
Robust Estimation Under Heterogeneous Corruption Rates

Syomantak Chaudhuri
University of California, Berkeley

Jerry Li
University of Washington

Thomas A. Courtade
University of California, Berkeley

Abstract

We study the problem of robust estimation under heterogeneous corruption rates, where each sample may be independently corrupted with a known but non-identical probability. This setting arises naturally in distributed and federated learning, crowdsourcing, and sensor networks, yet existing robust estimators typically assume uniform or worst-case corruption, ignoring structural heterogeneity. For mean estimation for multivariate bounded distributions and univariate gaussian distributions, we give tight minimax rates for all heterogeneous corruption patterns. For multivariate gaussian mean estimation and linear regression, we establish the minimax rate for squared error up to a factor of $\sqrt{d}$, where d is the dimension. Roughly, our findings suggest that samples beyond a certain corruption threshold may be discarded by the optimal estimators – this threshold is determined by the empirical distribution of the corruption rates given.

1 Introduction

In traditional statistics, we typically rely on the assumption that our data is generated in a “nice”, i.i.d. manner from some population distribution. Robust statistics can be seen as a relaxation of these assumptions, aiming to ensure meaningful performance even when the data that has been corrupted by a small fraction of arbitrary, but potentially adversarially chosen, outliers. First proposed in seminal work by statisticians such as Huber, Tukey, and Anscombe more than fifty years ago [1–4], this problem remains highly relevant to this day, and has received considerable recent attention from the statistics, machine learning, and theoretical computer science community, see e.g. [5–20]. A full survey of this line of work is beyond the scope of this paper, see e.g. [21] for a more comprehensive overview.

There are a number of related ways in the literature to model such outliers. For instance, one classical setting is known as the *Huber contamination model* [22]. Here, there is a “nice” distribution D , typically assumed to be from some well-behaved class of distributions, and we receive n samples $Z_1, \dots, Z_n$ from a distribution $D' = (1 - \epsilon)D + \epsilon N$ for some arbitrary noise distribution N , and some error parameter $\epsilon > 0$. In other words, to generate a sample point Z_i , we generate a clean sample from D with probability $(1 - \epsilon)$, and we sample an outlier from a noise distribution with probability ϵ . There are also a number of other, more powerful, adversaries considered in the literature, for instance, which allow the adversary to choose the corrupted points adaptively based on the clean samples; see e.g. [23–25] for a more detailed discussion. However, by now, at least for many basic estimation tasks such as mean estimation, tight (or nearly tight) rates are known for many parametric families of distributions under all of these corruption models.

But, all of these models share an unfortunate drawback: namely, they all essentially assume a homogeneous prior on the error rates across the dataset. In the Huber contamination model, for

example, every sample is corrupted with the same probability ϵ . In practice, this is usually not the case: we often obtain data from various heterogeneous sources, and consequently we should have different priors on the cleanliness of different data points in our dataset. As an example, many large-scale biomedical datasets are agglomerated in a distributed fashion over different institutions and at different points in time, and it is unreasonable to assume that the rate of outliers ought to be similar over the various sub-parts of the data. But because all existing robust statistics techniques assume a uniform noise prior, it is unclear whether can obtain the optimal statistical rates for such datasets, or what these rates even are in the first place.

Motivated by these considerations, we propose a natural generalization of the Huber contamination model with heterogeneous corruption rates, where every data point is corrupted with a potentially different error rate. Formally:

Definition 1 (λ -contamination). Let n be the number of samples, and let $\lambda = (\lambda_1, \dots, \lambda_n) \in [0, 1]^n$. Let P be some distribution. We say a dataset $Z = (Z_1, \dots, Z_n)$ is a λ -contaminated set of samples from P if $Z_i = (1 - B_i)X_i + B_i\tilde{X}_i$, where:

- $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$ is a set of clean samples from P ,
- $B_i \sim \text{Bern}(\lambda_i)$ are independent Bernoulli random variables for $i = 1, \dots, n$, and
- The outliers $(\tilde{X}_1, \dots, \tilde{X}_n)$ are sampled from some joint distribution conditioned on the realizations of $X_1, \dots, X_n$ and $B_1, \dots, B_n$.

When a set of samples is generated this way, we denote this by $Z \sim_\lambda P$.

In other words, we assume that the probability that the i -th sample is corrupted is λ_i , for different values of λ_i . We pause to make a couple of remarks on this definition. First, when all the λ_i are the same, this is essentially the standard Huber contamination model, except we allow that the outliers may be chosen adaptively. Second, we will assume that we know λ exactly. However, our upper bounds also naturally generalize to the setting where we only have approximate knowledge of the λ_i , i.e., we just need a constant factor approximation $\hat{\lambda}_i$ of the corruption rate for our results to hold (i.e., $\lambda_i \leq c\hat{\lambda}_i$). Such upper bounds on the corruption rates can often be estimated from past performance in practice. Finally, just as in standard robust statistics, there are many similar but not equivalent definitions we can consider here, for instance, corresponding to oblivious contamination or strong corruption [21, 23–25]. Our proposed error upper and lower bounds hold for both adaptive and non-adaptive adversaries. We leave the exploration between the differences between these definitions as an interesting future direction.

From a more conceptual point of view, the main challenge in this heterogeneous setting is how to effectively “mix” the information from data points with different noise levels. As an illuminating example, one important special case of the λ -contamination model is *semi-verified* learning [7, 9], where there is a (very) small amount of trusted data, and a large amount of potentially very noisy data, which corresponds to the setting where $\lambda_i = 0$ for a small number of i , and $\lambda_i = \gamma$ for some γ close to 1 for the rest of the data. In general, robust estimation is impossible if $\lambda_i \geq 1/2$ for all i , however, it turns out that in the semi-verified setting, one can obtain consistent estimators by combining the information from the two sets of samples [7]. This example illustrates the main technical challenge of the general heterogeneous setting. To obtain the tight rates, one must correctly identify how to incorporate information from the noisier samples into the information from cleaner ones, and to do so smoothly as a function of λ_i .

1.1 Our results

We establish lower and upper bounds for the statistical rates for a number of fundamental estimation tasks in the λ -contaminated setting. We first consider robust mean estimation in the heterogeneous setting, one of the most important and well-studied problems in robust statistics. For any distribution P , let $\mu_P = \mathbb{E}_{X \sim P}[X]$ be the expectation of P . We will measure the effectiveness of our estimators using the following minimax metrics:

Definition 2 (Minimax, Minimax PAC rates for heterogeneous robust mean estimation [26, 27]). Let $\mathcal{D}$ be a set of distributions over $\mathbb{R}^d$, and let $\lambda \in [0, 1]^n$. The *minimax rate for λ -corrupted mean estimation* for the class $\mathcal{D}$ is defined to be

$$L(\lambda, \mathcal{D}) = \inf_M \sup_{P \in \mathcal{D}} \mathbb{E}_{Z \sim_\lambda P} [\|M(Z) - \mu_P\|_2^2] , \quad (1)$$

where the infimum is taken over all estimators M . The *minimax PAC rate for λ -corrupted mean estimation* for the class $\mathcal{D}$ is defined to be

$$L_{\text{PAC}}(\lambda, \mathcal{D}, \delta) = \inf_M \sup_{P \in \mathcal{D}} \inf \left\{ t \in [0, \infty) : \Pr_{\mathbf{Z} \sim \lambda P} [\|M(\mathbf{Z}) - \mu_P\|_2^2 \geq t] \leq \delta \right\}. \quad (2)$$

Readers can refer to Ma et al. [26] for more details on the minimax PAC rate. Stated simply, the minimax PAC rate is simply the smallest rate so that the probability that the estimator exceeds $L_{\text{PAC}}(\lambda, \mathcal{D}, \delta)$ is small. Since by standard techniques, we can boost the error probability δ , we will typically focus on the setting where δ is a small constant, and we will let $L_{\text{PAC}}(\lambda, \mathcal{D}) := L_{\text{PAC}}(\lambda, \mathcal{D}, 1/5)$. We use this notion because it is necessary in the robust setting. In settings where the unknown mean could be unbounded, no estimator can achieve finite expected error, even in the standard Huber contamination model. This is because all samples may be corrupted with exponentially small probability, in which case the error could be arbitrarily large.

Mean Estimation for Bounded Distributions: We first consider the class of bounded, multivariate distributions. For this setting, we show:

Theorem 1. *Let $\mathcal{D}_r^b$ be the set of all distributions on $\mathbb{R}^d$ supported on the l_2 ball of radius r . Then,*

$$L(\lambda, \mathcal{D}_r^b) \simeq r^2 f(\lambda, 1), \text{ where } f(\lambda, k) = \min_{t \in [0, 1]} \left(\frac{k}{|\{i : \lambda_i \leq t\}|} + t^2 \right). \quad (3)$$

Moreover, the optimal estimator can be implemented in nearly-linear time.

The function $f(\lambda, k)$ can be thought of as a measure of the “effective” error rate of the dataset, and will play a crucial role in many of our results going forward. Indeed, Theorem 1 implies that the optimal robust mechanism chooses a corruption level t and only utilizes the data having corruption rate below t , i.e., there is no improvement by considering the other samples up to constant factors.

Mean Estimation for Gaussians: Our next result concerns heterogeneous mean estimation for Gaussians. Our main result can be summarized as follows:

Theorem 2 (informal, see Theorem 4). *Let $\mathcal{D}_d^N$ the set of all multivariate Gaussian distributions on $\mathbb{R}^d$ with identity covariance. Suppose that $f(\lambda, d) = O(1)$, then $d^{-1/2} f(\lambda, d) \lesssim L_{\text{PAC}}(\lambda, \mathcal{D}_d^N) \lesssim f(\lambda, d)$.*

In other words, we show that, in the regime where there are sufficiently many relatively clean samples, $f(\lambda, d)$ dictates the minimax rate, up to a $\sqrt{d}$ factor. Note that this is arguably the most interesting regime from a statistical perspective, as it is the only regime where the recovered Gaussian has non-trivial statistical overlap with the true Gaussian. We note that our lower bound technique also yields non-trivial bounds in more general settings, however, they are somewhat more difficult to interpret—see the supplementary material for more discussion.

We pause to make a couple of remarks on this result. First, in constant dimensions, our bound is tight up to constant factors, so our bound is tight in this regime. Second, we note that naïve multi-dimensional techniques such as coordinate-wise methods would lose a $\Theta(d)$ factor in the squared ℓ_2 -error, so our bound represents a polynomial improvement over baseline methods.

Linear Regression: Finally, we turn our attention to Gaussian design linear regression with heterogeneous corruptions. We assume the uncorrupted covariates are Gaussian with covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ and noise rate σ^2 . Then, for any choice of regression coefficients $\beta \in \mathbb{R}^d$, we assume that the clean data is of the form $(W, Y) \in \mathbb{R}^{d+1}$ where $W \sim \mathcal{N}(0, \Sigma)$, and conditioned on W , $Y \sim \mathcal{N}(W^T \beta, \sigma^2)$. We let $\mathcal{D}(\Sigma, \sigma^2)$ the family of distributions over (W, Y) of this form. For any $P \in \mathcal{D}(\Sigma, \sigma^2)$, we let β_P denote its associated regression coefficients. In analogy to Equation (2), we define the minimax PAC rate of squared excess risk under heterogeneous corruptions to be

$$L_{\text{reg}}(\lambda, \mathcal{D}(\Sigma, \sigma^2), \delta) = \inf_M \sup_{P \in \mathcal{D}(\Sigma, \sigma^2)} \inf \left\{ t \in [0, \infty) : \Pr_{\mathbf{Z} \sim \lambda P} [\|M(\mathbf{Z}) - \beta_P\|_\Sigma^2 > t] \leq \delta \right\}, \quad (4)$$

and as before, we let $L_{\text{reg}}(\lambda, \mathcal{D}(\Sigma, \sigma^2)) := L_{\text{reg}}(\lambda, \mathcal{D}(\Sigma, \sigma^2), 1/5)$. Here $\|x\|_\Sigma = x^T \Sigma x$, which is known to be the natural affine-invariant measure of error for linear regression [28]. Note that this

is the standard joint contamination model [10, 29], where both the covariate and the target can be jointly corrupted by the adversary, i.e., $Z_i = (W_i(1 - B_i) + B_i\tilde{W}_i, Y_i(1 - B_i) + B_i\tilde{Y}_i)$. This is opposed to the simpler target contamination model also considered in the literature where only the target Y can be corrupted [30].

Our main result in this setting is as follows:

Theorem 3 (informal, see Theorem 5). *Suppose that $f(\lambda, d) = O(1)$, then $d^{-1/2}f(\lambda, d) \lesssim L_{\text{reg}}(\lambda, \mathcal{D}(\Sigma, \sigma^2)) \lesssim f(\lambda, d)$.*

In other words, as before, we establish the tight rate up to a factor of $O(\sqrt{d})$.

Upper bound techniques For all four problems, the upper bounds in the theorem statements can be achieved by a thresholding estimator – set an appropriate threshold t and take the optimal robust mean estimator with homogeneous corruption rate t . Thus, at least for bounded distributions and Gaussians in a constant number of dimensions, our results show that there is no significant benefit in collecting data with higher corruption rate beyond a certain threshold.

Per-sample reweighting In addition to this simple thresholding method, we also give a family of more refined estimators that find an optimal per-sample weighting scheme, that match the theoretical guarantees of the simple thresholding estimator, but which has a number of additional advantages over it. First, as we discuss below, these methods seem to sometimes perform better in practice in preliminary synthetic evaluations. Second, in the high-dimensional settings, the thresholding-based methods have error rates which seem to plateau, resulting a the $\sqrt{d}$ gap in the upper and lower bounds. We believe resolving this gap is a very interesting open question. As a first step, we derive natural heterogeneous variants of Tukey depth [4] and regression depth [31] using these per-sample reweighting methods. We believe that these methods may allow us to bridge this gap, by leveraging the higher corruption points that the thresholding-based methods ignore to boost the accuracy. To reiterate, we propose the reweighted estimators as a possible avenue to improve the upper bounds but the estimators used to obtain the minimax upper bounds are based on the thresholding estimators.

Lower bound techniques A key technical challenge that separates our analysis from that of standard homogeneous robust statistics analysis is that of capturing the heterogeneity in the lower bound. The standard approach for the lower bound in homogeneous robust statistics is to add a term due to the corruption rate to the non-robust lower bound [32]. For example, in robust d -dimensional Gaussian mean estimation under identity covariance with ϵ corruption, the minimax rate is of the order $\Theta(\frac{d}{n} + \epsilon^2)$, where n is the number of samples. The term $\frac{d}{n}$ is obtained from classical statistics literature while the ϵ^2 term captures the fact that two distributions with mean within l_2 distance of ϵ can not be distinguished due to the ϵ contamination – proved via Le Cam’s method [28]. Such a two-staged approach is unsuitable for our setting since the number of samples n can be made arbitrarily large by artificially appending samples with $\lambda = 1$. Le Cam’s method captures the difficulty of the problem in terms of corruption rate but fails to capture the dimensionality of the problem, while Assouad’s method can capture the difficulty of the problem in higher dimension, but seems to have difficulty capturing the power of an adversary. Thus, the challenge lies in constructing tight lower bounds that incorporate both the terms jointly. Readers may refer to our lower bound construction in Appendix C.2 for more details.

Experimental evaluations While we emphasize that the main contribution of this work is theoretical, in the supplementary material, we also perform some preliminary synthetic evaluations to validate the effectiveness of our methods. In the bounded and univariate Gaussian settings, we demonstrate that both the thresholding-based methods as well as the per-sample reweighting methods outperform baselines from the standard homogeneous robust statistics literature. Our results also demonstrate that in some settings, the per-sample reweighting methods also yield improvements over the threshold-based methods in practice.

1.2 Related Works

As discussed previously, the robust statistics literature has typically focused on a homogeneous corruption rate. In terms of heterogeneity, Charikar et al. [7] introduced the notion of semi-verified

learning where a small amount of data is sampled from the true distribution in conjunction with a dataset having only α fraction of the data uncorrupted ($\alpha \ll 1$). The semi-verified learning problem is closely related to the problem of list-decodable estimation introduced by Balcan et al. [33]; list-decodable learning aims to recover a list of possible values of the quantity being estimated with the guarantee that one of entries in the list is close to the true value. Typical results in semi-verified learning lead of an error that depends on α but does not typically scale with the number of clean samples. There has been a line of work extending their results [9, 13, 14, 19, 34].

Robustness and privacy are known to be deeply related [35–38]. In the privacy literature there has been a line of work on optimal mechanisms under heterogeneous privacy demands [39–41]. However, these works only focus on bounded univariate setting, and do not seem to imply anything directly for the heterogeneous corruption setting we consider.

Organization: We define the problem of mean estimation for bounded distributions in Section 2, and provide a general proof sketch to obtain the minimax rate. In Section 3, we present our results on Gaussian mean estimation for both univariate and multivariate setting, followed by results on excess risk minimization for Gaussian linear regression in Section 4. We present our thoughts on future works in Section 5. Some illustrative experiments can be found in Appendix A.

2 Mean Estimation for Bounded Distributions

In this section, we prove Theorem 1.

2.1 Upper Bound

To obtain an upper bound on $L(\lambda, \mathcal{D}_r^b)$, we shall restrict our attention to the class of estimators of the form $\sum_{i=1}^n w_i Z_i$ for $w \in \Delta_n$. Via standard bias-variance decomposition,

$$\mathbb{E} \left[\left\| \sum_{i=1}^n w_i Z_i - \mu_P \right\|_2^2 \right] = \left\| \sum_{i=1}^n w_i \mathbb{E}[Z_i] - \mu_P \right\|_2^2 + \mathbb{E} \left[\left\| \sum_{i=1}^n w_i (Z_i - \mathbb{E}[Z_i]) \right\|_2^2 \right]. \quad (5)$$

Denoting $\mathbb{E}[\tilde{X}_i | B_i = 1] = \mu_{Q_i}$, the bias term can be upper bounded as

$$\left\| \sum_{i=1}^n w_i (\mathbb{E}[Z_i] - \mu_P) \right\|_2^2 = \left\| \sum_{i=1}^n w_i \lambda_i (\mu_{Q_i} - \mu_P) \right\|_2^2 \leq 4r^2 (w^T \lambda)^2, \quad (6)$$

using Jensen’s inequality and the fact that $\|\mu_{Q_i} - \mu_P\|_2 \leq 2r$. The variance term needs to be bounded with more care since $\tilde{X}$ are not independent can dependent on $\mathbf{X}, \mathbf{B}$. It can be upper bounded as $\mathbb{E} \left[\left\| \sum_{i=1}^n w_i (Z_i - \mathbb{E}[Z_i]) \right\|_2^2 \right] \leq 7r^2 \|w\|_2^2 + 16r^2 (w^T \lambda)^2$ (see Appendix B.1). Thus, we get the upper bound

$$\mathbb{E} \left[\left\| \sum_{i=1}^n w_i Z_i - \mu_P \right\|_2^2 \right] \leq 7r^2 (\|w\|_2^2 + 3(w^T \lambda)^2) \quad \forall w \in \Delta_n. \quad (7)$$

Taking minimum over $w \in \Delta_n$,

$$L(\lambda, \mathcal{D}_r^b) \leq 7r^2 \min_{w \in \Delta_n} \|w\|_2^2 + 3(w^T \lambda)^2. \quad (8)$$

Since Slater’s condition holds, by KKT we obtain the solution of the above to be of form $w_i = (\beta - 3(w^T \lambda) \lambda_i)_+$, where $(x)_+ = \max\{x, 0\}$, for suitable β that ensures $w \in \Delta_n$. Exploiting this structure, we obtain a near-linear time algorithm to find the exact minimizer of $\|w\|_2^2 + c(w^T \lambda)^2$ presented in Algorithm 1. For a vector λ , the notation λ_a^b denotes the sub-vector $(\lambda_a, \dots, \lambda_b)$ for $a \leq b$. The proof of correctness is presented in Appendix B.2.

Simpler Upper Bound: While Algorithm 1 recovers the exact minimizer for (8), we also discuss a significantly simpler upper bound – take the mean of the samples with corruption probability less

Algorithm 1 Robust Mean Estimation for Bounded Distributions

Input: $\mathbf{Z}$ and corresponding λ , with λ assumed to be in ascending order

$n \leftarrow \text{LENGTH}(\lambda)$

$k \leftarrow 1$

while $k \leq n$ **do**

if $\lambda_{k+1} < \frac{1+c\|\lambda_1^k\|_2^2}{c\|\lambda_1^k\|_1}$ **then**

$k \leftarrow k + 1$

else

break

end if

end while

$\beta \leftarrow \frac{1+c\|\lambda_1^k\|_2^2}{k(1+c\|\lambda_1^k\|_2^2)-c\|\lambda_1^k\|_1^2}, \alpha \leftarrow \frac{c\|\lambda_1^k\|_1\beta}{1+c\|\lambda_1^k\|_2^2}.$

$w_1^k \leftarrow \beta - \alpha\lambda_1^k, w_{k+1}^n \leftarrow \mathbf{0}$

return $\sum_{i=1}^n w_i Z_i$

than t for a hyperparameter t . Let $N(t) = |\{i : \lambda_i \leq t\}|$. Therefore, for any $t \in [0, 1]$, we get the upper bound from (7)

$$L(\lambda, r) \lesssim r^2 \left(\frac{1}{N(t)} + t^2 \right). \quad (9)$$

We can similarly take a minimum over t to get an upper bound on $L(\lambda, \mathcal{D}_r^b)$. This simpler upper bound suffices for proving minimax optimality.

2.2 Lower Bound

We use Le Cam's two-point method to obtain the lower bound. Fix some $0 \leq \delta \leq \frac{1}{4}$. Let $P_0 = re_1(2\text{Ber}(\frac{1}{2} - \delta) - 1)$ and $P_1 = re_1(2\text{Ber}(\frac{1}{2} + \delta) - 1)$, where e_1 is a unit vector; we have $\|\mu_{P_1} - \mu_{P_0}\|_2^2 = 4r^2\delta^2$.

Under distribution P_0 , consider the strategy such that adversary tries to ensure that the i -th sample is as close to $re_1(2\text{Ber}(\frac{1}{2}) - 1)$ as possible in mean. Let $Q_i = re_1(2\text{Ber}(\gamma_i) - 1)$ with

$$\gamma_i = \begin{cases} \frac{1}{2} - \delta + \frac{\delta}{\lambda_i} & \text{if } \lambda_i \geq \frac{2\delta}{1+2\delta}, \\ \frac{1}{2} - \delta & \text{else,} \end{cases} \quad (10)$$

then the adversary simply samples $\tilde{Y}_i$ independently from Q_i . Thus, the adversary can simulate¹ the distribution $re_1(2\text{Ber}(\frac{1}{2}) - 1)$ when $\lambda_i \geq \frac{2\delta}{1+2\delta}$. If $\lambda_i < \frac{2\delta}{1+2\delta}$, then the adversary does not perturb the distribution at all, or equivalently, samples $\tilde{X}_i$ from P_0 independently. Thus, the proposed adversarial strategy results in Z_i drawn independently from $re_1(2\text{Ber}(\frac{1}{2} - \epsilon_i) - 1)$, where

$$\epsilon_i := \delta \mathbb{I} \left\{ \lambda_i < \frac{2\delta}{1+2\delta} \right\}. \quad (11)$$

Similar arguments hold when the underlying distribution is P_1 .

Let $n(t) = |\{i : \lambda_i < t\}|$. Using Le Cam's method, we show in Appendix B.3

$$L(\lambda, r) \geq r^2\delta^2 \left(1 - \sqrt{6\delta^2 n(2\delta)} \right) \quad \forall \delta \in \left[0, \frac{1}{4} \right] \quad (12)$$

2.3 Proving Minimax Optimality

We outline how to pick a 'good' δ . For $x \in [0, \frac{1}{4}]$, let $h(x) = N(2x) - \frac{1}{12x^2}$. Note that h is monotone, piece-wise continuous, and $\exists \alpha > 0$ such that $h(\alpha) < 0$. We use the notation $h(x^-)$ to denote the left limit of h at x ; note that $h(x^-) = n(2x) - \frac{1}{12x^2}$.

¹By simulate, we mean the mixture distribution matches the claimed distribution.

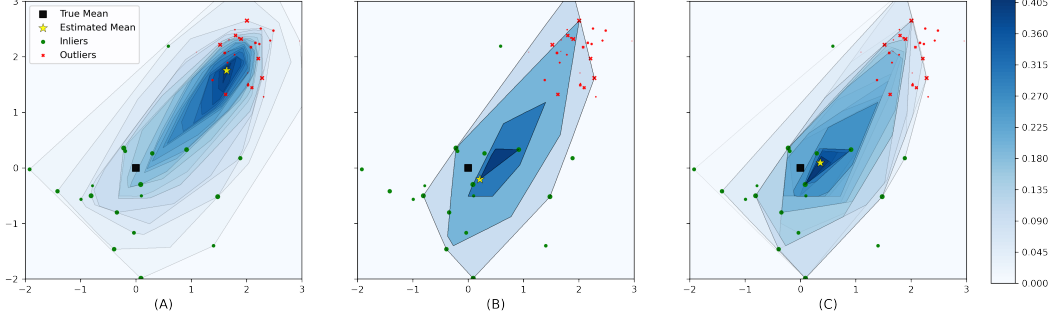


Figure 1: Plot of weighted Tukey depth (see (17)) visualized for three different weighing schemes. (A) is computed with the standard uniform weights $w_i = \frac{1}{n}$, (B) is computed with $w_i = \frac{\mathbb{I}\{\lambda_i \leq t\}}{|\{j: \lambda_j \leq t\}|}$ using the value of t from (21), and (C) is computed with weights given by Algorithm 1. For the dataset, the true underlying distribution is $\mathcal{N}((0, 0), I)$, and λ is sampled i.i.d. Points are contaminated by replacing them with samples from $\mathcal{N}((2, 2), I/5)$. The samples are marked in red ‘x’ if they were contaminated; the size of the markers for each point is proportional to $1 - \lambda_i$. The estimated mean, the point with maximum depth, is marked with a yellow star.

Consider the case that $h(\frac{1}{4}) < 0$, i.e., $|\{i : \lambda_i \leq \frac{1}{2}\}| < \frac{4}{3}$. This conditions correspond to extremely high levels of corruption and unsurprisingly, we show the minimax rate to be $\Theta(r^2)$. Plug-in $\delta = \frac{1}{4}$ in (12) to get $L(\lambda, r) \gtrsim r^2$. We can get matching upper bound in (9) by setting $t = 1$.

Alternately, if $h(\frac{1}{4}) > 0$ then since $h(\cdot)$ is a *cadlag* function and $h(\alpha) < 0$, there exists $\delta_* \in (0, \frac{1}{4}]$ such that $h(\delta_*) \geq 0$ and $h(\delta_*^-) \leq 0$, i.e., $N(2\delta_*) \geq \frac{1}{12\delta_*^2}$ and $n(2\delta_*) \leq \frac{1}{12\delta_*^2}$. Plug-in δ_* in (12) to get

$$L(\lambda, r) \geq r^2 \delta_*^2 (1 - \sqrt{6\delta_*^2 n(2\delta_*)}) \quad (13)$$

$$\geq r^2 \delta_*^2 \left(1 - \sqrt{\frac{1}{2}}\right) \simeq r^2 \delta_*^2. \quad (14)$$

In the upper bound of (9), set $t = 2\delta_*$ to get

$$L(\lambda, r) \lesssim r^2 \frac{1}{N(2\delta_*)} + r^2 \delta_*^2 \quad (15)$$

$$\leq 12r^2 \delta_*^2 + r^2 \delta_*^2 \simeq r^2 \delta_*^2, \quad (16)$$

proving minimax optimality of the proposed schemes in (8) and (9) up to constant factors.

3 Mean Estimation for Gaussian Distributions

In this section, we describe our results for the problem of mean estimation for multivariate Gaussian distributions at a high level. For conciseness, we defer the complete proofs to the appendix.

Upper Bound: We propose a weighted version of the Tukey median [4]. Define the weighted depth of a point $\eta \in \mathbb{R}^d$ for a dataset $\mathbf{Z} = (Z_1, \dots, Z_n)$ and $w \in \Delta_n$ as

$$D_w(\eta, \mathbf{Z}) = \min_{v \in \mathbb{S}_d} \sum_{i=1}^n w_i \mathbb{I}\{v^T (Z_i - \eta) \geq 0\}, \quad (17)$$

where $\mathbb{S}_d = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. The weighted Tukey median is defined as

$$\hat{\mu}_{\text{TM}}(\mathbf{Z}, w) := \arg \max_{\eta \in \mathbb{R}^d} D_w(\eta, \mathbf{Z}). \quad (18)$$

The standard Tukey depth (i.e., $w = 1/n$) of a point η is the minimum number of samples in a closed half-space with η at the boundary. Thus, depth of η is high if it is surrounded by samples in all

directions. Our weighted version allows us to decrease the sensitivity to samples which are more likely to be corrupted. In Figure 1, we illustrate the Tukey depth map for three different weighing schemes which demonstrates how using the weights w can be used to leverage the heterogeneity. Proposition 1 establishes an upper bound on the performance of the weighted Tukey median.

Proposition 1 (Upper Bound for Gaussian Distributions). *For all $w \in \Delta_n$ satisfying $(w^T \lambda)^2 + d\|w\|_2^2 \leq c$ for some universal constants c ,*

$$\sup_{P \in \mathcal{D}_d^N} \Pr_{Z \sim \lambda P} [\|\hat{\mu}_{\text{TM}}(Z, w) - \mu_P\|_2^2 \geq c' ((w^T \lambda)^2 + d\|w\|_2^2)] \leq 1/5, \quad (19)$$

for some universal constant c' .

The proof of Proposition 1 (see Appendix C.1) involves proving that the depth of any point sufficiently far from the true mean is going to be lower than the depth of the true mean in a uniform convergence sense. Due to the weighted nature of the depth definition, we modify results from empirical process theory to adapt to the weights and get tight upper bound.

Let $N(t) = |\{i : \lambda_i \leq t\}|$ and denote the weights $w(t)$ such that $w(t)_i = \frac{\mathbb{I}\{\lambda_i \leq t\}}{N(t)}$, i.e., we consider the Tukey median estimator obtained by only considering samples with corruption less than t . Define

$$\hat{\mu}_S(Z, t) = \hat{\mu}_{\text{TM}}(Z, w(t)). \quad (20)$$

By Proposition 1, the following holds for some universal constant c'

$$\sup_{P \in \mathcal{D}_d^N} \Pr_{Z \sim \lambda P} \left[\|\hat{\mu}_S(Z, t) - \mu_P\|_2^2 \geq c' \left(t^2 + \frac{d}{N(t)} \right) \right] \leq 1/5, \quad (21)$$

$\forall t$ such that $t^2 + \frac{d}{N(t)} \leq c$ for universal constants c .

3.1 Lower Bound and Minimax Rate

Univariate Case We begin our exposition by considering our result for $d = 1$, i.e., univariate Gaussian distributions. We use Le Cam's method to obtain the lower bound [28]; Le Cam's method lower bounds the estimation problem by a hypothesis testing problem where the optimal test is entirely determined by the distribution of data from the different hypotheses. The two hypotheses we consider are $\mathcal{N}(\delta, 1)$ and $\mathcal{N}(-\delta, 1)$ for δ to be chosen later. The adversary tries to ensure that the resulting corrupted distribution under either hypotheses is the same, rendering those sample useless for hypothesis testing.

If the corruption rate is greater than $\frac{1}{2}$ for a sample, then the adversary can trivially 'simulate' the distribution $\frac{1}{2}(\mathcal{N}(\delta, 1) + \mathcal{N}(-\delta, 1))$. Alternately, when δ is small, adversary can simulate the distribution $\max\{\mathcal{N}(\delta, 1), \mathcal{N}(-\delta, 1)\}/Z(\delta)$ instead, where $Z(\delta)$ is an appropriate normalizing constant. The adversary can simulate this distribution when the corruption rate is greater than $1 - e^{-\delta}$. Thus, only the samples in the set $\{i : \lambda_i < \min\{\frac{1}{2}, 1 - e^{-\delta}\}\}$ are relevant for hypothesis testing, and it remains to choose a value of δ judiciously.

Based on our generic multivariate bounds of Theorem 4 later, we have the following Corollary 1 for the univariate case.

Corollary 1 (Minimax Rate for Univariate Gaussian Distributions). *Suppose $\min_{t \in [0, 1]} \left(t^2 + \frac{1}{N(t)} \right) \leq c$ for some universal constants c , then*

$$L_{\text{PAC}}(\lambda, \mathcal{D}_1^N) \simeq \min_{t \in [0, 1]} \left(\frac{1}{N(t)} + t^2 \right). \quad (22)$$

Moreover, the estimator $\hat{\mu}_S(Z, w(t^*))$, with $t^* = \arg \min_{t \in [0, 1]} \left(t^2 + \frac{1}{N(t)} \right)$, achieves this error.

Thus the proposed weighted Tukey median scheme is optimal in the univariate case – demonstrating that beyond a certain corruption threshold determined by $N(\cdot)$, there is no way to leverage the more contaminated samples up to universal constants.

Multivariate Case In the multivariate setting, our upper and lower bounds (see Appendix C.2) are off by a multiplicative factor of $\sqrt{d}$. This gap can possibly be attributed by our rather relaxed handling of the heterogeneity since the proof technique we consider can recover the minimax rate of $\frac{d}{n} + \epsilon^2$ in the homogeneous setting where $\lambda_i = \epsilon \forall i$ (see Appendix F).

Theorem 4. Suppose $\min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$ for some universal constants c , then

$$\frac{1}{\sqrt{d}} \min_{t \in [0,1]} \left(\frac{d}{N(t)} + t^2 \right) \lesssim L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_1^N) \lesssim \min_{t \in [0,1]} \left(\frac{d}{N(t)} + t^2 \right). \quad (23)$$

Moreover, the estimator $\hat{\mu}_S(\mathbf{Z}, w(t^*))$, with $t^* = \arg \min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right)$, achieves the upper bound.

The proof of Theorem 4 can be found in Appendix C.3. As Theorem 4 suggests, it might be possible to leverage the heterogeneity beyond just discarding samples with a higher threshold and this might reduce the squared-error by a factor of $\sqrt{d}$. However, as stated before, this gap might be an artifact of the analysis rather than a true phenomenon. We leave this as an open question.

4 Linear Regression

We now consider the problem of linear regression under Gaussian covariates. The proofs for this section can be found in Appendix D.

Background: Under no corruptions, the minimax rate is known to be $\frac{\sigma^2 d}{n}$, i.e., independent of Σ as long as Σ is non-singular – the lack of information in directions with low eigenvalues of Σ is offset by the $\|\cdot\|_\Sigma$ distance. In the homogeneous robust regression setting with ϵ corruption, the minimax rate is of the order $\frac{\sigma^2 d}{n} + \sigma^2 \epsilon^2$ [31].

For the upper bound, we consider a weighted version of Tukey median adapted to regression [31, 42]. Define the weighted regression depth of a point $\eta \in \mathbb{R}^d$ for a dataset $\mathbf{Z}$ with $Z_i = (\hat{W}_i, \hat{Y}_i)$ and $w \in \Delta_n$ as

$$D_w(\eta, \mathbf{Z}) = \min_{v \in \mathbb{S}^d} \sum_{i=1}^n w_i \mathbb{I}\{(\hat{Y}_i - \eta^T \hat{W}_i)(v^T \hat{W}_i) \geq 0\}. \quad (24)$$

The regression depth of a point $\eta \in \mathbb{R}^d$ is high if for all half-spaces, the sign of the residual errors using η are distributed equally in every direction. The weighted Tukey regression coefficient is defined as

$$\hat{\beta}_{\text{TC}}(\mathbf{Z}, w) := \arg \max_{\eta \in \mathbb{R}^d} D_w(\eta, \mathbf{Z}). \quad (25)$$

Proposition 2 (Upper Bound for Regression). For all $w \in \Delta_n$ satisfying $(w^T \lambda)^2 + d \|w\|_2^2 \leq c$ for some universal constants c ,

$$\sup_{P \in \mathcal{D}(\Sigma, \sigma^2)} \Pr_{\mathbf{Z} \sim \lambda P} \left[\|\hat{\beta}_{\text{TC}}(\mathbf{Z}, w) - \beta_P\|_\Sigma^2 \geq c' ((w^T \lambda)^2 + d \|w\|_2^2) \right] \leq 1/5, \quad (26)$$

for some universal constant c' .

Like (20), we can similarly consider a thresholding method – define $\hat{\beta}_S(\mathbf{Z}, t) = \hat{\beta}_{\text{TM}}(\mathbf{Z}, w(t))$. The lower bound construction is similar in spirit to Theorem 4. Based on our upper and lower bounds, we present Theorem 5.

Theorem 5 (Minimax Rate for Linear Regression). Suppose $\min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$ for some universal constants c , then

$$\frac{\sigma^2}{\sqrt{d}} \min_{t \in [0,1]} \left(\frac{d}{N(t)} + t^2 \right) \lesssim L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{D}(\Sigma, \sigma^2)) \lesssim \sigma^2 \min_{t \in [0,1]} \left(\frac{d}{N(t)} + t^2 \right). \quad (27)$$

Moreover, the estimator $\hat{\beta}_S(\mathbf{Z}, w(t^*))$, with $t^* = \arg \min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right)$, achieves the upper bound.

Note that when $\lambda_i = \epsilon \forall i$, we can recover the homogeneous minimax rate of $\frac{\sigma^2 d}{n} + \sigma^2 \epsilon^2$ using the more refined lower bound in Appendix F.

5 Discussion and Future Work

There exists several avenues of future work. Perhaps most pressing issue would be to address the $\sqrt{d}$ gap in our upper and lower bounds for multivariate Gaussian mean estimation and linear regression. In this regard, it seems that neither our upper nor lower bound in Theorem 4 can be tight for all choices of λ . Indeed, the lower bound in Equation (23) is not tight for homogeneous corruption (although this can be avoided with our more sophisticated lower bound technique, see Appendix F), and on the other hand, the upper bound fails to be tight in the aforementioned regime of semi-verified learning, where the upper bound would suggest that the error in the estimator should diverge as $d \rightarrow \infty$, whereas existing upper bounds due to [7, 9] already demonstrate that dimension independent rates are possible. This suggests that fundamentally new measures of robustness are needed to fully characterize the complexity of learning under heterogeneous errors in the high-dimensional setting. The heterogeneous robust estimation problem can also be studied under different corruption models such as total-variation distance or Kullback-Leibler divergence corruption.

Acknowledgments and Disclosure of Funding

S.C. acknowledges the support of AI Policy Hub, UC Berkeley.

References

- [1] Peter J Huber. The 1972 wald lecture robust statistics: A review. *The Annals of Mathematical Statistics*, pages 1041–1067, 1972.
- [2] Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, pages 799–821, 1973.
- [3] Frank J Anscombe. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960.
- [4] J. W. Tukey. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, 2:523–531, 1975.
- [5] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664, 2016. doi: 10.1109/FOCS.2016.85.
- [6] Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674, 2016. doi: 10.1109/FOCS.2016.76.
- [7] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017*, page 47–60, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450345286. doi: 10.1145/3055399.3055491. URL <https://doi.org/10.1145/3055399.3055491>.
- [8] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017.
- [9] Michela Meister and Gregory Valiant. A data prism: Semi-verified learning in the small-alpha regime. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1530–1546. PMLR, 06–09 Jul 2018. URL <https://proceedings.mlr.press/v75/meister18a.html>.
- [10] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019.

- [11] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32, 2019.
- [12] Sam Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *Advances in Neural Information Processing Systems*, 33:11902–11912, 2020.
- [13] Ilias Diakonikolas, Daniel Kane, and Daniel Kongsgaard. List-decodable mean estimation via iterative multi-filtering. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9312–9323. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6933b5648c59d618bbb30986c84080fe-Paper.pdf.
- [14] Prasad Raghavendra and Morris Yau. *List Decodable Learning via Sum of Squares*, pages 161–180. 2020. doi: 10.1137/1.9781611975994.10. URL <https://epubs.siam.org/doi/abs/10.1137/1.9781611975994.10>.
- [15] Jules Depersin and Guillaume Lecué. Robust sub-gaussian estimation of a mean vector in nearly linear time. *The Annals of Statistics*, 50(1):511–536, 2022.
- [16] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M Kane, Pravesh K Kothari, and Santosh S Vempala. Robustly learning mixtures of k arbitrary gaussians. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1234–1247, 2022.
- [17] Jules Depersin and Guillaume Lecué. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probability Theory and Related Fields*, 183(3): 997–1025, 2022.
- [18] Allen Liu and Ankur Moitra. Robustly learning general mixtures of gaussians. *Journal of the ACM*, 70(3):1–53, 2023.
- [19] Shiwei Zeng and Jie Shen. Semi-verified pac learning from the crowd. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 505–522. PMLR, 25–27 Apr 2023. URL <https://proceedings.mlr.press/v206/zeng23a.html>.
- [20] Roberto I Oliveira and Zoraida F Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5):1953–1977, 2024.
- [21] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.
- [22] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.
- [23] Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. On the power of adaptivity in statistical adversaries. In *Conference on Learning Theory*, pages 5030–5061. PMLR, 2022.
- [24] Clément Canonne, Samuel B Hopkins, Jerry Li, Allen Liu, and Shyam Narayanan. The full landscape of robust mean testing: Sharp separations between oblivious and adaptive contamination. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2159–2168. IEEE, 2023.
- [25] Guy Blanc and Gregory Valiant. Adaptive and oblivious statistical adversaries are equivalent. *arXiv preprint arXiv:2410.13548*, 2024.
- [26] Tianyi Ma, Kabir A Verchand, and Richard J Samworth. High-probability minimax lower bounds. *arXiv preprint arXiv:2406.13447*, 2024.
- [27] Syomantak Chaudhuri and Thomas A. Courtade. Private estimation when data and privacy demands are correlated, 2025. URL <https://arxiv.org/abs/2407.11274>.

- [28] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.
- [29] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.
- [30] Takayuki Kawashima and Hironori Fujisawa. Robust regression against heavy heterogeneous contamination. *Metrika*, 86(4):421–442, 2023.
- [31] Chao Gao. Robust regression via multivariate regression depth. *Bernoulli*, 26(2):1139 – 1170, 2020. doi: 10.3150/19-BEJ1144.
- [32] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under Huber’s contamination model. *The Annals of Statistics*, 46(5):1932 – 1960, 2018. doi: 10.1214/17-AOS1607. URL <https://doi.org/10.1214/17-AOS1607>.
- [33] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. A discriminative framework for clustering via similarity functions. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC ’08, page 671–680, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580470. doi: 10.1145/1374376.1374474. URL <https://doi.org/10.1145/1374376.1374474>.
- [34] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 1047–1060, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355599. doi: 10.1145/3188745.3188758. URL <https://doi.org/10.1145/3188745.3188758>.
- [35] Kristian Georgiev and Samuel Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *Advances in neural information processing systems*, 35: 6829–6842, 2022.
- [36] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2023.
- [37] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1121–1146. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/asi23b.html>.
- [38] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC ’09, page 371–380, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585062. doi: 10.1145/1536414.1536466. URL <https://doi.org/10.1145/1536414.1536466>.
- [39] Rachel Cummings, Hadi Elzayn, Emmanouil Pountourakis, Vasilis Gkatzelis, and Juba Ziani. Optimal data acquisition with privacy-aware agents. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023, Raleigh, NC, USA, February 8-10, 2023*, pages 210–224. IEEE, 2023. doi: 10.1109/SATML54575.2023.00023. URL <https://doi.org/10.1109/SaTML54575.2023.00023>.
- [40] Alireza Fallah, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Optimal and differentially private data acquisition: Central and local mechanisms. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC ’22, page 1141, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391504. doi: 10.1145/3490486.3538329. URL <https://doi.org/10.1145/3490486.3538329>.
- [41] Syomantak Chaudhuri, Konstantin Miagkov, and Thomas A. Courtade. Mean estimation under heterogeneous privacy demands. *IEEE Transactions on Information Theory*, 71(2):1362–1375, 2025. doi: 10.1109/TIT.2024.3511498.

- [42] Peter J. Rousseeuw and Mia Hubert. Regression depth. *Journal of the American Statistical Association*, 94(446):388–402, Jun 1999. doi: 10.1080/01621459.1999.10474129.
- [43] Luc Devroye and Gábor Lugosi. Combinatorial methods in density estimation. In *Springer Series in Statistics*, 2001. URL <https://api.semanticscholar.org/CorpusID:118160190>.
- [44] Renjie Liao. Notes on rademacher complexity. https://www.cs.toronto.edu/~rjliao/notes/Notes_on_Rademacher_Complexity.pdf. Accessed 10-05-2025.
- [45] Patrick Rebeschini. Covering numbers bounds for rademacher complexity. chaining. <https://www.stats.ox.ac.uk/~rebeschini/teaching/AFoL/22/material/lecture05.pdf>. Accessed 13-05-2025.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Claims are theoretical and their proofs are provided in the appendix.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We make the assumptions clear. For example, our methods are not useful without knowing the λ vector. We also make it clear that our estimators are not tractable in high dimension.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Formal proofs are presented in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide all the high level details to understand the experiments. Exact hyper-parameters used are not specified for ease of exposition but can be found in the linked codebase. The experiments are fairly straightforward to implement (less than 200 lines of code).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Attached Jupyter notebooks allows reviewers to verify the plots without writing any extra code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the necessary details to understand and interpret the result are contained in the paper. Auxiliary details like the specific choice of hyper-parameters are included in the codebase.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes] .

Justification: Some notion of error/variance is captured in the graphs in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Experiments performed at not computationally expensive and can be performed on most laptops on CPU (without GPU).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: No violation of the Code of Ethics were committed.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The work is more theoretical in nature and no immediate significant impact on the society is foreseen by the authors.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No real-world datasets are used in this work.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We did not use any external assets. Prior research is properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We introduce no new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Not relevant for this work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer:[NA]

Justification: Not relevant.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM-usage was restricted to editing the paper and help in speeding up the development of the codebase.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

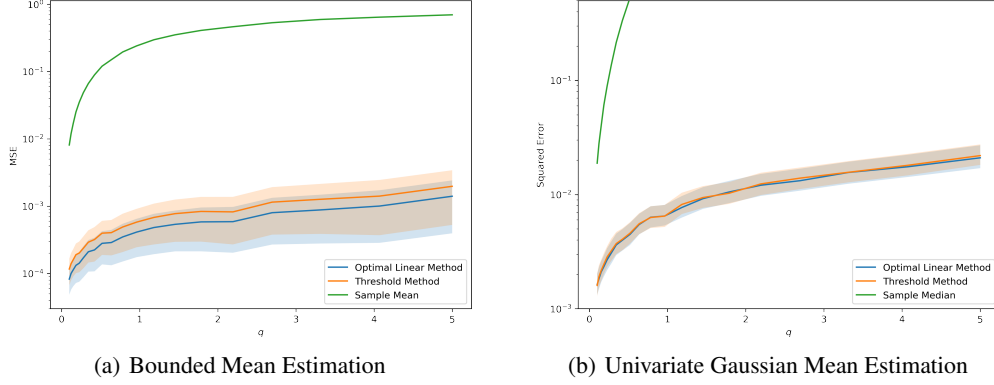


Figure 2: Mean estimation algorithms for (a) bounded distributions and (b) univariate Gaussian distributions. The x-axis is a proxy for degree of contamination of the model.

A Experiments

We investigate the practical performance of our algorithms and compare it with baseline in Figure 2. For bounded mean estimation, we use the sample mean as the baseline since it is the minimax robust estimator in the homogeneous setting. For univariate Gaussian mean estimation, we use the sample median, equivalent to Tukey median in one dimension, as the baseline for the same reason.

We set $n = 10^4$ and for a fixed value q , we sample the corruption rates λ i.i.d. from the distribution with cdf given by $F(t) = 1 - (1 - t)^q$. As q increases we can expect a higher corruption rate. Fixing this sampled λ , we sample the dataset 10^4 times. For bounded distribution, we plot the mean squared-error and the corresponding standard deviations over the trials at each value of q considered. For the Gaussian distribution, we plot the empirical $\frac{4}{5}$ -th quantile of the squared-error along with $\frac{15}{20}$ -th and $\frac{17}{20}$ -th quantiles over the trials. For the bounded distribution, we choose $r = 1$ and choose the true underlying distribution to be the point mass at 0, and the corrupted values to be 1. For univariate Gaussian distribution, we fix the true distribution to be $\mathcal{N}(0, 1)$ and the corrupted values sampled i.i.d. from $\mathcal{N}(100, 1)$.

Optimal linear method in the plots refer to the reweighing scheme proposed while threshold method refers to the special case of reweighing that discards samples above a certain corruption threshold and performs standard homogeneous robust estimation on the sub-sampled dataset.

While unclear for the Gaussian distribution, the reweighing does seem to provide marginal improvement over thresholding method. Further investigation is required to establish whether reweighing may pose significant advantages in high dimensions.

B Bounded Mean Estimation: Proofs

B.1 Variance Upper Bound

Using $E\|X + Y\|_2^2 \leq 2E\|X\|_2^2 + 2E\|Y\|_2^2$, we get

$$\mathbb{E} \left[\left\| \sum_{i=1}^n w_i (Z_i - \mathbb{E}[Z_i]) \right\|_2^2 \right] = \mathbb{E} \left[\left\| \sum_{i=1}^n w_i ((1 - B_i)X_i - (1 - \lambda_i)\mu_P) + \sum_{i=1}^n w_i (B_i\tilde{X}_i - \lambda_i\mu_{Q_i}) \right\|_2^2 \right] \quad (28)$$

$$\leq 2\mathbb{E} \left[\left\| \sum_{i=1}^n w_i ((1 - B_i)X_i - (1 - \lambda_i)\mu_P) \right\|_2^2 \right] + 2\mathbb{E} \left[\left\| \sum_{i=1}^n w_i (B_i\tilde{X}_i - \lambda_i\mu_{Q_i}) \right\|_2^2 \right] \quad (29)$$

Since $\{(1 - B_i)X_i - (1 - \lambda_i)\mu_P\}$ are independent random variables and $\|(1 - B_i)X_i\|_2 \leq r$, we can use the crude variance bound $E\|(1 - B_i)X_i - (1 - \lambda_i)\mu_P\|_2^2 \leq r^2$ to obtain

$$\mathbb{E} \left[\left\| \sum_{i=1}^n w_i ((1 - B_i)(X_i - \mu_P) - (1 - \lambda_i)\mu_P) \right\|_2^2 \right] \leq r^2 \|w\|_2^2. \quad (30)$$

Inspecting the other term, we use the law of total variance by conditioning on $\mathbf{B}$. In particular,

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{i=1}^n w_i (B_i \tilde{X}_i - \lambda_i \mu_{Q_i}) \right\|_2^2 \right] &= \mathbb{E} \left[\left\| \sum_{i=1}^n w_i (B_i (\tilde{X}_i - \mu_{Q_i}) + \mu_{Q_i} (B_i - \lambda_i)) \right\|_2^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| \sum_{i=1}^n w_i B_i (\tilde{X}_i - \mu_{Q_i}) \right\|_2^2 \right] + 2\mathbb{E} \left[\left\| \sum_i w_i \mu_{Q_i} (B_i - \lambda_i) \right\|_2^2 \right]. \end{aligned} \quad (31)$$

For the first term in (32), use Jensen's inequality as

$$\mathbb{E} \left[\left\| \sum_{i=1}^n w_i B_i (\tilde{X}_i - \mu_{Q_i}) \right\|_2^2 \right] = \mathbb{E} \left[\mathbb{E} \left[\left\| \sum_{i=1}^n w_i B_i (\tilde{X}_i - \mu_{Q_i}) \right\|_2^2 \middle| \mathbf{B} \right] \right] \quad (33)$$

$$\leq 4r^2 \mathbb{E} \left[\left(\sum_{i=1}^n w_i B_i \right)^2 \right] \quad (34)$$

$$= 4r^2 \sum_i w_i^2 \lambda_i (1 - \lambda_i) + 4r^2 (w^T \lambda)^2 \quad (35)$$

$$\leq r^2 \|w\|_2^2 + 4r^2 (w^T \lambda)^2. \quad (36)$$

For the second term in (32), using the fact that $\|\mu_{Q_i}\|_2 \leq r$, using Jensen's inequality we get

$$\mathbb{E} \left[\left\| \sum_i w_i \mu_{Q_i} (B_i - \lambda_i) \right\|_2^2 \right] \leq r^2 \mathbb{E} \left[\left(\sum_i w_i (B_i - \lambda_i) \right)^2 \right] \leq \frac{r^2}{4} \|w\|_2^2. \quad (37)$$

Combining the above, obtain

$$\mathbb{E} \left[\left\| \sum_{i=1}^n w_i (Z_i - \mathbb{E}[Z_i]) \right\|_2^2 \right] \leq 7r^2 \|w\|_2^2 + 16r^2 (w^T \lambda)^2 \quad (38)$$

B.2 Upper Bound Solution

To solve

$$\min_{w \in \Delta_n} \|w\|^2 + c(w^T \lambda)^2, \quad (39)$$

consider the Lagrangian $\mathcal{L}(w, \beta, \gamma) = \|w\|_2^2 + c(w^T \lambda)^2 + 2\beta(1 - \sum_i w_i) - \sum_i 2\gamma_i w_i$. KKT condition on the Lagrangian leads to

$$w_i = \beta - c(w^T \lambda) \lambda_i + \gamma_i \quad \forall i, \quad (40)$$

where $\gamma_i w_i = 0, \gamma_i \geq 0 \quad \forall i$. Thus, we can equivalently write

$$w_i = (\beta - c(w^T \lambda) \lambda_i)_+. \quad (41)$$

Notice that w_i are decreasing in λ_i thus, order the indices such that $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Note that since this is a strictly convex objective with a convex compact constraint set we are guaranteed a *unique* solution w .

Let m such that $w_i > 0 \forall i \leq m$ and $w_i = 0 \forall i > m$; if no such m exists then it is understood $m = n$. Since $\sum w_i = 1$, use (41) to obtain the condition

$$m\beta - c(w^T \lambda) \|\lambda_1^m\|_1 = 1. \quad (42)$$

Noting that $w^T \lambda = \sum_{i=1}^m w_i \lambda_i$, we can use (41) to obtain

$$w^T \lambda = \beta \|\lambda_1^m\|_1 - c w^T \lambda \|\lambda_1^m\|_2^2. \quad (43)$$

Solving for $w^T \lambda$ and substituting in (42), obtain $\beta = \frac{1+c\|\lambda_1^m\|_2^2}{k(1+c\|\lambda_1^m\|_2^2)-c\|\lambda_1^m\|_1^2}$, and

$$w_i = \frac{1+c\|\lambda_1^m\|_2^2}{m(1+c\|\lambda_1^m\|_2^2)-c\|\lambda_1^m\|_1^2} \left(1 - c\lambda_i \frac{\|\lambda_1^m\|_1}{1+c\|\lambda_1^m\|_2^2}\right)_+ \quad \forall i \in [n]. \quad (44)$$

Thus, the problem of solving for the weights has been reduced to identifying the index k after which the weights are zero. This is precisely what Algorithm 1 does. In particular, $m+1 = \min\{j : w_j = 0\}$ by definition and $w_{m+1} = 0 \Leftrightarrow \lambda_{m+1} \geq \frac{1+c\|\lambda_1^m\|_2^2}{c\|\lambda_1^m\|_1}$ by (44). Therefore, if the loop in Algorithm 1 runs without termination till index $k = m$, then it will correctly terminate at $k = m$ since $\lambda_{m+1} \geq \frac{1+c\|\lambda_1^m\|_2^2}{c\|\lambda_1^m\|_1}$.

Thus, we need to show that the algorithm does not terminate before $k = m$ to prove correctness.

Assume the contrary that it terminates at $k = p < m$, i.e., $\lambda_{p+1} \geq \frac{1+c\|\lambda_1^p\|_2^2}{c\|\lambda_1^p\|_1}$. Observe that

$$\lambda_{p+1} \geq \frac{1+c\|\lambda_1^p\|_2^2}{c\|\lambda_1^p\|_1} \Leftrightarrow \lambda_{p+1} \geq \frac{1+c\|\lambda_1^{p+1}\|_2^2}{c\|\lambda_1^{p+1}\|_1} \quad (45)$$

$$\Rightarrow \lambda_{p+2} \geq \frac{1+c\|\lambda_1^{p+1}\|_2^2}{c\|\lambda_1^{p+1}\|_1}, \quad (46)$$

where (46) follows since $\lambda(s)$ are indexed in non-decreasing order. Extending this argument, we get $\lambda_m \geq \frac{1+c\|\lambda_1^m\|_2^2}{c\|\lambda_1^m\|_1}$. Note since $w_m > 0$, we have

$$\lambda_m < \frac{1+c\|\lambda_1^m\|_2^2}{c\|\lambda_1^m\|_1} \quad (47)$$

by (44) – a contradiction. This proves that the proposed algorithm solves for w correctly.

B.3 Lower Bound

By Le Cam's method,

$$L(\lambda, r) \geq r^2 \delta^2 \left(1 - \text{TV} \left(\otimes_{i=1}^n \text{Ber} \left(\frac{1}{2} - \epsilon_i \right), \otimes_{i=1}^n \text{Ber} \left(\frac{1}{2} + \epsilon_i \right) \right) \right), \quad (48)$$

$$= r^2 \delta^2 \left(1 - \sqrt{\frac{1}{2} \sum_{i=1}^n \text{KL} \left(\text{Ber} \left(\frac{1}{2} - \epsilon_i \right), \text{Ber} \left(\frac{1}{2} + \epsilon_i \right) \right)} \right) \quad (49)$$

$$= r^2 \delta^2 \left(1 - \sqrt{6 \sum_{i=1}^n \epsilon_i^2} \right), \quad (50)$$

where we used $2\epsilon \log \frac{1+2\epsilon}{1-2\epsilon} \leq 12\epsilon^2 \forall \epsilon \in [0, \frac{1}{4}]$. Let $n(t) = |\{i : \lambda_i < t\}|$, then we obtain

$$L(\lambda, r) \geq r^2 \delta^2 \left(1 - \sqrt{6\delta^2 n \left(\frac{2\delta}{1+2\delta} \right)} \right) \quad (51)$$

$$\geq r^2 \delta^2 \left(1 - \sqrt{6\delta^2 n(2\delta)} \right) \quad \forall \delta \in \left[0, \frac{1}{4}\right] \quad (52)$$

C Mean Estimation for Gaussian Distributions: Proofs

C.1 Upper Bound

Recall

$$D_w(\eta, \mathbf{Z}) = \min_{v \in \mathbb{S}_d} \sum_{i=1}^n w_i \mathbb{I}\{v^T (Z_i - \eta) \geq 0\}, \quad (53)$$

$$\hat{\mu}_{\text{TM}}(\mathbf{Z}, w) := \arg \max_{\eta \in \mathbb{R}^d} D_w(\eta, \mathbf{Z}). \quad (54)$$

Let $G = \{i : B_i = 0\}$ and $B = [n] \setminus G$. Note that $Z_i = X_i$ for $i \in G$. The depth of the true mean is lower bounded as

$$D_w(\mu, \mathbf{Z}) \geq \min_{v \in \mathbb{S}_d} \sum_{i \in G} w_i \mathbb{I}\{(X_i - \mu)^T v \geq 0\}. \quad (55)$$

Define the class of indicator functions $\mathcal{F}_\mu = \{f_v(x) = \mathbb{I}\{(x - \mu)^T v \geq 0\} | v \in \mathbb{S}_d\}$. Note that $E[f(X)] = \frac{1}{2} \forall f \in \mathcal{F}_\mu$. With some abuse of notation, let $w(G) = \{w_i | i \in G\}$. By Proposition 4, we have with probability at least $1 - \frac{\delta}{4}$

$$\min_{f \in \mathcal{F}_\mu} \sum_{i \in G} w_i \left(f(X_i) - \frac{1}{2} \right) \geq -62 \|w(G)\|_2 \sqrt{\text{VC}(\mathcal{F}_\mu)} - \|w(G)\|_2 \sqrt{\frac{\log 4/\delta}{2}} \quad (56)$$

$$\geq -\|w\|_2 \left(62\sqrt{d} + \sqrt{\frac{\log 4/\delta}{2}} \right), \quad (57)$$

where we used $\text{VC}(\mathcal{F}_\mu) = d$; readers may refer to [43, Corollary 4.2.2] for VC dimension of homogeneous half-space classifiers. Further, with probability at least $1 - \delta/4$, by McDiarmid's inequality [28]

$$\sum_{i \in G} w_i = \sum_{i=1}^n w_i \mathbb{I}\{B_i = 0\} \quad (58)$$

$$\geq \sum_{i=1}^n w_i (1 - \lambda_i) - \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}} \quad (59)$$

$$= 1 - w^T \lambda - \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}}. \quad (60)$$

Thus, with probability at least $1 - \delta/2$,

$$D_w(\mu, \mathbf{Z}) \geq \frac{1}{2} - \frac{w^T \lambda}{2} - \|w\|_2 \left(62\sqrt{d} + \sqrt{\frac{9 \log 4/\delta}{8}} \right). \quad (61)$$

Next, we show that depth of any point far away from the true mean is low. For any $\eta \in \mathbb{R}^d$ such that $\|\eta - \mu\|_2 \geq r = \Phi^{-1}(\frac{1}{2} + \alpha)$, let $v_\eta = \frac{\eta - \mu}{\|\eta - \mu\|_2}$. We shall set the value of $\alpha > 0$ later.

$$\sup_{\eta: \|\eta - \mu\|_2 \geq r} D_w(\eta, \mathbf{Z}) \leq \sum_{i \in B} w_i + \sup_{\eta: \|\eta - \mu\|_2 \geq r} \sum_{i=1}^n w_i \mathbb{I}\{(X_i - \eta)^T v_\eta \geq 0\}. \quad (62)$$

Define the class of indicator functions $\mathcal{G}_\mu = \{f_\eta(x) = \mathbb{I}\{(x - \eta)^T v_\eta \geq 0\} | \|\eta - \mu\|_2 \geq r\}$. Since $\mathbb{E}[\mathbb{I}\{(X - \eta)^T v_\eta \geq 0\}] = \Phi(-\|\eta - \mu\|_2)$, we have $E[f(X)] \leq \frac{1}{2} - \alpha \forall f \in \mathcal{G}_\mu$.

Now, note that $\mathcal{G}_\mu \subseteq \{f_\eta(x) = \mathbb{I}\{(x - \eta)^T v_\eta \geq 0\} | \eta \in \mathbb{R}^d\}$. By reparameterizing x , we have $\text{VC}(\{f_\eta(x) = \mathbb{I}\{(x - \eta)^T v_\eta \geq 0\} | \eta \in \mathbb{R}^d\}) = \text{VC}(\{f_\eta(x) = \mathbb{I}\{(x - \eta)^T \eta \geq 0\} | \eta \in \mathbb{R}^d\})$.

Observe that $\text{VC}(\{f_\eta(x) = \mathbb{I}\{(x - \eta)^T \eta \geq 0\} | \eta \in \mathbb{R}^d\}) \leq \text{VC}(\{f_{\eta,v}(x) = \mathbb{I}\{(x - \eta)^T v \geq 0\} | \eta, v \in \mathbb{R}^d\}) = d + 1$. Thus, $\text{VC}(\mathcal{G}_\mu) \leq d + 1 \leq 2d$.

Thus, by Proposition 4, with probability at least $1 - \delta/4$,

$$\sup_{\eta: \|\eta - \mu\|_2 \geq r} D_w(\eta, \mathbf{Z}) \leq \sum_{i \in B} w_i + \sup_{g \in \mathcal{G}_\mu} \sum_{i=1}^n w_i g(X_i) \quad (63)$$

$$\leq \sum_{i \in B} w_i + \frac{1}{2} - \alpha + \|w\|_2 \left(62\sqrt{2d} + \sqrt{\frac{\log 4/\delta}{2}} \right). \quad (64)$$

Again, by McDiarmid's inequality, with probability at least $1 - \delta/4$,

$$\sum_{i \in B} w_i \leq w^T \lambda + \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}}. \quad (65)$$

Thus, with probability at least $1 - \delta/2$, we have

$$\sup_{\eta: \|\eta - \mu\|_2 \geq r} D_w(\eta, \mathbf{Z}) \leq \frac{1}{2} - \alpha + w^T \lambda + \|w\|_2 \left(88\sqrt{d} + 2\sqrt{\frac{\log 4/\delta}{2}} \right). \quad (66)$$

Combining (61) and (66), picking $\alpha = \frac{3}{2}w^T \lambda + \|w\|_2 \left(150\sqrt{d} + 3.5\sqrt{\frac{\log 4/\delta}{2}} \right)$ ensures that no point η such that $\|\mu - \eta\|_2 \geq \Phi^{-1}(\frac{1}{2} + \alpha)$ can be returned by the Tukey median estimator. Thus, with probability at least $1 - \delta$, we have

$$\|\hat{\mu}_{\text{TM}} - \mu\|_2 \leq \Phi^{-1} \left(\frac{1}{2} + \alpha \right) \quad (67)$$

$$\leq 3\alpha, \quad (68)$$

using the identity $\Phi^{-1}(\frac{1}{2} + x) \leq 3x \forall x \in [0, \frac{1}{3}]$. The above upper bound is valid as long as $\alpha = \frac{3}{2}w^T \lambda + \|w\|_2 \left(150\sqrt{d} + 3.5\sqrt{\frac{\log 4/\delta}{2}} \right) < \frac{1}{3}$.

Setting $\delta = 1/5$, ensuring $\frac{3}{2}w^T \lambda + \|w\|_2 \left(150\sqrt{d} + 4.3 \right) < \frac{3}{2}w^T \lambda + 155\|w\|_2 \sqrt{d} \leq \frac{1}{3}$ suffices.

Thus, summarizing, let $g(w, \lambda) = \frac{3}{2}w^T \lambda + 155\|w\|_2 \sqrt{d}$. For w and λ such that $g(w, \lambda) \leq \frac{1}{3}$, we have the guarantee that with probability at least $\frac{4}{5}$,

$$\|\hat{\mu}_{\text{TM}} - \mu\|_2 \leq 3g(w, \lambda). \quad (69)$$

To reduce the above condition to the simpler form stated in the main paper, note that

$$g(w, \lambda)^2 \leq \left(\frac{3}{2}w^T \lambda + 155\|w\|_2 \sqrt{d} \right)^2 \quad (70)$$

$$\leq \left(155w^T \lambda + 155\|w\|_2 \sqrt{d} \right)^2 \quad (71)$$

$$\leq 2 \times 155^2 \left((w^T \lambda)^2 + d\|w\|_2^2 \right). \quad (72)$$

Ensuring the above is less than $\frac{1}{9}$ suffices. Thus, $\forall w \in \Delta_n$ such that $(w^T \lambda)^2 + d\|w\|_2^2 \leq \frac{1}{432450}$, we have

$$\sup_{P \in \mathcal{D}_d^N} \Pr_{\mathbf{Z} \sim \lambda^P} [\|\hat{\mu}_{\text{TM}}(\mathbf{Z}, w) - \mu_P\|_2^2 \geq 432450 \left((w^T \lambda)^2 + d\|w\|_2^2 \right)] \leq 1/5 \quad (73)$$

Correspondingly, the threshold based estimator satisfies the following $\forall t \in [0, 1]$ such that $t^2 + \frac{d}{N(t)} \leq \frac{1}{432450}$, we have

$$\sup_{P \in \mathcal{D}_d^N} \Pr_{\mathbf{Z} \sim \lambda^P} \left[\|\hat{\mu}_S(\mathbf{Z}, w(t)) - \mu_P\|_2^2 \geq 432450 \left(t^2 + \frac{d}{N(t)} \right) \right] \leq 1/5 \quad (74)$$

C.2 Lower Bound

We provide a lower bound for Gaussian mean estimation in $\mathbb{R}^d$ in this section. For a vector $v \in \mathbb{R}^{\sqrt{d}}$, let $e(v)$ denote a vector in $\mathbb{R}^d$ with v as the first $\sqrt{d}$ elements and the last $d - \sqrt{d}$ elements equal to 0. Define the parameterized distribution $P_\delta(\tau) = \mathcal{N}(\delta e(\tau), I)$ for $\delta > 0$ to be specified later, and let $\mathcal{P}_\delta = \{P_\delta(\tau) | \tau \in \{-1, 1\}^{\sqrt{d}}\}$.

Note: The supremum in our minimax definition is over both the true distribution and the adversarial strategy. We shall specify the adversarial strategy later for our lower bound argument.

Note that $L_{\text{PAC}}(\lambda, \mathcal{D}_d^N) \geq L_{\text{PAC}}(\lambda, \mathcal{P}_\delta) \forall \delta$ and thus, we shall lower bound $L_{\text{PAC}}(\lambda, \mathcal{P}_\delta)$.

We begin by lower bounding

$$L_{\mathbb{E}}(\lambda, \mathcal{P}_\delta) = \inf_M \sup_{P \in \mathcal{P}_\delta} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} [\|M(\mathbf{Z}) - \mu_P\|_2^2]. \quad (75)$$

For a vector τ , let τ'^j denote the vector such that $\tau_i = \tau'_i \forall i \neq j$ and $\tau_j = -\tau'_j$. Using Assouad's lower bound technique [28],

$$\inf_M \sup_{P \in \mathcal{P}_\delta} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} [\|M(\mathbf{Z}) - \mu_P\|_2^2] \geq \inf_M \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \mathbb{E} [\|M(\mathbf{Z}) - \delta e(\tau)\|_2^2] \quad (76)$$

$$= \inf_M \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \sum_{j=1}^{\sqrt{d}} \mathbb{E} [|M(\mathbf{Z})_j - \delta e(\tau)_j|^2] \quad (77)$$

$$= \inf_M \sum_{j=1}^{\sqrt{d}} \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \mathbb{E} [|M(\mathbf{Z})_j - \delta e(\tau)_j|^2] \quad (78)$$

$$\geq \inf_M \sum_{j=1}^{\sqrt{d}} \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \mathbb{E} [|M(\mathbf{Z})_j - \delta \tau_j|^2]. \quad (79)$$

Now, note that

$$\sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \mathbb{E} [|M(\mathbf{Z})_j - \delta \tau_j|^2] = \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \frac{\mathbb{E} [|M(\mathbf{Z})_j - \delta \tau_j|^2] + \mathbb{E} [|M(\mathbf{Z})_j - \delta \tau'_j|^2]}{2}, \forall j \quad (80)$$

Thus, we get

$$L_{\mathbb{E}}(\lambda, \mathcal{P}_\delta) \geq \inf_M \sum_{j=1}^{\sqrt{d}} \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \frac{\mathbb{E} [|M(\mathbf{Z})_j - \delta \tau_j|^2] + \mathbb{E} [|M(\mathbf{Z})_j - \delta \tau'_j|^2]}{2} \quad (81)$$

$$\geq \sum_{j=1}^{\sqrt{d}} \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \inf_M \frac{\mathbb{E} [|M(\mathbf{Z})_j - \delta \tau_j|^2] + \mathbb{E} [|M(\mathbf{Z})_j - \delta \tau'_j|^2]}{2} \quad (82)$$

$$\geq \sum_{j=1}^{\sqrt{d}} \frac{1}{2^{\sqrt{d}}} \sum_{\tau \in \{-1, 1\}^{\sqrt{d}}} \delta^2 (1 - \text{TV}(P_{\mathbf{Z} \sim_\lambda P_\delta(\tau)}, P_{\mathbf{Z} \sim_\lambda P_\delta(\tau'^j)})), \quad (83)$$

where the last line follows by Le Cam's method and the notation $P_{\mathbf{Z} \sim_\lambda P_\delta(\tau)}$ refers to the distribution of $\mathbf{Z}$ when the true distribution is $P_\delta(\tau)$. We shall now specify a particular adversarial strategy to lower bound the above.

Adversarial strategy motivation: Consider a particular sample with corruption rate λ and let the underlying true distribution be $P_\delta(\tau)$. Denote the perturbed sample as $Z(\tau)$ and the outlier to be $\tilde{X}(\tau)$. Note that $Z(\tau) \sim (1 - \lambda)P_{X(\tau)} + \lambda P_{\tilde{X}(\tau)}$. For any particular clean sample $X(\tau) \sim P_\delta(\tau)$, the adversary's goal is to ensure the sample $Z(\tau)$ contains no information about τ . One possible way

is that the adversary can try to ensure $Z(\tau) \sim \frac{\max_{\tau'} P(\tau')}{T}$, where the maximum is taken pointwise over the pdf and T is a normalizing constant. Observe the identity

$$P(\tau) + \left(\max_{\tau' \neq \tau} P(\tau') - P(\tau) \right)_+ = \max_{\tau'} P(\tau'), \quad (84)$$

where $(\cdot)_+ := \max\{\cdot, 0\}$ is pointwise over the pdf. First, note that

$$\int_{x \in \mathbb{R}^d} \left(\max_{\tau' \neq \tau} P_\delta(\tau')(x) - P_\delta(\tau)(x) \right)_+ dx = T - 1,$$

where $P_\delta(\tau)(x)$ is understood to be the pdf of $P_\delta(\tau)$ at x . Thus, $\frac{(\max_{\tau' \neq \tau} P(\tau') - P(\tau))_+}{T-1}$ is a valid pdf, and for $\lambda = \frac{T-1}{T}$, we have $(1 - \lambda)P(\tau) + \lambda \frac{(\max_{\tau' \neq \tau} P(\tau') - P(\tau))_+}{T-1} = \frac{\max_{\tau'} P(\tau')}{T}$. Thus, for any $\lambda \geq 1 - \frac{1}{T}$, there exists a way for the adversary to make the distribution of $Z(\tau) \sim \frac{\max_{\tau'} P(\tau')}{T}$, rendering the sample useless for identifying τ .

Now, we find an upper bound on T in terms of δ .

$$T = \int_{x \in \mathbb{R}^d} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{\|x - \delta e(\tau)\|_2^2}{2}} dx \quad (85)$$

$$= \int_{x \in \mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} e^{-\frac{\|x - \delta e(\tau)\|_2^2}{2}} dx \quad (86)$$

$$= \int_{x' \in \mathbb{R}^{\sqrt{d}}} \frac{1}{\sqrt{(2\pi)^{\sqrt{d}}}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} e^{-\frac{\|x' - \delta \tau\|_2^2}{2}} dx' \quad (87)$$

$$= \left[\int_{x'' \in \mathbb{R}} \frac{1}{\sqrt{(2\pi)}} \max_{\tau \in \{-1, 1\}} e^{-\frac{\|x'' - \delta \tau\|_2^2}{2}} dx'' \right]^{\sqrt{d}} \quad (88)$$

$$= \left[2 \int_{y \in \mathbb{R}: y \geq 0} \frac{1}{\sqrt{2\pi}} e^{-\frac{\|y - \delta\|_2^2}{2}} dy \right]^{\sqrt{d}} \quad (89)$$

$$= [2\Phi(\delta)]^{\sqrt{d}}, \quad (90)$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Using $2\Phi(x) \leq 1 + x$ and $1 + x \leq e^x$, we obtain

$$T \leq e^{\delta\sqrt{d}}. \quad (91)$$

Thus, if $\lambda \geq 1 - e^{-\delta\sqrt{d}}$ then the adversary can ensure that $Z(\tau) \sim \frac{\max_{\tau'} P(\tau')}{T}$ and contains no information about τ .

Adversarial strategy: If for sample i , $\lambda_i \geq 1 - e^{-\delta\sqrt{d}}$, then independently sample $\tilde{X}(\tau) \sim \beta \left(\frac{\max_{\tau' \neq \tau} P_\delta(\tau') - P_\delta(\tau)}{T-1} \right)_+ + (1 - \beta) \frac{\max_{\tau'} P_\delta(\tau')}{T}$, where $\beta = \frac{(T-1)(1-\lambda)}{\lambda}$. The constraint $\lambda_i \geq 1 - e^{-\delta\sqrt{d}}$ ensures $\beta \in [0, 1]$ and the above is a valid distribution.

Thus, $Z(\tau) \sim \frac{\max_{\tau'} P_\delta(\tau')}{T}$. If $\lambda_i < 1 - e^{-\delta\sqrt{d}}$, then adversary does no corruption, or equivalently, independently sample $\tilde{X}(\tau) \sim P_\delta(\tau)$ so that $Z(\tau) \sim P_\delta(\tau)$.

Thus, using Pinsker's inequality in (83), obtain

$$L_{\mathbb{E}}(\boldsymbol{\lambda}, \mathcal{P}_{\delta}) \geq \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 \left(1 - \sqrt{\frac{1}{2} \text{KL}(P_{\mathbf{Z} \sim \lambda} P_{\delta}(\tau), P_{\mathbf{Z} \sim \lambda} P_{\delta}(\tau'^j))} \right) \quad (92)$$

$$= \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 \left(1 - \sqrt{\frac{1}{2} \text{KL}(\otimes_{i=1}^n P_{Z_i}(\tau), \otimes_{i=1}^n P_{Z_i}(\tau'^j))} \right) \quad (93)$$

$$= \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 \left(1 - \sqrt{\frac{1}{2} \sum_{i: \lambda_i < 1-e^{-\delta\sqrt{d}}} \text{KL}(P_{\delta}(\tau), P_{\delta}(\tau'^j))} \right) \quad (94)$$

$$= \sqrt{d} \delta^2 \left(1 - \sqrt{\delta^2 n (1 - e^{-\delta\sqrt{d}})} \right) \quad \forall \delta \quad (95)$$

Let δ_* be such that

$$n(1 - e^{-\delta_*\sqrt{d}}) \leq \frac{1}{64\delta_*^2}, \text{ and } N(1 - e^{-\delta_*\sqrt{d}}) \geq \frac{1}{64\delta_*^2}. \quad (96)$$

Substituting δ_* in (95),

$$\frac{7}{8} \sqrt{d} \delta_*^2 \leq \inf_M \sup_{P \in \mathcal{P}_{\delta_*}} \mathbb{E}_{\mathbf{Z} \sim \lambda P} [\|M(\mathbf{Z}) - \mu_P\|_2^2]. \quad (97)$$

For a random variable K , let $Q(K, \alpha) := \inf\{t : \Pr[K \geq t] \leq (1-\alpha)\}$, i.e., $Q(\cdot, \alpha)$ is the α -quantile of the random variable. Note that for a random variable K , $\mathbb{E}K \leq Q(K, x)x + (1-x)\text{ess-sup}K$ $\forall x \in [0, 1]$, where ess-sup denotes essential supremum. Further, note that in the minimax term $\inf_M \sup_{P \in \mathcal{P}_{\delta_*}}$, we can restrict ourselves to estimators M which output in $[-\delta_*, \delta_*]^{\sqrt{d}} \times \{0\}^{d-\sqrt{d}}$ almost surely – otherwise the error can be reduced by projected to this region. Thus $\text{ess-sup}\|M(\mathbf{Z}) - \mu_P\|_2^2 \leq 4\sqrt{d}\delta_*^2$ for any estimator M and any distribution P . Thus, combining the above observations, we have

$$\inf_M \sup_{P \in \mathcal{P}_{\delta_*}} \mathbb{E}_{\mathbf{Z} \sim \lambda P} \|M(\mathbf{Z}) - \mu_P\|_2^2 \leq \inf_M \sup_{P \in \mathcal{P}_{\delta_*}} \frac{4}{5} Q\left(\|M(\mathbf{Z}) - \mu_P\|_2^2, \frac{4}{5}\right) + \frac{4\sqrt{d}\delta_*^2}{5}. \quad (98)$$

Using (97), we get

$$\frac{7}{8} \sqrt{d} \delta_*^2 \leq \frac{4}{5} \inf_M \sup_{P \in \mathcal{P}_{\delta_*}} Q\left(\|M(\mathbf{Z}) - \mu_P\|_2^2, \frac{4}{5}\right) + \frac{4\sqrt{d}\delta_*^2}{5} \quad (99)$$

$$= \frac{4}{5} L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{P}_{\delta_*}) + \frac{4\sqrt{d}\delta_*^2}{5} \quad (100)$$

$$\implies \frac{3}{40} \sqrt{d} \delta_*^2 \leq L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{P}_{\delta_*}) \leq L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^{\mathcal{N}}). \quad (101)$$

C.3 Minimax Optimality

For proving Theorem 4, we shall use the weighing $w_i = \frac{\mathbb{I}\{\lambda_i \leq 1-e^{-\delta_*\sqrt{d}}\}}{N(1-e^{-\delta_*\sqrt{d}})}$ in our upper bound. From (74), $\forall t \in [0, 1]$ such that $t^2 + \frac{d}{N(t)} \leq c$ for some universal constant c , we have

$$\|\hat{\mu}_S(\mathbf{Z}, w(t)) - \mu\|_2^2 \lesssim \left(t^2 + \frac{d}{N(t)} \right). \quad (102)$$

Thus, if $\min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$ and let the minimum value be attained at t^* , then we have

$$\|\hat{\mu}_S(\mathbf{Z}, w(t^*)) - \mu\|_2^2 \lesssim \left(t^{*2} + \frac{d}{N(t^*)} \right) \quad (103)$$

$$\leq \left(1 - e^{-\delta_* \sqrt{d}} \right)^2 + \frac{d}{N \left(1 - e^{-\delta_* \sqrt{d}} \right)} \quad (104)$$

$$\leq d\delta_*^2 + 64d\delta_*^2 \quad (105)$$

$$\simeq d\delta_*^2, \quad (106)$$

where we used (96) in (105). Combining with (101), when $\min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$, we have

$$\sqrt{d}\delta_*^2 \lesssim L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^N) \lesssim \min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \leq d\delta_*^2. \quad (107)$$

Thus, when $\min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$,

$$\frac{1}{\sqrt{d}} \min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right) \lesssim L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^N) \lesssim \min_{t \in [0,1]} \left(t^2 + \frac{d}{N(t)} \right). \quad (108)$$

D Linear Regression: Proofs

D.1 Upper Bound

Recall

$$D_w(\eta, \mathbf{Z}) = \min_{v \in \mathbb{S}_d} \sum_{i=1}^n w_i \mathbb{I}\{(\hat{Y}_i - \eta^T \hat{W}_i)(v^T \hat{W}_i) \geq 0\}, \quad (109)$$

$$\hat{\beta}_{\text{TC}}(\mathbf{Z}, w) := \arg \max_{\eta \in \mathbb{R}^d} D_w(\eta, \mathbf{Z}). \quad (110)$$

Let the true underlying regression coefficient be β , i.e., $W \sim \mathcal{N}(0, \Sigma)$ and conditioned on W , $Y \sim \mathcal{N}(\beta^T W, \sigma^2)$. Let $G = \{i : B_i = 0\}$, $B = [n] \setminus G$, and Let $w(G) = \{w_i | i \in G\}$. Note that $Z_i = (W_i, Y_i)$ for $i \in G$. The depth of the true coefficient is lower bounded as

$$D_w(\beta, \mathbf{Z}) \geq \min_{v \in \mathbb{S}_d} \sum_{i \in G} w_i \mathbb{I}\{(Y_i - \beta^T W_i)(v^T W_i) \geq 0\}. \quad (111)$$

Define the class of indicator functions $\mathcal{F}_\beta = \{f_v(w, y) = \mathbb{I}\{(y - w^T \beta)(v^T w) \geq 0\} | v \in \mathbb{S}_d\}$. Note that $E[f(Z)] = \frac{1}{2} \forall f \in \mathcal{F}_\beta$. By Proposition 4, we have with probability at least $1 - \frac{\delta}{4}$

$$\min_{f \in \mathcal{F}_\beta} \sum_{i \in G} w_i \left(f(X_i) - \frac{1}{2} \right) \geq -62 \|w(G)\|_2 \sqrt{\text{VC}(\mathcal{F}_\beta)} - \|w(G)\|_2 \sqrt{\frac{\log 4/\delta}{2}} \quad (112)$$

$$\geq -\|w\|_2 \left(62\sqrt{d} + \sqrt{\frac{\log 4/\delta}{2}} \right), \quad (113)$$

where we used $\text{VC}(\mathcal{F}_\beta) = d$. To see this, notice $\mathbb{I}\{(y - w^T \beta)(v^T w) \geq 0\} = \mathbb{I}\{v^T (w(y - w^T \beta)) \geq 0\} = \mathbb{I}\{v^T \tilde{w} \geq 0\}$, where $\tilde{w} = w(y - w^T \beta) \in \mathbb{R}^d$. Thus, it is equal to the VC dimension of homogeneous half-space classifiers, which is d [43, Corollary 4.2.2]. Further, with probability at

least $1 - \delta/4$, by McDiarmid's inequality [28]

$$\sum_{i \in G} w_i = \sum_{i=1}^n w_i \mathbb{I}\{B_i = 0\} \quad (114)$$

$$\geq \sum_{i=1}^n w_i (1 - \lambda_i) - \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}} \quad (115)$$

$$= 1 - w^T \lambda - \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}}. \quad (116)$$

Thus, with probability at least $1 - \delta/2$,

$$D_w(\beta, \mathbf{Z}) \geq \frac{1}{2} - \frac{w^T \lambda}{2} - \|w\|_2 \left(62\sqrt{d} + 1.5\sqrt{\frac{\log 4/\delta}{2}} \right). \quad (117)$$

Next, we show that depth of any point far away from the coefficient is low. For any $\eta \in \mathbb{R}^d$ such that $\|\eta - \mu\|_\Sigma \geq r$, let $v_\eta = \frac{\eta - \mu}{\|\eta - \mu\|_2}$. We shall set the value of $r > 0$ later.

$$\sup_{\eta: \|\eta - \mu\|_2 \geq r} D_w(\eta, \mathbf{Z}) \leq \sum_{i \in B} w_i + \sup_{\eta: \|\eta - \mu\|_2 \geq r} \sum_{i=1}^n w_i \mathbb{I}\{(Y_i - \eta^T W_i) v_\eta^T W_i \geq 0\}. \quad (118)$$

Again, by McDiarmid's inequality, with probability at least $1 - \delta/4$,

$$\sum_{i \in B} w_i \leq w^T \lambda + \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}}. \quad (119)$$

Define the class of indicator functions $\mathcal{G}_\beta = \{f_\eta(w, y) = \mathbb{I}\{(y - \eta^T w)(v_\eta^T w) \geq 0\} | \|\eta - \mu\|_\Sigma \geq r\}$. Thus, by Proposition 4, with probability at least $1 - \delta/4$,

$$\sup_{f \in \mathcal{G}_\beta} \sum_{i=1}^n w_i (f(Z_i) - \mathbb{E}[f(Z_i)]) \leq 62\|w\|_2 \sqrt{\text{VC}(\mathcal{G}_\beta)} + \|w\|_2 \sqrt{\frac{\log 4/\delta}{2}} \quad (120)$$

Now, note that

$$\mathbb{E}[f(Z)] = \Pr[(Y - \eta^T W)W^T(\eta - \beta) \geq 0] \quad (121)$$

Using $W \sim \mathcal{N}(0, \Sigma)$ and $Y|W \sim \mathcal{N}(\beta^T W, \sigma^2)$, we get that

$$(Y - \eta^T W)W^T(\eta - \beta) \sim M(\zeta - M) \quad (122)$$

where $M = W^T(\eta - \beta) \sim \mathcal{N}(0, \|\eta - \beta\|_\Sigma^2)$ and $\zeta \sim \mathcal{N}(0, \sigma^2)$ are independent. Letting $T_1, T_2 \sim_{iid} \mathcal{N}(0, 1)$,

$$\Pr[(Y - \eta^T W)W^T(\eta - \beta) \geq 0] = \Pr[M(\zeta - M) \geq 0] \quad (123)$$

$$= \Pr[\zeta \geq M | M \geq 0] \quad (124)$$

$$= \Pr[\sigma T_2 \geq \|\eta - \beta\|_\Sigma T_1 | T_1 \geq 0] \quad (125)$$

$$= 1 - \Pr\left[T_2 \leq \frac{\|\eta - \beta\|_\Sigma}{\sigma} T_1 | T_1 \geq 0\right] \quad (126)$$

$$= 1 - \left(\frac{1}{2} + \frac{1}{\pi} \arctan \frac{\|\eta - \beta\|_\Sigma}{\sigma}\right) \quad (127)$$

$$= \frac{1}{2} - \frac{1}{\pi} \arctan \frac{\|\eta - \beta\|_\Sigma}{\sigma} \quad (128)$$

Thus,

$$\mathbb{E}[f(Z)] \leq \frac{1}{2} - \frac{1}{\pi} \arctan \frac{r}{\sigma} \quad \forall f \in \mathcal{G}_\beta. \quad (129)$$

Thus, with probability at least $1 - \delta/2$, and using VC dimension bound presented in Proposition 3,

$$\sup_{\eta: \|\eta - \mu\|_{\Sigma} \geq r} D_w(\eta, \mathbf{Z}) \leq \frac{1}{2} - \frac{1}{\pi} \arctan \frac{r}{\sigma} + w^T \lambda + \|w\|_2 \left(2\sqrt{\frac{\log 4/\delta}{2}} + 879\sqrt{d} \right). \quad (130)$$

Combining with (117), setting

$$r = \sigma \tan \left\{ \pi \left[\frac{3}{2} w^T \lambda + \|w\|_2 \left(3.5\sqrt{\frac{\log 4/\delta}{2}} + 941\sqrt{d} \right) \right] \right\} \quad (131)$$

ensures that the estimator $\hat{\beta}_{\text{TC}}$ satisfies

$$\|\hat{\beta}_{\text{TC}} - \beta\|_{\Sigma} \leq r \quad (132)$$

with probability at least $1 - \delta$ as long as $\left[\frac{3}{2} w^T \lambda + \|w\|_2 \left(3.5\sqrt{\frac{\log 4/\delta}{2}} + 941\sqrt{d} \right) \right] \leq \frac{2}{5}$.

Substitute $\delta = \frac{1}{5}$ and let $g(w, \lambda) = \frac{3}{2} w^T \lambda + 946\|w\|_2 \sqrt{d}$. Then, for w such that $g(w, \lambda) \leq \frac{2}{5}$, we have with probability at least $\frac{4}{5}$,

$$\|\hat{\beta}_{\text{TC}} - \beta\|_{\Sigma} \leq \sigma \tan \pi g(w, \lambda) \leq 8\sigma g(w, \lambda), \quad (133)$$

where we used the identity $\tan \pi x \leq 8x \forall x \in [0, 2/5]$.

To reduce the above condition to the simpler form stated in the main paper, note that

$$\left(\frac{3}{2} w^T \lambda + 946\|w\|_2 \sqrt{d} \right)^2 \leq \left(946 w^T \lambda + 946\|w\|_2 \sqrt{d} \right)^2 \quad (134)$$

$$\leq 2 \times 946^2 ((w^T \lambda)^2 + d\|w\|_2^2). \quad (135)$$

Ensuring the above is less than $\frac{4}{25}$ suffices.

Thus, $\forall w \in \Delta_n$ such that $(w^T \lambda)^2 + d\|w\|_2^2 \leq \frac{1}{11186450}$, we have

$$\sup_{P \in \mathcal{D}_d^N} \Pr_{\mathbf{Z} \sim_{\lambda} P} \left[\|\hat{\beta}_{\text{TC}}(\mathbf{Z}, w) - \mu_P\|_{\Sigma}^2 \geq 114549248\sigma^2 ((w^T \lambda)^2 + d\|w\|_2^2) \right] \leq 1/5 \quad (136)$$

Correspondingly, the threshold based estimator satisfies the following $\forall t \in [0, 1]$ such that $t^2 + \frac{d}{N(t)} \leq \frac{1}{11186450}$, we have

$$\sup_{P \in \mathcal{D}_d^N} \Pr_{\mathbf{Z} \sim_{\lambda} P} \left[\|\hat{\mu}_{\text{S}}(\mathbf{Z}, w(t)) - \mu_P\|_{\Sigma}^2 \geq 114549248\sigma^2 \left(t^2 + \frac{d}{N(t)} \right) \right] \leq 1/5 \quad (137)$$

Proposition 3. $\text{VC}(\mathcal{G}_{\beta}) \leq 200d$.

Proof. Recall $\mathcal{G}_{\beta} = \{f_{\eta}(w, y) = \mathbb{I}\{(y - \eta^T w)(v_{\eta}^T w) \geq 0\} | \|\eta - \mu\|_{\Sigma} \geq r\}$, where $v_{\eta} = \frac{\eta - \beta}{\|\eta - \beta\|_2}$. Let $\mathcal{G}_{\beta}^+ = \{f_{\eta}(w, y) = \mathbb{I}\{(y - \eta^T w)(v_{\eta}^T w) \geq 0\} | \eta \in \mathbb{R}^d, \eta \neq \mu\}$. Note that for the purposes of VC dimension calculation, by re-parameterizing y and η , we can write

$$\mathcal{G}_{\beta}^+ = \{g_{\eta}(w, y) = \mathbb{I}\{(y - \eta^T w)(\eta^T w) \geq 0\} | \eta \in \mathbb{R}^d, \eta \neq 0\}. \quad (138)$$

We now switch to region notation instead of a functional notation. Let $G_{\eta} = \{(w, y) | g_{\eta}(w, y) = 1\}$ and define the following

- $H_{\eta}^{1+} = \{(w, y) | y - \eta^T w \geq 0\}$,
- $H_{\eta}^{1-} = \{(w, y) | y - \eta^T w \leq 0\}$,
- $H_{\eta}^{2+} = \{(w, y) | \eta^T w \geq 0\}$,

- $H_\eta^{2-} = \{(w, y) | \eta^T w \leq 0\}.$

Note that $G_\eta = (H_\eta^{1+} \cap H_\eta^{2+}) \cup (H_\eta^{1-} \cap H_\eta^{2-})$. Define $\mathcal{G} = \{G_\eta | \eta \in \mathbb{R}^d, \eta \neq 0\}$, $\mathcal{H}^+ = \{H_\eta^{1+} \cap H_\eta^{2+} | \eta \in \mathbb{R}^d, \eta \neq 0\}$ and $\mathcal{H}^- = \{H_\eta^{1-} \cap H_\eta^{2-} | \eta \in \mathbb{R}^d, \eta \neq 0\}$.

We use the following property related to VC dimension [43]:

$$\text{VC}(\{A \cap B | A \in \mathcal{A}, B \in \mathcal{B}\}) \leq 10 \max\{\text{VC}(\mathcal{A}), \text{VC}(\mathcal{B})\}. \quad (139)$$

The above holds for union as well.

Noting that $\mathcal{G} \subseteq \{G = H_1 \cup H_2 | H_1 \in \mathcal{H}^+, H_2 \in \mathcal{H}^-\}$. Define $T = \max\{\text{VC}(\mathcal{H}^+), \text{VC}(\mathcal{H}^-)\}$ then $\text{VC}(\mathcal{G}_\beta) = \text{VC}(\mathcal{G}) \leq 10T$. Similarly, by writing $\mathcal{H}^+ \subseteq \{H_\eta^{1+} \cap H_{\eta'}^{2+} | \eta, \eta' \in \mathbb{R}^d, \eta, \eta' \neq 0\}$, we have $\text{VC}(\mathcal{H}^+) \leq 10 \max\{\text{VC}(\{H_\eta^{1+}\}), \text{VC}(\{H_\eta^{2+}\})\} = 10(d+1)$, where we used $\text{VC}(\{H_\eta^{1+}\}) = d+1$ and $\text{VC}(\{H_\eta^{2+}\}) = d$. Similarly, $\text{VC}(\mathcal{H}^-) \leq 10(d+1)$.

Thus, $\text{VC}(\mathcal{G}_\beta) \leq 100(d+1) \leq 200d$. $\square$

D.2 Lower Bound

The lower bound argument we present is similar to Appendix C.2.

We begin by noting that when true regression coefficient is β then $(W, Y) \sim \mathcal{N}\left(0, \begin{bmatrix} \Sigma & \Sigma\beta \\ \beta^T \Sigma & \beta^T \Sigma \beta + \sigma^2 \end{bmatrix}\right)$.

For a vector $v \in \mathbb{R}^{\sqrt{d}}$, let $e(v)$ denote a vector in $\mathbb{R}^d$ with v as the first $\sqrt{d}$ elements and the last $d - \sqrt{d}$ elements equal to 0. Let $\beta(\tau) = \delta \Sigma^{-\frac{1}{2}} e(\tau)$ for $\delta > 0$ to be specified later. For $\tau \in \{-1, 1\}^{\sqrt{d}}$, define the $Y|W \sim \mathcal{N}(W^T \beta(\tau), \sigma^2)$. In other words, for $\tau \in \{-1, 1\}^{\sqrt{d}}$, $(W, Y) \sim P_\delta(\tau)$ where

$$P_\delta(\tau) = \mathcal{N}\left(0, \begin{bmatrix} \Sigma & \Sigma\beta(\tau) \\ \beta(\tau)^T \Sigma & \beta(\tau)^T \Sigma \beta(\tau) + \sigma^2 \end{bmatrix}\right)$$

and let $\mathcal{P}_\delta = \{P_\delta(\tau) | \tau \in \{-1, 1\}^{\sqrt{d}}\}$.

We have $L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{D}(\Sigma, \sigma^2)) \geq L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{P}_\delta) \forall \delta$ and thus, we shall lower bound $L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{P}_\delta)$.

We begin by lower bounding

$$L_{\mathbb{E}}(\boldsymbol{\lambda}, \mathcal{P}_\delta) = \inf_M \sup_{P \in \mathcal{P}_\delta} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} [\|M(\mathbf{Z}) - \beta_P\|_\Sigma^2]. \quad (140)$$

For a vector τ , let τ^{ij} denote the vector such that $\tau_i = \tau_i^{ij} \forall i \neq j$ and $\tau_j = -\tau_j^{ij}$. For an estimation $M(\mathbf{Z})$, define $N(\mathbf{Z}) = \Sigma^{-\frac{1}{2}} M(\mathbf{Z})$ – note that this is a bijective map. Using Assouad's lower bound

technique [28],

$$\inf_M \sup_{P \in \mathcal{P}_\delta} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} [\|M(\mathbf{Z}) - \beta_P\|_\Sigma^2] \geq \inf_M \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \mathbb{E} [\|M(\mathbf{Z}) - \beta(\tau)\|_\Sigma^2] \quad (141)$$

$$= \inf_M \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \mathbb{E} [\|N(\mathbf{Z}) - \delta e(\tau)\|_2^2] \quad (142)$$

$$= \inf_N \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \mathbb{E} [\|N(\mathbf{Z}) - \delta e(\tau)\|_2^2] \quad (143)$$

$$= \inf_N \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \sum_{j=1}^d \mathbb{E} [|N(\mathbf{Z})_j - \delta e(\tau)_j|^2] \quad (144)$$

$$= \inf_N \sum_{j=1}^d \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \mathbb{E} [|N(\mathbf{Z})_j - \delta e(\tau)_j|^2] \quad (145)$$

$$\geq \inf_N \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j|^2]. \quad (146)$$

Now, note that

$$\sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j|^2] = \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \frac{\mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j|^2] + \mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j'^j|^2]}{2} \quad \forall j \quad (147)$$

Thus, we get

$$L_{\mathbb{E}}(\lambda, \mathcal{P}_\delta) \geq \inf_N \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \frac{\mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j|^2] + \mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j'^j|^2]}{2} \quad (148)$$

$$\geq \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \inf_N \frac{\mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j|^2] + \mathbb{E} [|N(\mathbf{Z})_j - \delta \tau_j'^j|^2]}{2} \quad (149)$$

$$\geq \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 (1 - \text{TV}(P_{\mathbf{Z} \sim_\lambda P_\delta(\tau)}, P_{\mathbf{Z} \sim_\lambda P_\delta(\tau'^j)})), \quad (150)$$

where the last line follows by Le Cam's method and the notation $P_{\mathbf{Z} \sim_\lambda P_\delta(\tau)}$ refers to the distribution of $\mathbf{Z}$ when the true distribution is $P_\delta(\tau)$. We shall now specify a particular adversarial strategy to lower bound the above.

Adversarial strategy motivation: Readers can refer to the motivation in Appendix C.2 for more details on the main idea behind the strategy described. We need to find an upper bound on the normalizing constant T for the pdf $\frac{\max_\tau P_\delta(\tau)}{T}$. Let

$$S(\tau) = \begin{bmatrix} \Sigma & \Sigma \beta(\tau) \\ \beta(\tau)^T \Sigma & \beta(\tau)^T \Sigma \beta(\tau) + \sigma^2 \end{bmatrix}$$

Thus,

$$T = \int_{w \in \mathbb{R}^d, y \in \mathbb{R}} \max_{\tau \in \{-1,1\}^{\sqrt{d}}} \frac{1}{\sqrt{(2\pi)^{d+1} |S(\tau)|}} e^{-\frac{\|(w,y)\|_{S(\tau)}^2}{2}} dw dy \quad (151)$$

By Schur's formula, $|S(\tau)| = |\Sigma| \sigma^2$ and by block inversion formula,

$$S(\tau)^{-1} = \begin{bmatrix} \Sigma^{-1} + \frac{\beta(\tau)\beta(\tau)^T}{\sigma^2} & -\frac{\beta(\tau)}{\sigma^2} \\ -\frac{\beta(\tau)^T}{\sigma^2} & \frac{1}{\sigma^2} \end{bmatrix}.$$

Performing change of variable $(w, y) = G(x, y)$ where $G = \begin{bmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & 1 \end{bmatrix}$, we obtain

$$T = \int_{x \in \mathbb{R}^d, y \in \mathbb{R}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} \frac{1}{\sqrt{(2\pi)^{d+1} |\Sigma| \sigma^2}} e^{-\frac{(x, y)^T G^T S(\tau)^{-1} G (x, y)}{2}} \begin{vmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & 1 \end{vmatrix} dx dy \quad (152)$$

Noting that

$$G^T S(\tau)^{-1} G = \begin{bmatrix} 1 + \frac{\delta^2}{\sigma^2} e(\tau) e(\tau)^T & -\frac{\delta}{\sigma^2} e(\tau) \\ -\frac{\delta}{\sigma^2} e(\tau)^T & \frac{1}{\sigma^2} \end{bmatrix},$$

we get

$$T = \int_{x \in \mathbb{R}^d, y \in \mathbb{R}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} \frac{1}{\sqrt{(2\pi)^{d+1} \sigma^2}} e^{-\frac{\|x\|_2^2}{2} - \frac{(\delta e(\tau)^T x - y)^2}{2\sigma^2}} dx dy \quad (153)$$

$$= \int_{x \in \mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d}} e^{-\|x\|_2^2/2} \int_{y \in \mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} e^{-\frac{(\delta e(\tau)^T x - y)^2}{2\sigma^2}} dy dx \quad (154)$$

$$= \int_{x \in \mathbb{R}^{\sqrt{d}}} \frac{1}{\sqrt{(2\pi)^{\sqrt{d}}}} e^{-\|x\|_2^2/2} \int_{y \in \mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} e^{-\frac{(\delta \tau^T x - y)^2}{2\sigma^2}} dy dx \quad (155)$$

Let $X \sim \mathcal{N}(0, I)$ be a random variable in $\mathbb{R}^{\sqrt{d}}$. We can write

$$T = \mathbb{E}_X \left[\int_{y \in \mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \max_{\tau \in \{-1, 1\}^{\sqrt{d}}} e^{-\frac{(\delta \tau^T X - y)^2}{2\sigma^2}} dy \right] \quad (156)$$

Note that $\max_{\tau \in \{-1, 1\}^{\sqrt{d}}} \delta \tau^T X = \delta \|X\|_1$. Thus,

$$\max_{\tau \in \{-1, 1\}^{\sqrt{d}}} e^{-\frac{(\delta \tau^T X - y)^2}{2\sigma^2}} \begin{cases} = e^{-(y - \delta \|X\|_1)^2 / 2\sigma^2} & y \geq \delta \|X\|_1, \\ = e^{-(y + \delta \|X\|_1)^2 / 2\sigma^2} & y \leq -\delta \|X\|_1, \\ \leq 1 & \text{else.} \end{cases}$$

Using the above upper bound, we obtain

$$T \leq \mathbb{E}_X \left[1 + \frac{2\delta \|X\|_1}{\sqrt{2\pi\sigma^2}} \right] \quad (157)$$

$$= 1 + \frac{2\delta \mathbb{E}[\|X\|_1]}{\sqrt{2\pi\sigma^2}} \quad (158)$$

$$= 1 + \frac{2\delta \sqrt{d}}{\pi \sigma} \quad (159)$$

$$\leq e^{\frac{2\delta \sqrt{d}}{\pi \sigma}}. \quad (160)$$

Thus, if $\lambda \geq 1 - e^{-\frac{2\delta \sqrt{d}}{\pi \sigma}}$ then the adversary can ensure that $Z(\tau) \sim \frac{\max_{\tau'} P(\tau')}{T}$ and contains no information about τ .

Adversarial strategy: If for sample i , $\lambda_i \geq 1 - e^{-\frac{2\delta \sqrt{d}}{\pi \sigma}}$, then independently sample $\tilde{X}(\tau) \sim \beta \left(\frac{\max_{\tau' \neq \tau} P_\delta(\tau') - P_\delta(\tau)}{T-1} \right)_+ + (1 - \beta) \frac{\max_{\tau'} P_\delta(\tau')}{T}$, where $\beta = \frac{(T-1)(1-\lambda)}{\lambda}$. The constraint $\lambda_i \geq 1 - e^{-\frac{2\delta \sqrt{d}}{\pi \sigma}}$ ensures $\beta \in [0, 1]$ and the above is a valid distribution.

Thus, $Z(\tau) \sim \frac{\max_{\tau'} P_\delta(\tau')}{T}$. If $\lambda_i < 1 - e^{-\frac{2\delta \sqrt{d}}{\pi \sigma}}$, then adversary does no corruption, or equivalently, independently sample $\tilde{X}(\tau) \sim P_\delta(\tau)$ so that $Z(\tau) \sim P_\delta(\tau)$.

Thus, using Pinsker's inequality in (150), obtain

$$L_{\mathbb{E}}(\boldsymbol{\lambda}, \mathcal{P}_{\delta}) \geq \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 \left(1 - \sqrt{\frac{1}{2} \text{KL}(P_{\mathbf{Z} \sim_{\lambda} P_{\delta}(\tau)}, P_{\mathbf{Z} \sim_{\lambda} P_{\delta}(\tau'^j)})} \right) \quad (161)$$

$$= \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 \left(1 - \sqrt{\frac{1}{2} \text{KL}(\otimes_{i=1}^n P_{Z_i(\tau)}, \otimes_{i=1}^n P_{Z_i(\tau'^j)})} \right) \quad (162)$$

$$= \sum_{j=1}^{\sqrt{d}} \frac{1}{2\sqrt{d}} \sum_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta^2 \left(1 - \sqrt{\frac{1}{2} \sum_{i: \lambda_i < 1 - e^{-\frac{2\delta\sqrt{d}}{\pi\sigma}}} \text{KL}(P_{\delta}(\tau), P_{\delta}(\tau'^j))} \right) \quad (163)$$

$$= \sqrt{d} \delta^2 \left(1 - \sqrt{\frac{\delta^2}{\sigma^2} n(1 - e^{-\frac{2\delta\sqrt{d}}{\pi\sigma}})} \right) \quad \forall \delta \quad (164)$$

where (164) follows since

$$\text{KL}(P_{\delta}(\tau), P_{\delta}(\tau'^j)) = \mathbb{E}_{W \sim \mathcal{N}(0, \Sigma)} \text{KL}(\mathcal{N}(W^T \beta(\tau), \sigma^2), \mathcal{N}(W^T \beta(\tau'^j), \sigma^2)) \quad (165)$$

$$= \mathbb{E}_W \frac{\delta^2 (\tau - \tau'^j)^T \Sigma^{-\frac{1}{2}} W W^T \Sigma^{-\frac{1}{2}} (\tau - \tau'^j)}{2\sigma^2} \quad (166)$$

$$= \frac{\delta^2 \|\tau - \tau'^j\|_2^2}{2\sigma^2} \quad (167)$$

$$= \frac{2\delta^2}{\sigma^2}. \quad (168)$$

Replacing $\delta \leftarrow \delta/\sigma$ in (164), we get

$$L_{\mathbb{E}}(\boldsymbol{\lambda}, \mathcal{P}_{\sigma\delta}) \geq \sigma^2 \sqrt{d} \delta^2 \left(1 - \sqrt{\delta^2 n(1 - e^{-\frac{2\delta\sqrt{d}}{\pi}})} \right) \quad \forall \delta, \quad (169)$$

Let δ_* be such that

$$n(1 - e^{-\frac{2\delta_*\sqrt{d}}{\pi}}) \leq \frac{1}{64\delta_*^2}, \quad \text{and} \quad N(1 - e^{-\frac{2\delta_*\sqrt{d}}{\pi}}) \geq \frac{1}{64\delta_*^2}. \quad (170)$$

Substituting δ_* in (169),

$$\frac{7}{8} \sqrt{d} \sigma^2 \delta_*^2 \leq \inf_M \sup_{P \in \mathcal{P}_{\sigma\delta_*}} \mathbb{E}_{\mathbf{Z} \sim_{\lambda} P} [\|M(\mathbf{Z}) - \beta_P\|_{\Sigma}^2]. \quad (171)$$

Note that in the minimax term $\inf_M \sup_{P \in \mathcal{P}_{\sigma\delta_*}}$, we can restrict ourselves to estimators M which output in $V = \{\Sigma^{-\frac{1}{2}} v | v \in [-\delta_*\sigma, \delta_*\sigma]^{\sqrt{d}} \times \{0\}^{d-\sqrt{d}}\}$ almost surely – otherwise the error can be reduced by projected to this region. Thus $\text{ess-sup} \|M(\mathbf{Z}) - \beta_P\|_{\Sigma}^2 \leq 4\sqrt{d}\delta_*^2\sigma^2$ for any estimator M and any distribution P . Thus, combining the above observations, we have

$$\inf_M \sup_{P \in \mathcal{P}_{\sigma\delta_*}} \mathbb{E}_{\mathbf{Z} \sim_{\lambda} P} \|M(\mathbf{Z}) - \mu_P\|_{\Sigma}^2 \leq \inf_M \sup_{P \in \mathcal{P}_{\sigma\delta_*}} \frac{4}{5} Q\left(\|M(\mathbf{Z}) - \mu_P\|_{\Sigma}^2, \frac{4}{5}\right) + \frac{4\sqrt{d}\delta_*^2\sigma^2}{5}. \quad (172)$$

Using (171), we get

$$\frac{7}{8} \sqrt{d} \delta_*^2 \sigma^2 \leq \frac{4}{5} \inf_M \sup_{P \in \mathcal{P}_{\sigma\delta_*}} Q\left(\|M(\mathbf{Z}) - \mu_P\|_{\Sigma}^2, \frac{4}{5}\right) + \frac{4\sqrt{d}\delta_*^2\sigma^2}{5} \quad (173)$$

$$= \frac{4}{5} L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{P}_{\sigma\delta_*}) + \frac{4\sqrt{d}\delta_*^2\sigma^2}{5} \quad (174)$$

$$\Rightarrow \frac{3}{40} \sqrt{d} \delta_*^2 \sigma^2 \leq L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{P}_{\sigma\delta_*}) \leq L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{D}(\Sigma, \sigma^2)). \quad (175)$$

D.3 Minimax Optimality

For proving Theorem 5, we shall use the weighing $w_i = \frac{\mathbb{I}\{\lambda_i \leq 1 - e^{-\frac{2\delta_* \sqrt{d}}{\pi}}\}}{N(1 - e^{-\frac{2\delta_* \sqrt{d}}{\pi}})}$ in our upper bound.

From (137), $\forall t \in [0, 1]$ such that $t^2 + \frac{d}{N(t)} \leq c$ for some universal constant c , we have

$$\|\hat{\beta}_S(\mathbf{Z}, w(t)) - \mu\|_{\Sigma}^2 \lesssim \sigma^2 \left(t^2 + \frac{d}{N(t)} \right). \quad (176)$$

Thus, if $\min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$ and let the minimum value be attained at t^* , then we have

$$\|\hat{\beta}_S(\mathbf{Z}, w(t^*)) - \mu\|_{\Sigma}^2 \lesssim \sigma^2 \left(t^{*2} + \frac{d}{N(t^*)} \right) \quad (177)$$

$$\leq \sigma^2 \left(1 - e^{-\frac{2\delta_* \sqrt{d}}{\pi}} \right)^2 + \sigma^2 \frac{d}{N \left(1 - e^{-\frac{2\delta_* \sqrt{d}}{\pi}} \right)} \quad (178)$$

$$\leq \sigma^2 \frac{4d\delta_*^2}{\pi^2} + 64\sigma^2 d\delta_*^2 \quad (179)$$

$$\simeq \sigma^2 d\delta_*^2, \quad (180)$$

where we used (170) in (179). Combining with (175), when $\min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$, we have

$$\sqrt{d}\sigma^2\delta_*^2 \lesssim L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{D}(\Sigma, \sigma^2)) \lesssim \sigma^2 \min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)} \right) \leq d\sigma^2\delta_*^2. \quad (181)$$

Thus, when $\min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)} \right) \leq c$,

$$\frac{\sigma^2}{\sqrt{d}} \min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)} \right) \lesssim L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{D}(\Sigma, \sigma^2)) \lesssim \sigma^2 \min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)} \right). \quad (182)$$

E Weighted Generalization Bound

We refer the readers to [43] for an introduction to VC dimension, covering numbers, and packing numbers.

Define the weighted empirical Rademacher complexity to be $\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i f(X_i) \mid \mathbf{X} \right]$. We first generalize Massart's lemma to weighted Rademacher complexity.

Lemma 1 (Weighted Massart's Lemma). *Assume $|\mathcal{F}|$ is finite and let*

$$B_w = \max_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^n w_i^2 f(X_i)}, \quad (183)$$

then

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq B_w \sqrt{2 \ln |\mathcal{F}|} \quad (184)$$

Proof.

$$e^{s\mathcal{R}_w(\mathcal{F} \circ \mathbf{X})} = e^{s\mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i f(X_i) \mid \mathbf{X}\right]} \quad (185)$$

$$\leq \mathbb{E}\left[e^{s \sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i f(X_i)} \mid \mathbf{X}\right] \quad (\text{conditional Jensen's inequality}) \quad (186)$$

$$\leq \mathbb{E}\left[\sum_{f \in \mathcal{F}} e^{s \sum_i w_i \sigma_i f(X_i)} \mid \mathbf{X}\right] \quad (187)$$

$$= \sum_{f \in \mathcal{F}} \prod_{i=1}^n \mathbb{E}\left[e^{s w_i \sigma_i f(X_i)} \mid \mathbf{X}\right] \quad (188)$$

$$\leq \sum_{f \in \mathcal{F}} \prod_{i=1}^n e^{s^2 w_i^2 f(X_i)^2 / 2} \quad (\text{by Hoeffding's lemma}) \quad (189)$$

$$\leq |\mathcal{F}| e^{s^2 B_w^2 / 2}. \quad (190)$$

Thus,

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq \frac{\ln |\mathcal{F}|}{s} + \frac{s B_w^2}{2} \quad \forall s > 0. \quad (191)$$

Setting $s = \frac{\sqrt{2 \ln |\mathcal{F}|}}{B_w}$, we get the claimed upper bound. $\square$

Lemma 2 (Weighted Rademacher Complexity Bound). $\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq 31 \|w\|_2 \sqrt{\text{VC}(\mathcal{F})} \forall \mathbf{X}$, where $\text{VC}(\mathcal{F})$ is the Vapnik–Chervonenkis dimension of $\mathcal{F}$.

Proof. Much of this proof follows the analysis of Liao [44], Rebeschini [45], adapted to the weighted version of our problem.

In the context of this proof, for the realized $\mathbf{X}$ and for $f \in \mathcal{F}$, define

$$\|f\|_w = \frac{\sqrt{\sum_{i=1}^n w_i^2 f(X_i)^2}}{\|w\|_2}. \quad (192)$$

Let $\max_{f \in \mathcal{F}} \|f\|_w \leq c$. For family of functions that we consider, $c = 1$.

Let $\mathcal{F}_j \subseteq \mathcal{F}$ be the minimal $\epsilon_j = \frac{c}{2^j}$ cover of $\mathcal{F}$ with respect to $\|\cdot\|_w$, i.e., $\mathcal{F}_j$ is the set of least cardinality such that for any $f \in \mathcal{F}$, $\exists f' \in \mathcal{F}_j$ such that $\|f - f'\|_w \leq \epsilon_j$. Let $|\mathcal{F}_j| = C(\mathcal{F}, \epsilon_j, \|\cdot\|_w)$.

For any $f \in \mathcal{F}$, let $f_j(f) \in \mathcal{F}_j$ be such that $\|f - f_j(f)\|_w \leq \epsilon_j$. Denoting $\mathbb{E}[\cdot \mid \mathbf{X}]$ as $\mathbb{E}_\sigma[\cdot]$, we have for any $m \geq 1$,

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i (f(X_i) - f_m(f)(X_i)) + \sum_{j=1}^m (f_j(f)(X_i) - f_{j-1}(f)(X_i)) \right] \quad (193)$$

$$\leq \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i (f(X_i) - f_m(f)(X_i)) \right] + \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i \sum_{j=1}^m f_j(f)(X_i) - f_{j-1}(f)(X_i) \right]. \quad (194)$$

Notice that

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i (f(X_i) - f_m(f)(X_i)) \right] \leq \sup_{f \in \mathcal{F}} \sum_i w_i |f(X_i) - f_m(f)(X_i)| \quad (195)$$

$$\leq \sup_{f \in \mathcal{F}} \sqrt{n} \|w\|_2 \|f - f_m(f)\|_w \quad (196)$$

$$\leq \sqrt{n} \|w\|_2 \epsilon_m. \quad (197)$$

Now, for the second term, by Lemma 1, we have

$$\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i \sum_{j=1}^m f_j(f)(X_i) - f_{j-1}(f)(X_i) \leq \sum_{j=1}^m \sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i (f_j(f)(X_i) - f_{j-1}(f)(X_i)). \quad (198)$$

Thus, using Lemma 1,

$$\mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i \sum_{j=1}^m f_j(f)(X_i) - f_{j-1}(f)(X_i) \right] \leq \quad (199)$$

$$\sum_{j=1}^m \|w\|_2 \sup_{f \in \mathcal{F}} \|f_j(f) - f_{j-1}(f)\|_w \sqrt{2 \ln |\{f_j(f) - f_{j-1}(f) | f \in \mathcal{F}\}|}. \quad (200)$$

Note that $|\{f_j(f) - f_{j-1}(f) | f \in \mathcal{F}\}| \leq |\mathcal{F}_j| |\mathcal{F}_{j-1}| \leq 2 \ln |\mathcal{F}_j|$. Further, by triangle inequality,

$$\|f_j(f) - f_{j-1}(f)\|_w \leq \|f_j(f) - f\|_w + \|f - f_{j-1}(f)\|_w \quad (201)$$

$$\leq \epsilon_j + \epsilon_{j-1} \quad (202)$$

$$= 3\epsilon_j \quad (203)$$

$$= 6(\epsilon_j - \epsilon_{j+1}). \quad (204)$$

Thus,

$$\mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i \sum_{j=1}^m f_j(f)(X_i) - f_{j-1}(f)(X_i) \right] \leq 12 \|w\|_2 \sum_{j=1}^m (\epsilon_j - \epsilon_{j+1}) \sqrt{\ln |\mathcal{F}_j|}. \quad (205)$$

Combining the above bounds,

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq \|w\|_2 \left\{ 2\sqrt{n}\epsilon_{m+1} + 12 \sum_{j=1}^m (\epsilon_j - \epsilon_{j+1}) \sqrt{\ln C(\mathcal{F}, \epsilon_j, \|\cdot\|_w)} \right\} \quad (206)$$

$$\leq \|w\|_2 \left\{ 2\sqrt{n} \frac{c}{2^{m+1}} + 12 \int_{c/2^{m+1}}^{c/2} \sqrt{\ln C(\mathcal{F}, x, \|\cdot\|_w)} dx \right\}. \quad (207)$$

For any $\epsilon \in (0, \frac{c}{2}]$, $\exists m \in \mathbb{Z}_{\geq 1}$ such that $\frac{c}{2^{m+1}} \leq \epsilon \leq \frac{c}{2^m}$. Thus,

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq \|w\|_2 \left\{ 2\sqrt{n}\epsilon + 12 \int_{\epsilon/2}^{c/2} \sqrt{\ln C(\mathcal{F}, x, \|\cdot\|_w)} dx \right\}. \quad (208)$$

Taking $\epsilon \rightarrow 0$, obtain

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq 12 \|w\|_2 \int_0^{c/2} \sqrt{\ln C(\mathcal{F}, x, \|\cdot\|_w)} dx. \quad (209)$$

Setting $c = 1$ and using Lemma 3 obtain

$$\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) \leq 12 \sqrt{\text{VC}(\mathcal{F})} \|w\|_2 \int_0^{c/2} \sqrt{\frac{10}{x^2} \ln \frac{2e}{x^2}} dx \quad (210)$$

$$\leq 12 \sqrt{\text{VC}(\mathcal{F})} \|w\|_2 \int_0^{c/2} \sqrt{\ln \frac{20}{x^4}} dx \quad \text{using } \ln x \leq x/e \forall x \geq e \quad (211)$$

$$\leq 31 \sqrt{\text{VC}(\mathcal{F})} \|w\|_2. \quad (212)$$

□

Lemma 3. $C(\mathcal{F}, x, \|\cdot\|_w) \leq \left[\frac{10}{x^2} \ln \frac{2e}{x^2} \right]^{\text{VC}(\mathcal{F})}$

The readers can refer to the proof of Lemma 3 presented in [43, 45] which generalizes to our modified distance $\|\cdot\|_w$.

Proposition 4. Let $\mathcal{F}$ be a family of $0 - 1$ valued functions and let $X_1, \dots, X_n \stackrel{iid}{\sim} P$. For a fixed vector $w \in \mathbb{R}^d$, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i (f(X_i) - \mathbb{E}[f(X)]) \right\} \leq 62 \|w\|_2 \sqrt{\text{VC}(\mathcal{F})} + \|w\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}}, \quad (213)$$

Proof. By McDiarmid's inequality, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i (f(X_i) - \mathbb{E}[f(X)]) \right\} \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i (f(X_i) - \mathbb{E}[f(X)]) \right\} \right] + \|w\|_2 \sqrt{\frac{\log \frac{1}{\delta}}{2}}. \quad (214)$$

Let $X'_1, \dots, X'_n \stackrel{iid}{\sim} P$ then

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i (f(X_i) - \mathbb{E}[f(X)]) \right\} \right] = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} \left[\sum_i w_i (f(X_i) - f(X'_i)) \mid \mathbf{X} \right] \right\} \right] \quad (215)$$

$$\leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i (f(X_i) - f(X'_i)) \right\} \right] \quad (216)$$

by Jensen's inequality. Let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher random variables then

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i (f(X_i) - f(X'_i)) \right\} \right] = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left\{ \sum_i w_i \sigma_i (f(X_i) - f(X'_i)) \right\} \right] \quad (217)$$

$$\leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i f(X_i) + \sup_{f \in \mathcal{F}} \sum_i w_i (-\sigma_i) f(X'_i) \right] \quad (218)$$

$$= 2 \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i f(X_i) \right] \quad (219)$$

$$= 2 \mathbb{E} [\mathcal{R}_w(\mathcal{F} \circ \mathbf{X})]. \quad (220)$$

The result of Lemma 2 completes the proof. $\square$

F LeCam-Assouad Interpolation Lower Bound

In this section, we provide an improved lower bound over those in Appendix C.2 and Appendix D.2, while using similar ideas.

Fix a value of $t \in [d]$ and $m = \frac{d}{t}$ such that they are whole numbers. Roughly speaking, we shall parametrize the mean of the d -dimensional Gaussian distributions in the hypotheses by vectors in $\{-\delta, \delta\}^t$ with repeating each element m times. Choosing $t = 1$ leads to Le Cam's method while $t = d$ leads to Assouad's method.

For a vector $v \in \mathbb{R}^t$, let $e_t(v) \in \mathbb{R}^d$ with

$$e_t(v)_i = v_{\lceil i/t \rceil} \quad \forall i \in [d], \quad (221)$$

where $\lceil \cdot \rceil$ denotes the ceiling function. For convenience, we drop the subscript and denote $e_t(v)$ by $e(v)$. Define the parameterized distribution $P_\delta(\tau) = \mathcal{N}(\delta e(\tau), I)$ for $\delta > 0$ to be specified later, and let $\mathcal{P}_\delta = \{P_\delta(\tau) \mid \tau \in \{-1, 1\}^t\}$.

Note: The supremum in our minimax definition is over both the true distribution and the adversarial strategy. We shall specify the adversarial strategy later for our lower bound argument.

Note that $L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^N) \geq L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{P}_\delta) \forall \delta$ and thus, we shall lower bound $L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{P}_\delta)$.

We begin by lower bounding

$$L_{\mathbb{E}}(\boldsymbol{\lambda}, \mathcal{P}_\delta) = \inf_M \sup_{P \in \mathcal{P}_\delta} \mathbb{E}_{\mathbf{Z} \sim \lambda P} [\|M(\mathbf{Z}) - \mu_P\|_2^2]. \quad (222)$$

For a vector τ , let τ'^j denote the vector such that $\tau_i = \tau'_i \forall i \neq j$ and $\tau_j = -\tau'_j$. Using Assouad's lower bound technique [28],

$$\inf_M \sup_{P \in \mathcal{P}_\delta} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} [\|\mathbf{M}(\mathbf{Z}) - \mu_P\|_2^2] \geq \inf_M \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \mathbb{E} [\|\mathbf{M}(\mathbf{Z}) - \delta e(\tau)\|_2^2] \quad (223)$$

$$= \inf_M \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \sum_{j=1}^d \mathbb{E} [|M(\mathbf{Z})_j - \delta e(\tau)_j|^2] \quad (224)$$

$$= \inf_M \sum_{k=1}^m \sum_{i=1}^t \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau_i|^2] \quad (225)$$

$$\geq \sum_{k=1}^m \inf_M \sum_{i=1}^t \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau_i|^2] \quad (226)$$

Now, note that $\forall i, k$

$$\sum_{\tau \in \{-1,1\}^t} \mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau_i|^2] = \sum_{\tau \in \{-1,1\}^t} \frac{\mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau_i|^2] + \mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau'_i|^2]}{2} \quad (227)$$

Thus, we get

$$L_{\mathbb{E}}(\lambda, \mathcal{P}_\delta) \geq \sum_{k=1}^m \inf_M \sum_{i=1}^t \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \frac{\mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau_i|^2] + \mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau'_i|^2]}{2} \quad (228)$$

$$\geq \sum_{k=1}^m \sum_{i=1}^t \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \inf_M \frac{\mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau_i|^2] + \mathbb{E} [|M(\mathbf{Z})_{(i-1)t+k} - \delta \tau'_i|^2]}{2} \quad (229)$$

$$\geq \sum_{k=1}^m \sum_{i=1}^t \frac{1}{2^t} \sum_{\tau \in \{-1,1\}^t} \delta^2 (1 - \text{TV}(P_{\mathbf{Z} \sim_\lambda P_\delta(\tau)}, P_{\mathbf{Z} \sim_\lambda P_\delta(\tau'_i)})), \quad (230)$$

where the last line follows by Le Cam's method and the notation $P_{\mathbf{Z} \sim_\lambda P_\delta(\tau)}$ refers to the distribution of $\mathbf{Z}$ when the true distribution is $P_\delta(\tau)$.

Adversarial strategy motivation: Consider a particular sample with corruption rate λ and let the underlying true distribution be $P_\delta(\tau)$. Denote the perturbed sample as $Z(\tau)$ and the outlier to be $\tilde{X}(\tau)$. Note that $Z(\tau) \sim (1 - \lambda)P_{X(\tau)} + \lambda P_{\tilde{X}(\tau)}$. For any particular clean sample $X(\tau) \sim P_\delta(\tau)$, the adversary's goal is to ensure the sample $Z(\tau)$ contains no information about τ . One possible way is that the adversary can try to ensure $Z(\tau) \sim \frac{\max_{\tau' \neq \tau} P(\tau')}{T}$, where the maximum is taken pointwise over the pdf and T is a normalizing constant. Observe the identity

$$P(\tau) + \left(\max_{\tau' \neq \tau} P(\tau') - P(\tau) \right)_+ = \max_{\tau'} P(\tau'), \quad (231)$$

where $(\cdot)_+ := \max\{\cdot, 0\}$ is pointwise over the pdf. First, note that

$$\int_{x \in \mathbb{R}^d} \left(\max_{\tau' \neq \tau} P_\delta(\tau')(x) - P_\delta(\tau)(x) \right)_+ dx = T - 1,$$

where $P_\delta(\tau)(x)$ is understood to be the pdf of $P_\delta(\tau)$ at x . Thus, $\frac{(\max_{\tau' \neq \tau} P(\tau') - P(\tau))_+}{T-1}$ is a valid pdf, and for $\lambda = \frac{T-1}{T}$, we have $(1 - \lambda)P(\tau) + \lambda \frac{(\max_{\tau' \neq \tau} P(\tau') - P(\tau))_+}{T-1} = \frac{\max_{\tau'} P(\tau')}{T}$. Thus, for

any $\lambda \geq 1 - \frac{1}{T}$, there exists a way for the adversary to make the distribution of $Z(\tau) \sim \frac{\max_{\tau'} P(\tau')}{T}$, rendering the sample useless for identifying τ .

Now, we find an upper bound on T in terms of δ . Let $\mathbf{1}$ denote a vector of size m with all elements equal to 1.

$$T = \int_{x \in \mathbb{R}^d} \max_{\tau \in \{-1, 1\}^t} \frac{1}{\sqrt{(2\pi)^d}} e^{-\frac{\|x - \delta e(\tau)\|_2^2}{2}} dx \quad (232)$$

$$= \int_{x \in \mathbb{R}^d} \frac{1}{\sqrt{(2\pi)^d}} \max_{\tau \in \{-1, 1\}^t} e^{-\frac{\|x - \delta e(\tau)\|_2^2}{2}} dx \quad (233)$$

$$= \left[\int_{x' \in \mathbb{R}^m} \frac{1}{\sqrt{(2\pi)^m}} \max_{\tau \in \{-1, 1\}} e^{-\frac{\|x' - \delta \tau \mathbf{1}\|_2^2}{2}} dx' \right]^t. \quad (234)$$

Now note that

$$\int_{x' \in \mathbb{R}^m} \frac{1}{\sqrt{(2\pi)^m}} \max_{\tau \in \{-1, 1\}} e^{-\frac{\|x' - \delta \tau \mathbf{1}\|_2^2}{2}} dx' = 1 + \text{TV}(\mathcal{N}(-\delta \mathbf{1}, I), \mathcal{N}(\delta \mathbf{1}, I)), \quad (235)$$

by total variation definition. Using Pinsker's inequality and KL divergence between multivariate Gaussians, we obtain

$$\text{TV}(\mathcal{N}(-\delta \mathbf{1}, I), \mathcal{N}(\delta \mathbf{1}, I)) \leq \sqrt{\frac{\text{KL}(\mathcal{N}(-\delta \mathbf{1}, I), \mathcal{N}(\delta \mathbf{1}, I))}{2}} \quad (236)$$

$$= \delta \sqrt{m} \quad (237)$$

Combining (234), (235) and (237), and using $m = \frac{d}{t}$

$$T \leq (1 + \delta \sqrt{m})^t \quad (238)$$

$$\leq e^{\delta \sqrt{td}}. \quad (239)$$

Thus, if $\lambda \geq 1 - e^{-\delta \sqrt{td}}$ then the adversary can ensure that $Z(\tau) \sim \frac{\max_{\tau'} P(\tau')}{T}$ and contains no information about τ .

Adversarial strategy: If for sample i , $\lambda_i \geq 1 - e^{-\delta \sqrt{td}}$, then independently sample $\tilde{X}(\tau) \sim \beta \left(\frac{\max_{\tau' \neq \tau} P_\delta(\tau') - P_\delta(\tau)}{T-1} \right)_+ + (1 - \beta) \frac{\max_{\tau'} P_\delta(\tau')}{T}$, where $\beta = \frac{(T-1)(1-\lambda)}{\lambda}$. The constraint $\lambda_i \geq 1 - e^{-\delta \sqrt{td}}$ ensures $\beta \in [0, 1]$ and the above is a valid distribution.

Thus, $Z(\tau) \sim \frac{\max_{\tau'} P_\delta(\tau')}{T}$. If $\lambda_i < 1 - e^{-\delta \sqrt{td}}$, then adversary does no corruption, or equivalently, independently sample $\tilde{X}(\tau) \sim P_\delta(\tau)$ so that $Z(\tau) \sim P_\delta(\tau)$.

Thus, using Pinsker's inequality in (230), obtain

$$L_{\mathbb{E}}(\lambda, \mathcal{P}_\delta) \geq \sum_{k=1}^m \sum_{i=1}^t \frac{1}{2t} \sum_{\tau \in \{-1, 1\}^t} \delta^2 \left(1 - \sqrt{\frac{1}{2} \text{KL}(P_{\mathbf{Z} \sim \lambda} P_\delta(\tau), P_{\mathbf{Z} \sim \lambda} P_\delta(\tau^{i_i}))} \right) \quad (240)$$

$$= \sum_{k=1}^m \sum_{i=1}^t \frac{1}{2t} \sum_{\tau \in \{-1, 1\}^t} \delta^2 \left(1 - \sqrt{\frac{1}{2} \text{KL}(\otimes_{j=1}^n P_{Z_j(\tau)}, \otimes_{j=1}^n P_{Z_i(\tau^{i_i})})} \right) \quad (241)$$

$$= \sum_{k=1}^m \sum_{i=1}^t \frac{1}{2t} \sum_{\tau \in \{-1, 1\}^t} \delta^2 \left(1 - \sqrt{\frac{1}{2} \sum_{j: \lambda_j < 1 - e^{-\delta \sqrt{dt}}} \text{KL}(P_\delta(\tau), P_\delta(\tau^{i_i}))} \right) \quad (242)$$

$$= d\delta^2 \left(1 - \sqrt{\frac{d\delta^2}{t} n(1 - e^{-\delta \sqrt{dt}})} \right) \forall \delta \quad (243)$$

Let δ_* be such that

$$n(1 - e^{-\delta_* \sqrt{dt}}) \leq \frac{t}{64d\delta_*^2}, \text{ and } N(1 - e^{-\delta_* \sqrt{dt}}) \geq \frac{t}{64d\delta_*^2}. \quad (244)$$

Substituting δ_* in (243),

$$\frac{7}{8}d\delta_*^2 \leq \inf_M \sup_{P \in \mathcal{P}_{\delta_*}} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} [\|M(\mathbf{Z}) - \mu_P\|_2^2]. \quad (245)$$

For a random variable K , let $Q(K, \alpha) := \inf\{t : \Pr[K \geq t] \leq (1 - \alpha)\}$, i.e., $Q(\cdot, \alpha)$ is the α -quantile of the random variable. Note that for a random variable K , $\mathbb{E}K \leq Q(K, x)x + (1 - x)\text{ess-sup}K$ $\forall x \in [0, 1]$, where ess-sup denotes essential supremum. Further, note that in the minimax term $\inf_M \sup_{P \in \mathcal{P}_{\delta_*}}$, we can restrict ourselves to estimators M which output in $[-\delta_*, \delta_*]^d$ almost surely – otherwise the error can be reduced by projected to this region. Thus $\text{ess-sup}\|M(\mathbf{Z}) - \mu_P\|_2^2 \leq 4d\delta_*^2$ for any estimator M and any distribution P . Thus, combining the above observations, we have

$$\inf_M \sup_{P \in \mathcal{P}_{\delta_*}} \mathbb{E}_{\mathbf{Z} \sim_\lambda P} \|M(\mathbf{Z}) - \mu_P\|_2^2 \leq \inf_M \sup_{P \in \mathcal{P}_{\delta_*}} \frac{4}{5}Q\left(\|M(\mathbf{Z}) - \mu_P\|_2^2, \frac{4}{5}\right) + \frac{4d\delta_*^2}{5}. \quad (246)$$

Using (245), we get

$$\frac{7}{8}d\delta_*^2 \leq \frac{4}{5} \inf_M \sup_{P \in \mathcal{P}_{\delta_*}} Q\left(\|M(\mathbf{Z}) - \mu_P\|_2^2, \frac{4}{5}\right) + \frac{4d\delta_*^2}{5} \quad (247)$$

$$= \frac{4}{5}L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{P}_{\delta_*}) + \frac{4d\delta_*^2}{5} \quad (248)$$

$$\Rightarrow \frac{3}{40}d\delta_*^2 \leq L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{P}_{\delta_*}) \leq L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^{\mathcal{N}}). \quad (249)$$

We remark that the lower bound presented here appears to be different from that of Appendix C.2. However, the exact difference is hard to comment on since the term δ_* is implicitly defined in the two bounds. We do however remark that we can reconstruct the same $\sqrt{d}$ gap in upper and lower bound using this method as demonstrated below.

F.1 On Minimax Optimality

In the upper bound, use the weighing $w_i = \frac{\mathbb{I}\{\lambda_i \leq 1 - e^{-\delta_* \sqrt{dt}}\}}{N(1 - e^{-\delta_* \sqrt{dt}})}$. From (74), $\forall v \in [0, 1]$ such that $v^2 + \frac{d}{N(v)} \leq c$ for some universal constant c , we have

$$\|\hat{\mu}_S(\mathbf{Z}, w(v)) - \mu\|_2^2 \lesssim \left(v^2 + \frac{d}{N(v)}\right). \quad (250)$$

Thus, if $\min_{v \in [0, 1]} \left(v^2 + \frac{d}{N(v)}\right) \leq c$ and let the minimum value be attained at v^* , then we have

$$\|\hat{\mu}_S(\mathbf{Z}, w(v^*)) - \mu\|_2^2 \lesssim \left(v^{*2} + \frac{d}{N(v^*)}\right) \quad (251)$$

$$\leq \left(1 - e^{-\delta_* \sqrt{dt}}\right)^2 + \frac{d}{N(1 - e^{-\delta_* \sqrt{dt}})} \quad (252)$$

$$\leq dt\delta_*^2 + \frac{64d^2\delta_*^2}{t} \quad (253)$$

$$\simeq d^{1.5}\delta_*^2 \text{ (setting } t = \sqrt{d}\text{)} \quad (254)$$

where we used (244) in (253). Combining with (249), when $\min_{v \in [0, 1]} \left(v^2 + \frac{d}{N(v)}\right) \leq c$, we have

$$d\delta_*^2 \lesssim L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^{\mathcal{N}}) \lesssim \min_{v \in [0, 1]} \left(v^2 + \frac{d}{N(v)}\right) \leq d^{1.5}\delta_*^2. \quad (255)$$

Thus, when $\min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)}\right) \leq c$,

$$\frac{1}{\sqrt{d}} \min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)}\right) \lesssim L_{\text{PAC}}(\boldsymbol{\lambda}, \mathcal{D}_d^{\mathcal{N}}) \lesssim \min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)}\right). \quad (256)$$

F.2 Benefits of this Lower Bound

The lower bound presented in Appendix C.2 does not match the standard lower bound when all the corruptions rates are equal, i.e., the homogeneous robust estimation problem. When $\lambda = \lambda \mathbf{1}$, from (95) in Appendix C.2, we obtain

$$L_{\mathbb{E}}(\lambda \mathbf{1}, \mathcal{P}_{\delta}) \geq \sqrt{d} \delta^2 \left(1 - \sqrt{n \delta^2 \mathbb{I} \left\{ \sqrt{d} \delta < \log \left(\frac{1}{1-\lambda} \right) \right\}} \right). \quad (257)$$

Setting $\delta_* = \max \left\{ \frac{1}{\sqrt{2n}}, \frac{1}{\sqrt{d}} \log \left(\frac{1}{1-\lambda} \right) \right\}$, obtain the lower bound

$$L_{\mathbb{E}}(\lambda \mathbf{1}, \mathcal{P}_{\delta_*}) \gtrsim \frac{\sqrt{d}}{n} + \frac{\lambda^2}{\sqrt{d}}, \quad (258)$$

and thus,

$$L_{\text{PAC}}(\lambda \mathbf{1}, \mathcal{D}_d^{\mathcal{N}}) \gtrsim \frac{\sqrt{d}}{n} + \frac{\lambda^2}{\sqrt{d}}. \quad (259)$$

Using the lower bound presented of (243), we can fix this issue. Take $t = 1$ and set $\delta_* = \max \left\{ \frac{1}{\sqrt{2dn}}, \frac{1}{\sqrt{d}} \log \left(\frac{1}{1-\lambda} \right) \right\}$ to get the lower bound

$$L_{\mathbb{E}}(\lambda \mathbf{1}, \mathcal{P}_{\delta_*}) \gtrsim \frac{1}{n} + \lambda^2. \quad (260)$$

Now take $t = d$ and set $\delta_* = \max \left\{ \frac{1}{\sqrt{2n}}, \frac{1}{d} \log \left(\frac{1}{1-\lambda} \right) \right\}$ to get the lower bound

$$L_{\mathbb{E}}(\lambda \mathbf{1}, \mathcal{P}_{\delta_*}) \gtrsim \frac{d}{n} + \frac{\lambda^2}{d}. \quad (261)$$

Combining the two cases, obtain the bound

$$L_{\mathbb{E}}(\lambda \mathbf{1}, \mathcal{P}_{\delta_*}) \gtrsim \frac{d}{n} + \lambda^2, \quad (262)$$

and thus, we get the right lower bound of

$$L_{\text{PAC}}(\lambda \mathbf{1}, \mathcal{D}_d^{\mathcal{N}}) \gtrsim \frac{d}{n} + \lambda^2. \quad (263)$$

Similar steps can be used for Gaussian linear regression.