



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

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

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

814 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
 815 can use the crude variance bound $E\|(1 - B_i)X_i - (1 - \lambda_i)\mu_P\|_2^2 \leq r^2$ to obtain

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

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

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

(32)

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

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

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

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

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

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

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

819 Combining the above, obtain

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

820 B.2 Upper Bound Solution

821 To solve

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

822 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
 823 condition on the Lagrangian leads to

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

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

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

825 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
 826 since this is a strictly convex objective with a convex compact constraint set we are guaranteed a
 827 *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

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

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

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

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

845 C Mean Estimation for Gaussian Distributions: Proofs

846 C.1 Upper Bound

847 Recall

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

848

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

849 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
850 lower bounded as

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

851 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
852 $E[f(X)] = \frac{1}{2} \forall f \in \mathcal{F}_\mu$. With some abuse of notation, let $w(G) = \{w_i | i \in G\}$. By Proposition 4,
853 we have with probability at least $1 - \frac{\delta}{4}$

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

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

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

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

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

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

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

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

858 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
859 $\|\eta - \mu\|_2 \geq r = \Phi^{-1}(\frac{1}{2} + \alpha)$, let $v_\eta = \frac{\eta - \mu}{\|\eta - \mu\|_2}$. We shall set the value of $\alpha > 0$ later.

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

860 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
861 $\mathbb{E}[\mathbb{I}\{(X - \eta)^T v_\eta \geq 0\}] = \Phi(-\|\eta - \mu\|_2)$, we have $E[f(X)] \leq \frac{1}{2} - \alpha \forall f \in \mathcal{G}_\mu$.

862 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
863 $\text{VC}(\{f_\eta(x) = \mathbb{I}\{(x - \eta)^T v_\eta \geq 0\} | \eta \in \mathbb{R}^d\}) = \text{VC}(\{f_\eta(x) = \mathbb{I}\{(x - \eta)^T \eta \geq 0\} | \eta \in \mathbb{R}^d\})$.

864 Observe that $\text{VC}(\{f_\eta(x) = \mathbb{I}\{(x - \eta)^T \eta \geq 0\} | \eta \in \mathbb{R}^d\}) \leq \text{VC}(\{f_{\eta,v}(x) = \mathbb{I}\{(x - \eta)^T v \geq$
865 $0\} | \eta, v \in \mathbb{R}^d\}) = d + 1$. Thus, $\text{VC}(\mathcal{G}_\mu) \leq d + 1 \leq 2d$.

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

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

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

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

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

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

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

869 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
870 point η such that $\|\mu - \eta\|_2 \geq \Phi^{-1}(\frac{1}{2} + \alpha)$ can be returned by the Tukey median estimator. Thus,
871 with probability at least $1 - \delta$, we have

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

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

872 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

$$873 \alpha = \frac{3}{2}w^T \lambda + \|w\|_2 \left(150\sqrt{d} + 3.5\sqrt{\frac{\log 4/\delta}{2}} \right) < \frac{1}{3}.$$

874 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.

875 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
876 have the guarantee that with probability at least $\frac{4}{5}$,

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

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

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

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

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

878 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
879 have

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

880 Correspondingly, the threshold based estimator satisfies the following $\forall t \in [0, 1]$ such that $t^2 +$
881 $\frac{d}{N(t)} \leq \frac{1}{432450}$, we have

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

882 C.2 Lower Bound

883 We provide a lower bound for Gaussian mean estimation in \mathbb{R}^d in this section. For a vector $v \in \mathbb{R}^{\sqrt{d}}$,
 884 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
 885 0. Define the distribution parameterized distribution $P_\delta(\tau) = \mathcal{N}(\delta e(\tau), I)$ for $\delta > 0$ to be specified
 886 later, and let $\mathcal{P}_\delta = \{P_\delta(\tau) | \tau \in \{-1, 1\}^{\sqrt{d}}\}$.

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

889 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 $L_{\text{PAC}}(\lambda, \mathcal{P}_\delta)$.

890 We begin by lower bounding

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

891 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
 892 lower bound technique [28],

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

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

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

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

893 Now, note that

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

894 Thus, we get

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

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

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

895 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
 896 of \mathbf{Z} when the true distribution is $P_\delta(\tau)$. We shall now specify a particular adversarial strategy to
 897 lower bound the above.

898 **Adversarial strategy motivation:** Consider a particular sample with corruption rate λ and let the
 899 underlying true distribution be $P_\delta(\tau)$. Denote the perturbed sample as $Z(\tau)$ and the outlier to be
 900 $\tilde{X}(\tau)$. Note that $Z(\tau) \sim (1 - \lambda)P_{X(\tau)} + \lambda P_{\tilde{X}(\tau)}$. For any particular clean sample $X(\tau) \sim P_\delta(\tau)$,
 901 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 $Z(\tau) \sim P_\delta(\tau)$.

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

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

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

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

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

920 Let δ_* be such that

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

921 Substituting δ_* in (95),

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

922 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
 923 of the random variable. Note that for a random variable K , $\mathbb{E}K \leq Q(K, x)x + (1-x)\text{ess-sup}K$
 924 $\forall x \in [0, 1]$, where ess-sup denotes essential supremum. Further, note that in the minimax term
 925 $\inf_M \sup_{P \in \mathcal{P}_{\delta_*}}$, we can restrict ourselves to estimators M which output in $[-\delta_*, \delta_*]^{\sqrt{d}} \times \{0\}^{d-\sqrt{d}}$
 926 almost surely – otherwise the error can be reduced by projected to this region. Thus $\text{ess-sup}\|M(\mathbf{Z}) -$
 927 $\mu_P\|_2^2 \leq 4\sqrt{d}\delta_*^2$ for any estimator M and any distribution P . Thus, combining the above observations,
 928 we have

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

929 Using (97), we get

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

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

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

930 C.3 Minimax Optimality

931 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
 932 (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]

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

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

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

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

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

950 Next, we show that depth of any point far away from the coefficient is low. For any $\eta \in \mathbb{R}^d$ such that
 951 $\|\eta - \mu\|_\Sigma \geq r$, let $v_\eta = \frac{\eta - \mu}{\|\eta - \mu\|_2}$. We shall set the value of $r > 0$ later.

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

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

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

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

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

955 Now, note that

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

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

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

957 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 $T_1, T_2 \sim_{iid}$
 958 $\mathcal{N}(0, 1)$,

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

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

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

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

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

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

959 Thus,

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

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

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

961 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

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

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

991 D.2 Lower Bound

992 The lower bound argument we present is similar to Appendix [C.2](#).

993 We begin by noting that when true regression coefficient is β then $(W, Y) \sim$
 994 $\mathcal{N}\left(0, \begin{bmatrix} \Sigma & \Sigma\beta \\ \beta^T \Sigma & \beta^T \Sigma \beta + \sigma^2 \end{bmatrix}\right).$

995 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 $d - \sqrt{d}$
 996 elements equal to 0. Let $\beta(\tau) = \delta \Sigma^{-\frac{1}{2}} e(\tau)$ for $\delta > 0$ to be specified later. For $\tau \in \{-1, 1\}^{\sqrt{d}}$,
 997 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

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

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

999 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 $L_{\text{reg}}(\boldsymbol{\lambda}, \mathcal{P}_\delta)$.

1000 We begin by lower bounding

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

1001 For a vector τ , let τ'^j denote the vector such that $\tau_i = \tau'_i{}^j \forall i \neq j$ and $\tau_j = -\tau'_j{}^j$. For an estimation
 1002 $M(\mathbf{Z})$, define $N(\mathbf{Z}) = \Sigma^{-\frac{1}{2}} M(\mathbf{Z})$ – note that this is a bijective map. Using Assouad’s lower bound
 1003 technique [\[28\]](#),

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

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

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

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

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

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

1004 Now, note that

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

1005 Thus, we get

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

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

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

1006 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
 1007 of \mathbf{Z} when the true distribution is $P_{\delta}(\tau)$. We shall now specify a particular adversarial strategy to
 1008 lower bound the above.

1009 **Adversarial strategy motivation:** Readers can refer to the motivation in Appendix C.2 for more
 1010 details on the main idea behind the strategy described. We need to find an upper bound on the
 1011 normalizing constant T for the pdf $\frac{\max_{\tau} P_{\delta}(\tau)}{T}$. Let

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

1012 Thus,

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

1013 By Schur's formula, $|S(\tau)| = |\Sigma| \sigma^2$ and by block inversion formula,

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

1014 Performing change of variable $(w, y) = G(x, y)$ where $G = \begin{bmatrix} \Sigma^{\frac{1}{2}} & 0 \\ 0 & 1 \end{bmatrix}$, we obtain

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

1015 Noting that

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

1016 we get

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

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

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

1017 Let $X \sim \mathcal{N}(0, I)$ be a random variable in $\mathbb{R}^{\sqrt{d}}$. We can write

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

1018 Note that $\max_{\tau \in \{-1,1\}^{\sqrt{d}}} \delta \tau^T X = \delta \|X\|_1$. Thus,

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

1019 Using the above upper bound, we obtain

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

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

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

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

1020 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
1021 information about τ .

1022 **Adversarial strategy:** If for sample i , $\lambda_i \geq 1 - e^{-\frac{2\delta \sqrt{d}}{\pi\sigma}}$, then independently sample $\tilde{X}(\tau) \sim$
1023 $\beta \left(\frac{\max_{\tau' \neq \tau} P_\delta(\tau') - P_\delta(\tau)}{T-1} \right)_+ + (1 - \beta) \frac{\max_{\tau'} P_\delta(\tau')}{T}$, where $\beta = \frac{(T-1)(1-\lambda)}{\lambda}$. The constraint $\lambda_i \geq$
1024 $1 - e^{-\frac{2\delta \sqrt{d}}{\pi\sigma}}$ ensures $\beta \in [0, 1]$ and the above is a valid distribution.

1025 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,
1026 independently sample $\tilde{X}(\tau) \sim P_\delta(\tau)$ so that $Z(\tau) \sim P_\delta(\tau)$.

1027 Thus, using Pinsker's inequality in (150), obtain

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

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

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

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

1028 where (164) follows since

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

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

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

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

1029 Replacing $\delta \leftarrow \delta/\sigma$ in (164), we get

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

1030 Let δ_* be such that

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

1031 Substituting δ_* in (169),

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

1032 Note that in the minimax term $\inf_M \sup_{P \in \mathcal{P}_{\sigma\delta_*}}$, we can restrict ourselves to estimators M which
 1033 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
 1034 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
 1035 and any distribution P . Thus, combining the above observations, we have

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

1036 Using (171), we get

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

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

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

1037 D.3 Minimax Optimality

1038 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.

1039 From (137), $\forall t \in [0, 1]$ such that $t^2 + \frac{d}{N(t)} \leq c$ for some universal constant c , we have

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

1040 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

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

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

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

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

1041 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

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

1042 Thus, when $\min_{t \in [0, 1]} \left(t^2 + \frac{d}{N(t)}\right) \leq c$,

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

1043 E Weighted Generalization Bound

1044 We refer the readers to [43] for an introduction to VC dimension, covering numbers, and packing
1045 numbers.

1046 Define the weighted empirical Rademacher complexity to be $\mathcal{R}_w(\mathcal{F} \circ \mathbf{X}) =$
1047 $\mathbb{E} [\sup_{f \in \mathcal{F}} \sum_i w_i \sigma_i f(X_i) | \mathbf{X}]$. We first generalize Massart's lemma to weighted Rademacher
1048 complexity.

1049 **Lemma 1** (Weighted Massart's Lemma). *Assume $|\mathcal{F}|$ is finite and let*

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

1050 *then*

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

Proof.

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

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

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

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

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

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

1051 Thus,

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

1052 Setting $s = \frac{\sqrt{2 \ln |\mathcal{F}|}}{B_w}$, we get the claimed upper bound. \square

1053 **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
1054 $\text{VC}(\mathcal{F})$ is the Vapnik–Chervonenkis dimension of \mathcal{F} .

1055 *Proof.* Much of this proof follows the analysis of Liao [44], Rebeschini [45], adapted to the weighted
1056 version of our problem.

1057 In the context of this proof, for the realized \mathbf{X} and for $f \in \mathcal{F}$, define

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

1058 Let $\max_{f \in \mathcal{F}} \|f\|_w \leq c$. For family of functions that we consider, $c = 1$.

1059 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
1060 cardinality such that for any $f \in \mathcal{F}$, $\exists f' \in \mathcal{F}_j$ such that $\|f - f'\|_w \leq \epsilon_j$. Let $|\mathcal{F}_j| = C(\mathcal{F}, \epsilon_j, \|\cdot\|_w)$.

1061 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 $\mathbb{E}[\cdot | \mathbf{X}]$ as $\mathbb{E}_\sigma[\cdot]$, we have
 1062 for any $m \geq 1$,

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

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

1063 Notice that

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

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

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

1064 Now, for the second term, by Lemma [1](#) we have

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

1065 Thus, using Lemma [1](#)

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

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

1066 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,

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

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

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

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

1067 Thus,

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

1068 Combining the above bounds,

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

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

1069 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,

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

1070 Taking $\epsilon \rightarrow 0$, obtain

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

1071 Setting $c = 1$ and using Lemma 3 obtain

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

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

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

1072

□

1073 **Lemma 3.** $C(\mathcal{F}, x, \|\cdot\|_w) \leq \left[\frac{10}{x^2} \ln \frac{2e}{x^2}\right]^{\text{VC}(\mathcal{F})}$

1074 The readers can refer to the proof of Lemma 3 presented in [43, 45] which generalizes to our modified
1075 distance $\|\cdot\|_w$.

1076 **Proposition 4.** Let \mathcal{F} be a family of 0 – 1 valued functions and let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$. For a fixed
1077 vector $w \in \mathbb{R}^d$, with probability at least $1 - \delta$,

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

1078 *Proof.* By McDiarmid's inequality, with probability at least $1 - \delta$,

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

1079 Let $X'_1, \dots, X'_n \stackrel{\text{iid}}{\sim} P$ then

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

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

1080 by Jensen's inequality. Let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher random variables then

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

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

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

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

1081 The result of Lemma 2 completes the proof. □