n^{\mathrm{K}} - \theta_0| \leq 40\sigma b$ when $\sqrt{\frac{\log(4/\delta)}{nq(1-\epsilon)}} < b \leq 2\sqrt{\frac{\log(nq(1-\epsilon))\log(4/\delta)}{(nq(1-\epsilon))^{31/36}}}$.

As our third case, suppose that $2\sqrt{\frac{\log(nq(1-\epsilon))\log(4/\delta)}{\{nq(1-\epsilon)\}^{31/36}}} < b \leq 1/2$. Let $a := \frac{9b}{\sqrt{\frac{5}{36}\log(nq(1-\epsilon))}}$, so that $a \geq 18\sqrt{\frac{\log(4/\delta)}{\frac{5}{36}(nq(1-\epsilon))^{31/36}}}$. By the assumption (19), we have $b \leq \frac{5\log(nq(1-\epsilon))}{432}$, so $a \leq 7b/a$. Therefore,

$$\begin{aligned}
a\cdot\phi(a+b/a) &\geq 18\sqrt{\frac{\log(4/\delta)}{\frac{5}{36}\{nq(1-\epsilon)\}^{31/36}}}\cdot\phi(8b/a) \\
&= 18\sqrt{\frac{\log(4/\delta)}{\frac{5}{36}\{nq(1-\epsilon)\}^{31/36}}}\cdot\frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{40\log\big(nq(1-\epsilon)\big)}{729}\right\}
\end{aligned}$$

$$= \sqrt{\frac{5832\log(4/\delta)}{5\pi\{nq(1-\epsilon)\}^{2831/2916}}} \geq \sqrt{\left(1+\frac{\epsilon}{q(1-\epsilon)}\right)\cdot\frac{36\log(4/\delta)}{nq(1-\epsilon)}},$$

where the final inequality holds since $b \leq \frac{5\log(nq(1-\epsilon))}{432}$. Hence, by (58), we have on $\mathcal{E}$ that $|\widehat{\theta}_n^{\mathrm{K}} - \theta_0| \leq \frac{108\sigma b}{\sqrt{5\log(nq(1-\epsilon))}}$ when $2\sqrt{\frac{\log(nq(1-\epsilon))\log(4/\delta)}{(nq(1-\epsilon))^{31/36}}} < b \leq 1/2$.

Finally, consider the case where $1/2 < b \leq \frac{5\log(nq(1-\epsilon))}{432}$ (when this interval is not vacuous). Then by (57) and Lemma 24 we have that on $\mathcal{E}$,

$$\begin{aligned}|\widehat{\theta}_n^{\mathrm{K}} - \theta_0| &\leq 2\inf\{a \geq 0 : f_{\mathrm{K},b}(a) \geq 2r_n\} \\ &\leq 2\inf\Bigg\{a \geq 0 : q(1-\epsilon)\cdot\Phi\left(\frac{a}{\sigma} - \frac{2\sigma b}{a}\right) - \{q(1-\epsilon)+\epsilon\}\cdot\Phi\left(-\frac{a}{\sigma} - \frac{2\sigma b}{a}\right) \\ &\qquad\qquad \geq \sqrt{\frac{36\{q(1-\epsilon)+\epsilon\}\log(4/\delta)}{n}}\Bigg\}. \end{aligned} \tag{59}$$

Letting $a := \frac{6\sigma b}{\sqrt{\log(nq(1-\epsilon))}}$, we have

$$\begin{aligned}&q(1-\epsilon)\cdot\Phi\left(\frac{a}{\sigma} - \frac{2\sigma b}{a}\right) - \{q(1-\epsilon)+\epsilon\}\cdot\Phi\left(-\frac{a}{\sigma} - \frac{2\sigma b}{a}\right) \\ &\overset{(i)}{\geq} \frac{q(1-\epsilon)}{\left(-\frac{a}{\sigma}+\frac{2\sigma b}{a}\right) + \left(-\frac{a}{\sigma}+\frac{2\sigma b}{a}\right)^{-1}}\cdot\phi\left(-\frac{a}{\sigma}+\frac{2\sigma b}{a}\right) - \frac{q(1-\epsilon)+\epsilon}{\frac{a}{\sigma}+\frac{2\sigma b}{a}}\cdot\phi\left(\frac{a}{\sigma}+\frac{2\sigma b}{a}\right) \\ &\overset{(ii)}{\geq} \left(\frac{a}{\sigma}+\frac{2\sigma b}{a}\right)^{-1}\frac{1}{\sqrt{2\pi}}\cdot\Bigg\{q(1-\epsilon)\exp\left(-\frac{a^2}{2\sigma^2} - \frac{2\sigma^2 b^2}{a^2} + 2b\right) \\ &\qquad\qquad - \{q(1-\epsilon)+\epsilon\}\exp\left(-\frac{a^2}{2\sigma^2} - \frac{2\sigma^2 b^2}{a^2} - 2b\right)\Bigg\} \\ &\overset{(iii)}{\geq} \left(\frac{a}{\sigma}+\frac{2\sigma b}{a}\right)^{-1}\frac{4\epsilon}{\sqrt{2\pi}}\cdot\exp\left(-\frac{a^2}{2\sigma^2} - \frac{2\sigma^2 b^2}{a^2}\right) \\ &\overset{(iv)}{\geq} \frac{2}{\sqrt{\log\big(nq(1-\epsilon)\big)}}\cdot\frac{4\epsilon}{\sqrt{2\pi}}\cdot\big(nq(1-\epsilon)\big)^{-5/72} \\ &\overset{(v)}{\geq} \sqrt{\frac{36\{q(1-\epsilon)+\epsilon\}\log(4/\delta)}{n}} = 2r_n.\end{aligned}$$

Here, (i) follows from the Mills ratio bound $\phi(x)/(x+x^{-1}) \leq \Phi(-x) \leq \phi(x)/x$ for $x > 0$; (ii) follows since $1/2 < b \leq \frac{\log(nq(1-\epsilon))}{36}$ implies $\left(-\frac{a}{\sigma}+\frac{2\sigma b}{a}\right) + \left(-\frac{a}{\sigma}+\frac{2\sigma b}{a}\right)^{-1} \leq \frac{a}{\sigma}+\frac{2\sigma b}{a}$; (iii) follows by substituting the definition of b; (iv) follows since, by assumption $b \leq \frac{5\log(nq(1-\epsilon))}{432}$, so $\frac{a}{\sigma} \leq \frac{\sqrt{\log(nq(1-\epsilon))}}{6}$; and (v) follows from the assumptions that $b > 1/2$, so $q(1-\epsilon) < 3\epsilon$, and moreover $\delta \geq 4\exp\left(-\frac{\{nq(1-\epsilon)\}^{31/36}}{6400\log(nq(1-\epsilon))}\right)$. Hence, by (59), we have on $\mathcal{E}$ that $|\widehat{\theta}_n^{\mathrm{K}} - \theta_0| \leq \frac{12\sigma b}{\sqrt{\log(nq(1-\epsilon))}}$ when $1/2 < b \leq \frac{5\log(nq(1-\epsilon))}{432}$. Combining all four cases yields the desired result. □

D.1.4 Proof of Theorem 8

Lemma 25. *Let $\epsilon \in [0,1)$, $\pi \in \mathcal{P}\big(\{\emptyset,[d]\}\big)$, $P \in \mathcal{P}(\mathbb{R}^d)$, $R \in \mathcal{R}_{\emptyset,[d]}(P,\epsilon,\pi)$ and $v \in \mathbb{R}^d$. Suppose that $X \sim P$, $Z \sim R$ and define $Z^{(v)} := v^\top Z \cdot \mathbb{1}_{\{Z \in \mathbb{R}^d\}} + \star \cdot \mathbb{1}_{\{Z \notin \mathbb{R}^d\}}$ for $v \in \mathbb{R}^d$. Then, writing $P^{(v)} := \mathsf{Law}(v^\top X)$ and $R^{(v)} := \mathsf{Law}(Z^{(v)})$, we have $R^{(v)} \in \mathcal{R}\big(P^{(v)}, \epsilon, \pi([d])\big)$.*

Proof. Let $q := \pi([d])$. We have $\mathsf{Law}(Z) = (1-\epsilon)\mathsf{Law}(X \circledast \Omega^{(1)}) + \epsilon \mathsf{Law}(X \circledast \Omega^{(2)})$ where $\Omega^{(1)} \perp\!\!\!\perp X$ and $\mathbb{P}(\Omega^{(1)} = \mathbf{1}_{[d]}) = q = 1 - \mathbb{P}(\Omega^{(1)} = 0)$ and where $\Omega^{(2)}$ takes values in $\{0, \mathbf{1}_{[d]}\}$. By properties of disintegrations (see Section G), we may define $m^{(v)} : \mathbb{R} \to [0,1]$ by $m^{(v)}(y) := \mathbb{P}(\Omega^{(2)} = \mathbf{1}_{[d]} \,|\, v^\top X = y)$. We also let $\mu^{(v)}$ be a σ-finite measure on $\mathbb{R}$ such that $P^{(v)} \ll \mu^{(v)}$ and let $p^{(v)} := \mathrm{d}P^{(v)}/\mathrm{d}\mu^{(v)}$. Finally, define $g : \mathbb{R}_\star \to [0,\infty)$ by

$$g(z) := \begin{cases} q(1-\epsilon)p^{(v)}(z) + \epsilon m^{(v)}(z) p^{(v)}(z) & \text{if } z \in \mathbb{R} \\ 1 - q(1-\epsilon) - \epsilon \int_{\mathbb{R}} m^{(v)}(y) p^{(v)}(y)\,\mathrm{d}\mu^{(v)}(y) & \text{if } z = \star. \end{cases}$$

Then, for $A \in \mathcal{B}(\mathbb{R})$, we have

$$\begin{aligned} \int_A g(z)\,\mathrm{d}\mu_\star^{(v)}(z) &= q(1-\epsilon)\mathbb{P}(v^\top X \in A) + \epsilon \mathbb{P}\big(\{\Omega^{(2)} = \mathbf{1}_{[d]}\} \cap \{v^\top X \in A\}\big) \\ &= (1-\epsilon)\mathbb{P}\big(v^\top(X \circledast \Omega^{(1)}) \in A\big) + \epsilon \mathbb{P}\big(v^\top (X \circledast \Omega^{(2)}) \in A\big) \\ &= \mathbb{P}(Z^{(v)} \in A) = R^{(v)}(A). \end{aligned}$$

It follows that $R^{(v)} \ll \mu_\star^{(v)}$, with Radon–Nikodym derivative g. Hence, by Proposition 2, we have $R^{(v)} \in \mathcal{R}(P^{(v)}, \epsilon, q)$. □

Proof of Theorem 8. We have $\mathsf{Law}(Z_1) = (1-\epsilon)\mathsf{Law}(X_1 \circledast \Omega_1^{(1)}) + \epsilon \mathsf{Law}(X_1 \circledast \Omega_1^{(2)})$ where $X_1 \sim \mathsf{N}_d(\theta_0, \Sigma)$, $\Omega_1^{(1)} \perp\!\!\!\perp X_1$ and $\mathbb{P}(\Omega_1^{(1)} = \mathbf{1}_{[d]}) = q$. Define $m : \mathbb{R} \to [0,1]$ by $m(y) := \mathbb{P}(\Omega_1^{(2)} = \mathbf{1}_{[d]} \,|\, v^\top X_1 = y)$. We claim that the distribution $R^{(v)}$ of $Z_1^{(v)}$ is absolutely continuous with respect to $\lambda_\star$, with Radon–Nikodym derivative

$$\frac{\mathrm{d}R^{(v)}}{\mathrm{d}\lambda_\star}(z) = \begin{cases} q(1-\epsilon)\phi_{(v^\top\theta_0, v^\top\Sigma v)}(z) + \epsilon m(z)\phi_{(v^\top\theta_0, v^\top\Sigma v)}(z) & \text{if } z \in \mathbb{R} \\ 1 - q(1-\epsilon) - \epsilon \int_{\mathbb{R}} m(y)\phi_{(v^\top\theta_0, v^\top\Sigma v)}(y)\,\mathrm{d}\lambda(y) & \text{if } z = \star. \end{cases}$$

To see this, it suffices to observe that for $A \in \mathcal{B}(\mathbb{R})$, we have

$$\int_A \frac{\mathrm{d}R^{(v)}}{\mathrm{d}\lambda_\star}(z)\,\mathrm{d}\lambda_\star(z) = q(1-\epsilon)\mathbb{P}(v^\top X_1 \in A) + \epsilon \mathbb{P}\big(\{\Omega_1^{(2)} = \mathbf{1}_{[d]}\} \cap \{v^\top X_1 \in A\}\big) = \mathbb{P}(Z_1^{(v)} \in A).$$

The claim therefore follows, so by Proposition 2 we have $R^{(v)} \in \mathcal{R}\big(\mathsf{N}(v^\top\theta_0, v^\top\Sigma v), \epsilon, q\big)$. We deduce from Theorem 7 and a union bound that

$$\max_{v \in \mathcal{N}} \big(\widehat{\theta}_n^{\mathrm{K}}(v) - v^\top\theta_0\big)^2 \lesssim C_{n,q,\epsilon,\delta/(4\cdot 9^d)} \bigg\{ \frac{\|\Sigma\|_{\mathrm{op}}\big(d + \log(4/\delta)\big)}{nq(1-\epsilon)} + \frac{\|\Sigma\|_{\mathrm{op}} \log^2\big(1 + \frac{4\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)} \bigg\}, \quad (60)$$

with probability at least $1-\delta$. Next, since any $v \in \mathbb{S}^{d-1}$ can be written as $v = v_1 + v_2$, where $v_1 \in \mathcal{N}$ and $\|v_2\|_2 \le 1/4$, we have

$$\|\widehat{\theta}_n^{\mathrm{MK}} - \theta_0\|_2 = \sup_{v \in \mathbb{S}^{d-1}} |v^\top\widehat{\theta}_n^{\mathrm{MK}} - v^\top\theta_0| \le \max_{v \in \mathcal{N}} |v^\top\widehat{\theta}_n^{\mathrm{MK}} - v^\top\theta_0| + \frac{1}{4} \cdot \|\widehat{\theta}_n^{\mathrm{MK}} - \theta_0\|_2,$$

so

$$\|\widehat{\theta}_n^{\mathrm{MK}} - \theta_0\|_2 \leq \frac{4}{3} \cdot \max_{v \in \mathcal{N}} |v^\top \widehat{\theta}_n^{\mathrm{MK}} - v^\top \theta_0|.$$

Hence,

$$\begin{aligned}
\|\widehat{\theta}_n^{\mathrm{MK}} - \theta_0\|_2^2 &\leq 2 \max_{v \in \mathcal{N}} \big(v^\top \widehat{\theta}_n^{\mathrm{MK}} - \widehat{\theta}_n^{\mathrm{K}}(v) + \widehat{\theta}_n^{\mathrm{K}}(v) - v^\top \theta_0\big)^2 \\
&\leq 4 \max_{v \in \mathcal{N}} \big(v^\top \widehat{\theta}_n^{\mathrm{MK}} - \widehat{\theta}_n^{\mathrm{K}}(v)\big)^2 + 4 \max_{v \in \mathcal{N}} \big(v^\top \theta_0 - \widehat{\theta}_n^{\mathrm{K}}(v)\big)^2 \\
&\leq 8 \max_{v \in \mathcal{N}} \big(v^\top \theta_0 - \widehat{\theta}_n^{\mathrm{K}}(v)\big)^2 \\
&\lesssim C_{n,q,\epsilon,\delta/(4 \cdot 9^d)} \left\{ \frac{\|\Sigma\|_{\mathrm{op}} \big(d + \log(4/\delta)\big)}{nq(1-\epsilon)} + \frac{\|\Sigma\|_{\mathrm{op}} \log^2\big(1 + \frac{4\epsilon}{q(1-\epsilon)}\big)}{\log\big(nq(1-\epsilon)\big)} \right\},
\end{aligned}$$

with probability at least $1 - \delta$, where the third inequality follows from the definition of $\widehat{\theta}_n^{\mathrm{MK}}$, and the last inequality follows from (60). □

D.1.5 Proof of Lemma 9

Proof of Lemma 9. We first show that, for any $R \in \mathcal{R}(P, \epsilon, q)$, we have

$$d_{\mathrm{K}}(\widehat{R}_n, R) = \max_{i \in \{0\} \cup [m]} \left\{ \left| \frac{i}{n} - R\big((-\infty, Z_{(i)})\big) \right| \vee \left| \frac{i}{n} - R\big((-\infty, Z_{(i+1)})\big) \right| \right\}.$$

To this end, fix $i \in \{0\} \cup [m]$. Then, since $\widehat{R}_n\big((-\infty, t)\big) = i/n$ for $t \in [Z_{(i)}, Z_{(i+1)}) \cap \mathbb{R}$, $t \mapsto R\big((-\infty, t]\big)$ is increasing on this interval and since $R \ll \lambda_\star$ by Proposition 2, we have

$$\begin{aligned}
\sup_{t \in [Z_{(i)}, Z_{(i+1)}) \cap \mathbb{R}} \big|\widehat{R}_n\big((-\infty, t]\big) - R\big((-\infty, t]\big)\big| &= \left| \frac{i}{n} - R\big((-\infty, Z_{(i)})\big) \right| \vee \lim_{t \nearrow Z_{(i+1)}} \left| \frac{i}{n} - R\big((-\infty, t)\big) \right| \\
&= \left| \frac{i}{n} - R\big((-\infty, Z_{(i)})\big) \right| \vee \left| \frac{i}{n} - R\big((-\infty, Z_{(i+1)})\big) \right|.
\end{aligned}$$

Hence

$$\sup_{t \in \mathbb{R}} \big|\widehat{R}_n\big((-\infty, t]\big) - R\big((-\infty, t]\big)\big| = \max_{i \in \{0\} \cup [m]} \left\{ \left| \frac{i}{n} - R\big((-\infty, Z_{(i)})\big) \right| \vee \left| \frac{i}{n} - R\big((-\infty, Z_{(i+1)})\big) \right| \right\}.$$

Now, by Proposition 2, for $0 \leq V_1 \leq \ldots \leq V_{m+1} \leq 1$, there exists $R \in \mathcal{R}(P, \epsilon, q)$ such that $V_i = R\big((-\infty, Z_{(i)}]\big)$ for $i \in [m]$ and $V_{m+1} = R\big((-\infty, \infty)\big)$ if and only if $(V_1, \ldots, V_{m+1})^\top \in \mathcal{V}$. The claim then follows. □

D.2 Proofs from Section 4.2

D.2.1 Proof of Theorem 10

The proof of Theorem 10 relies on the following preliminary result, which controls the bias.

Proposition 26. *Let $\theta_0 \in \mathbb{R}$, $\epsilon \in [0, 1)$, $q \in (0, 1]$ and $\sigma > 0$.*

(a) Let $r \geq 2$, $P \in \mathcal{P}_{L^r}(\theta_0, \sigma^2)$ and $Z \sim R \in \mathcal{R}(P, \epsilon, q)$. Then

$$\big\{\mathbb{E}(Z \,|\, Z \neq \star) - \theta_0\big\}^2 \leq \sigma^2 \cdot \left\{ \left(\frac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \left(\frac{\epsilon}{q(1-\epsilon)} \right)^{2/r} \right\}.$$

(b) Let $r \geq 1$, $P \in \mathcal{P}_{\psi_r}(\theta_0, \sigma^2)$ and $Z \sim R \in \mathcal{R}(P, \epsilon, q)$. Then

$$\big\{\mathbb{E}(Z \,|\, Z \neq \star) - \theta_0\big\}^2 \leq \sigma^2 \cdot \left\{ 4\left(\frac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)} \right) \right\}.$$

Proof. Let $\kappa := \frac{\epsilon}{q(1-\epsilon)}$. By translation invariance, we may assume without loss of generality that $\theta_0 = 0$ throughout the proof.

(a) Let μ be a measure on $\mathbb{R}$ such that $P \ll \mu$ and let $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$, then by Proposition 2, we have

$$\frac{\mathrm{d}R}{\mathrm{d}\mu_\star}(z) = \begin{cases} q(1-\epsilon) \cdot p(z) + \epsilon \cdot m(z)p(z) & \text{if } z \in \mathbb{R} \\ 1 - q(1-\epsilon) - \epsilon \int_{\mathbb{R}} m(x)p(x)\,\mathrm{d}\mu(x) & \text{if } z = \star, \end{cases} \tag{61}$$

for some Borel measurable function $m : \mathbb{R} \to [0,1]$. Therefore,

$$\begin{aligned} \big|\mathbb{E}(Z \,|\, Z \neq \star)\big| &= \frac{\big|q(1-\epsilon) \cdot \int_{\mathbb{R}} xp(x)\,\mathrm{d}\mu(x) + \epsilon \cdot \int_{\mathbb{R}} xm(x)p(x)\,\mathrm{d}\mu(x)\big|}{q(1-\epsilon) + \epsilon \int_{\mathbb{R}} m(x)p(x)\,\mathrm{d}\mu(x)} \\ &= \frac{\epsilon \cdot \big|\mathbb{E}_P\{Xm(X)\}\big|}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}} \leq \frac{\epsilon \cdot \sigma \cdot \big\{\mathbb{E}_P\big(m^{r/(r-1)}(X)\big)\big\}^{1-1/r}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}}, \end{aligned}$$

where the second equality follows from the assumption that $\theta_0 = 0$, and where the inequality follows from Hölder's inequality and the fact that $\mathbb{E}_P(|X|^r)^{1/r} \leq \sigma$. On the one hand, since $\big\{\mathbb{E}_P\big(m^{r/(r-1)}(X)\big)\big\}^{1-1/r} \leq 1$ and $\mathbb{E}_P\{m(X)\} \geq 0$, we have

$$\frac{\epsilon \cdot \big\{\mathbb{E}_P\big(m^{r/(r-1)}(X)\big)\big\}^{1-1/r}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}} \leq \kappa. \tag{62}$$

On the other hand, since $m(x) \in [0,1]$, we have $m^{r/(r-1)}(x) \leq m(x)$ for all $x \in \mathbb{R}$ and thus $\big\{\mathbb{E}_P\big(m^{r/(r-1)}(X)\big)\big\}^{1-1/r} \leq \big\{\mathbb{E}_P\big(m(X)\big)\big\}^{1-1/r} =: t$. Therefore,

$$\begin{aligned} \frac{\epsilon \cdot \big\{\mathbb{E}_P\big(m^{r/(r-1)}(X)\big)\big\}^{1-1/r}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}} &\leq \frac{\epsilon t}{q(1-\epsilon) + \epsilon t^{r/(r-1)}} \\ &\leq \sup_{t' \geq 0} \frac{\epsilon t'}{q(1-\epsilon) + \epsilon (t')^{r/(r-1)}} \overset{(i)}{=} \frac{\epsilon \cdot \{(r-1)q(1-\epsilon)/\epsilon\}^{1-1/r}}{q(1-\epsilon) + (r-1)q(1-\epsilon)} \\ &\leq (r-1)^{-1/r}\kappa^{1/r} \leq \kappa^{1/r}, \end{aligned} \tag{63}$$

where (i) follows from the fact that the function $t' \mapsto \frac{\epsilon t'}{q(1-\epsilon) + \epsilon (t')^{r/(r-1)}}$ is maximised when $t' = \{(r-1)q(1-\epsilon)/\epsilon\}^{1-1/r}$. Combining (62) and (63), we deduce that

$$\big|\mathbb{E}(Z \,|\, Z \neq \star)\big| \leq \frac{\epsilon \cdot \sigma \cdot \big\{\mathbb{E}_P\big(m^{r/(r-1)}(X)\big)\big\}^{1-1/r}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}} \leq \sigma(\kappa \wedge \kappa^{1/r}),$$

as desired.

(b) Let $Q \in \mathcal{P}(\mathbb{R})$ such that $Q \ll P$. By the variational characterisation of Kullback–Leibler divergence (e.g. Boucheron, Lugosi and Massart, 2013, Corollary 4.15),

$$\mathbb{E}_{X\sim Q}\big(g(X)\big) \le \mathrm{KL}(Q,P) + \log \mathbb{E}_{X\sim P}\big(e^{g(X)}\big), \tag{64}$$

for all Borel measurable functions $g : \mathbb{R} \to [0,\infty)$. Now take Q to be the conditional distribution of Z given $\{Z \neq \star\}$. Let μ and p be as in the proof of *(a)*, so that (61) holds for some Borel measurable function $m : \mathbb{R} \to [0,1]$. Therefore, for all $x \in \mathbb{R}$,

$$\frac{\mathrm{d}Q}{\mathrm{d}\mu}(x) = \frac{q(1-\epsilon)\cdot p(x) + \epsilon \cdot m(x)p(x)}{q(1-\epsilon) + \epsilon \cdot \int_{\mathbb{R}} m(y)p(y)\,\mathrm{d}\mu(y)}.$$

Hence $Q \ll P$ and

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \in \left[1 - \frac{\epsilon}{q(1-\epsilon)+\epsilon},\, 1 + \frac{\epsilon}{q(1-\epsilon)}\right], \tag{65}$$

for all $x \in \mathbb{R}$, from which we deduce that

$$\mathrm{KL}(Q,P) = \int_{\mathbb{R}} \log\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}Q \le \log(1+\kappa). \tag{66}$$

Taking $g(\cdot) = |\cdot|^r/\sigma^r$ and combining (64) and (66) yields

$$\begin{aligned}\mathbb{E}\big(|Z|^r/\sigma^r \,\big|\, Z \neq \star\big) &\le \log(1+\kappa) + \log \mathbb{E}_{X\sim P}\big\{\exp\big(|X|^r/\sigma^r\big)\big\} \\ &\le \log(1+\kappa) + \log 2 = \log(2+2\kappa),\end{aligned} \tag{67}$$

where the second inequality follows since $P \in \mathcal{P}_{\psi_r}(\theta_0, \sigma^2)$ and since $\theta_0 = 0$ by assumption. Thus,

$$\big|\mathbb{E}(Z \,|\, Z \neq \star)\big| \le \mathbb{E}\big(|Z| \,\big|\, Z \neq \star\big) \le \mathbb{E}\big(|Z|^r \,\big|\, Z \neq \star\big)^{1/r} \le \sigma \log^{1/r}(2+2\kappa), \tag{68}$$

where the second inequality follows from the conditional version of Jensen's inequality and the third inequality follows from (67). Moreover, by Götze, Sambale and Sinulis (2021, Lemma A.2), we have $\mathrm{Var}_{X\sim P}(X)^{1/2} \le 2\big(\frac{2}{re}\big)^{1/r}\sigma \le 2\sigma$ for $r \ge 1$. Hence $P \in \mathcal{P}_{L^2}(0, 4\sigma^2)$, so we can apply part *(a)* of the theorem to obtain

$$\big|\mathbb{E}\big(Z \,\big|\, Z \neq \star\big)\big| \le 2\sigma\kappa. \tag{69}$$

Combining (68) and (69) proves part *(b)*. □

Proof of Theorem 10. Let $\kappa := \frac{\epsilon}{q(1-\epsilon)}$.

(a) Let μ and p be as in the proof of Proposition 26*(a)*, so that (61) holds for some Borel measurable function $m : \mathbb{R} \to [0,1]$. On the one hand, since $m(X) \in [0,1]$, we have

$$\begin{aligned}\mathrm{Var}(Z_1 \,|\, Z_1 \neq \star) &= \mathrm{Var}(Z_1 - \theta_0 \,|\, Z_1 \neq \star) \le \mathbb{E}\big\{(Z_1-\theta_0)^2 \,|\, Z_1 \neq \star\big\} \\ &= \frac{\int_{\mathbb{R}} (x-\theta_0)^2\{q(1-\epsilon)p(x) + \epsilon m(x)p(x)\}\,\mathrm{d}\mu(x)}{q(1-\epsilon) + \epsilon\int_{\mathbb{R}} m(x)p(x)\,\mathrm{d}\mu(x)}\end{aligned}$$

$$\leq \sigma^2 + \frac{\epsilon \cdot \mathbb{E}_P\{(X-\theta_0)^2 m(X)\}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}} \leq (1+\kappa)\sigma^2. \tag{70}$$

On the other hand, for $r > 2$, we have by Hölder's inequality that

$$\begin{aligned}
\mathrm{Var}(Z_1 \,|\, Z_1 \neq \star) &\leq \sigma^2 + \frac{\epsilon \cdot \mathbb{E}_P\{(X-\theta_0)^2 m(X)\}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}} \leq \left[1 + \frac{\epsilon \cdot \left\{\mathbb{E}_P\big(m^{r/(r-2)}(X)\big)\right\}^{1-2/r}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}}\right]\sigma^2 \\
&\leq \left[1 + \frac{\epsilon \cdot \left\{\mathbb{E}_P\big(m(X)\big)\right\}^{1-2/r}}{q(1-\epsilon) + \epsilon \cdot \mathbb{E}_P\{m(X)\}}\right]\sigma^2 \leq \sup_{t' \geq 0}\left(1 + \frac{\epsilon t'}{q(1-\epsilon) + \epsilon (t')^{r/(r-2)}}\right)\sigma^2 \\
&= \left\{1 + \frac{2}{r}\left(\frac{r-2}{2}\right)^{1-2/r} \cdot \kappa^{2/r}\right\}\sigma^2 \leq \left\{1 + \left(\frac{2}{r}\right)^{2/r}\kappa^{2/r}\right\}\sigma^2 \leq (1+\kappa^{2/r})\sigma^2,
\end{aligned} \tag{71}$$

where the equality follows since the supremum is attained when $t' = \big(\frac{(r-2)q(1-\epsilon)}{2\epsilon}\big)^{1-2/r}$. Combining (70) and (71) yields that for $r \geq 2$,

$$\mathrm{Var}(Z_1 \,|\, Z_1 \neq \star) \leq (1+\kappa^{2/r})\sigma^2. \tag{72}$$

By Lemma 38*(b)*, the event $\mathcal{E}_0 := \{|\mathcal{D}| \geq nq(1-\epsilon)/2\}$ has $\mathbb{P}(\mathcal{E}_0) \geq 1 - \delta/2$, since $nq(1-\epsilon) \geq 8\log(2/\delta)$. Moreover, writing

$$\mathcal{E}_1 := \left\{\big(\widehat{\theta}_n - \mathbb{E}(Z_1 | Z_1 \neq \star)\big)^2 \leq 48e(1+\kappa^{2/r})\frac{\sigma^2 \log(2e/\delta)}{nq(1-\epsilon)}\right\},$$

we have by Lemma 48 and (72) that $\mathbb{P}\big(\mathcal{E}_1 \,\big|\, |\mathcal{D}| = s\big) \geq 1 - \delta/2$ for $s \geq nq(1-\epsilon)/2$, so

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_1) &= \mathbb{E}\big\{\mathbb{P}\big(\mathcal{E}_0 \cap \mathcal{E}_1 \,\big|\, |\mathcal{D}|\big)\big\} = \sum_{s=\lceil nq(1-\epsilon)/2\rceil}^{n} \mathbb{P}\big(\mathcal{E}_1 \,\big|\, |\mathcal{D}| = s\big)\mathbb{P}(|\mathcal{D}| = s) \\
&\geq \Big(1 - \frac{\delta}{2}\Big) \sum_{s=\lceil nq(1-\epsilon)/2\rceil}^{n} \mathbb{P}(|\mathcal{D}| = s) = \Big(1 - \frac{\delta}{2}\Big)\mathbb{P}(\mathcal{E}_0) \geq \Big(1 - \frac{\delta}{2}\Big)^2 \geq 1 - \delta.
\end{aligned} \tag{73}$$

On the event $\mathcal{E}_0 \cap \mathcal{E}_1$, we have by Proposition 26*(a)* that

$$\begin{aligned}
\big(\widehat{\theta}_n - \theta_0\big)^2 &\leq 2\big\{\widehat{\theta}_n - \mathbb{E}(Z_1 \,|\, Z_1 \neq \star)\big\}^2 + 2\big\{\mathbb{E}(Z_1 \,|\, Z_1 \neq \star) - \theta_0\big\}^2 \\
&\leq 96e \cdot \frac{\sigma^2 \log(2e/\delta)}{nq(1-\epsilon)} + 96e\kappa^{2/r} \cdot \frac{\sigma^2 \log(2e/\delta)}{nq(1-\epsilon)} + 2\sigma^2(\kappa^2 \wedge \kappa^{2/r}) \\
&\leq 192e \cdot \frac{\sigma^2 \log(2e/\delta)}{nq(1-\epsilon)} + (12e+2)\sigma^2(\kappa^2 \wedge \kappa^{2/r}),
\end{aligned}$$

where the final inequality follows by considering separately the cases $\kappa \leq 1$ and $\kappa > 1$, and in the second case noting that $\frac{\log(2e/\delta)}{nq(1-\epsilon)} \leq 1/8$ by assumption. Combining this with (73) establishes the result.

(b) Let μ and p be as in the proof of Proposition 26*(a)*, so that (61) holds for some Borel measurable function $m : \mathbb{R} \to [0,1]$. Then, for integers $\ell \geq 2$, we have

$$\big\{\mathbb{E}\big(|Z_1 - \theta_0|^\ell \,|\, Z_1 \neq \star\big)\big\}^{1/\ell} = \left(\frac{\int_{\mathbb{R}} |x-\theta_0|^\ell \{q(1-\epsilon)p(x) + \epsilon m(x)p(x)\}\,\mathrm{d}\mu(x)}{q(1-\epsilon) + \epsilon\int_{\mathbb{R}} m(x)p(x)\,\mathrm{d}\mu(x)}\right)^{1/\ell}$$

$$\leq \left(\frac{\{q(1-\epsilon)+\epsilon\}\mathbb{E}_{X\sim P}(|X-\theta_0|^\ell)}{q(1-\epsilon)} \right)^{1/\ell} \overset{(i)}{\lesssim} \sqrt{1+\kappa}\cdot\sigma\ell,$$

where (i) is true since $X \sim P \in \mathcal{P}_{\psi_r}(\theta_0,\sigma^2) \subseteq \mathcal{P}_{\psi_1}(\theta_0,\sigma^2)$ by Lemma 39, so $\big(\mathbb{E}_{X\sim P}|X-\theta_0|^\ell\big)^{1/\ell} \lesssim \sigma\ell$ by Vershynin (2018, Proposition 2.7.1). Moreover, by the Cauchy–Schwarz inequality and (70), $\mathbb{E}\big(|Z_1-\theta_0|\,|\,Z_1\neq\star\big) \leq \big\{\mathbb{E}\big(|Z_1-\theta_0|^2\,|\,Z_1\neq\star\big)\big\}^{1/2} \lesssim \sigma\sqrt{1+\kappa}$. Hence, by Vershynin (2018, Proposition 2.7.1) again, conditional on $\{Z_1\neq\star\}$, we have $\|Z_1-\theta_0\|_{\psi_1} \lesssim \sigma\sqrt{1+\kappa}$. Then, by Vershynin (2018, Lemma 2.7.10), we have, conditional on $\{Z_1\neq\star\}$, that

$$\big\|Z_1 - \mathbb{E}(Z_1\,|\,Z_1\neq\star)\big\|_{\psi_1} \lesssim \sigma\sqrt{1+\kappa}. \tag{74}$$

Recall the definition of the event $\mathcal{E}_0$ from the proof of *(a)*, and observe that $\mathbb{P}(\mathcal{E}_0) \geq 1-\delta/4$ by Lemma 38*(b)*. Now, similarly to (73), by Bernstein's inequality (Vershynin, 2018, Corollary 2.8.3) and since $\frac{\log(8/\delta)}{nq(1-\epsilon)} \leq 1/8$, there exists a universal constant $C_2>0$ such that the event

$$\mathcal{E}_2 := \left\{ \left(\frac{\sum_{i\in\mathcal{D}}\{Z_i - \mathbb{E}(Z_1\,|\,Z_1\neq\star)\}}{|\mathcal{D}|} \right)^2 \leq C_2(1+\kappa)\frac{\sigma^2\log(8/\delta)}{nq(1-\epsilon)} \right\}$$

satisfies $\mathbb{P}(\mathcal{E}_0\cap\mathcal{E}_2) \geq 1-\delta/2$. Moreover, on $\mathcal{E}_0\cap\mathcal{E}_2$, by Proposition 26*(b)*,

$$\begin{aligned}
\big(\widehat{\theta}_n-\theta_0\big)^2 &\lesssim \left(\frac{\sum_{i\in\mathcal{D}}\{Z_i - \mathbb{E}(Z_1\,|\,Z_1\neq\star)\}}{|\mathcal{D}|} \right)^2 + \big\{\mathbb{E}(Z_1\,|\,Z_1\neq\star)-\theta_0\big\}^2 \\
&\lesssim (1+\kappa)\cdot\frac{\sigma^2\log(8/\delta)}{nq(1-\epsilon)} + \sigma^2\kappa^2 \lesssim \frac{\sigma^2\log(8/\delta)}{nq(1-\epsilon)} + \sigma^2\left(\frac{\log(8/\delta)}{nq(1-\epsilon)}\right)^2 + \sigma^2\kappa^2 \\
&\lesssim \frac{\sigma^2\log(8/\delta)}{nq(1-\epsilon)} + \sigma^2\kappa^2,
\end{aligned} \tag{75}$$

where the penultimate inequality follows from the inequality $ab \leq \frac{a^2+b^2}{2}$ for $a,b\in\mathbb{R}$, and the final inequality follows from the assumption $\frac{\log(8/\delta)}{nq(1-\epsilon)} \leq 1/8$.

Next, let $Q\in\mathcal{P}(\mathbb{R})$ be such that $Q \ll P$. Then the variational characterisation of χ^2-divergence (e.g., Polyanskiy and Wu, 2024, Example 7.4) yields that

$$2\mathbb{E}_{X\sim Q}\{g(X)\} \leq 1 + \chi^2(Q,P) + \mathbb{E}_{X\sim P}\{g^2(X)\}, \tag{76}$$

for all Borel measurable $g:\mathbb{R}\to[0,\infty)$. We first consider the case $r>1$. Now take Q to be the conditional distribution of Z_1 given $\{Z_1\neq\star\}$, so that, by the representation in (65), we have

$$\chi^2(Q,P) = \int_{\mathbb{R}}\left(\frac{\mathrm{d}Q}{\mathrm{d}P}-1\right)^2\mathrm{d}P \leq \kappa^2.$$

Thus, taking $g: x\mapsto \exp\big\{\lambda(x-\theta_0)\big\}$ in (76) and applying Lemma 40 yields that

$$2\mathbb{E}\big[\exp\big\{\lambda(Z_1-\theta_0)\big\}\;\big|\;Z_1\neq\star\big] \leq 1+\kappa^2+2\exp\big\{(2\sigma\lambda)^{r/(r-1)}\big\},$$

for all $\lambda>0$. Hence, for $s\in[n]$,

$$\log\mathbb{E}\left\{\exp\left(\frac{\lambda}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}(Z_i-\theta_0)\right)\;\middle|\;|\mathcal{D}|=s\right\} \leq s\log\left\{\frac{1}{2}\left[1+\kappa^2+2\exp\left\{\left(\frac{2\sigma\lambda}{s}\right)^{r/(r-1)}\right\}\right]\right\}$$

$$\leq s\biggl\{\log(1+\kappa^2)+\log 2+\biggl(\frac{2\sigma\lambda}{s}\biggr)^{r/(r-1)}\biggr\}, \quad (77)$$

where the final inequality follows from the fact that $\log\bigl(\frac{a+b}{2}\bigr) \leq \log a + \log b$ for all $a, b \geq 1$. Then, applying a Chernoff bound gives that for every $t \geq 0$ and $s \in [n]$,

$$\mathbb{P}\bigl(\widehat{\theta}_n - \theta_0 \geq t \;\big|\; |\mathcal{D}| = s\bigr) = \mathbb{P}\biggl(\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}(Z_i - \theta_0) \geq t \;\biggm|\; |\mathcal{D}| = s\biggr) \leq \exp\bigl(-\psi^*(t)\bigr),$$

where

$$\begin{aligned}\psi^*(t) &:= \sup_{\lambda>0}\biggl\{\lambda t - s\log(1+\kappa^2) - s\log 2 - \frac{(2\sigma\lambda)^{r/(r-1)}}{s^{1/(r-1)}}\biggr\} \\ &= \frac{st^r}{(2\sigma)^r}\cdot\biggl(\frac{r-1}{r}\biggr)^{r-1}\cdot\frac{1}{r} - s\log(2+2\kappa^2) \geq \frac{st^r}{(2\sigma)^r}\cdot\frac{1}{er} - s\log(2+2\kappa^2),\end{aligned}$$

since the supremum over $\lambda \in (0,\infty)$ is attained at $\lambda^* := \bigl(\frac{r-1}{r}\cdot\frac{ts^{1/(r-1)}}{(2\sigma)^{r/(r-1)}}\bigr)^{r-1}$. By replacing $Z_i - \theta_0$ with $-(Z_i - \theta_0)$ for $i \in [n]$, we deduce that for every $t \geq 0$,

$$\mathbb{P}\bigl(|\widehat{\theta}_n - \theta_0| \geq t \;\big|\; |\mathcal{D}| = s\bigr) \leq 2\exp\biggl\{-\frac{st^r}{(2\sigma)^r}\cdot\frac{1}{er} + s\log(2+2\kappa^2)\biggr\}.$$

Hence, defining the event

$$\mathcal{E}_3 := \biggl\{(\widehat{\theta}_n - \theta_0)^2 \leq \biggl(\frac{(2\sigma)^r er\log(8/\delta)}{nq(1-\epsilon)/2} + (2\sigma)^r er\log(2+2\kappa^2)\biggr)^{2/r}\biggr\},$$

and proceeding in a similar fashion to (73), we deduce that $\mathbb{P}(\mathcal{E}_0\cap\mathcal{E}_3) \geq 1-\delta/2$. Moreover, on $\mathcal{E}_0\cap\mathcal{E}_3$, we have

$$\begin{aligned}(\widehat{\theta}_n - \theta_0)^2 &\leq \biggl\{\frac{(2\sigma)^r er\log(8/\delta)}{nq(1-\epsilon)/2} + (2\sigma)^r er\log(2+2\kappa^2)\biggr\}^{2/r} \\ &\leq \biggl\{\frac{(2\sigma)^r er}{4} + (2\sigma)^r er\log(2+2\kappa^2)\biggr\}^{2/r} \leq \biggl\{\frac{3}{2}\cdot(2\sigma)^r er\log(2+2\kappa^2)\biggr\}^{2/r} \\ &\leq \bigl\{3\cdot(2\sigma)^r er\log(2+2\kappa)\bigr\}^{2/r} \leq (9e\sigma)^2\log^{2/r}(2+2\kappa). \quad (78)\end{aligned}$$

Thus, on $\mathcal{E}_0\cap\mathcal{E}_2\cap\mathcal{E}_3$, which satisfies $\mathbb{P}(\mathcal{E}_0\cap\mathcal{E}_2\cap\mathcal{E}_3) \geq 1-\delta$, we combine (75) and (78) to obtain the desired result for $r > 1$.

Finally, we consider the case where $r = 1$. By Zhivotovskiy (2024, Lemma 2.5), (76) yields, with the same choice of Q and g, that

$$2\mathbb{E}\bigl[\exp\bigl\{\lambda(Z_1 - \theta_0)\bigr\} \;\big|\; Z_1 \neq \star\bigr] \leq 1 + \kappa^2 + \exp\bigl\{(2\sigma\lambda)^2\bigr\},$$

for all $|\lambda| \leq \frac{1}{2\sigma}$. Hence, by a similar argument to the $r > 1$ case,

$$\log\mathbb{E}\biggl\{\exp\biggl(\frac{\lambda}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}(Z_i - \theta_0)\biggr) \;\biggm|\; |\mathcal{D}| = s\biggr\} \leq s\biggl\{\log(1+\kappa^2) + \biggl(\frac{2\sigma\lambda}{s}\biggr)^2\biggr\}, \quad (79)$$

for $s \in [n]$ and $|\lambda| \le \frac{s}{2\sigma}$. Then, applying a Chernoff bound yields

$$\mathbb{P}\big(|\widehat{\theta}_n - \theta_0| \ge t \;\big|\; |\mathcal{D}| = s\big) = \mathbb{P}\bigg(\Big|\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}(Z_i - \theta_0)\Big| \ge t \;\bigg|\; |\mathcal{D}| = s\bigg) \le 2\exp\big(-\psi^*(t)\big),$$

where

$$\psi^*(t) := \sup_{0<\lambda\le\frac{s}{2\sigma}} \bigg\{\lambda t - s\log(1+\kappa^2) - \frac{(2\sigma\lambda)^2}{s}\bigg\}.$$

Taking $t := 8\sigma\log(2+2\kappa^2)$ and $\lambda = s/(2\sigma)$ yields, for $s \ge nq(1-\epsilon)/2$, that

$$\mathbb{P}\big(|\widehat{\theta}_n - \theta_0| \ge t \;\big|\; |\mathcal{D}| = s\big) \le 2\exp\big\{-3s\log(2+2\kappa^2) + s\big\} \le 2\exp(-s) \le \frac{\delta}{4},$$

where the final inequality follows from the assumption $nq(1-\epsilon) \ge 8\log(8/\delta)$. Therefore, letting $\mathcal{E}_4 := \big\{|\widehat{\theta}_n - \theta_0| < 8\sigma\log(2+2\kappa^2)\big\}$, we have $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_4) \ge 1 - \delta/2$. Moreover, on $\mathcal{E}_0 \cap \mathcal{E}_4$,

$$(\widehat{\theta}_n - \theta_0)^2 < 64\sigma^2\log^2(2+2\kappa^2) \le 256\sigma^2\log^2(2+2\kappa). \tag{80}$$

Thus, on $\mathcal{E}_0 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, which satisfies $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_2 \cap \mathcal{E}_4) \ge 1 - \delta$, we combine (75) and (80) to obtain the desired result for $r = 1$. □

D.2.2 Proof of Theorem 11

For $\theta \in \mathbb{R}$ and $K > 0$, define

$$\mathcal{P}_{\mathrm{b}}(\theta, K) := \Big\{P \in \mathcal{P}(\mathbb{R}) : \mathbb{E}_P(X) = \theta, P \text{ is supported on an interval of length at most } K\Big\}. \tag{81}$$

Proof of Theorem 11. (a) Define $a := \frac{q(1-\epsilon)}{q(1-\epsilon)+\epsilon} \in (0,1]$ and $b := \frac{\sigma}{2}\cdot a^{-1/r} > 0$. Let $X_1 \sim P_1$ and $X_2 \sim P_2$ be random variables satisfying

$$X_1 = \begin{cases} -b & \text{with probability } \frac{1}{a+1} \\ b & \text{with probability } \frac{a}{a+1} \end{cases} \quad \text{and} \quad X_2 = \begin{cases} -b & \text{with probability } \frac{a}{a+1} \\ b & \text{with probability } \frac{1}{a+1}. \end{cases}$$

Then $\theta_1 := \mathbb{E}(X_1) = -\frac{(1-a)b}{a+1}$ and $\theta_2 := \mathbb{E}(X_2) = \frac{(1-a)b}{a+1}$. Moreover,

$$\mathbb{E}\big(|X_1 - \theta_1|^r\big) = \frac{(2ab)^r + a(2b)^r}{(a+1)^{r+1}} \le a\cdot(2b)^r = \sigma^r,$$

where we have used the fact that $a^r + a \le a(a+1) \le a(a+1)^{r+1}$; by symmetry, $\mathbb{E}\big(|X_2 - \theta_2|^r\big) \le \sigma^r$. Consequently, $P_1 \in \mathcal{P}_{L^r}(\theta_1, \sigma^2)$ and $P_2 \in \mathcal{P}_{L^r}(\theta_2, \sigma^2)$. Now define $R_0 \in \mathcal{P}(\mathbb{R}_\star)$ by

$$R_0(\{-b\}) := \frac{q(1-\epsilon)}{a+1} =: R_0(\{b\}) \quad \text{and} \quad R_0(\{\star\}) := 1 - R_0(\{-b\}) - R_0(\{b\}) \in [0,1).$$

By Proposition 2, $R_0 \in \mathcal{R}(P_1, \epsilon, q) \cap \mathcal{R}(P_2, \epsilon, q)$. Therefore, by Ma, Verchand and Samworth (2024, Theorem 4 and Lemma 5),

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \frac{(\theta_2 - \theta_1)^2}{4} = \biggl\{\frac{(1-a)b}{a+1}\biggr\}^2 = \frac{\sigma^2}{4} \cdot \biggl(\frac{\epsilon}{2q(1-\epsilon)+\epsilon}\biggr)^2 \biggl(\frac{q(1-\epsilon)+\epsilon}{q(1-\epsilon)}\biggr)^{2/r}$$
$$\geq \frac{\sigma^2}{36} \cdot \biggl\{\biggl(\frac{\epsilon}{q(1-\epsilon)}\biggr)^2 \wedge \biggl(\frac{\epsilon}{q(1-\epsilon)}\biggr)^{2/r}\biggr\},$$

where the final bound is obtained by considering separately the cases $\epsilon \leq q(1-\epsilon)$ and $\epsilon > q(1-\epsilon)$. This proves the second term in the lower bound.

For the first term in the lower bound, we observe that $\mathcal{P}_{\mathrm{b}}(\theta, \sigma) \subseteq \mathcal{P}_{L^r}(\theta, \sigma^2)$ for all $r \geq 2$. We therefore obtain the desired conclusion by choosing the contamination distribution $Q \in \mathcal{P}(\mathbb{R}_\star)$ such that $Q\bigl(\{\star\}\bigr) = 1$ and applying Proposition 46*(b)*.

(b) Define $P_1, P_2 \in \mathcal{P}(\mathbb{R})$ with Lebesgue densities p_1, p_2 respectively as in Lemma 27, so that $P_1 \in \mathcal{P}_{\psi_r}\bigl(\mathbb{E}_{P_1}(X_1), \sigma^2\bigr)$ and $P_2 \in \mathcal{P}_{\psi_r}\bigl(\mathbb{E}_{P_2}(X_2), \sigma^2\bigr)$. Further, with $b > 0$ defined as in Lemma 27, define $R_1 \in \mathcal{P}(\mathbb{R}_\star)$ through its Radon–Nikodym derivative

$$\frac{\mathrm{d}R_1}{\mathrm{d}\lambda_\star}(z) := \begin{cases} q(1-\epsilon) \cdot p_1(z) & \text{if } z \in (-\infty, b) \\ \bigl\{q(1-\epsilon)+\epsilon\bigr\} \cdot p_1(z) & \text{if } z \in [b, \infty) \\ 1 - q(1-\epsilon) \cdot \int_{-\infty}^b p_1(x)\,\mathrm{d}x - \bigl\{q(1-\epsilon)+\epsilon\bigr\} \cdot \int_b^\infty p_1(x)\,\mathrm{d}x & \text{if } z = \star, \end{cases}$$

so that, by Proposition 2, $R_1 \in \mathcal{R}(P_1, \epsilon, q) \cap \mathcal{R}(P_2, \epsilon, q)$. Therefore, by Ma, Verchand and Samworth (2024, Theorem 4 and Lemma 5),

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \frac{\bigl\{\mathbb{E}_{P_2}(X_2) - \mathbb{E}_{P_1}(X_1)\bigr\}^2}{4}. \tag{82}$$

Now, writing $\sigma_0 := \sigma / C_0$,

$$\begin{aligned} \mathbb{E}_{P_2}(X_2) - \mathbb{E}_{P_1}(X_1) &= \frac{\epsilon}{q(1-\epsilon)} \int_b^\infty x p_1(x)\,\mathrm{d}x - \frac{\epsilon}{q(1-\epsilon)+\epsilon} \int_0^b x p_1(x)\,\mathrm{d}x \\ &\overset{(i)}{=} \frac{\epsilon}{q(1-\epsilon)} \biggl\{b e^{-(b/\sigma_0)^r} + \int_b^\infty e^{-(x/\sigma_0)^r}\,\mathrm{d}x\biggr\} \\ &\quad - \frac{\epsilon}{q(1-\epsilon)+\epsilon} \biggl\{-b e^{-(b/\sigma_0)^r} + \int_0^b e^{-(x/\sigma_0)^r}\,\mathrm{d}x\biggr\} \\ &\overset{(ii)}{=} \biggl(\frac{\epsilon}{q(1-\epsilon)} + \frac{\epsilon}{q(1-\epsilon)+\epsilon}\biggr) \cdot \frac{q(1-\epsilon)}{2q(1-\epsilon)+\epsilon} \cdot \sigma_0 \log^{1/r}\biggl(2 + \frac{\epsilon}{q(1-\epsilon)}\biggr) \\ &\quad + \frac{\epsilon}{q(1-\epsilon)} \int_b^\infty e^{-(x/\sigma_0)^r}\,\mathrm{d}x - \frac{\epsilon}{q(1-\epsilon)+\epsilon} \int_0^b e^{-(x/\sigma_0)^r}\,\mathrm{d}x, \end{aligned} \tag{83}$$

where (i) follows from integration by parts, (ii) follows by substituting the definition of b. Now let $h(t) := \frac{1}{t}\int_0^t e^{-(x/\sigma_0)^r}\,\mathrm{d}x$. Then

$$h'(t) = \frac{t e^{-(t/\sigma_0)^r} - \int_0^t e^{-(x/\sigma_0)^r}\,\mathrm{d}x}{t^2} \leq 0,$$

so h is a decreasing function.

First consider the case where $\frac{\epsilon}{q(1-\epsilon)} \geq e^{2^r} - 2$ or equivalently $\epsilon \geq \frac{\{\exp(2^r)-2\}q}{1+\{\exp(2^r)-2\}q}$, so that $\log^{1/r}\big(2 + \frac{\epsilon}{q(1-\epsilon)}\big) \geq 2$ and

$$h(b) \leq h(2\sigma_0) = \frac{\int_0^2 e^{-x^r}\,\mathrm{d}x}{2} = \frac{\int_0^1 e^{-x^r}\,\mathrm{d}x + \int_1^2 e^{-x^r}\,\mathrm{d}x}{2} \leq \frac{1+e^{-1}}{2}.$$

Hence, by (83),

$$\begin{aligned}
\mathbb{E}_{P_2}(X_2) - \mathbb{E}_{P_1}(X_1) &\geq \left(\frac{\epsilon}{q(1-\epsilon)} + \frac{\epsilon}{q(1-\epsilon)+\epsilon}\right) \cdot \frac{q(1-\epsilon)}{2q(1-\epsilon)+\epsilon} \cdot \sigma_0 \log^{1/r}\left(2 + \frac{\epsilon}{q(1-\epsilon)}\right) \\
&\quad - \frac{\epsilon}{q(1-\epsilon)+\epsilon} \cdot \frac{1+e^{-1}}{2} \cdot \sigma_0 \log^{1/r}\left(2 + \frac{\epsilon}{q(1-\epsilon)}\right) \\
&= \frac{1-e^{-1}}{2} \cdot \frac{\epsilon}{q(1-\epsilon)+\epsilon} \cdot \sigma_0 \log^{1/r}\left(2 + \frac{\epsilon}{q(1-\epsilon)}\right) \\
&\geq \frac{1-e^{-1}}{8} \cdot \frac{\sigma}{C_0} \cdot \log^{1/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)}\right).
\end{aligned}$$

Therefore, by (82), when $\epsilon \geq \frac{\{\exp(2^r)-2\}q}{1+\{\exp(2^r)-2\}q}$,

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \gtrsim \sigma^2 \log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)}\right). \tag{84}$$

Next consider the case where $\frac{\epsilon}{q(1-\epsilon)} \leq 1$, or equivalently $\epsilon \leq q/(1+q)$. Define $P_3, P_4 \in \mathcal{P}(\mathbb{R})$ by

$$P_3\left(\left\{-\frac{\sigma}{4}\right\}\right) := \frac{q(1-\epsilon)}{2q(1-\epsilon)+\epsilon} =: P_4\left(\left\{\frac{\sigma}{4}\right\}\right), \; P_3\left(\left\{\frac{\sigma}{4}\right\}\right) := \frac{q(1-\epsilon)+\epsilon}{2q(1-\epsilon)+\epsilon} =: P_4\left(\left\{-\frac{\sigma}{4}\right\}\right).$$

Thus $P_3 \in \mathcal{P}_{\psi_r}\big(\mathbb{E}_{P_3}(X_3), \sigma^2\big)$ and $P_4 \in \mathcal{P}_{\psi_r}\big(\mathbb{E}_{P_4}(X_4), \sigma^2\big)$. Further define $R_2 \in \mathcal{P}(\mathbb{R}_\star)$ by

$$\begin{aligned}
R_2\left(\left\{-\frac{\sigma}{4}\right\}\right) &:= \frac{q(1-\epsilon)\{q(1-\epsilon)+\epsilon\}}{2q(1-\epsilon)+\epsilon} =: R_2\left(\left\{\frac{\sigma}{4}\right\}\right), \\
R_2(\{\star\}) &:= 1 - \frac{2q(1-\epsilon)\{q(1-\epsilon)+\epsilon\}}{2q(1-\epsilon)+\epsilon}.
\end{aligned}$$

By Proposition 2, $R_2 \in \mathcal{R}(P_3, \epsilon, q) \cap \mathcal{R}(P_4, \epsilon, q)$. Therefore, by Ma, Verchand and Samworth (2024, Theorem 4 and Lemma 5), when $\epsilon \leq \frac{q}{1+q}$,

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \frac{\big\{\mathbb{E}_{P_3}(X_3) - \mathbb{E}_{P_4}(X_4)\big\}^2}{4} \geq \frac{\sigma^2}{16}\left(\frac{\epsilon}{2q(1-\epsilon)+\epsilon}\right)^2 \geq \frac{\sigma^2}{144}\left(\frac{\epsilon}{q(1-\epsilon)}\right)^2. \tag{85}$$

Combining (84) and (85) yields that when $\epsilon \leq \frac{q}{1+q}$ or $\epsilon \geq \frac{\{\exp(2^r)-2\}q}{1+\{\exp(2^r)-2\}q}$,

$$\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \gtrsim \sigma^2 \cdot \left\{\left(\frac{\epsilon}{q(1-\epsilon)}\right)^2 \wedge \log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)}\right)\right\}. \tag{86}$$

Further observe that $\mathcal{R}(P, \frac{q}{1+q}, q) \subseteq \mathcal{R}(P, \epsilon, q)$ for all $P \in \mathcal{P}(\mathbb{R})$ when $\epsilon > \frac{q}{1+q}$. Thus, by (86), we deduce that when $\frac{q}{1+q} < \epsilon < \frac{\{\exp(2^r)-2\}q}{1+\{\exp(2^r)-2\}q}$,

$$\begin{aligned}\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) &\gtrsim \sigma^2 \cdot \left\{ \left(\frac{\frac{q}{1+q}}{q(1-\frac{q}{1+q})} \right)^2 \wedge \log^{2/r}\left(2 + \frac{\frac{2q}{1+q}}{q(1-\frac{q}{1+q})}\right) \right\} \\ &= \sigma^2 \gtrsim \sigma^2 \cdot \left\{ \left(\frac{\epsilon}{q(1-\epsilon)} \right)^2 \wedge \log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)}\right) \right\}, \end{aligned} \tag{87}$$

where the last inequality follows from the fact that when $\epsilon < \frac{\{\exp(2^r)-2\}q}{1+\{\exp(2^r)-2\}q}$,

$$\log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)}\right) < \log^{2/r}\big(2 + 2(e^{2^r} - 2)\big) \le \log^{2/r}\big(e^{2^r + \log 2}\big) \le (2 \cdot 2^r)^{2/r} = 2^{2/r} \cdot 2 \le 8.$$

Combining (86) and (87) yields the second term in the lower bound.

For the first term in the lower bound, we observe that $\mathcal{P}_{\mathrm{b}}(\theta, \sigma/2) \subseteq \mathcal{P}_{\psi_r}(\theta, \sigma^2)$ for all $r \ge 1$. We therefore obtain the desired conclusion by choosing the contamination distribution $Q \in \mathcal{P}(\mathbb{R}_\star)$ such that $Q\big(\{\star\}\big) = 1$ and applying Proposition 46*(b)*. □

Lemma 27. *Let $\epsilon \in [0, 1)$, $q \in (0, 1]$, $\sigma > 0$ and $r \ge 1$. There exists a universal constant $C_0 > 0$ such that if $X_1 \sim P_1 \in \mathcal{P}(\mathbb{R})$ and $X_2 \sim P_2 \in \mathcal{P}(\mathbb{R})$ have Lebesgue densities p_1 and p_2 respectively, where $p_1(x) := \frac{r x^{r-1}}{(\sigma/C_0)^r} e^{-(C_0 x/\sigma)^r} \mathbb{1}_{\{x \ge 0\}}$ and*

$$p_2(x) := \begin{cases} \frac{q(1-\epsilon)}{q(1-\epsilon)+\epsilon} \cdot p_1(x) & \text{if } x < b \\ \frac{q(1-\epsilon)+\epsilon}{q(1-\epsilon)} \cdot p_1(x) & \text{if } x \ge b \end{cases} \quad \text{with} \quad b := \frac{\sigma}{C_0} \log^{1/r}\left(2 + \frac{\epsilon}{q(1-\epsilon)}\right),$$

then $\|X_1 - \mathbb{E}X_1\|_{\psi_r} \vee \|X_2 - \mathbb{E}X_2\|_{\psi_r} \le \sigma$.

Proof. Since $\mathbb{P}(|X_1| \ge x) = e^{-(C_0 x/\sigma)^r}$ for all $x \ge 0$, we have by Vershynin (2018, Proposition 2.7.1) that $\|X_1^r\|_{\psi_1} \le C_1(\sigma/C_0)^r$ for some universal constant $C_1 > 0$, so $\|X_1\|_{\psi_r} \le C_1^{1/r}\sigma/C_0 \le (C_1 \vee 1)\sigma/C_0$. Then, by Götze, Sambale and Sinulis (2021, Lemma A.3), we have

$$\|X_1 - \mathbb{E}X_1\|_{\psi_r} \le \left\{1 + \left(\frac{2}{(re)^{1/r}\log 2}\right)^{1/r}\right\}(C_1 \vee 1)\frac{\sigma}{C_0} \le 4(C_1 \vee 1)\frac{\sigma}{C_0}.$$

Turning to X_2, first observe that p_2 is a Lebesgue density, since

$$\int_{\mathbb{R}} p_2(x)\,\mathrm{d}x = \frac{q(1-\epsilon)}{q(1-\epsilon)+\epsilon}\{1 - e^{-(C_0 b/\sigma)^r}\} + \frac{q(1-\epsilon)+\epsilon}{q(1-\epsilon)} e^{-(C_0 b/\sigma)^r} = 1.$$

Now, for $x \ge 0$, we have

$$\begin{aligned}\mathbb{P}(X_2 - b \ge x) = \frac{q(1-\epsilon)+\epsilon}{q(1-\epsilon)} \cdot \mathbb{P}(X_1 \ge b + x) &\le \frac{q(1-\epsilon)+\epsilon}{q(1-\epsilon)} \cdot e^{-(C_0 b/\sigma)^r - (C_0 x/\sigma)^r} \\ &= \frac{q(1-\epsilon)+\epsilon}{2q(1-\epsilon)+\epsilon} \cdot e^{-(C_0 x/\sigma)^r} \le e^{-(C_0 x/\sigma)^r}.\end{aligned}$$

Define $a \coloneqq \frac{q(1-\epsilon)}{q(1-\epsilon)+\epsilon} \in (0,1]$, so that $b = (\sigma/C_0)\log^{1/r}\big(\frac{1+a}{a}\big)$. For $x \in [0,b]$, we have

$$\mathbb{P}(X_2 - b \le -x) = a \cdot \mathbb{P}(X_1 \le b - x) \le a \le \frac{2a}{1+a} = 2e^{-(C_0 b/\sigma)^r} \le 2e^{-(C_0 x/\sigma)^r}.$$

For $x > b$, we have $\mathbb{P}(X_2 - b \le -x) = 0$. Combining these inequalities, we obtain $\mathbb{P}(|X_2 - b| \ge x) \le 3e^{-(C_0 x/\sigma)^r}$ for all $x \ge 0$. Therefore, by Vershynin (2018, Proposition 2.7.1), we deduce[7] that $\|(X_2 - b)^r\|_{\psi_1} \le C_2(\sigma/C_0)^r$ for some universal constant $C_2 > 0$, so

$$\|X_2 - b\|_{\psi_r} \le C_2^{1/r}\frac{\sigma}{C_0} \le (C_2 \vee 1)\frac{\sigma}{C_0}.$$

By Götze, Sambale and Sinulis (2021, Lemma A.3) again, $\|X_2 - \mathbb{E}X_2\|_{\psi_r} \le 4(C_2 \vee 1)\sigma/C_0$. Finally, taking $C_0 \coloneqq 4C_1 \vee 4C_2 \vee 4$ completes the proof. □

D.2.3 Proof of Theorem 12

The following proposition, which is analogous to Proposition 26 in the univariate case, will be used in the proof of Theorem 12.

Proposition 28. *Let $\theta_0 \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}_{++}^{d\times d}$, $\epsilon \in [0,1)$, $\delta \in (0,1]$, $\pi \in \mathcal{P}\big(\{\emptyset, [d]\}\big)$ and $q \coloneqq \pi([d])$.*

(a) Let $r \ge 2$, $P \in \mathcal{P}_{d,L^r}(\theta_0, \Sigma)$ and $Z \sim R \in \mathcal{R}_{\emptyset,[d]}(P, \epsilon, \pi)$. Then

$$\big\|\mathbb{E}(Z \,|\, Z \in \mathbb{R}^d) - \theta_0\big\|_2^2 \le \|\Sigma\|_{\mathrm{op}}\left\{\left(\frac{\epsilon}{q(1-\epsilon)}\right)^2 \wedge \left(\frac{\epsilon}{q(1-\epsilon)}\right)^{2/r}\right\}.$$

(b) Let $r \ge 1$, $P \in \mathcal{P}_{d,\psi_r}(\theta_0, \Sigma)$, and $Z \sim R \in \mathcal{R}_{\emptyset,[d]}(P, \epsilon, \pi)$. Then

$$\big\|\mathbb{E}(Z \,|\, Z \in \mathbb{R}^d) - \theta_0\big\|_2^2 \le \|\Sigma\|_{\mathrm{op}}\left\{4\left(\frac{\epsilon}{q(1-\epsilon)}\right)^2 \wedge \log^{2/r}\left(2 + \frac{2\epsilon}{q(1-\epsilon)}\right)\right\}.$$

Proof. Let $\kappa \coloneqq \frac{\epsilon}{q(1-\epsilon)}$, $X \sim P$, $v \in \mathbb{S}^{d-1}$, $Z^{(v)} \coloneqq v^\top Z \cdot \mathbb{1}_{\{Z\in\mathbb{R}^d\}} + \star \cdot \mathbb{1}_{\{Z\notin\mathbb{R}^d\}}$, $R^{(v)} \coloneqq \mathsf{Law}(Z^{(v)})$ and $P^{(v)} \coloneqq \mathsf{Law}(v^\top X)$. By Lemma 25, we have $R^{(v)} \in \mathcal{R}(P^{(v)}, \epsilon, q)$.

(a) Since $P^{(v)} \in \mathcal{P}_{L^r}(v^\top\theta_0, v^\top\Sigma v)$ we have by Proposition 26*(a)* that

$$\begin{aligned}
\big\|\mathbb{E}(Z \,|\, Z \in \mathbb{R}^d) - \theta_0\big\|_2^2 &= \sup_{v\in\mathbb{S}^{d-1}}\big\{v^\top\mathbb{E}(Z|Z \in \mathbb{R}^d) - v^\top\theta_0\big\}^2 \\
&= \sup_{v\in\mathbb{S}^{d-1}}\big\{\mathbb{E}(Z^{(v)}|Z^{(v)} \ne \star) - v^\top\theta_0\big\}^2 \\
&\le \sup_{v\in\mathbb{S}^{d-1}} v^\top\Sigma v \cdot (\kappa^2 \wedge \kappa^{2/r}) \\
&= \|\Sigma\|_{\mathrm{op}}(\kappa^2 \wedge \kappa^{2/r}),
\end{aligned}$$

as required.

(b) We now have $P^{(v)} \in \mathcal{P}_{\psi_r}(v^\top\theta_0, v^\top\Sigma v)$, so the proof is the same as part *(a)*, except that we use Proposition 26*(b)* instead. □

[7] Note that in Vershynin (2018, Proposition 2.7.1*(a)*), the condition is that $\mathbb{P}(|X| \ge t) \le 2\exp(-t/K_1)$ for all $t \ge 0$. However, the result is still true if we replace the factor 2 by 3. See for example, Vershynin (2018, Proposition 2.5.2) for the proof strategy.

Proof of Theorem 12. Let $\kappa \coloneqq \frac{\epsilon}{q(1-\epsilon)}$.

(a) For $v \in \mathbb{S}^{d-1}$, we have by the same argument as the proof of (72) that

$$\mathrm{Var}(v^\top Z_1 \,|\, Z_1 \in \mathbb{R}^d) \le (1+\kappa^{2/r}) v^\top \Sigma v.$$

Therefore, writing $\Gamma \coloneqq \mathrm{Cov}(Z_1 \,|\, Z_1 \in \mathbb{R}^d) \in \mathcal{S}_+^{d\times d}$, we have

$$\|\Gamma\|_{\mathrm{op}} \le (1+\kappa^{2/r})\|\Sigma\|_{\mathrm{op}} \quad \text{and} \quad \mathrm{tr}(\Gamma) \le (1+\kappa^{2/r})\,\mathrm{tr}(\Sigma). \tag{88}$$

By Lemma 38*(b)*, the event $\mathcal{E}_0 \coloneqq \{|\mathcal{D}| \ge nq(1-\epsilon)/2\}$ has $\mathbb{P}(\mathcal{E}_0) \ge 1-\delta/2$, since $nq(1-\epsilon) \ge 8\log(2/\delta)$. Moreover, writing

$$\mathcal{E}_1 \coloneqq \left\{ \left\|\widehat{\theta}_n - \mathbb{E}(Z_1 \,|\, Z_1 \in \mathbb{R}^d)\right\|_2^2 \le C_1 \cdot \frac{\mathrm{tr}(\Gamma) + \|\Gamma\|_{\mathrm{op}} \log(2/\delta)}{nq(1-\epsilon)} \right\},$$

we have by Depersin and Lecué (2022b, Theorem 2.1) that when $C_1 > 0$ is a sufficiently large universal constant, we have $\mathbb{P}\big(\mathcal{E}_1 \,\big|\, |\mathcal{D}| = s\big) \ge 1-\delta/2$ for $s \ge nq(1-\epsilon)/2$, so $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_1) \ge 1-\delta$. On the event $\mathcal{E}_0 \cap \mathcal{E}_1$, we have by (88) and Proposition 28*(a)* that

$$\begin{aligned}
\|\widehat{\theta}_n - \theta_0\|_2^2 &\le 2\left\|\widehat{\theta}_n - \mathbb{E}(Z_1 \,|\, Z_1 \in \mathbb{R}^d)\right\|_2^2 + 2\left\|\mathbb{E}(Z_1 \,|\, Z_1 \in \mathbb{R}^d) - \theta_0\right\|_2^2 \\
&\lesssim \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}} \log(2/\delta)}{nq(1-\epsilon)} + \frac{\mathbf{r}(\Sigma) + \log(2/\delta)}{nq(1-\epsilon)} \cdot \|\Sigma\|_{\mathrm{op}} \kappa^{2/r} + \|\Sigma\|_{\mathrm{op}}(\kappa^2 \wedge \kappa^{2/r}) \\
&\lesssim \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}} \log(2/\delta)}{nq(1-\epsilon)} + \|\Sigma\|_{\mathrm{op}}(\kappa^2 \wedge \kappa^{2/r}),
\end{aligned}$$

where the final inequality follows by considering separately the cases $\kappa \le 1$ and $\kappa > 1$, and in the second case noting that $\mathbf{r}(\Sigma) \le Cnq(1-\epsilon)$ and $\log(2/\delta) \le nq(1-\epsilon)/C$.

(b) For $v \in \mathbb{S}^{d-1}$, we have by the same argument as in the proof of (74) that conditional on $\{Z_1 \in \mathbb{R}^d\}$,

$$\left\|v^\top Z_1 - \mathbb{E}(v^\top Z_1 \,|\, Z_1 \in \mathbb{R}^d)\right\|_{\psi_1} \le \sqrt{(1+\kappa) v^\top \Sigma v},$$

so that $Z_1 \,|\, \{Z_1 \in \mathbb{R}^d\} \in \mathcal{P}_{d,\psi_1}\big(\mathbb{E}(v^\top Z_1 \,|\, Z_1 \in \mathbb{R}^d), (1+\kappa)\Sigma\big)$. By Lemma 38*(b)*, the event $\mathcal{E}_0 \coloneqq \{|\mathcal{D}| \ge nq(1-\epsilon)/2\}$ satisfies $\mathbb{P}(\mathcal{E}_0) \ge 1-\delta/4$ since $nq(1-\epsilon) \ge 8\log(8/\delta)$. Moreover, writing

$$\mathcal{E}_2 \coloneqq \left\{ \left\|\widehat{\theta}_n - \mathbb{E}(Z_1 \,|\, Z_1 \in \mathbb{R}^d)\right\|_2^2 \le 48(1+\kappa) \cdot \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}} \log(8/\delta)}{nq(1-\epsilon)} \right\},$$

we have by Lemma 42 (a consequence of the PAC–Bayes lemma) that $\mathbb{P}\big(\mathcal{E}_2 \,\big|\, |\mathcal{D}| = s\big) \ge 1-\delta/4$ for $s \ge nq(1-\epsilon)/2$, so $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_2) \ge 1-\delta/2$. On $\mathcal{E}_0 \cap \mathcal{E}_2$, we have by Proposition 28*(b)* that

$$\begin{aligned}
\|\widehat{\theta}_n - \theta_0\|_2^2 &\le 2\left\|\widehat{\theta}_n - \mathbb{E}(Z_1 \,|\, Z_1 \in \mathbb{R}^d)\right\|_2^2 + 2\left\|\mathbb{E}(Z_1 \,|\, Z_1 \in \mathbb{R}^d) - \theta_0\right\|_2^2 \\
&\lesssim \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}} \log(8/\delta)}{nq(1-\epsilon)} + \|\Sigma\|_{\mathrm{op}} \kappa \cdot \frac{\mathbf{r}(\Sigma) + \log(8/\delta)}{nq(1-\epsilon)} + \|\Sigma\|_{\mathrm{op}} \kappa^2 \\
&\lesssim \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}} \log(8/\delta)}{nq(1-\epsilon)} + \|\Sigma\|_{\mathrm{op}} \kappa^2,
\end{aligned} \tag{89}$$

where the final inequality follows by considering separately the cases $\kappa \le 1$ and $\kappa > 1$, and in the second case noting that $\mathbf{r}(\Sigma) \le nq(1-\epsilon)$ and $\log(8/\delta) \le nq(1-\epsilon)/8$.

For the last term in the upper bound, we first consider the case where $r > 1$. For $w \in \mathbb{R}^d$, we have by the same argument as in the proof of (77) that

$$\log \mathbb{E}\big\{\exp\big(\lambda w^\top (Z_1 - \theta_0)\big) \,\big|\, Z_1 \in \mathbb{R}^d\big\} \le \log(1+\kappa^2) + \log 2 + \big(2\lambda\sqrt{w^\top \Sigma w}\big)^{r/(r-1)}, \tag{90}$$

for all $\lambda > 0$. Let $\beta := \mathbf{r}(\Sigma)$, let μ denote the distribution of $\mathsf{N}_d(0, \beta^{-1}\Sigma)$ and for $u \in \Sigma^{1/2}\mathbb{S}^{d-1}$, let ρ_u denote the conditional distribution of Y given $\big\{\|Y-u\|_2 \le 2\|\Sigma\|_{\mathrm{op}}^{1/2}\big\}$, where $Y \sim \mathsf{N}_d(u, \beta^{-1}\Sigma)$. By Chebychev's inequality,

$$\mathbb{P}\big(\|Y-u\|_2 \ge 2\|\Sigma\|_{\mathrm{op}}^{1/2}\big) \le \frac{\mathrm{tr}(\Sigma)}{4\beta\|\Sigma\|_{\mathrm{op}}} = \frac{1}{4}.$$

Hence, by the third displayed equation of Zhivotovskiy (2024, p. 11), we have

$$\mathrm{KL}(\rho_u, \mu) = \log\biggl(\frac{1}{\mathbb{P}\big(\|Y-u\|_2 \le 2\|\Sigma\|_{\mathrm{op}}^{1/2}\big)}\biggr) + \frac{\beta}{2} \le 2\log 2 + \frac{\mathbf{r}(\Sigma)}{2}.$$

Fix $u \in \Sigma^{1/2}\mathbb{S}^{d-1}$, let $v \in \mathbb{R}^d$ be such that $\|v-u\|_2 \le 2\|\Sigma\|_{\mathrm{op}}^{1/2}$, and for $\lambda > 0$, define $f_\lambda : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by $f_\lambda(x,y) := \lambda y^\top \Sigma^{-1/2}(x-\theta_0)$. Then, since $\|v\|_2 \le 3\|\Sigma\|_{\mathrm{op}}^{1/2}$, we have by (90) that

$$\log \mathbb{E}_{Z\sim R}\big(e^{f_\lambda(Z,v)} \,\big|\, Z \in \mathbb{R}^d\big) \le \log(2+2\kappa^2) + \big(6\lambda\|\Sigma\|_{\mathrm{op}}^{1/2}\big)^{r/(r-1)},$$

so $\mathbb{E}_{\xi_u\sim\rho_u}\big\{\log \mathbb{E}_{Z\sim R}\big(e^{f_\lambda(Z,\xi_u)} \,\big|\, Z \in \mathbb{R}^d\big)\big\} \le \log(2+2\kappa^2) + \big(6\lambda\|\Sigma\|_{\mathrm{op}}^{1/2}\big)^{r/(r-1)}$. Therefore, for $s \ge nq(1-\epsilon)/2$, by the PAC–Bayes lemma (Lemma 41), conditional on $|\mathcal{D}| = s$, we have with probability at least $1-\delta/4$ that

$$\begin{aligned}
\biggl\|\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} Z_i - \theta_0\biggr\|_2 &= \sup_{u\in\Sigma^{1/2}\mathbb{S}^{d-1}} \frac{1}{\lambda|\mathcal{D}|}\sum_{i\in\mathcal{D}} \mathbb{E}_{\xi_u\sim\rho_u} f_\lambda(Z_i, \xi_u) \\
&\le \inf_{\lambda>0}\biggl\{\frac{\log(2+2\kappa^2)}{\lambda} + \big(6\|\Sigma\|_{\mathrm{op}}^{1/2}\big)^{r/(r-1)}\lambda^{1/(r-1)} + \frac{\mathbf{r}(\Sigma)/2 + 2\log(4/\delta)}{s\lambda}\biggr\} \\
&\overset{(i)}{\le} 12\|\Sigma\|_{\mathrm{op}}^{1/2}\biggl\{\log(2+2\kappa^2) + \frac{\mathbf{r}(\Sigma)/2 + 2\log(4/\delta)}{s}\biggr\}^{1/r} \\
&\overset{(ii)}{\le} 12\|\Sigma\|_{\mathrm{op}}^{1/2}\big\{\log(2+2\kappa^2)+2\big\}^{1/r} \lesssim \|\Sigma\|_{\mathrm{op}}^{1/2}\log^{1/r}(2+2\kappa),
\end{aligned}$$

where (i) follows by choosing $\lambda = \frac{1}{6\|\Sigma\|_{\mathrm{op}}^{1/2}}\big\{\log(2+2\kappa^2) + \frac{\mathbf{r}(\Sigma)/2+2\log(4/\delta)}{s}\big\}^{(r-1)/r}$ and (ii) follows from the assumptions that $nq(1-\epsilon) \ge \mathbf{r}(\Sigma)$ and $\delta \ge 8\exp\big(-nq(1-\epsilon)/8\big)$. Hence, there exists a universal constant $C_1 > 0$ such that the event

$$\mathcal{E}_3 := \biggl\{\biggl\|\frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} Z_i - \theta_0\biggr\|_2^2 \le C_1\|\Sigma\|_{\mathrm{op}}\log^{2/r}(2+2\kappa)\biggr\}, \tag{91}$$

satisfies $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_3) \geq 1 - \delta/2$. Thus, on the event $\mathcal{E}_0 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, which has probability at least $1 - \delta$, we combine (89) and (91) to obtain the desired result for $r > 1$.

Finally, we consider the case where $r = 1$. For $w \in \mathbb{R}^d$, we have by the same argument as the proof of (79) that

$$\log \mathbb{E}\big\{\exp\big(\lambda w^\top (Z_1 - \theta_0)\big) \,\big|\, Z_1 \in \mathbb{R}^d\big\} \leq \log(1 + \kappa^2) + \log 2 + \big(2\lambda\sqrt{w^\top \Sigma w}\big)^2,$$

for $|\lambda| \leq \frac{1}{2}\|\Sigma\|_{\mathrm{op}}^{-1/2} \leq \frac{1}{2\sqrt{w^\top \Sigma w}}$. Hence, for $s \geq nq(1-\epsilon)/2$, by following the same proof as the $r > 1$ case above, we deduce that, conditional on $|\mathcal{D}| = s$, we have with probability at least $1 - \delta/4$ that

$$\begin{aligned}
\bigg\| \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} Z_i - \theta_0 \bigg\|_2 &\leq \inf_{\lambda \in (0, \frac{1}{2}\|\Sigma\|_{\mathrm{op}}^{-1/2}]} \bigg\{ \frac{\log(2 + 2\kappa^2)}{\lambda} + \big(6\|\Sigma\|_{\mathrm{op}}^{1/2}\big)^2 \lambda + \frac{\mathbf{r}(\Sigma)/2 + 2\log(4/\delta)}{s\lambda} \bigg\} \\
&\overset{(i)}{\leq} 2\|\Sigma\|_{\mathrm{op}}^{1/2} \bigg\{ \log(2 + 2\kappa^2) + 9 + \frac{\mathbf{r}(\Sigma)/2 + 2\log(4/\delta)}{s} \bigg\} \\
&\overset{(ii)}{\leq} 2\|\Sigma\|_{\mathrm{op}}^{1/2} \big\{ \log(2 + 2\kappa^2) + 11 \big\} \lesssim \|\Sigma\|_{\mathrm{op}}^{1/2} \log(2 + 2\kappa),
\end{aligned}$$

where (i) follows by choosing $\lambda = \frac{1}{2}\|\Sigma\|_{\mathrm{op}}^{-1/2}$, and (ii) follows from the assumptions that $nq(1-\epsilon) \geq \mathbf{r}(\Sigma)$ and $\delta \geq 8\exp\big(-nq(1-\epsilon)/8\big)$. Hence, there exists a universal constant $C_2 > 0$ such that the event

$$\mathcal{E}_4 := \bigg\{ \bigg\| \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} Z_i - \theta_0 \bigg\|_2^2 \leq C_2 \|\Sigma\|_{\mathrm{op}} \log^2(2 + 2\kappa) \bigg\}, \tag{92}$$

satisfies $\mathbb{P}(\mathcal{E}_0 \cap \mathcal{E}_4) \geq 1 - \delta/2$. Thus, on the event $\mathcal{E}_0 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, which has probability at least $1 - \delta$, we combine (89) and (92) to obtain the desired result for $r = 1$. □

E Proofs from Section 5

E.1 Proof of Lemma 13

Proof of Lemma 13. (a) Let (v_m) be a sequence in $\mathbb{S}^{d-1}$ with

$$\mathbb{P}\big(|X_1^\top v_m| > \gamma\big) \searrow \inf_{v \in \mathbb{S}^{d-1}} \mathbb{P}\big(|X_1^\top v| > \gamma\big)$$

as $m \to \infty$. Then by compactness of $\mathbb{S}^{d-1}$, there exists a subsequence (v_{m_k}), as well as $v_* \in \mathbb{S}^{d-1}$, for which $v_{m_k} \to v_*$ as $k \to \infty$. But then $|X_1^\top v_{m_k}| \xrightarrow{d} |X_1^\top v_*|$ as $k \to \infty$, so by, e.g., van der Vaart (1998, Lemma 2.2),

$$\mathbb{P}\big(|X_1^\top v_*| > \gamma\big) \leq \liminf_{k \to \infty} \mathbb{P}\big(|X_1^\top v_{m_k}| > \gamma\big) = \inf_{v \in \mathbb{S}^{d-1}} \mathbb{P}\big(|X_1^\top v| > \gamma\big).$$

It follows that the infimum in the definition of β is attained.

If $\beta = 0$ for all $\gamma > 0$, then for every $\gamma > 0$ we can find $v_*(\gamma) \in \mathbb{S}^{d-1}$ with $\mathbb{P}\big(|X_1^\top v_*(\gamma)| > \gamma\big) = 0$. Writing $v_m := v_*(1/m)$, there exist integers $1 \leq m_1 < m_2 < \ldots$ and $v_{**} \in \mathbb{S}^{d-1}$ with

$v_{m_k} \to v_{**}$ as $k \to \infty$. Since $|X_1^\top v_{m_k}| - 1/m_k \xrightarrow{d} |X_1^\top v_{**}|$ as $k \to \infty$ we have by van der Vaart (1998, Lemma 2.2) again that

$$\mathbb{P}\big(|X_1^\top v_{**}| > 0\big) \le \liminf_{k\to\infty} \mathbb{P}\bigg(|X_1^\top v_{m_k}| > \frac{1}{m_k}\bigg) = 0.$$

But then, defining the hyperplane $H := \{x \in \mathbb{R}^d : x^\top v_{**} = 0\}$, we have $P(H) = 1$.

(b) The claim is equivalent to showing that there exists a universal constant $c > 0$ such that if $\frac{d+\log(1/\delta)}{n} \le c\beta^2$, then, with probability at least $1 - \delta$,

$$\sup_{v\in\mathbb{S}^{d-1}} -\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{|X_i^\top v|>\gamma\}} \le -2\beta.$$

To establish this, let $\mathcal{H} := \{x \mapsto -\mathbb{1}_{\{|x^\top v|>\gamma\}} : v \in \mathbb{S}^{d-1}\}$. Then

$$\begin{aligned}\sup_{v\in\mathbb{S}^{d-1}} -\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{|X_i^\top v|>\gamma\}} + 3\beta &\le \sup_{v\in\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n \big\{-\mathbb{1}_{\{|X_i^\top v|>\gamma\}} + \mathbb{P}(|X_i^\top v| > \gamma)\big\} \\ &= \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \big\{h(X_i) - \mathbb{E}h(X_i)\big\} =: V, \end{aligned} \tag{93}$$

where the first inequality follows since $\mathbb{P}(|X_i^\top v| > \gamma) \ge 3\beta$ for all $v \in \mathbb{S}^{d-1}$. By the bounded differences inequality (e.g., Boucheron, Lugosi and Massart, 2013, Theorem 6.2), with probability at least $1 - \delta$,

$$V \le \mathbb{E}(V) + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{94}$$

For a collection $\mathcal{H}_1$ of binary-valued functions, we write $\mathrm{VC}(\mathcal{H}_1)$ for its Vapnik–Chervonenkis dimension. For $v \in \mathbb{R}^d$ and $b \in \mathbb{R}$, define $g_{v,b} : \mathbb{R}^d \to \mathbb{R}$ by $g_{v,b}(x) := x^\top v + b$, and define the vector space $\mathcal{G} := \{g_{v,b} : v \in \mathbb{R}^d, b \in \mathbb{R}\}$. Now let $\mathcal{H}' := \{x \mapsto -\mathbb{1}_{\{g(x)>0\}} : g \in \mathcal{G}\}$, which by Wainwright (2019, Proposition 4.20) satisfies $\mathrm{VC}(\mathcal{H}') \le \dim(\mathcal{G}) = d + 1$. Then

$$\mathcal{H} = \big\{x \mapsto -\mathbb{1}_{\{g_{v,-\gamma}(x)>0\}\cup\{g_{-v,-\gamma}(x)>0\}} : v \in \mathbb{S}^{d-1}\big\} \subseteq \big\{x \mapsto -\mathbb{1}_{\{g_1(x)>0\}\cup\{g_2(x)>0\}} : g_1, g_2 \in \mathcal{G}\big\}.$$

Hence, by Blumer et al. (1989, Lemma 3.2.3), we have $\mathrm{VC}(\mathcal{H}) \le 4\log_2(6)\mathrm{VC}(\mathcal{H}') \le 11d+11$. We deduce by Vershynin (2018, Theorem 8.3.23) that there exists a universal constant $C_1 > 0$ such that $\mathbb{E}(V) \le C_1\sqrt{\frac{d+1}{n}}$. Thus, by (93) and (94) we conclude that there exists a universal constant $C_2 > 0$ such that with probability at least $1 - \delta$,

$$\sup_{v\in\mathbb{S}^{d-1}} -\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{|X_i^\top v|>\gamma\}} + 3\beta \le \mathbb{E}(V) + \sqrt{\frac{\log(1/\delta)}{2n}} \le C_2\sqrt{\frac{d+\log(1/\delta)}{n}} \le \beta,$$

where the final inequality follows by choosing $c := 1/C_2^2$ and using the assumption that $\frac{d+\log(1/\delta)}{n} \le c\beta^2$. This proves the claim. □

E.2 Proof of Theorem 14

We begin with some preliminary lemmas.

Lemma 29. *Consider the setting of Theorem 14. There exists a universal constant $C > 0$ such that for $\delta \in (0,1]$, we have with probability at least $1-\delta$ conditional on $X_1 = x_1, \dots, X_n = x_n$ that*

$$\sup_{\theta \in \mathbb{R}^d} d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta}, R_{n,\theta}) \le C\sqrt{\frac{d + \log(1/\delta)}{n}}.$$

Proof. First define

$$V := \sup_{\theta \in \mathbb{R}^d} d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta}, R_{n,\theta}) = \sup_{\theta \in \mathbb{R}^d} \sup_{A \in \mathcal{A}^{\mathrm{sym}}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i - x_i^\top \theta \in A\}} - \frac{1}{n}\sum_{i=1}^n \widetilde{R}_{i,\theta}(A) \right|.$$

Then, by the bounded differences inequality (e.g., Boucheron, Lugosi and Massart, 2013, Theorem 6.2), with probability at least $1-\delta$,

$$V \le \mathbb{E}(V) + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{95}$$

Now define

$$\mathcal{G} := \Big\{ g : \mathbb{R}^d \times \mathbb{R}_\star \to \mathbb{R} \text{ s.t. } g(x,z) = (z - x^\top\theta - t)\mathbb{1}_{\{z \neq \star\}} + \mathbb{1}_{\{z=\star\}} \text{ for some } \theta \in \mathbb{R}^d,\, t \in \mathbb{R} \Big\},$$

and define $\mathcal{H}_+ := \big\{(x,z) \mapsto \mathbb{1}_{\{g(x,z)\le 0\}} : g \in \mathcal{G}\big\}$ and $\mathcal{H}_- := \big\{(x,z) \mapsto \mathbb{1}_{\{g(x,z)\le 0\}} : g \in -\mathcal{G}\big\}$. Then

$$V = \sup_{h \in \mathcal{H}_+ \cup \mathcal{H}_-} \left| \frac{1}{n}\sum_{i=1}^n \big\{ h(x_i, Z_i) - \mathbb{E}h(x_i, Z_i) \big\} \right|.$$

Since $\mathcal{G}$ is a vector space of functions with $\dim(\mathcal{G}) = d+1$, by Mohri, Rostamizadeh and Talwalkar (2018, Exercise 3.24(b)) and Wainwright (2019, Proposition 4.20), we deduce that $\mathrm{VC}(\mathcal{H}_+ \cup \mathcal{H}_-) \le \mathrm{VC}(\mathcal{H}_+) + \mathrm{VC}(\mathcal{H}_-) + 1 \le 2\dim(\mathcal{G}) + 1 \le 2d+3$. Therefore, applying Vershynin (2018, Theorem 8.3.23) yields that $\mathbb{E}(V) \le C'\sqrt{\frac{2d+3}{n}}$ for some universal constant $C' > 0$. Combining this with (95) proves the desired result. □

Lemma 30. *Consider the setting of Theorem 14, and assume that $\theta \neq \theta_0$. Then, writing $a := \frac{1}{2}\|\theta_0 - \theta\|_2 > 0$ and $b := \frac{1}{2}\log\big(1 + \frac{4(1-\beta q(1-\epsilon))}{\beta q(1-\epsilon)}\big)$, we have*

$$d_{\mathrm{K}}^{\mathrm{sym}}\big(R_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}\big) \ge \beta q(1-\epsilon)\Phi\left(\frac{a\gamma}{\sigma} - \frac{2\sigma b}{a\gamma}\right) - \Phi\left(-\frac{a\gamma}{\sigma} - \frac{2\sigma b}{a\gamma}\right) =: f_{\mathrm{K},b}(a),$$

where $f_{\mathrm{K},b} : (0,\infty) \to (0,\infty)$ is strictly increasing and continuous.

Proof. By Assumption 1, we may assume without loss of generality that there exists $\mathcal{T}_+ \subseteq [n]$ such that $|\mathcal{T}_+| \geq \beta n$ and $-x_i^\top(\theta_0 - \theta) \geq 2a\gamma$. By Proposition 2, for $i \in \mathcal{T}_+$ and $t \in \mathbb{R}$, we have

$$\begin{aligned}\widetilde{R}_{i,\theta}\big((-\infty, t]\big) &\geq q(1-\epsilon)\Phi_{(0,\sigma)}\big(t - x_i^\top(\theta_0 - \theta)\big) \geq q(1-\epsilon)\Phi_{(0,\sigma)}(t + 2a\gamma) \\ &= q(1-\epsilon)\Phi_{(-2a\gamma,\sigma)}(t).\end{aligned}$$

Moreover, by Proposition 2 again for $R_0 \in \mathcal{R}_0^{\mathrm{Lin}}$ and $t \in \mathbb{R}$, we have $R_0\big((-\infty, t]\big) \leq \Phi_{(0,\sigma)}(t)$. Therefore,

$$\begin{aligned}d_{\mathrm{K}}^{\mathrm{sym}}\big(R_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}\big) &\geq \inf_{R_0 \in \mathcal{R}_0^{\mathrm{Lin}}} \sup_{t \in \mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^n \widetilde{R}_{i,\theta}\big((-\infty, t]\big) - R_0\big((-\infty, t]\big) \right\} \\ &\geq \inf_{R_0 \in \mathcal{R}_0^{\mathrm{Lin}}} \sup_{t \in \mathbb{R}} \left\{ \frac{1}{n}\sum_{i \in \mathcal{T}_+} \widetilde{R}_{i,\theta}\big((-\infty, t]\big) - R_0\big((-\infty, t]\big) \right\} \\ &\geq \sup_{t \in \mathbb{R}} \left\{ \beta q(1-\epsilon)\Phi_{(-2a\gamma,\sigma)}(t) - \Phi_{(0,\sigma)}(t) \right\} \\ &\geq \beta q(1-\epsilon)\Phi\left(\frac{a\gamma}{\sigma} - \frac{2\sigma b}{a\gamma}\right) - \Phi\left(-\frac{a\gamma}{\sigma} - \frac{2\sigma b}{a\gamma}\right) = f_{\mathrm{K},b}(a),\end{aligned}$$

where the final inequality follows by choosing $t = -\frac{2\sigma^2 b}{a\gamma} - a\gamma$. The function $f_{\mathrm{K},b}$ is continuous as a composition of continuous functions, and the fact that it is strictly increasing follows as in the proof of Lemma 24, setting (ϵ, q) therein as $(\bar{\epsilon}, \bar{q})$, with $\bar{\epsilon} := 1 - \beta q(1-\epsilon)$ and $\bar{q} := 1$. □

Proof of Theorem 14. Let $\delta \in (0, 1]$ and for the universal constant $C > 0$ from Lemma 29, define the event

$$\mathcal{E} := \left\{ \sup_{\theta \in \mathbb{R}^d} d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta}, R_{n,\theta}) \leq C\sqrt{\frac{d + \log(1/\delta)}{n}} \right\}.$$

By Lemma 29, satisfies $\mathbb{P}(\mathcal{E} \,|\, X_1 = x_1, \ldots, X_n = x_n) \geq 1 - \delta$, and from now on, we will work on the event $\mathcal{E}$. Recalling that $R_{n,\theta_0} \in \mathcal{R}_0^{\mathrm{Lin}}$, we have

$$d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta_0}, \mathcal{R}_0^{\mathrm{Lin}}) \leq d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta_0}, R_{n,\theta_0}) \leq C\sqrt{\frac{d + \log(1/\delta)}{n}}.$$

Moreover, if $\theta \in \mathbb{R}^d$ satisfies $d_{\mathrm{K}}^{\mathrm{sym}}(R_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}) > 2C\sqrt{\frac{d+\log(1/\delta)}{n}}$, then

$$\begin{aligned}d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}) &\geq d_{\mathrm{K}}^{\mathrm{sym}}(R_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}) - d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta}, R_{n,\theta}) \\ &> C\sqrt{\frac{d + \log(1/\delta)}{n}} \geq d_{\mathrm{K}}^{\mathrm{sym}}(\widehat{R}_{n,\theta_0}, \mathcal{R}_0^{\mathrm{Lin}}),\end{aligned}$$

so $\widehat{\theta}_n^{\mathrm{K}} \neq \theta$. Therefore, with b and $f_{\mathrm{K},b}$ as defined in Lemma 30, we deduce that

$$\|\widehat{\theta}_n^{\mathrm{K}} - \theta_0\|_2 \leq \sup\left\{ \|\theta - \theta_0\|_2 : \theta \in \mathbb{R}^d,\, d_{\mathrm{K}}^{\mathrm{sym}}(R_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}) \leq 2C\sqrt{\frac{d + \log(1/\delta)}{n}} \right\}$$

$$\leq 2\inf\Bigg\{a>0 : \beta q(1-\epsilon)\Phi\bigg(\frac{a\gamma}{\sigma}-\frac{2\sigma b}{a\gamma}\bigg)-\Phi\bigg(-\frac{a\gamma}{\sigma}-\frac{2\sigma b}{a\gamma}\bigg) \geq 2C\sqrt{\frac{d+\log(1/\delta)}{n}}\Bigg\}, \quad (96)$$

where the second inequality follows since by Lemma 30, $d_{\mathrm{K}}^{\mathrm{sym}}(R_{n,\theta}, \mathcal{R}_0^{\mathrm{Lin}}) \geq f_{\mathrm{K},b}\big(\frac{\|\theta-\theta_0\|_2}{2}\big)$ and $f_{\mathrm{K},b}$ is a strictly increasing and continuous function. Letting $a = \frac{6\sigma b}{\gamma\sqrt{\log n}}$, we have by our assumption on b that $2\sigma b/(a\gamma) - a\gamma/\sigma = \sqrt{\log n}/3 - 6b/\sqrt{\log n} > 0$, so

$$\begin{aligned}
&\beta q(1-\epsilon)\Phi\bigg(\frac{a\gamma}{\sigma}-\frac{2\sigma b}{a\gamma}\bigg)-\Phi\bigg(-\frac{a\gamma}{\sigma}-\frac{2\sigma b}{a\gamma}\bigg)\\
&\overset{(i)}{\geq} \frac{\beta q(1-\epsilon)}{\big(-\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\big)+\big(-\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\big)^{-1}}\cdot\phi\bigg(-\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\bigg)-\frac{1}{\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}}\cdot\phi\bigg(\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\bigg)\\
&\overset{(ii)}{\geq} \bigg(\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\bigg)^{-1}\frac{1}{\sqrt{2\pi}}\bigg\{\beta q(1-\epsilon)\exp\bigg(-\frac{a^2\gamma^2}{2\sigma^2}-\frac{2\sigma^2b^2}{a^2\gamma^2}+2b\bigg)\\
&\qquad\qquad -\exp\bigg(-\frac{a^2\gamma^2}{2\sigma^2}-\frac{2\sigma^2b^2}{a^2\gamma^2}-2b\bigg)\bigg\}\\
&\overset{(iii)}{\geq} \bigg(\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\bigg)^{-1}\frac{3}{2\sqrt{2\pi}}\cdot\exp\bigg(-\frac{a^2\gamma^2}{2\sigma^2}-\frac{2\sigma^2b^2}{a^2\gamma^2}\bigg)\\
&\overset{(iv)}{\geq} \frac{1}{\sqrt{\log n}}\cdot n^{-5/72} \overset{(v)}{\geq} 2C\sqrt{\frac{d+\log(1/\delta)}{n}}
\end{aligned}$$

where (i) follows from the Mills ratio bound $\phi(x)/(x+x^{-1}) \leq \Phi(-x) \leq \phi(x)/x$ for $x>0$; (ii) follows since $\frac{1}{2} \leq b \leq \frac{\log n}{36}$ implies $\big(-\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\big)+\big(-\frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}\big)^{-1} \leq \frac{a\gamma}{\sigma}+\frac{2\sigma b}{a\gamma}$; (iii) follows by substituting the definition of b and using the fact that $\beta q(1-\epsilon) \leq 1/2$; (iv) follows since $b \leq \frac{\log n}{36}$ implies $\frac{a\gamma}{\sigma} \leq \frac{\sigma b}{a\gamma}$; and (v) follows from the assumption that $\frac{n^{31/36}}{\log n} \geq C_1\big\{d+\log(1/\delta)\big\}$ and taking $C_1 := 4C^2$. Therefore, with probability at least $1-\delta$, we have

$$\|\widehat{\theta}_n^{\mathrm{K}}-\theta_0\|_2^2 \leq \frac{144\sigma^2b^2}{\gamma^2\log n} \leq \frac{36\sigma^2\log^2\big(1+\frac{4(1-\beta q(1-\epsilon))}{\beta q(1-\epsilon)}\big)}{\gamma^2\log\big(nq(1-\epsilon)\big)},$$

as required. □

F Auxiliary lemmas

If $\mathcal{Z}$ is a topological space, then we define the embedding $\phi_{\mathcal{Z}} : C_{\mathrm{b}}(\mathcal{Z}) \to \mathcal{M}(\mathcal{Z})^*$ by $\phi_{\mathcal{Z}}(f)(\mu) := \mu(f)$. If $\mathcal{Z}$ is a locally compact Hausdorff space, then a Borel measure μ on $\mathcal{Z}$ is *regular* if $\mu(E) = \inf\{\mu(U) : U \supseteq E, U \text{ open}\}$ and $\mu(E) = \sup\{\mu(K) : K \subseteq E, K \text{ compact}\}$ for every Borel subset E of $\mathcal{Z}$.

Lemma 31. *Let $\mathcal{Z}$ be a locally compact Hausdorff space in which every open set is σ-compact. Then $\phi_{\mathcal{Z}}$ embeds $C_{\mathrm{b}}(\mathcal{Z})$ into a subspace of $\mathcal{M}(\mathcal{Z})^*$ that separates points.*

Proof. If $f, g \in C_{\mathrm{b}}(\mathcal{Z})$ and $\lambda_1, \lambda_2 \in \mathbb{R}$, then $\phi_{\mathcal{Z}}(\lambda_1 f + \lambda_2 g) = \lambda_1 \phi_{\mathcal{Z}}(f) + \lambda_2 \phi_{\mathcal{Z}}(g)$, so $\phi_{\mathcal{Z}}$ embeds $C_{\mathrm{b}}(\mathcal{Z})$ into a subspace of $\mathcal{M}(\mathcal{Z})^*$.

Let μ and μ' be two distinct measures in $\mathcal{M}(\mathcal{Z})$ and define $\nu \coloneqq \mu - \mu' \in \mathcal{M}(\mathcal{Z})$. By the the Jordan decomposition theorem (Folland, 1999, Theorem 3.3), we can write $\nu = \nu_+ - \nu_-$ where $\nu_+, \nu_- \in \mathcal{M}_+(\mathcal{Z})$ are supported on disjoint measurable sets $P, N \subseteq \mathcal{Z}$ respectively. Since $\nu \neq 0$, there exists a Borel set $B \subseteq \mathcal{Z}$ and $\epsilon > 0$ such that either $\nu_+(B \cap P) \geq \epsilon$ or $\nu_-(B \cap N) \geq \epsilon$. Without loss of generality, we assume the former. By Folland (1999, Theorem 7.8), ν_+ and ν_- are regular measures, so there exists a compact set $K \subseteq \mathcal{Z}$ and an open set $U \subseteq \mathcal{Z}$ such that $K \subseteq B \cap P \subseteq U$ and $\nu_+(U \setminus K) + \nu_-(U \setminus K) \leq \epsilon/2$. By Urysohn's lemma for locally compact Hausdorff spaces (Folland, 1999, Lemma 4.32), there exists a continuous function $f : \mathcal{Z} \to [0, 1]$ such that $f(K) = \{1\}$, $f(U^c) = \{0\}$. Observe that

$$\nu(f) \geq \nu_+(K) - \nu_-(U \setminus P) \geq \nu_+(B \cap P) - \big(\nu_+(U \setminus K) + \nu_-(U \setminus K)\big) \geq \epsilon/2.$$

Consequently, $f \in C_{\mathrm{b}}(\mathcal{Z})$ separates μ and μ' as desired. □

If (X, τ) and (Y, σ) are topological spaces, we write $\tau \otimes \sigma$ for the product topology on the Cartesian product $X \times Y$, i.e. $\tau \otimes \sigma$ is the coarsest topology for which the projections $(x, y) \mapsto x$ and $(x, y) \mapsto y$ are continuous.

Lemma 32. *If X and Y are real vector spaces and X' and Y' are subspaces of X^* and Y^*, then the map $\iota : X' \times Y' \to (X \times Y)^*$ given by $\iota(f, g)(x, y) \coloneqq f(x) + g(y)$ embeds $X' \times Y'$ as a subspace of $(X \times Y)^*$. Furthermore, if $\tau(X; X')$, $\tau(Y; Y')$ and $\tau\big(X \times Y; \iota(X' \times Y')\big)$ denote the weak topologies generated by X', Y' and $\iota(X' \times Y')$ on X, Y and $X \times Y$ respectively, then $\tau(X; X') \otimes \tau(Y; Y') = \tau\big(X \times Y; \iota(X' \times Y')\big)$.*

Proof. To check that ι embeds $X' \times Y'$ as a subspace of $(X \times Y)^*$, we only need to verify the bilinearity of the map $((x, y), (f, g)) \mapsto f(x) + g(y)$ on $(X \times Y) \times (X' \times Y')$, which is true since X, Y, X', Y' are vector spaces and (X, X'), (Y, Y') are dual pairs.

For the second claim, let $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ be projection maps defined by $\pi_X(x, y) \coloneqq x$ and $\pi_Y(x, y) \coloneqq y$. By the definition of the product topology, $\tau(X; X') \otimes \tau(Y; Y')$ is the coarsest topology on $X \times Y$ under which both π_X and π_Y are continuous. Also, $\tau\big(X \times Y; \iota(X' \times Y')\big)$ is the coarsest topology on $X \times Y$ under which $\iota(f, g)$ is continuous for all $f \in X'$ and $g \in Y'$. Hence the desired result is equivalent to the statement that for any topology $\mathcal{T}$ on $X \times Y$, the functions $\pi_X : (X \times Y, \mathcal{T}) \to (X, \tau(X; X'))$ and $\pi_Y : (X \times Y, \mathcal{T}) \to (Y, \tau(Y; Y'))$ are continuous if and only if $\iota(f, g) : (X \times Y, \mathcal{T}) \to \mathbb{R}$ is continuous for all $(f, g) \in X' \times Y'$.

The 'only if' direction is true since for any $(f, g) \in X' \times Y'$, $\iota(f, g) = f \circ \pi_X + g \circ \pi_Y$ is the sum of compositions of continuous functions, and hence continuous. For the 'if' direction, we assume that $\iota(f, g)$ is continuous for all $(f, g) \in X' \times Y'$; by symmetry we only need to check that π_X is continuous. Taking g to be the zero map, we have $\iota(f, 0)(x, y) = f(\pi_X(x, y))$, so $f \circ \pi_X$ is continuous for every f. Open sets in $(X, \tau(X; X'))$ are unions of sets in $\{f^{-1}(U) : f \in X', U \text{ open in } \mathbb{R}\}$. Since $f \circ \pi_X$ is continuous, we have

$$\pi_X^{-1}\big(f^{-1}(U)\big) = (f \circ \pi_X)^{-1}(U)$$

is open for every $f \in X'$ and U open in $\mathbb{R}$. Therefore, π_X is continuous as desired, and this establishes the lemma. □

Lemma 33. *Let X, Y, Z be topological spaces and equip $Y \times Z$ with the product topology. Then $f : X \to Y$ and $g : X \to Z$ are continuous if and only if $h : x \mapsto \big(f(x), g(x)\big)$ is a continuous function from X to $Y \times Z$.*

Proof. By definition of the product topology, the projection maps $\pi_Y : Y \times Z \to Y$ and $\pi_Z : Y \times Z \to Z$ defined by $\pi_Y(y, z) \coloneqq y$ and $\pi_Z(y, z) \coloneqq z$ are continuous. This proves the 'if' direction since $f = \pi_Y \circ h$ and $g = \pi_Z \circ h$. For the 'only if' direction, we observe that open sets in $Y \times Z$ are unions of sets of the form $U \times V$ for U open in Y and V open in Z. Since f and g are continuous, $h^{-1}(U \times V) = f^{-1}(U) \cap g^{-1}(V)$ is open in X, so h is continuous as desired. □

Recall that if A_1 and A_2 are sets, then the *disjoint union* of A_1 and A_2 is defined by $A_1 \sqcup A_2 := \{(a, 1) : a \in A_1\} \cup \{(a, 2) : a \in A_2\}$. Moreover, if (A_1, τ_1) and (A_2, τ_2) are topological spaces, then $A_1 \sqcup A_2$ can be endowed with the *disjoint union topology*, given by $\big\{(U_1 \times \{1\}) \cup (U_2 \times \{2\}) : U_1 \in \tau_1,\, U_2 \in \tau_2\big\}$. In the special case where A_1 and A_2 are disjoint subsets of a topological space $(\mathcal{X}, \tau)$, the second argument of elements in $A_1 \sqcup A_2$ becomes redundant, so we can identify $A_1 \sqcup A_2$ with $A_1 \cup A_2$, and we may write the disjoint union topology simply as $\{U_1 \cup U_2 : U_1 \in A_1 \cap \tau, U_2 \in A_2 \cap \tau\}$.

Lemma 34. *Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_d$ be topological spaces, and let $\mathcal{Z} \coloneqq \prod_{j=1}^d \mathcal{Z}_j$ be the product space equipped with the product topology. Let $S \subseteq [d]$ be non-empty and let $\mathcal{Z}_S \coloneqq \prod_{j \in S} \mathcal{Z}_j$ be the product space equipped with the product topology.*

(a) If $U \subseteq \mathcal{Z}$ is open, then the set $U_S \coloneqq \{x_S : x \in U\}$ is open in $\mathcal{Z}_S$.

(b) If $K \subseteq \mathcal{Z}$ is compact, then the set $K_S \coloneqq \{x_S : x \in K\}$ is compact in $\mathcal{Z}_S$.

Proof. (a) We can write $U = \bigcup_{i \in I} U^{(i)}$ for some index set I, where $U^{(i)} = \prod_{j=1}^d U_j^{(i)}$ and $U_j^{(i)}$ is open in $\mathcal{Z}_j$ for all $i \in I$, $j \in [d]$. Hence $U_S = \bigcup_{i \in I} U_S^{(i)}$ where $U_S^{(i)} = \prod_{j \in S} U_j^{(i)}$, so U_S is open in $\mathcal{Z}_S$.

(b) For any open cover $\{U_S^{(i)}\}_{i \in I}$ of K_S, define $U^{(i)} \coloneqq \{x \in \mathcal{Z} : x_S \in U_S^{(i)}\}$ for $i \in I$. Note that $U^{(i)}$ is open in $\mathcal{Z}$ for $i \in I$, as it is the pre-image of an open set under a projection map (which is continuous, by definition of the product topology). Thus, $\{U^{(i)}\}_{i \in I}$ is an open cover of K, which has a finite subcover $I_0 \subseteq I$ since K is compact. Therefore, $\{U_S^{(i)}\}_{i \in I_0}$ is also a finite subcover of K_S, so K_S is compact in $\mathcal{Z}_S$. □

Lemma 35. *Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be topological spaces.*

(a) If $\mathcal{X}_1$ and $\mathcal{X}_2$ are Hausdorff, then $\mathcal{X}_1 \times \mathcal{X}_2$ is Hausdorff in the product topology and $\mathcal{X}_1 \sqcup \mathcal{X}_2$ is Hausdorff in the disjoint union topology.

(b) If $\mathcal{X}_1$ and $\mathcal{X}_2$ are locally compact, then $\mathcal{X} \times \mathcal{X}_2$ is locally compact in the product topology and $\mathcal{X}_1 \sqcup \mathcal{X}_2$ is locally compact in the disjoint union topology.

Proof. (a) The first statement follows from Munkres (2014, Theorem 19.4). For the second statement, let $(x_1, j_1), (x_2, j_2) \in \mathcal{X}_1 \sqcup \mathcal{X}_2$ be distinct, where $j_1, j_2 \in \{1, 2\}$, and for $\ell \in \{1, 2\}$, we have $x_\ell \in \mathcal{X}_\ell$. If $j_1 = j_2$, then the result follows from the fact that $\mathcal{X}$ and $\mathcal{X}_2$ are Hausdorff. Otherwise, we can separate the two points using the open sets $\mathcal{X}_1 \times \{1\}$ and $\mathcal{X}_2 \times \{2\}$.

(b) For the first statement, if $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, then we can find compact neighbourhoods $K_j \subseteq \mathcal{X}_j$ of x_j for $j \in \{1, 2\}$. Then by Tychonov's theorem (Munkres, 2014, Theorem 37.3), $K_1 \times K_2$ is a compact neighbourhood of (x_1, x_2). For the second statement, if $(x, j) \in \mathcal{X}_1 \sqcup \mathcal{X}_2$, then we can find a compact subset $K \subseteq \mathcal{X}_j$ containing x. Then $K \times \{j\}$ is a compact subset of $\mathcal{X}_1 \sqcup \mathcal{X}_2$ containing (x, j). □

Lemma 36. *Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be locally compact Hausdorff spaces, and let $\mathcal{X} := \prod_{j=1}^d \mathcal{X}_j$. Then $\mathcal{X}$, $\mathcal{X}_\star$ and $\mathcal{X} \times 2^{[d]}$ are locally compact Hausdorff spaces. Moreover, if every open set in $\mathcal{X}$ is σ-compact, then $\mathcal{X}_\star$ and $\mathcal{X} \times 2^{[d]}$ also have this property.*

Proof. The fact that $\mathcal{X}$ is a locally compact Hausdorff space follows from Lemma 35. Moreover, the singleton space $\{\star\}$ as well as the space $2^{[d]}$ endowed with the discrete topology are both locally compact Hausdorff spaces. We observe that $\mathcal{X}$, $\mathcal{X}_\star$ and $\mathcal{X} \times 2^{[d]}$ can be generated from $\mathcal{X}_1, \ldots, \mathcal{X}_d$, $\{\star\}$, $2^{[d]}$ via a combination of product space and disjoint union operations. Hence the first result follows from Lemma 35.

To check that every open set in $\mathcal{X}_\star = \bigsqcup_{S \in 2^{[d]}} \mathcal{X}^{(S)}$ is σ-compact, observe that for any open set $U \subseteq \mathcal{X}_\star$, we can write $U = \bigcup_{S \subseteq [d]} U^{(S)}$ where $U^{(S)} := U \cap \mathcal{X}^{(S)}$. Therefore, it suffices to show that for every $S \subseteq [d]$, any open set $U \subseteq \mathcal{X}^{(S)}$ is σ-compact. Let $U_S := \{a_S : a \in U\}$ and let $V := \{x \in \mathcal{X} : x_S \in U_S\}$. Then V is open in $\mathcal{X}$ since it is the pre-image of a projection of an open set, so we can write $V = \bigcup_{i=1}^\infty K(i)$, where $K(i)$ is compact in $\mathcal{X}$ for each i. Moreover, $U_S = \bigcup_{i=1}^\infty K(i)_S$, and $K(i)_S$ is compact in $\mathcal{X}_S$ by Lemma 34*(b)*. We claim that $K(i)^{(S)} := \{z \in \mathcal{X}_\star : z_S \in K(i)_S, z_j = \star \; \forall j \notin S\}$ is compact. To see this, consider any open cover $\{U(j)\}_{j \in J}$ of $K(i)^{(S)}$, where, without loss of generality, we assume that $U(j) \subseteq \mathcal{X}^{(S)}$ for all $j \in J$. Writing $U(j)_S := \{a_S : a \in U(j)\}$ for $j \in J$, we have by Lemma 34*(a)* that $\{U(j)_S\}_{j \in J}$ forms an open cover of $K(i)_S$. We can therefore find a finite subcover $\{U(j)_S\}_{j \in J_0}$ of $K(i)_S$, so that $\{U(j)\}_{j \in J_0}$ forms a finite subcover of $K(i)^{(S)}$. We deduce that $U = \bigcup_{i=1}^\infty K(i)^{(S)}$ is a countable union of compact sets.

To show that every open set in $\mathcal{X} \times 2^{[d]}$ is σ-compact, observe that since $2^{[d]}$ is finite, any open set in $\mathcal{X} \times 2^{[d]}$ is of the form $\bigcup_{S \subseteq [d]} \{U(S) \times S\}$, where $U(S)$ is open in $\mathcal{X}$. Since each $U(S)$ is σ-compact and S is finite (hence compact), it follows that each set $U(S) \times S$ is σ-compact, and hence $\bigcup_{S \subseteq [d]} \{U(S) \times S\}$ is σ-compact. □

Lemma 37. *If $(\mathcal{X}_1, \tau_1), \ldots, (\mathcal{X}_d, \tau_d)$ are Polish spaces, then the Cartesian product space $\mathcal{X}_\star := \prod_{j=1}^d \mathcal{X}_{j,\star}$ equipped with the product topology is also a Polish space.*

Proof. A finite (or even countable) Cartesian product of Polish spaces is Polish (Kechris, 2012, Proposition 3.3), so it suffices to show that $(\mathcal{X}_{j,\star}, \tau_{j,\star})$ is Polish for each $j \in [d]$, where $\tau_{j,\star} := \tau_j \cup \{A \cup \{\star\} : A \in \tau_j\}$. Now fix $j \in [d]$, and find a countable dense subset $\{x_n\}_{n=1}^\infty$ of $\mathcal{X}_j$. Then $\{\star\} \cup \{x_n\}_{n=1}^\infty$ is a countable dense subset of $\mathcal{X}_{j,\star}$, so $\mathcal{X}_{j,\star}$ is separable. Now find a metric d on $\mathcal{X}_j$ such that d generates the topology τ_j and $(\mathcal{X}_j, d)$ is complete. Define the standard bounded metric $\bar{d}$ by $\bar{d}(x, y) := d(x, y) \wedge 1$ for $x, y \in \mathcal{X}_j$. Then, by Munkres (2014, Theorem 20.1), $\bar{d}$ also generates the topology τ_j. Define a metric d' on $\mathcal{X}_{j,\star}$ by $d'(x, y) := \bar{d}(x, y)$ for $x, y \in \mathcal{X}_j$, $d'(x, \star) := 2$ for $x \in \mathcal{X}_j$, and $d'(\star, \star) := 0$. Letting $\tau'_{j,\star}$ denote the topology on $\mathcal{X}_{j,\star}$ generated by d', we first show that $\tau'_{j,\star} = \tau_{j,\star}$. On the one hand, since $\{\star\} \in \tau'_{j,\star}$ and $\tau_j \subseteq \tau'_{j,\star}$, we have $\tau_{j,\star} \subseteq \tau'_{j,\star}$. On the other hand, if $x_0 \in \mathcal{X}_{j,\star}$, $r \geq 0$ and

$A \coloneqq \{x \in \mathcal{X}_{j,\star} : d'(x, x_0) < r\}$ denotes an open ball in $\tau'_{j,\star}$, then

$$A = \begin{cases} \mathcal{X}_{j,\star} & \text{if } r > 2 \\ \{x \in \mathcal{X}_j : \bar{d}(x, x_0) < r\} & \text{if } r \leq 2 \text{ and } x_0 \in \mathcal{X}_j \\ \{\star\} & \text{if } r \leq 2 \text{ and } x_0 = \star. \end{cases}$$

We deduce that $A \in \tau_{j,\star}$, so since such open balls generate $\tau'_{j,\star}$, we have $\tau'_{j,\star} \subseteq \tau_{j,\star}$. Hence, d' generates the topology $\tau_{j,\star}$. Next, we show that $(\mathcal{X}_{j,\star}, d')$ is complete. Let $(z_n)_{n=1}^{\infty}$ be a Cauchy sequence in $\mathcal{X}_{j,\star}$, so there exists $N \in \mathbb{N}$ such that $d'(z_{n_1}, z_{n_2}) \leq 1/2$ for all $n_1, n_2 \geq N$. Therefore, either $z_n = \star$ for all $n \geq N$ or $z_n \in \mathcal{X}_j$ for all $n \geq N$. In the former case, $z_n \to \star$ as $n \to \infty$. In the latter case, $(z_n)_{n=N}^{\infty}$ is also a Cauchy sequence in $(\mathcal{X}_j, d)$ and hence by completeness of $(\mathcal{X}_j, d)$, it has a limit in $\mathcal{X}_j$. This shows that $(\mathcal{X}_{j,\star}, d')$ is complete and d' generates the topology $\tau_{j,\star}$, so $(\mathcal{X}_{j,\star}, \tau_{j,\star})$ is completely metrisable. Therefore, $(\mathcal{X}_{j,\star}, \tau_{j,\star})$ is a Polish space, as required. □

Lemma 38. *Suppose that* $(B_i)_{i\in[n]} \overset{\text{iid}}{\sim} \mathsf{Ber}(q)$.

(a) With probability at least $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^{n} B_i \leq 2q + \frac{\log(1/\delta)}{n}.$$

(b) If $q \geq \frac{8\log(1/\delta)}{n}$, *then with probability at least* $1 - \delta$,

$$\frac{1}{n}\sum_{i=1}^{n} B_i \geq \frac{q}{2}.$$

Proof. *(a)* By Bernstein's inequality (Boucheron, Lugosi and Massart, 2013, Theorem 2.10), we have with probability at least $1 - \delta$ that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} B_i &\leq q + \sqrt{\frac{2q(1-q)}{n}}\log^{1/2}(1/\delta) + \frac{1}{3n}\log(1/\delta) \\ &\leq \left(q^{1/2} + \frac{1}{\sqrt{2n}}\log^{1/2}(1/\delta)\right)^2 \leq 2q + \frac{1}{n}\log(1/\delta). \end{aligned}$$

(b) By the multiplicative Chernoff bound (McDiarmid, 1998, Theorem 2.3(c)) for the sum of independent Bernoulli random variables, we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} B_i \leq \frac{q}{2}\right) \leq \exp(-nq/8) \leq \delta,$$

where the final inequality follows from the assumption that $q \geq \frac{8\log(1/\delta)}{n}$. □

Lemma 39. *Let* $0 < r_1 \leq r_2$. *Then* $\mathcal{P}_{\psi_{r_2}}(\theta_0, \sigma^2) \subseteq \mathcal{P}_{\psi_{r_1}}(\theta_0, \sigma^2)$.

Proof. Let $X \sim P \in \mathcal{P}_{\psi_{r_2}}(\theta_0, \sigma^2)$. Then

$$2 \geq \mathbb{E}\left\{\exp\left(\frac{|X-\theta_0|^{r_2}}{\sigma^{r_2}}\right)\right\} \geq \left[\mathbb{E}\left\{\exp\left(\frac{|X-\theta_0|^{r_1}}{\sigma^{r_1}}\right)\right\}\right]^{r_2/r_1},$$

by Jensen's inequality. Thus $\mathbb{E}\exp\bigl(|X-\theta_0|^{r_1}/\sigma^{r_1}\bigr) \leq 2$, so $P \in \mathcal{P}_{\psi_{r_1}}(\theta_0, \sigma^2)$. □

Lemma 40. *Let $r > 1$, $\sigma > 0$ and $X \sim P \in \mathcal{P}_{\psi_r}(0, \sigma^2)$. Then*

$$\mathbb{E}\exp(\lambda X) \leq 2\exp\bigl\{(\sigma\lambda)^{r/(r-1)}\bigr\},$$

for all $\lambda > 0$.

Proof. Young's inequality states that whenever $p, q > 1$ are such that $1/p + 1/q = 1$, we have $ab \leq a^p/p + b^q/q$ for all $a, b \geq 0$. Hence

$$\lambda X \leq \lambda|X| \leq \frac{|X|^r}{r\sigma^r} + \frac{(\sigma\lambda)^{r/(r-1)}}{r/(r-1)} \leq \frac{|X|^r}{\sigma^r} + (\sigma\lambda)^{r/(r-1)}.$$

Therefore,

$$\mathbb{E}\exp(\lambda X) \leq \mathbb{E}\bigl\{\exp(|X|^r/\sigma^r)\bigr\} \cdot \exp\bigl\{(\sigma\lambda)^{r/(r-1)}\bigr\} \leq 2\exp\bigl\{(\sigma\lambda)^{r/(r-1)}\bigr\},$$

as required. □

Lemma 41 (PAC–Bayes lemma)**.** *Let $\mathcal{X}$ be a measurable space and let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \in \mathcal{P}(\mathcal{X})$. Let $\Xi \subseteq \mathbb{R}^d$ and $\mu \in \mathcal{P}(\Xi)$. Further let $f : \mathcal{X} \times \Xi \to \mathbb{R}$ be such that $\mathbb{E}_{X\sim P}(e^{f(X,\xi)}) < \infty$ for μ-almost all $\xi \in \Xi$. Then, for every $\delta \in (0, 1]$, we have with probability at least $1-\delta$ that*

$$\sup_{\rho \in \mathcal{P}(\Xi): \rho \ll \mu} \left\{ \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi\sim\rho} f(X_i, \xi) - \mathbb{E}_{\xi\sim\rho} \log\bigl\{\mathbb{E}_{X\sim P}(e^{f(X,\xi)})\bigr\} - \frac{\mathrm{KL}(\rho, \mu) + \log(1/\delta)}{n} \right\} \leq 0,$$

where, for instance, $\mathbb{E}_{\xi\sim\rho} f(X_i, \xi) \coloneqq \int_\Xi f(X_i, v)\,\mathrm{d}\rho(v)$.

Proof. See, for example, Zhivotovskiy (2024, Lemma 2.1). □

The following lemma provides a concentration result for the sample mean of independent and identically distributed sub-exponential random vectors. The proof strategy follows that of Zhivotovskiy (2024, Proposition 3.1), who considered the case $n = 1$.

Lemma 42. *Let $\theta_0 \in \mathbb{R}^d$, $\Sigma \in \mathcal{S}_{++}^{d\times d}$, $\delta \in (0, 1]$ and $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \in \mathcal{P}_{d,\psi_1}(\theta_0, \Sigma)$. Assume further that $\delta \geq 2e^{-n/3}$. Then with probability at least $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i - \theta_0\right\|_2^2 \leq 24 \cdot \frac{\mathrm{tr}(\Sigma) + \|\Sigma\|_{\mathrm{op}} \log(2/\delta)}{n}.$$

Proof. Let $\beta \coloneqq 2\log(2/\delta)$, let μ denote the distribution of $\mathsf{N}_d(0, \beta^{-1}\Sigma)$ and for $u \in \Sigma^{1/2}\mathbb{S}^{d-1}$, let ρ_u denote the conditional distribution of Y given $\{\|Y-u\|_2 \le \sqrt{2\beta^{-1}\operatorname{tr}(\Sigma)}\}$, where $Y \sim \mathsf{N}_d(u, \beta^{-1}\Sigma)$. By the computation of Zhivotovskiy (2024, p. 11), we have

$$\mathrm{KL}(\rho_u, \mu) \le \frac{\beta}{2} + \log 2 \le 2\log\Big(\frac{2}{\delta}\Big).$$

Now, let $v \in \mathbb{R}^d$ be such that $\|v-u\|_2 \le \sqrt{2\beta^{-1}\operatorname{tr}(\Sigma)}$, and for $\lambda \in \mathbb{R}$, define $f_\lambda : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ by $f_\lambda(x,y) \coloneqq \lambda y^\top \Sigma^{-1/2}(x-\theta_0)$. Then, for $X \sim P$ and $\lambda \in \mathbb{R}$, we have

$$\big\|v^\top\Sigma^{-1/2}(X-\theta_0)\big\|_{\psi_1} \le \|v\|_2 \le \|\Sigma\|_{\mathrm{op}}^{1/2} + \sqrt{2\beta^{-1}\operatorname{tr}(\Sigma)} =: R.$$

It follows by Zhivotovskiy (2024, Lemma 2.5) that

$$\log \mathbb{E}_{X\sim P}(e^{f_\lambda(X,v)}) = \log \mathbb{E}_{X\sim P}(e^{\lambda\cdot v^\top\Sigma^{-1/2}(X-\theta_0)}) \le 4\lambda^2R^2,$$

for all $|\lambda| \le \frac{1}{2R}$, so $\mathbb{E}_{\xi_u\sim\rho_u}\{\log \mathbb{E}_{X\sim P}(e^{f_\lambda(X,\xi_u)})\} \le 4\lambda^2R^2$ for all $|\lambda| \le \frac{1}{2R}$. The PAC–Bayes lemma (Lemma 41) then yields that with probability at least $1-\delta$,

$$\sup_{u\in\Sigma^{1/2}\mathbb{S}^{d-1}}\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi_u\sim\rho_u} f_\lambda(X_i,\xi_u) - \mathbb{E}_{\xi_u\sim\rho_u}\{\log \mathbb{E}_{X\sim P}(e^{f_\lambda(X,\xi_u)})\} - \frac{\mathrm{KL}(\rho_u,\mu)+\log(1/\delta)}{n}\right\} \le 0.$$

Therefore, we deduce that with probability at least $1-\delta$,

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n X_i - \theta_0\right\|_2 &= \sup_{u\in\Sigma^{1/2}\mathbb{S}^{d-1}} \frac{1}{n}\sum_{i=1}^n u^\top\Sigma^{-1/2}(X_i-\theta_0) \\
&= \sup_{u\in\Sigma^{1/2}\mathbb{S}^{d-1}} \frac{1}{n\lambda}\sum_{i=1}^n \mathbb{E}_{\xi_u\sim\rho_u} f_\lambda(X_i,\xi_u) \\
&\le \inf_{\lambda\in[0,\frac{1}{2R}]}\left\{4\lambda R^2 + \frac{3\log(2/\delta)}{n\lambda}\right\} \\
&\overset{(i)}{=} 2R\sqrt{\frac{3\log(2/\delta)}{n}} = 2\sqrt{\frac{3}{n}}\cdot\left\{\sqrt{\operatorname{tr}(\Sigma)} + \sqrt{\|\Sigma\|_{\mathrm{op}}\log(2/\delta)}\right\}.
\end{aligned}$$

where (i) follows by choosing $\lambda = \frac{1}{2R}\sqrt{\frac{3\log(2/\delta)}{n}}$, which is at most $\frac{1}{2R}$ since $\frac{3\log(2/\delta)}{n} \le 1$ by assumption. The final conclusion follows by squaring both sides of the inequality above and using the inequality $(a+b)^2 \le 2a^2+2b^2$ for $a, b \in \mathbb{R}$. □

G Background on disintegrations

Our definition of MAR relies on the decomposition of a probability measure on a product space into the marginal distribution on one coordinate and a family of conditional distributions on the other. This can be achieved via the notion of disintegration. Let $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ be measurable spaces, and let P be a probability measure on the product space $(\mathcal{X}\times\mathcal{Y}, \mathcal{A}\otimes\mathcal{B})$. Further, let μ denote the marginal distribution of P on $(\mathcal{X}, \mathcal{A})$. We say that $(P_x)_{x\in\mathcal{X}}$ is a *disintegration of P into conditional distributions on $\mathcal{Y}$* if

(a) P_x is a probability measure on $(\mathcal{Y}, \mathcal{B})$, for each $x \in \mathcal{X}$;

(b) $x \mapsto P_x(B)$ is an $\mathcal{A}$-measurable function, for every $B \in \mathcal{B}$;

(c) $P(A \times B) = \int_A P_x(B)\, d\mu(x)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$.

In our setting, P denotes the joint distribution of a random pair (X, Y), taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. We interpret P_x as the conditional distribution of Y given $X = x$, even though it may be the case that the conditioning event has probability zero. Going further, we also interpret P_X as the conditional distribution of Y given X. Indeed, we then have for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$ that

$$\begin{aligned}\mathbb{E}\big(P_X(B)\mathbb{1}_A(X)\big) = \int_A P_x(B)\, d\mu(x) = P(A \times B) &= \mathbb{P}(X \in A, Y \in B) \\ &= \mathbb{E}\big(\mathbb{1}_A(X)\mathbb{1}_B(Y)\big),\end{aligned}$$

so $\mathbb{P}(Y \in B \,|\, X) = \mathbb{E}\big(\mathbb{1}_B(Y) \,|\, X\big) = P_X(B)$ almost surely. The following result, which follows from Dudley (2018, Theorems 10.2.1 and 10.2.2), provides a sufficient condition for the existence of a disintegration and may be regarded as a generalisation of Fubini's theorem for probability measures on the product of Polish spaces.

Theorem 43. *Suppose that $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ are Polish spaces with their corresponding Borel σ-algebras. Let P be a probability distribution on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$, with μ denoting the marginal distribution of P on $(\mathcal{X}, \mathcal{A})$. Then there exists a disintegration $(P_x)_{x \in \mathcal{X}}$ of P into conditional distributions on $\mathcal{Y}$ with the property that*

$$\int_{\mathcal{X} \times \mathcal{Y}} g(x, y)\, dP(x, y) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} g(x, y)\, dP_x(y) \right) d\mu(x),$$

for every P-integrable function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Moreover, the disintegration $(P_x)_{x \in \mathcal{X}}$ of P is unique in the sense that if there exists another disintegration $(\widetilde{P}_x)_{x \in \mathcal{X}}$ of P into conditional distributions on $\mathcal{Y}$, then $\widetilde{P}_x = P_x$ for μ-almost every $x \in \mathcal{X}$.

In order to apply this result in our missing data context, recall the random pair (X, Ω') taking values in $\mathcal{X} \times \{0, 1\}^d$ from (2). For each $\omega \in \{0, 1\}^d$, we assume the existence of disintegrations $(P_{x \circledast \omega})_{x \in \mathcal{X}}$ of the joint distribution of $(X \circledast \omega, \Omega')$ into conditional distributions on $\{0, 1\}^d$ as well as $(P_x)_{x \in \mathcal{X}}$ of the joint distribution of (X, Ω') into conditional distributions on $\{0, 1\}^d$. The existence of these disintegrations is guaranteed by Theorem 43 when $\mathcal{X}_j$ is a Polish space for each $j \in [d]$, because it then follows from Lemma 37 and its proof that $\mathcal{X} := \prod_{j=1}^d \mathcal{X}_j$ and $\mathcal{X}_\star := \prod_{j=1}^d \mathcal{X}_{j,\star}$ are Polish. Formally then, the condition $\mathbb{P}(\Omega' = \omega \,|\, X = x) = \mathbb{P}(\Omega' = \omega \,|\, X \circledast \omega = x \circledast \omega)$ in (2) means that $P_x(\omega) = P_{x \circledast \omega}(\omega)$. In fact, since the MAR definition refers to a family of distributions of $X \circledast \Omega'$, we need these disintegrations for each possible joint distribution of (X, Ω') with $X \sim P$ and $\mathbb{P}(\Omega' = \mathbf{1}_S) = \pi(S)$ for all $S \subseteq [d]$ (such disintegrations are again guaranteed to exist by Theorem 43 when $\mathcal{X}_j$ is a Polish space for each $j \in [d]$).

H MCAR lower bounds for mean estimation

Recall the definition of an f-divergence $\mathrm{Div}_f(\cdot,\cdot)$ from (28). Lemma 45 below relates the f-divergence of two MCAR distributions on $\mathcal{X}_\star$ to a notion of average f-divergence given in Definition 44 below. For probability measures $P, Q \in \mathcal{P}(\mathcal{X})$, we let, for $S \subseteq [d]$, P_S and Q_S denote their respective marginal distributions on $\mathcal{X}_S$.

Definition 44. *Let $P, Q \in \mathcal{P}(\mathcal{X})$, let $\pi \in \mathcal{P}(2^{[d]})$ and let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. We define the* average f-divergence *between P and Q with respect to π to be*

$$\mathrm{ADiv}_f(P, Q; \pi) \coloneqq \sum_{S \subseteq [d]} \pi(S) \cdot \mathrm{Div}_f\big(P_S, Q_S\big),$$

where we adopt the convention that $\mathrm{Div}_f(P_S, Q_S) \coloneqq 0$ if $S = \emptyset$.

We will write $\mathrm{ATV}(\cdot,\cdot;\pi)$ and $\mathrm{AKL}(\cdot,\cdot;\pi)$ respectively for the average total variation distance and average Kullback–Leibler divergence with respect to π. It is worth noting that the average total variation distance is a pseudo-metric but not necessarily a metric on $\mathcal{P}(\mathcal{X})$; indeed, we have $\mathrm{ATV}(P, Q; \pi) = 0$ whenever P and Q have the same marginal distributions on the support of π.

The following lemma shows how an f-divergence between two MCAR distributions on $\mathcal{P}(\mathcal{X}_\star)$ can be computed as an average f-divergence on $\mathcal{P}(\mathcal{X})$ in the sense of Definition 44.

Lemma 45. *Let $P, Q \in \mathcal{P}(\mathcal{X})$ and let $\pi \in \mathcal{P}(2^{[d]})$. Then*

$$\mathrm{Div}_f\big(\mathsf{MCAR}_{(\pi,P)}, \mathsf{MCAR}_{(\pi,Q)}\big) = \mathrm{ADiv}_f(P, Q; \pi).$$

Proof of Lemma 45. Recall the definition of $\mathcal{X}^{(S)}$ and $\mathcal{X}_S$ from Section A. For $A \in \mathcal{B}(\mathcal{X}_\star)$, note that $A \cap \mathcal{X}^{(S)} \in \mathcal{B}(\mathcal{X}^{(S)})$ and define $(A \cap \mathcal{X}^{(S)})_S \coloneqq \big\{x_S : x \in A \cap \mathcal{X}^{(S)}\big\} \in \mathcal{B}(\mathcal{X}_S)$. Let $P^{(S)} \in \mathcal{P}(\mathcal{X}_\star)$ be defined as $P^{(S)}(A) \coloneqq P_S\big((A \cap \mathcal{X}^{(S)})_S\big)$ for $A \in \mathcal{B}(\mathcal{X}_\star)$, so that $P^{(S)}$ is supported on $\mathcal{X}^{(S)}$. For each $S \subseteq [d]$, we can apply the Lebesgue decomposition theorem to obtain the decomposition $P^{(S)} = P^{(S)}_{\mathrm{ac}} + P^{(S)}_{\mathrm{sing}}$ (with respect to $Q^{(S)}$). Then, with respect to $\mathsf{MCAR}_{(\pi,Q)}$,

$$\big(\mathsf{MCAR}_{(\pi,P)}\big)_{\mathrm{ac}} = \left(\sum_{S \subseteq [d]} \pi(S) \cdot P^{(S)}\right)_{\mathrm{ac}} = \sum_{S \subseteq [d]} \pi(S) \cdot P^{(S)}_{\mathrm{ac}}$$

and

$$\big(\mathsf{MCAR}_{(\pi,P)}\big)_{\mathrm{sing}} = \left(\sum_{S \subseteq [d]} \pi(S) \cdot P^{(S)}\right)_{\mathrm{sing}} = \sum_{S \subseteq [d]} \pi(S) \cdot P^{(S)}_{\mathrm{sing}}.$$

Hence, since $\mathcal{X}_\star = \sqcup_{S \subseteq [d]} \mathcal{X}^{(S)}$,

$$\begin{aligned}
&\mathrm{Div}_f\big(\mathsf{MCAR}_{(\pi,P)}, \mathsf{MCAR}_{(\pi,Q)}\big) \\
&\quad= \int_{\mathcal{X}_\star} f\left(\frac{\mathrm{d}\sum_{S \subseteq [d]} \pi(S) P^{(S)}_{\mathrm{ac}}}{\mathrm{d}\sum_{S \subseteq [d]} \pi(S) Q^{(S)}}\right) \sum_{S \subseteq [d]} \pi(S)\, \mathrm{d}Q^{(S)} + M_f \cdot \sum_{S \subseteq [d]} \pi(S) P^{(S)}_{\mathrm{sing}}(\mathcal{X}_\star) \\
&\quad= \sum_{S \subseteq [d]} \pi(S) \int_{\mathcal{X}^{(S)}} f\left(\frac{\mathrm{d}P^{(S)}_{\mathrm{ac}}}{\mathrm{d}Q^{(S)}}\right) \mathrm{d}Q^{(S)} + M_f \cdot \sum_{S \subseteq [d]} \pi(S) P^{(S)}_{\mathrm{sing}}(\mathcal{X}^{(S)}) = \mathrm{ADiv}_f(P, Q; \pi),
\end{aligned}$$

as desired. □

Very often, it is convenient to apply Pinsker's inequality to total variation distances, in order to control them via (more tractable) Kullback–Leibler divergences. We remark that in doing so directly to the left-hand side of Lemma 45, we obtain

$$\begin{aligned}\mathrm{TV}\big(\mathsf{MCAR}_{(\pi,P)}, \mathsf{MCAR}_{(\pi,Q)}\big) &\leq \frac{1}{2^{1/2}} \cdot \mathrm{KL}^{1/2}\big(\mathsf{MCAR}_{(\pi,P)}, \mathsf{MCAR}_{(\pi,Q)}\big) \\ &= \frac{1}{2^{1/2}} \mathrm{AKL}^{1/2}(P, Q; \pi) = \frac{1}{2^{1/2}} \bigg\{ \sum_{S \subseteq [d]} \pi(S) \cdot \mathrm{KL}\big(P_S, Q_S\big) \bigg\}^{1/2}.\end{aligned}$$

On the other hand, applying Pinsker's inequality to the right-hand side of Lemma 45 yields the bound

$$\mathrm{ATV}(P, Q; \pi) = \sum_{S \subseteq [d]} \pi(S) \cdot \mathrm{TV}\big(P_S, Q_S\big) \leq \frac{1}{2^{1/2}} \sum_{S \subseteq [d]} \pi(S) \cdot \mathrm{KL}^{1/2}\big(P_S, Q_S\big),$$

which is an improvement, by Jensen's inequality.

We now state two lower bounds in the MCAR setting, beginning with the univariate setting.

Proposition 46. *Let $n \in \mathbb{N}$, $q \in (0, 1]$ and $\Theta \coloneqq \mathbb{R}$.*

(a) Let $\sigma > 0$ and $\delta \in (0, 1/4]$. Then, writing $\mathcal{P}_\theta \coloneqq \big\{\mathsf{MCAR}^{\otimes n}_{(q, \mathsf{N}(\theta, \sigma^2))}\big\}$, we have

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, |\cdot|^2\big) \begin{cases} \geq \dfrac{\sigma^2 \log(1/\delta)}{20nq} & \text{if } \delta \geq \dfrac{(1-q)^n}{2} \\ = \infty & \text{if } \delta < \dfrac{(1-q)^n}{2}. \end{cases}$$

(b) Let $K > 0$ and $\delta \in (0, 1/4]$. Then, with $\mathcal{P}_{\mathrm{b}}(\theta, K)$ as in (81) and writing $\mathcal{P}_\theta \coloneqq \big\{\mathsf{MCAR}^{\otimes n}_{(q,P)} : P \in \mathcal{P}_{\mathrm{b}}(\theta, K)\big\}$, we have

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, |\cdot|^2\big) \begin{cases} \geq \dfrac{K^2 \log(1/\delta)}{80nq} & \text{if } \delta \geq \exp(-nq/2) \\ = \infty & \text{if } \delta < \dfrac{(1-q)^n}{2}. \end{cases}$$

Proof. *(a)* Let $\theta_1 \coloneqq 0$ and $\theta_2 \coloneqq \sigma\sqrt{\frac{1}{nq} \log\big(\frac{1}{4\delta(1-\delta)}\big)}$. By Lemma 45, we have

$$\begin{aligned}\mathrm{KL}\big(\mathsf{MCAR}^{\otimes n}_{(\pi, \mathsf{N}(\theta_1, \sigma^2))}, \mathsf{MCAR}^{\otimes n}_{(\pi, \mathsf{N}(\theta_2, \sigma^2))}\big) &= nq \cdot \mathrm{KL}\big(\mathsf{N}(\theta_1, \sigma^2), \mathsf{N}(\theta_2, \sigma^2)\big) \\ &= \frac{1}{2} \log\bigg(\frac{1}{4\delta(1-\delta)}\bigg) < \log\bigg(\frac{1}{4\delta(1-\delta)}\bigg).\end{aligned}$$

Therefore, by Ma, Verchand and Samworth (2024, Corollary 6 and Theorem 4), we deduce that for $\delta \in (0, 1/4]$,

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, |\cdot|^2\big) \geq \bigg(\frac{\theta_1 - \theta_2}{2}\bigg)^2 = \frac{\sigma^2 \log\big(\frac{1}{4\delta(1-\delta)}\big)}{4nq} \geq \frac{\sigma^2 \log(1/\delta)}{20nq}.$$

Moreover, for any $\theta_1, \theta_2 \in \mathbb{R}$, we have

$$\begin{aligned}&\mathrm{TV}\big(\mathsf{MCAR}^{\otimes n}_{(q,\mathsf{N}(\theta_1,\sigma^2))}, \mathsf{MCAR}^{\otimes n}_{(q,\mathsf{N}(\theta_2,\sigma^2))}\big)\\ &\quad= \sup_{A \in \mathcal{B}(\mathbb{R}^n_\star)\setminus\{\star\}^n} \Big\{\mathsf{MCAR}^{\otimes n}_{(q,\mathsf{N}(\theta_1,\sigma^2))}(A) - \mathsf{MCAR}^{\otimes n}_{(q,\mathsf{N}(\theta_2,\sigma^2))}(A)\Big\} \leq 1-(1-q)^n,\end{aligned}$$

where both steps follow since $\mathsf{MCAR}^{\otimes n}_{(q,\mathsf{N}(\theta_1,\sigma^2))}(\{\star\}^n) = \mathsf{MCAR}^{\otimes n}_{(q,\mathsf{N}(\theta_2,\sigma^2))}(\{\star\}^n) = (1-q)^n$. Therefore, by Ma, Verchand and Samworth (2024, Lemma 5), we have that $\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq (\theta_1-\theta_2)^2/4$ for $\delta < \frac{(1-q)^n}{2}$. The claim follows since θ_1, θ_2 were arbitrary.

(b) Define $P_1, P_2 \in \mathcal{P}(\mathbb{R})$ by

$$P_1(\{x\}) \coloneqq \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = K \end{cases} \quad \text{and} \quad P_2(\{x\}) \coloneqq \begin{cases} \frac{1-a}{2} & \text{if } x = 0 \\ \frac{1+a}{2} & \text{if } x = K, \end{cases}$$

where $a \coloneqq \sqrt{\frac{1}{nq}\log\big(\frac{1}{4\delta(1-\delta)}\big)} \leq \sqrt{\frac{\log(1/\delta)}{nq}} \leq 1/\sqrt{2}$ for $\delta \in [e^{-nq/2}, 1/4]$. Let $\theta_1 \coloneqq \mathbb{E}_{P_1}(X) = K/2$ and $\theta_2 \coloneqq \mathbb{E}_{P_2}(X) = (1+a)K/2$, so that $P_\ell \in \mathcal{P}_{\mathrm{b}}(\theta_\ell, K)$ for $\ell \in \{1,2\}$. Moreover, by Lemma 45,

$$\mathrm{KL}\big(\mathsf{MCAR}^{\otimes n}_{(q,P_1)}, \mathsf{MCAR}^{\otimes n}_{(q,P_2)}\big) = nq\mathrm{KL}(P_1,P_2) = \frac{nq}{2}\log\bigg(\frac{1}{1-a^2}\bigg) < nqa^2 = \log\bigg(\frac{1}{4\delta(1-\delta)}\bigg),$$

where the inequality follows because $\log\big(\frac{1}{1-x^2}\big) < 2x^2$ for $x \in (0, 1/\sqrt{2}]$. Hence, by Ma, Verchand and Samworth (2024, Corollary 6 and Theorem 4), we deduce that for $\delta \in [e^{-nq/2}, 1/4]$,

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, |\cdot|^2\big) \geq \bigg(\frac{\theta_1-\theta_2}{2}\bigg)^2 \geq \frac{K^2\log(1/\delta)}{80nq}.$$

Now let $\theta \in \mathbb{R}$, $P_1' \coloneqq \mathsf{Unif}[0,K]$ and $P_2' \coloneqq \mathsf{Unif}[\theta, \theta+K]$. Then by the same argument as in part *(a)*, we have

$$\mathrm{TV}\big(\mathsf{MCAR}^{\otimes n}_{(q,P_1')}, \mathsf{MCAR}^{\otimes n}_{(q,P_2')}\big) \leq 1-(1-q)^n.$$

Therefore, by Ma, Verchand and Samworth (2024, Lemma 5), we have that $\mathcal{M}(\delta, \mathcal{P}_\Theta, |\cdot|^2) \geq \theta^2/4$ for $\delta < \frac{(1-q)^n}{2}$. The claim follows since θ_1, θ_2 were arbitrary. □

Our next proposition lower bounds the minimax quantile for mean estimation in the multivariate Gaussian setting when the covariance matrix is diagonal.

Proposition 47. *Let $\delta \in (0, 1/4]$, $\Sigma = (\Sigma_{jk})_{j,k\in[d]} \in \mathcal{S}^{d\times d}_{++}$ be diagonal, $\pi \in \mathcal{P}(2^{[d]})$, and let $P_\theta \coloneqq \mathsf{N}(\theta, \Sigma)$ for $\theta \in \mathbb{R}^d$. Then, writing $\mathcal{P}_\theta \coloneqq \big\{\mathsf{MCAR}^{\otimes n}_{(\pi,P_\theta)}\big\}$, we have*

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \gtrsim \frac{\mathrm{tr}\big(\Sigma^{\mathrm{IPW}}\big)}{n} + \frac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}}\log(1/\delta)}{n}.$$

Proof. We consider two separate constructions to capture each of the terms in the lower bound. For the first, let $\mathcal{V} \coloneqq \{0,1\}^d$ and for each $v = (v_1, \ldots, v_d)^\top \in \mathcal{V}$, set $\theta_v = (\theta_{v,1}, \ldots, \theta_{v,d})^\top \coloneqq a \odot v$, where $a = (a_1, \ldots, a_d)^\top \in \mathbb{R}^d$ is given by $a_j \coloneqq \frac{4}{3}\sqrt{\Sigma_{jj}/(nq_j)}$

for $j \in [d]$. Define $\Theta_0 := \{\theta_v : v \in \mathcal{V}\}$, which has diameter $D := \frac{4}{3}\sqrt{\mathrm{tr}(\Sigma^{\mathrm{IPW}})/n}$. For any $v, v' \in \mathcal{V}$ that differ only in their jth coordinates, we have by Pinsker's inequality and Lemma 45 that

$$\begin{aligned}
\mathrm{TV}\big(\mathsf{MCAR}^{\otimes n}_{(\pi,P_{\theta_v})}, \mathsf{MCAR}^{\otimes n}_{(\pi,P_{\theta_{v'}})}\big) &\le \bigg\{\frac{n}{2} \cdot \mathrm{KL}\big(\mathsf{MCAR}_{(\pi,P_{\theta_v})}, \mathsf{MCAR}_{(\pi,P_{\theta_{v'}})}\big)\bigg\}^{1/2} \\
&= \bigg\{\frac{n}{2} \sum_{S \subseteq [d]} \pi(S) \cdot \mathrm{KL}\big((P_{\theta_v})_S, (P_{\theta_{v'}})_S\big)\bigg\}^{1/2} \\
&= \bigg\{\frac{n}{2} \sum_{S \subseteq [d]} \pi(S) \cdot \sum_{k \in S} \frac{(\theta_{v,k} - \theta_{v',k})^2}{2\Sigma_{kk}}\bigg\}^{1/2} \\
&= \bigg\{\frac{n}{4} \sum_{S \subseteq [d]: j \in S} \pi(S) \cdot \frac{a_j^2}{\Sigma_{jj}}\bigg\}^{1/2} = \frac{2}{3}.
\end{aligned}$$

Therefore, by Assouad's Lemma (e.g., Ma, Verchand and Samworth, 2024, Lemma 23),

$$\inf_{\widehat{\theta}_n \in \widehat{\Theta}_n} \sup_{\theta_0 \in \Theta_0} \mathbb{E}_{\mathsf{MCAR}^{\otimes n}_{(\pi,P_{\theta_0})}}\big(\|\widehat{\theta}_n - \theta\|_2^2\big) \ge \frac{4\,\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{27n}.$$

Applying Ma, Verchand and Samworth (2024, Theorem 8), with $\epsilon = 3/40$ therein, we deduce that for $\delta \in (0, 1/15]$,

$$\mathcal{M}_-\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \ge \frac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{100n}.$$

We then apply Ma, Verchand and Samworth (2024, Theorem 4 and Proposition 9), with $A = k = 2$ therein, to deduce that for $\delta \in (0, 1/4]$,

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \ge \frac{\mathrm{tr}(\Sigma^{\mathrm{IPW}})}{2^6 \cdot 3^2 \cdot 5^2 \cdot n}. \tag{97}$$

Our second construction involves just two distributions. Let $j_0 := \mathrm{sargmax}_{j \in [d]} \Sigma_{jj}/q_j$ and set $\theta_1 := 0$, $\theta_2 := \sqrt{\frac{\Sigma_{j_0 j_0}}{n q_{j_0}} \log\big(\frac{1}{4\delta(1-\delta)}\big)}\, e_{j_0}$. Then by Lemma 45,

$$\mathrm{KL}\big(\mathsf{MCAR}^{\otimes n}_{(\pi,P_{\theta_1})}, \mathsf{MCAR}^{\otimes n}_{(\pi,P_{\theta_2})}\big) = n \cdot \mathrm{AKL}\big(P_{\theta_1}, P_{\theta_2}; \pi\big) = \frac{1}{2}\log\bigg(\frac{1}{4\delta(1-\delta)}\bigg).$$

By Ma, Verchand and Samworth (2024, Theorem 4 and Corollary 6), we have for $\delta \in (0, 1/4]$ that

$$\mathcal{M}\big(\delta, \mathcal{P}_\Theta, \|\cdot\|_2^2\big) \ge \frac{\|\Sigma^{\mathrm{IPW}}\|_{\mathrm{op}} \log(1/\delta)}{20n}. \tag{98}$$

Finally, combining (97) and (98) yields the desired result. □

I Robust mean estimation algorithms for completely observed data

In this section, we briefly review some of the robust mean estimators.

I.1 Median of means

Given a sequence of real numbers $(a_i)_{i=1}^n$, let $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(n)}$ be the sorted version of $(a_i)_{i=1}^n$. Then the median of $(a_i)_{i=1}^n$ is defined as $\text{Median}(a_1, \ldots, a_n) \coloneqq a_{(\lfloor n/2 \rfloor + 1)}$.

Algorithm 2 Median of means

Input: Data $(x_i)_{i \in [n]} \in \mathbb{R}$ and the number of blocks $M \in \mathbb{N}$

Output: $\widehat{\theta}_n \in \mathbb{R}$

1: **function** MEDIAN_OF_MEANS($x_1, \ldots, x_n; M$)
2: Randomly partition $[n]$ into M disjoint subsets $(B_m)_{m=1}^M$ such that $\lfloor n/M \rfloor \leq |B_m| \leq \lceil n/M \rceil$ for all $m \in [M]$
3: **for** $m \in [M]$ **do**
4: $\bar{x}_m \leftarrow \frac{1}{|B_m|} \sum_{i \in B_m} x_i$
5: **end for**
6: $\widehat{\theta}_n \leftarrow \text{Median}(\bar{x}_1, \ldots, \bar{x}_M)$
7: **return** $\widehat{\theta}_n$
8: **end function**

Lemma 48 (Lerasle and Oliveira, 2011, Proposition 1)**.** *Let* $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \in \mathcal{P}_{L^2}(\theta_0, \sigma^2)$, $\delta \in \big[\exp(1 - n/2), 1\big]$, $M \coloneqq \lceil \log(1/\delta) \rceil$ *and* $\widehat{\theta}_n \coloneqq$ MEDIAN_OF_MEANS$(X_1, \ldots, X_n; M)$. *Then with probability at least* $1 - \delta$,

$$\big(\widehat{\theta}_n - \theta_0\big)^2 \leq \frac{24 e \sigma^2 \log(e/\delta)}{n}.$$

I.2 Univariate trimmed mean

For $\alpha \leq \beta \in \mathbb{R}$, we define $T_{\alpha,\beta} : \mathbb{R} \to [\alpha, \beta]$ by

$$T_{\alpha,\beta}(x) \coloneqq \begin{cases} \beta & \text{if } x \geq \beta \\ x & \text{if } \alpha < x < \beta \\ \alpha & \text{if } x \leq \alpha. \end{cases}$$

The univariate trimmed mean estimator is defined in Algorithm 3.

I.3 Robust Descent

Algorithms 4 and 5 below provide pseudo-code for the robust (block) descent algorithm of Depersin and Lecué (2022b, Algorithm 1). To describe the SOLVE_SDP function in line 7 of Algorithm 4, define

$$\Delta_M \coloneqq \left\{ (w_m)_{m=1}^M : 0 \leq w_m \leq \frac{10}{9M}, \sum_{m=1}^M w_m = 1 \right\},$$

Algorithm 3 Univariate trimmed mean

Input: Data $(x_i)_{i\in[n]} \in \mathbb{R}$, contamination parameter $\epsilon \in [0,1)$, and tolerance parameter $\delta \in (0,1]$

Output: $\widehat{\theta}_n \in \mathbb{R}$

1: **function** UNIVARIATE_TRIMMED_MEAN($x_1,\dots,x_n;\epsilon,\delta$)
2: Randomly partition $\{x_1,\dots,x_n\}$ into two disjoint sets $\{y_1,\dots,y_{n/2}\}$ and $\{z_1,\dots,z_{n/2}\}$
3: Arrange $\{z_1,\dots,z_{n/2}\}$ in increasing order $\widetilde{z}_1 \le \cdots \le \widetilde{z}_{n/2}$
4: $\eta \leftarrow 8\epsilon + 24\log(4/\delta)/n$
5: $\alpha \leftarrow \widetilde{z}_{n\eta/2}$, $\beta \leftarrow \widetilde{z}_{n(1-\eta)/2}$
6: $\widehat{\theta}_n \leftarrow \frac{2}{n}\sum_{i=1}^{n/2} T_{\alpha,\beta}(y_i)$
7: **return** $\widehat{\theta}_n$
8: **end function**

and let $\mathcal{V} \coloneqq \{V \in \mathcal{S}_+^{d\times d} : \mathrm{tr}(V) = 1\}$. The SOLVE_SDP$(\bar{x}_1,\dots,\bar{x}_M;\theta)$ function in line 7 of Algorithm 4 provides an approximate maximiser $\widehat{V} \in \mathcal{V}$ of the function $h_\theta : \mathcal{V} \to \mathbb{R}$ given by

$$h_\theta(V) \coloneqq \min_{(w_1,\dots,w_m)\in\Delta_M} \mathrm{tr}\Big\{V^\top \sum_{m=1}^{M} w_m\big(\bar{x}_m - \theta\big)\big(\bar{x}_m - \theta\big)^\top\Big\}. \tag{99}$$

A full description of SOLVE_SDP can be found in Depersin and Lecué (2022b, Section 4); see in particular their Algorithms 2 and 3.

In Algorithm 4 below, we invoke the notation that if $S \subseteq \mathbb{R}^d$ is compact and $f : S \to \mathbb{R}$ is continuous, then $\mathrm{sargmax}_{x\in S} f(x)$ denotes the smallest element of the argmax set in the lexicographic ordering.

Algorithm 4 Robust Block Descent

Input: $\bar{x}_1,\dots,\bar{x}_M \in \mathbb{R}^d$

Output: $\widehat{\theta}_n \in \mathbb{R}^d$

1: **function** ROBUST_BLOCK_DESCENT($\bar{x}_1,\dots,\bar{x}_M$)
2: $T \leftarrow \lceil \log(8\sqrt{d})/\log(10/9) \rceil$
3: **for** $j \in [d]$ **do**
4: $\widehat{\theta}_j^{(0)} \leftarrow \mathrm{Median}(\bar{x}_{1j},\dots,\bar{x}_{Mj})$
5: **end for**
6: **for** $t \in [T]$ **do**
7: $\widehat{V}^{(t)} \leftarrow$ SOLVE_SDP$(\bar{x}_1,\dots,\bar{x}_M;\widehat{\theta}^{(t-1)})$; see (99)
8: $\widehat{v}^{(t)} \leftarrow \mathrm{sargmax}_{u\in\mathbb{S}^{d-1}} u^\top \widehat{V}^{(t)} u$
9: $s^{(t)} \leftarrow -\mathrm{Median}\big((\bar{x}_m - \widehat{\theta}^{(t-1)})^\top \widehat{v}^{(t)} : m \in [M]\big)$
10: $\widehat{\theta}^{(t)} \leftarrow \widehat{\theta}^{(t-1)} - s^{(t)}\widehat{v}^{(t)}$
11: **end for**
12: $\widehat{\theta}_n \leftarrow \widehat{\theta}^{(T)}$
13: **return** $\widehat{\theta}_n$
14: **end function**

Algorithm 5 Robust Descent

Input: Data $(x_i)_{i\in[n]} \in \mathbb{R}^d$, contamination parameter $\epsilon \in [0,1)$, and tolerance parameter $\delta \in (0,1]$

Output: $\widehat{\theta}_n \in \mathbb{R}^d$

function ROBUST_DESCENT($x_1, \ldots, x_n; \epsilon, \delta$)
    $M \leftarrow \lceil 300\big(2\epsilon n + \log(2/\delta)\big) \vee 180{,}000 \log(2/\delta) \rceil \wedge n$
    Randomly draw M disjoint subsets $(B_m)_{m=1}^M$ from $[n]$, each with size $\lfloor n/M \rfloor$
    for $m \in [M]$ **do**
        $\bar{x}_m \leftarrow |B_m|^{-1} \sum_{i\in B_m} x_i$
    end for
    $\widehat{\theta}_n \leftarrow$ ROBUST_BLOCK_DESCENT($\bar{x}_1, \ldots, \bar{x}_M$)
    return $\widehat{\theta}_n$
end function