A Proofs

A.1 Proof of Lemma 1

Proof. Let $P = \mathcal{N}(0, \mathcal{I}_p)$ be the isotropic normal distribution. Let $R_P(\theta) = \mathbb{E}_{z\sim P}[\ell(\|z-\theta\|_2)]$, where $\ell : \mathbb{R} \mapsto \mathbb{R}$ is a convex loss, and let $\theta(P) = \operatorname{argmin}_\theta R_P(\theta)$ be the minimizer of the population risk. We assume that $\psi(\cdot) = \ell'(\cdot) < C$ is bounded. Note that when the derivative is unbounded, it is easy to argue that the corresponding risk will be non-robust. We also assumed that this risk is fisher-consistent for the Gaussian-distribution, *i.e.* $\theta(P) = 0$. For notational convenience, let $u(t) = \frac{\psi(t)}{t}$. Then,

$$\nabla R_P(\theta) = -\mathbb{E}_{z\sim P}\left[\underbrace{\frac{\psi(\|z-\theta\|_2)}{\|z-\theta\|_2}}_{u(\|z-\theta\|_2)}(z-\theta)\right].$$

As before, let $P_\epsilon = (1-\epsilon)P + \epsilon Q$. Then, we are interested in studying $\widehat{\theta}(P_\epsilon)$. To do this, by first order optimality, we know that $\theta(P_\epsilon)$ is a solution to the following equation:

$$(1-\epsilon)\nabla R_P(\theta(P_\epsilon)) + \epsilon \nabla R_Q(\theta(P_\epsilon)) = 0$$

First we calculate the derivative of $\theta(P_\epsilon)$ w.r.t. ϵ using the fixed point above. Taking derivative of the above equation w.r.t. ϵ

$$(1-\epsilon)\nabla^2 R_P(\theta(P_\epsilon))\dot{\theta}(P_\epsilon) - \nabla R_P(\theta(P_\epsilon)) + \epsilon\nabla^2 R_Q(\theta(P_\epsilon))\dot{\theta}(P_\epsilon) + \nabla R_Q(\theta(P_\epsilon)) = 0 \tag{8}$$

Under our assumption that ψ is continuous, we get that at $\epsilon = 0$,

$$\dot{\theta}(P_\epsilon)_{|\epsilon=0} = (-\nabla^2 R_P(\theta(P)))^{-1}\nabla R_Q(\theta(P)) \tag{9}$$

By fisher consistency of ℓ for $\mathcal{N}(0, \mathcal{I}_p)$, we have that $\theta(P) = 0$. Suppose that Q is a point mass distribution with all mass on θ_Q. Then, we have that,

$$\nabla R_Q(0) = -u(\|\theta_Q\|_2)\theta_Q$$

Our next step is to lower bound the operator norm of $-\nabla^2 R_P(\theta(P))$. To do this we show that for any unit vector $v \in \mathcal{S}^{p-1}$, $v^T(-\nabla^2 R_P(\theta(P)))v \leq \frac{C_2}{\sqrt{p}}$.

$$\nabla^2 R_P(\theta) = -\mathbb{E}_{z\sim P}\left[u(\|z-\theta\|_2)\mathcal{I}_p + \frac{u'(\|z-\theta\|_2)}{\|z-\theta\|_2}((z-\theta)(z-\theta)^T)\right]$$

Now, by definition $u(t) = \psi(t)/t$, so $u'(s) = (\psi'(s) - u(s))/s$. Plugging this above,

$$\nabla^2 R_P(\theta) = -E_{z\sim P}\left[u(\|z-\theta\|_2)(\mathcal{I}_p - \frac{(z-\theta)(z-\theta)^T)}{\|z-\theta\|_2^2}) + \frac{\psi'(\|z-\theta\|_2)}{\|z-\theta\|_2^2}(z-\theta)(z-\theta)^T))\right]$$

Hence, we get that

$$v^T\nabla^2 R_P(0)v = -\mathbb{E}_{z\sim N(0,I_p)}\left[u(\|z\|_2)(\|v\|_2^2 - (v^T(z/\|z\|_2))^2) + \psi'(\|z\|_2)(v^T(z/\|z\|_2))^2\right]$$

Further for Isotropic Gaussian, $\|z\|_2$ and $z/\|z\|_2$ are independent random variables. Also, since, $z/\|z\|_2$ is uniformly distributed on unit sphere, we get that $\mathbb{E}_{z\sim N(0,I)}[(v^T z/\|z\|_2)^2)] = \|v\|_2^2/p$.

$$(v^T(-\nabla^2 R_P(0))v) = \underbrace{\mathbb{E}_{z\sim N(0,I_p)}\left[u(\|z\|_2)\right](1-1/p)}_{\textbf{T1}} + \underbrace{\mathbb{E}_{z\sim N(0,I_p)}\left[\psi'(\|z\|_2)\right]/p}_{\textbf{T2}}$$

- **Controlling T1**

$$\begin{aligned}
\mathbb{E}_{z\sim N(0,I_p)}[u(\|z\|_2)] &= \mathbb{E}_{z\sim\mathcal{N}(0,I_p)}\left[\frac{\psi(\|z\|_2)}{\|z\|_2}\right] \\
&\leq \sqrt{C\mathbb{E}\frac{1}{\|z\|_2^2}} \\
&\leq \frac{\sqrt{C_1}}{\sqrt{p-2}}, \qquad (10)
\end{aligned}$$

  where we use that ψ is bounded by constant C. The last inequality is combination of Jensen's Inequality and plugging the mean of reciprocal of inverse chi-squared random variable (Bernardo and Smith, 2009).

- **Controlling T2.** Under our assumption that $\psi'(\cdot)$ exists and is bounded, we get that $T2 \leq \frac{C_1}{p}$ and can be ignored.

Hence, for large p, we get that $(v^T(-\nabla^2 R_P(0))v) \leq \sqrt{C_1/p}$. Now, if we put θ_Q at ∞, and use that $\psi(\infty) = C_1$, we get that,

$$\|\dot{\theta}(P_\epsilon)\|_2 = \psi(\|\theta_Q\|_2)\|\nabla^2 R_P(0)\frac{\theta_Q}{\|\theta_Q\|_2}\|_2 \geq C_2\sqrt{p}$$

$\square$

A.2 Proof of Lemma 2

Proof. Let $P = N(0, \mathcal{I}_p)$. Every subset of size $(1-\epsilon)n$ can be thought of as samples from a mixture distribution defined in (3), where the mixture proportion η, ranges from $[0, \epsilon/(1-\epsilon)]$. In the asymptotic setting of $n \mapsto \infty$, the empirical squared loss over each subset corresponds to the population risk with the sampling distribution as P_η. For a given contamination distribution Q, let $R_{P_\eta}(\theta) = \mathbb{E}_{x \sim P_\eta}\left[\|x - \theta\|_2^2\right]$ and let $\theta(P_\eta) \stackrel{\text{def}}{=} \text{argmin}_\theta R_{P_\eta}(\theta)$, then subset risk minimization returns,

$$\begin{aligned} &\widehat{\theta}_{\text{SRM}} = \theta(P_{\eta^*}) \\ &\text{where } \eta^* = \underset{\eta \in [0, \frac{\epsilon}{1-\epsilon}]}{\text{argmin}} R_{P_\eta}(\theta(P_\eta)) \end{aligned} \tag{11}$$

We are interested in bounding the bias of SRM *i.e.*

$$\sup_Q \|\widehat{\theta}_{\text{SRM}} - \theta^*\|_2$$

To do this, we know that for any contamination distribution Q, the solution of SRM necessarily satisfies the following conditions.

Condition 1: Local Stationarity. $\theta(P_\eta) = \text{argmin}_\theta R_{P_\eta}(\theta)$ is the minimizer of the risk with respect to a mixture distribution iff

$$\begin{aligned} \nabla R_{P_\eta}(\theta(P_\eta)) = (1-\eta)&\nabla R_{P_{\theta^*}}(\theta(P_\eta)) \\ &+ \eta \nabla R_Q(\theta(P_\eta)) = 0. \end{aligned} \tag{12}$$

Condition 2: Global Fit Optimality. $\widehat{\theta}_{\text{SRM}} = \theta(P_{\eta^*})$ is the global minimizer of the population risk over all mixture distributions iff

$$\begin{aligned} R_{P_{\eta^*}}(\theta(P_{\eta^*})) &= (1-\eta^*)R_{P_0}(\theta(P_{\eta^*})) + \eta^* R_Q(\theta(P_{\eta^*})) \\ &\leq R_{P_\eta}(\theta(P_\eta)) \quad \forall \eta \in \left[0, \frac{\epsilon}{1-\epsilon}\right] \end{aligned} \tag{13}$$

Using Conditions 1 and 2, we next derive the bias of SRM for mean estimation.

We make a few simple observations.

- **Observation 1.** For any distribution P, we have,

$$R_P(\theta) = \text{trace}\left(\Sigma(P)\right) + \|\theta - \mu(P)\|_2^2$$

- **Observation 2.** Condition 1 reduces to,

$$\mu(P_\eta) = \theta_\eta = (1-\eta)\mu(P) + \eta\mu(Q),$$

  where $\mu(\cdot)$ is the Expectation functional.

Lemma 9. *Under the mixture model in Equation (3), for the squared error, we have that,*

$$R_{P_\eta}(\theta_\eta) = \textit{trace}\left(\Sigma(P_\eta)\right) = (1-\eta)\textit{trace}\left(\Sigma(P^*)\right) + \eta\textit{trace}\left(\Sigma(Q)\right) + \eta(1-\eta)\|\mu(P^*) - \mu(Q)\|_2^2.$$

Now, from Lemma 9, we know that

$$R_{P_\eta}(\theta_\eta) = (1-\eta)\text{trace}\left(\Sigma(P)\right) + \eta\text{trace}\left(\Sigma(Q)\right) + \eta(1-\eta)\|\mu(P) - \mu(Q)\|_2^2$$

As a function of η, $R_{P_\eta}(\theta_\eta)$ is a concave quadratic function. Hence, it is always minimized at the end points of the interval $[0, \epsilon/(1-\epsilon)]$, which implies that $\eta^* \in \{0, \frac{\epsilon}{1-\epsilon}\}$.

Hence, we have that,

$$\widehat{\theta}_{\text{SRM}} = \begin{cases} \theta_{\frac{\epsilon}{1-\epsilon}}, & \text{if } R_{P_{\frac{\epsilon}{1-\epsilon}}}(\theta_{\frac{\epsilon}{1-\epsilon}}) \leq R_{P_0}(\theta_0). \\ \theta^*, & \text{otherwise.} \end{cases}$$

From Lemma 9, $R_{P_{\frac{\epsilon}{1-\epsilon}}}(\theta_{\frac{\epsilon}{1-\epsilon}}) \leq R_{P_0}(\theta_0)$ iff

$$\left(1 - \frac{\epsilon}{1-\epsilon}\right) \|\mu(P) - \mu(Q)\|_2^2 \leq \mathrm{trace}\left(\Sigma(P)\right) - \mathrm{trace}\left(\Sigma(Q)\right)$$

Moreover, from Observation 2, we have that,

$$\|\theta_{\frac{\epsilon}{1-\epsilon}} - \mu(P)\|_2 = \frac{\epsilon}{1-\epsilon}\|\mu(P) - \mu(Q)\|_2$$

Combining the above two, we get that,

$$\begin{aligned}\|\widehat{\theta}_{\mathrm{SRM}} - \mu(P)\|_2 = \left[\frac{\epsilon}{1-\epsilon}\|\mu(P) - \mu(Q)\|_2\right] .\mathbf{1}\Big\{\|\mu(P) - \mu(Q)\|_2^2 \leq \\ \left(\frac{1-\epsilon}{1-2\epsilon}\right)\left(\mathrm{trace}\left(\Sigma(P)\right) - \mathrm{trace}\left(\Sigma(Q)\right)\right)\Big\}.\end{aligned} \tag{14}$$

Equation 6 follows from it.

□

A.2.1 Proof of Lemma 9

Proof. We give two alternate proofs of the Lemma.

- Proof 1: This proceeds by expanding on the definition of risk.

$$\begin{aligned}R_{P_\eta}(\theta_\eta) &= E_{z\sim P_\eta}[\|z - \theta_\eta\|_2^2] \\ &= (1-\eta)E_{z\sim P_0}[\|z - \theta_\eta\|_2^2] + \eta E_{z\sim Q}[\|z - \theta_\eta\|_2^2] \quad \text{Expectation by conditioning.} \\ &= (1-\eta)\left[\mathrm{trace}\left(\Sigma(P^*)\right) + \|\theta_\eta - \mu(P^*)\|_2^2\right] \\ &+ \eta\left[\mathrm{trace}\left(\Sigma(Q)\right) + \|\theta_\eta - \mu(Q)\|_2^2\right] \quad \text{From Observation 1.}\end{aligned}$$

  Now, using Observation 2 we get that,

$$\|\theta_\eta - \mu(Q)\|_2 = (1-\eta)\|\mu(P^*) - \mu(Q)\|_2$$

$$\|\theta_\eta - \mu(P^*)\|_2 = \eta\|\mu(P^*) - \mu(Q)\|_2$$

  Plugging this into above, we get,

$$R_{P_\eta}(\theta_\eta) = (1-\eta)\mathrm{trace}\left(\Sigma(P^*)\right) + \eta\mathrm{trace}\left(\Sigma(Q)\right) + \|\mu(P^*) - \mu(Q)\|_2^2\left(\eta^2(1-\eta) + (1-\eta)^2\eta\right)$$

  which recovers the statement of the Lemma.

- Proof 2: This proceeds by Law of Total Variance, or the Law of Total Cummulants. We know that $R_{P_\eta} = \mathrm{trace}\left(\Sigma(P_\eta)\right)$. Let $Z \sim P_\eta$, and let $Y \sim \mathrm{Bernoulli}(1-\eta)$ be the indicator if the sample is from the true distribution. Then $Z|Y = 1 \sim P^*$, while $Z|Y = 0 \sim Q$.

$$\mathrm{trace}\left(\Sigma(P_\eta)\right) = \underbrace{(1-\eta)\mathrm{trace}\left(\Sigma(P^*)\right) + \eta\mathrm{trace}\left(\Sigma(Q)\right)}_{\mathrm{Var}(E[Z|Y])} + \underbrace{\eta(1-\eta)\|\mu(P^*) - \mu(Q)\|_2^2}_{E[\mathrm{Var}(Z|Y)]}.$$

□

A.3 Proof of Lemma 3

Proof. Let $P_\epsilon = (1-\epsilon)P^* + \epsilon Q$. Let I^* be the interval $\mu \pm \frac{\sigma}{\delta_1^{\frac{1}{2k}}}$, where $\mu = \mathbb{E}_{x\sim P^*}[x]$. Moreover for notational convenience, let $f_n(u,v) = \sqrt{u(1-u)}\sqrt{\frac{\log(2/v)}{n}} + \frac{2}{3}\frac{\log(2/v)}{n}$. Let $\hat{I} = [a,b]$ be the interval obtained using $\mathcal{Z}_1$, *i.e.* the shortest interval containing $n(1-(\delta_1+\epsilon+f_n(\epsilon+\delta_1,\delta_3)))$ points of $\mathcal{Z}_1$. Note that in the algorithm, we have $\delta_1 = \epsilon$, and $\delta_3 = \delta/4$. As a first step, we bound the length of $\hat{I}$ and show that $\hat{I}$ and I^* must necessarily intersect.

Claim 1. *Let $\hat{I}$ be the shortest interval containing $1-\delta_4$ fraction of points, where $\delta_4 = (\delta_1+\epsilon) + f_n(\epsilon+\delta_1,\delta_3)$. Further assume that $\delta_4 < \frac{1}{2}$. Then with probability at least $1-\delta_3$,*

$$\mathit{length}(\hat{I}) \le \mathit{length}(I^*) \le \frac{2\sigma}{\delta_1^{\frac{1}{2k}}},$$

Moreover, if $\delta_4 < \frac{1}{2}$, then $\hat{I}\cap I^* \ne \phi$, which implies

$$|z-\mu| \le \frac{4\sigma}{\delta_1^{\frac{1}{2k}}} \forall z \in \hat{I}$$

Proof. We first show that with probability at least $1-\delta_3$, I^* contains at least $n(1-\delta_4)$ points(Claim 5). Hence, since our algorithm chooses the shortest interval($\hat{I}$) containing $1-\delta_4$ fraction of points, length of $\hat{I}$ is less than length of I^*. Next, if δ_4 is less than $\frac{1}{2}$, then there are two intervals $\hat{I}$ and I^* respectively, which contain at least $n/2$ points. Hence, they must necessarily intersect. □

Next, we control the final error of our estimator. Let $|\hat{I}| = \sum_{z\in\mathcal{Z}_2} \mathbb{I}\left\{z_i \in \hat{I}\right\}$ be the number of points which lie in $\hat{I}$. Similarly, let $|\hat{I}_Q|$ and $|\hat{I}_{P^*}|$ number of points which lie in $\hat{I}$, which are distributed according to Q and P^* respectively.

$$\left|\frac{1}{|\hat{I}|}\sum_{x_i\in\hat{I}} x_i - \mu\right| \le \underbrace{\left|\frac{1}{|\hat{I}|}\sum_{\substack{x_i\in\hat{I}\\ x_i\sim Q}}(x_i-\mu)\right|}_{T1} + \underbrace{\left|\frac{1}{|\hat{I}|}\sum_{\substack{x_i\in\hat{I}\\ x_i\sim P^*}}(x_i-\mu)\right|}_{T2} \tag{15}$$

Control of T1. To control T1, we can write it as:

$$\begin{aligned} T1 &= \left|\frac{1}{|\hat{I}|}\sum_{\substack{x_i\in\hat{I}\\ x_i\sim Q}}(x_i-\mu)\right| \\ &\le \underbrace{\frac{|\hat{I}_Q|}{|\hat{I}|}}_{T1a}\underbrace{\max_{\substack{x_i\in\hat{I}\\ x_i\sim Q}}|x_i-\mu|}_{T1b} \end{aligned} \tag{16}$$

where $\hat{I}_Q$ is the number of points in $\hat{I}$ distributed according to Q. To control T1a, we use Bernsteins inequality. To control T1b, we use Claim 1. The claim below formally controls T1.

Claim 2. *Let $\hat{I}$ be the shortest interval containing $n(1-\delta_4)$ of the points, where $\delta_4 = (\delta_1+\epsilon) + f_n(\epsilon+\delta_1,\delta_3)$. Further assume that $\delta_4 < \frac{1}{2}$. Then, with probability at least $1-\delta_3-\delta_5$, we have that,*

$$T1 \le \frac{|\hat{I}_Q|}{|\hat{I}|}\max_{\substack{x_i\in\hat{I}\\ x_i\sim Q}}|x_i-\mu| \le \frac{\epsilon+f_n(\epsilon,\delta_5)}{1-\delta_4}\frac{4\sigma}{\delta_1^{1/2k}} \tag{17}$$

Proof. Using Bernstein's bound, we know that wp at least $1-\delta_5$,

$$|\hat{I}_Q| \le n(\epsilon + \sqrt{\epsilon(1-\epsilon)}\sqrt{\frac{\log(1/\delta_5)}{n}} + \frac{2}{3}\frac{\log(1/\delta_5)}{n}),$$

This follows from the fact that number of points drawn from Q which lie in $\hat{I}$ is less than the total number of points drawn according to Q. In Claim 1, we showed that when $\delta_4 < \frac{1}{2}$, then, with probability at least $1-\delta_3$, we get that $\hat{I} \cap I^* \neq \phi$, *i.e.* the intervals intersect, and that $length(\hat{I}) < length(I^*)$. Hence, we get,

$$\max_{\substack{x_i \in \hat{I} \\ x_i \sim Q}} |x_i - \mu| \le \frac{4\sigma}{\delta_1^{1/2k}}$$

□

Control of T2. To control T2, we write it as

$$T2 = \left| \frac{|\hat{I}_{P^*}|}{|\hat{I}|} \left[\frac{1}{|\hat{I}_{P^*}|} \sum_{\substack{x_i \in \hat{I} \\ x_i \sim P^*}} (x_i - \mu) \right] \right| \tag{18}$$

$$\le \frac{|\hat{I}_{P^*}|}{|\hat{I}|} \underbrace{\left| (\frac{1}{|\hat{I}_{P^*}|} \sum_{\substack{x_i \in \hat{I} \\ x_i \sim P^*}} x_i) - E[x|x \in \hat{I}, x \sim P^*] \right|}_{T2a} + \frac{|\hat{I}_{P^*}|}{|\hat{I}|} \underbrace{\left| E[x|x \in \hat{I}, x \sim P^*] - \mu \right|}_{T2b} \tag{19}$$

- **Control of T2a:** To bound the distance between the mean of the points from P^* within $\widehat{I}$ and $E[x|x \sim P^*, x \in \hat{I}]$, we will use Bernsteins bound(Lemma 10) for bounded random variables. We know that the random variables are in a bounded interval $b = length(\widehat{I}) \le \frac{\sigma}{\delta^{\frac{1}{2k}}}$, and that conditional variance of the random variables, when conditioned on them lying in $\hat{I}$ is controlled using Lemma 13. In particular, Lemma 13 shows that for any event E, which occurs with probability $P(E) \ge \frac{1}{2}$,

  $$E_{x \sim P^*}[(x - E[x|x \in E])^2 | x \in E] \le \sigma^2 / P(E).$$

  Using these arguments, we get that with probability at least $1-\delta_7$,

  $$T2a \le \sqrt{\frac{2\sigma^2(\log(3/\delta_7))}{P^*(\hat{I})|\hat{I}_{P^*}|}} + \frac{2\sigma}{\delta_1^{1/2k}} \frac{\log(3/\delta_7)}{|\hat{I}_{P^*}|}, \tag{20}$$

  where $P^*(\hat{I})$ is the probability that a random variable drawn according to P^* lies in $\hat{I}$.

- **Control of T2b:** To control $T2b$, we use the general mean shift lemma (Lemma 12), which controls how far the mean can move when conditioned on an event. We get that,

  $$T2b \le 2\sigma(P^*(\hat{I})^c)^{1-1/(2k)} \tag{21}$$

Combining the bounds in (20) and (21), we get

$$T2 \le 2\sigma(P^*(\hat{I})^c)^{1-1/(2k)} + \sqrt{\frac{2\sigma^2(\log(3/\delta_7))}{P^*(\hat{I})|\hat{I}_{P^*}|}} + \frac{2\sigma}{\delta_1^{1/2k}} \frac{\log(3/\delta_7)}{|\hat{I}_{P^*}|} \tag{22}$$

Combining the upper bound on T1 in (17) with (22), we get that with probability at least $1-\delta_3-\delta_5-\delta_6-\delta_7$

$$T1 + T2 \le \frac{\epsilon + f_n(\epsilon, \delta_5)}{1-\delta_4} \frac{4\sigma}{\delta_1^{1/2k}} + 2\sigma(P^*(\hat{I})^c)^{1-1/(2k)} + \sqrt{\frac{2\sigma^2(\log(3/\delta_7))}{P^*(\hat{I})|\hat{I}_{P^*}|}} + \frac{2\sigma}{\delta_1^{1/2k}} \frac{\log(3/\delta_7)}{|\hat{I}_{P^*}|}$$

We rearrange terms and use our assumption that ϵ is small enough that $\hat{I}_{P^*} \geq n/2$. We also plugin the upper bound on $(P^*(\hat{I})^c)^{1-1/(2k)}$ from Claim 3 and set $\delta_1 = \epsilon$, and $\delta_5 = \delta_6 = \delta_3 = \delta_7 = \delta/4$. Hence, we get that with probability at least $1 - \delta$

$$T1 + T2 \leq C_1 \sigma \epsilon^{1-1/2k} + C_2 \sigma (\frac{\log n}{n})^{1-\frac{1}{2k}} + C_3 \sigma \sqrt{\frac{\log(1/\delta)}{n}} + C_4 \sigma \frac{\log(1/\delta)}{n\epsilon^{\frac{1}{2k}}} \tag{23}$$

Since, we ensure that $\epsilon = \max(\epsilon, \frac{\log(1/\delta}{n})$ hence, $\frac{\log(1/\delta)}{n\epsilon^{\frac{1}{2k}}} \leq \epsilon^{1-\frac{1}{2k}}$. Note that our assumption of $\delta_4 < \frac{1}{2}$ boils down to ϵ being small enough such that $2\epsilon + \sqrt{\epsilon \frac{\log(4/\delta)}{n}} + \frac{\log(4/\delta)}{n} < \frac{1}{2}$. Hence, we recover the final statement of the theorem.

□

A.3.1 Auxillary Proofs

Claim 3. *Let $\hat{I}$ be the shorted interval containing $n(1-\delta_4)$ points from $\mathcal{Z}_1$. Let $P^*(\hat{I})$ is the probability that a random variable drawn according to P^* lies in $\hat{I}$. Then, there exists universal constants $C_1, C_2 > 0$ such that wp at least $1 - \delta_6$, we have that*

$$(P^*(\hat{I})^c)^{1-\frac{1}{2k}} \leq C_1 \epsilon^{1-\frac{1}{2k}} + C_2 \delta_1^{1-\frac{1}{2k}} + C_3 (\frac{\log n}{n})^{1-\frac{1}{2k}} + C_4 (\frac{\log(1/\delta_6)}{n})^{1-\frac{1}{2k}} + C_5 (\frac{\log(1/\delta_3)}{n})^{1-\frac{1}{2k}} \tag{24}$$

Proof. Note that $\hat{I}$ is obtained by choosing the shortest interval containing $n(1-\delta_4)$ points from $\mathcal{Z}_1$. We first bound $P_n^*(\hat{I})$, *i.e.* the empirical probability of samples distributed according to P^* which lie in $\hat{I}$. To do this, note that in $\mathcal{Z}_1$, number of points drawn from Q which lie in $\hat{I}$, say $\hat{n}_Q$ is less than the total number of points drawn according to Q. Using Bernstein's bound, we know that wp at least $1 - \delta_6$,

$$|\hat{n}_Q| \leq n(\epsilon + \sqrt{\epsilon(1-\epsilon)}\sqrt{\frac{\log(1/\delta_6)}{n}} + \frac{2}{3}\frac{\log(1/\delta_6)}{n})$$

Let $\hat{n}_{P^*}$ be the number of points in $\mathcal{Z}_1$, which are drawn from P^* and which lie in $\hat{I}$. Since $|\hat{n}_Q| + |\hat{n}_{P^*}| = |\hat{I}| = n(1-\delta_4)$, hence the above implies that with probability at least $1 - \delta_6$,

$$|\hat{n}_{P^*}| \geq n(1-\delta_4) - n(\epsilon + \sqrt{\epsilon(1-\epsilon)}\sqrt{\frac{\log(1/\delta_6)}{n}} + \frac{2}{3}\frac{\log(1/\delta_6)}{n}),$$

Note that $P_n^*(\hat{I}) = \frac{|\hat{n}_{P^*}|}{\sum_i \mathbb{I}\{x_i \sim P^*\}}$. Hence, we get that,

$$\begin{aligned} P_n^*(\hat{I}) &\geq \frac{|\hat{n}_{P^*}|}{n} \\ &\geq 1 - (\epsilon + \delta_4) - f_n(\epsilon, \delta_6) \end{aligned} \tag{25}$$

This implies that,

$$\begin{aligned} P_n^*(\hat{I})^c &\leq (\epsilon + \delta_4) + f_n(\epsilon, \delta_6) \\ &\leq 2\epsilon + \delta_1 + f_n(\epsilon, \delta_6) + f_n(\epsilon + \delta_1, \delta_3) \\ &\leq 4\epsilon + 2\delta_1 + C_1 \frac{\log(1/\delta_6)}{n} + C_2 \frac{\log(1/\delta_3)}{n} \end{aligned} \tag{26}$$

To finally bound the probability of a sample drawn from P^* to lie in $\hat{I}$, we use the relative deviations VC bound(Lemma 11), which gives us,

$$P^*(\hat{I})^c \leq \underbrace{P_n^*(\hat{I})^c}_{A_1} + 4\sqrt{(\frac{P_n^*(\hat{I})^c \log \mathcal{S}[2n]}{n}) + (\frac{P_n^*(\hat{I})^c \log(4/\delta_6)}{n})} + \frac{\log \mathcal{S}[2n]}{n} + \frac{\log(4/\delta_6)}{n} \tag{27}$$

where $\mathcal{S}[2n] = O(n^2)$. Using that $\sqrt{ab} \leq a + b, \forall a, b \geq 0$, we get that,

$$P^*(\hat{I})^c \leq C_1 P_n^*(\hat{I})^c + C_2(\frac{\log \mathcal{S}[2n]}{n} + \frac{\log(4/\delta_6)}{n}) \tag{28}$$

Hence, we get that,

$$(P^*(\hat{I})^c)^{1-\frac{1}{2k}} \leq C_1 \epsilon^{1-\frac{1}{2k}} + C_2 \delta_1^{1-\frac{1}{2k}} + C_3(\frac{\log n}{n})^{1-\frac{1}{2k}} + C_4(\frac{\log(1/\delta_6)}{n})^{1-\frac{1}{2k}} + C_5(\frac{\log(1/\delta_3)}{n})^{1-\frac{1}{2k}} \tag{29}$$

□

Claim 4. *Let $P^*(I^*)$ be the probability that a sample drawn according from P_ϵ is distributed according to P^* and lies in I^*.*

$$P^*(I^*) \geq (1-\epsilon)(1-\delta_1) = 1 - (\epsilon + \delta_1 - \epsilon\delta_1) \geq 1 - \underbrace{(\epsilon + \delta_1)}_{\delta_2} = 1 - \delta_2$$

Proof. For any $x \sim P_\epsilon$, define, $z_i = 1$ if $x \sim P^*$. Now, for any $x \sim P^*$, we know that, by chebyshevs we know that,

$$P(|x-\mu| \geq t) = P((x-\mu)^{2k} \geq t^{2k}) \leq E[(x-\mu)^{2k}]/t^{2k} \leq C_{2k}\sigma^{2k}/t^{2k}$$

Hence, we get that wp at least $1-\delta_1$, $x \in \mu \pm \sigma/(\delta_1)^{1/2k}$ □

The following claim lower bounds the empirical fraction of samples which are distributed according to P^* and lie in I^*, when n samples are drawn from P_ϵ.

Claim 5. *Let $P_n^*(I^*)$ be the empirical fraction of points which are distributed according to P^* and lie in I^*, when n samples are drawn from P_ϵ. Then, with probability at least $1-\delta_3$,*

$$P_n^*(I^*) \geq 1 - \underbrace{(\delta_2 + \sqrt{(\delta_2(1-\delta_2))}\sqrt{\frac{\log(1/\delta_3)}{n}} + \frac{2}{3}\frac{\log(1/\delta_3)}{n})}_{\delta_4 = (\delta_1+\epsilon)+f_n(\epsilon+\delta_1,\delta_3)},$$

Proof. This follows from Bernstein's inequality(Lemma 10). □

Lemma 10. *[Bernsteins bound,] Let $X \sim P^*$ be a scalar random variable such that $|X - E[x]| \leq b$ with variance σ^2. Then, given n samples $\{x_1, x_2, \ldots, x_n\} \sim P^*$, the empirical mean, $\bar{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is such that,*

$$P(|\bar{x_n} - E[x]| > t) \leq 2\exp(\frac{-nt^2}{2\sigma^2 + 2bt/3})$$

which can be equivalently re-written as. With probability at least $1-\delta$,

$$|\bar{x_n} - E[x]| \leq \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}} + \frac{2b\log(1/\delta)}{3n}$$

Lemma 11. *[Relative deviations, (Vapnik and Chervonenkis, 2015)] Let $\mathcal{F}$ be a function class consisting of binary functions f. Then, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}} |P(f) - P_n(f)| \leq 4\sqrt{P_n(f)\frac{\log(S_{\mathcal{F}}(2n)) + \log(4/\delta)}{n}} + C_1\frac{\log(S_{\mathcal{F}}(2n)) + \log(4/\delta)}{n},$$

where $S_{\mathcal{F}}(n) = \sup_{z_1,z_2,\ldots,z_n} |\{(f(z_1), f(z_2), \ldots, f(z_n)) : f \in \mathcal{F}\}|$ is the growth function, i.e. *the maximum number of ways into which n-points can be classified the function class.*

Lemma 12. *[General Mean shift, (Steinhardt, 2018)] Suppose that a distribution P^* has mean μ and variance σ^2 with bounded $2k^{th}$-moments. Then, for any event A which occurs with probability at least $1-\epsilon \geq \frac{1}{2}$,*

$$|\mu - E[x|A]| \leq 2\sigma\epsilon^{1-\frac{1}{2k}}$$

In particular, for just bounded second moments, we get that $|\mu - E[x|A]| \leq 2\sigma\sqrt{\epsilon}$.

Proof. For any event E, Let $\mathbb{I}\{E\}$ denote the indicator variable for E.

$$|E_{x\sim P^*}[x|E]-\mu|=\frac{|E_{x\sim P^*}((x-\mu)\mathbb{I}\{E\})|}{P(E)}\leq\frac{E[|x-\mu|^p]^{\frac{1}{p}}(E[\mathbb{I}\{E\}^q]^{1/q})}{P(E)}, \tag{30}$$

where $p,q>1$ are such that $1/p+1/q=1$. Put $p=2k$, we get,

$$|E_{x\sim P^*}[x|E]-\mu|\leq\frac{\sigma}{(P(E))^{1/2k}}$$

Now, we know that, $|E[X|A]-\mu|=\frac{1-P(A)}{P(A)}|E[X|A^c]-\mu|$. Putting $E=A^c$, we get,

$$|E[X|A]-\mu|\leq\frac{1-P(A)}{P(A)}\frac{\sigma}{(1-P(A))^{1/2k}}\leq 2\sigma\epsilon^{(1-\frac{1}{2k})}.$$

□

Lemma 13. *[Conditional Variance Bound] Suppose that a distribution P^* has mean μ and variance σ^2. Then, for any event A which occurs with probability at least $1-\epsilon$, the variance of the conditional distribution is bounded as:*

$$(E[(x-E[x|A])^2|A])\leq\frac{\sigma^2}{(1-\epsilon)}$$

Proof. Let $\mu_A=E[y|A]$, $d=\mu_A-\mu$. From Lemma 12, we know, $d\leq\sigma 2\sqrt{\epsilon}$. Observe the following,

$$\begin{aligned}
E[(y-\mu_A)^2|A]=E[(y-\mu-d)^2|A]&=E[((y-\mu)^2-2d(y-\mu)+d^2)|A] &(31)\\
&=E[(y-\mu)^2|A]-d^2 &(32)\\
&\leq E[(y-\mu)^2|A] &(33)\\
&\leq\frac{\sigma^2}{1-\epsilon}, &(34)
\end{aligned}$$

□

A.4 Proof of Lemma 5

Proof. For brevity, let $\widehat{\theta}_\delta = \underset{\theta}{\text{argmin}} \sup_{u \in \mathcal{N}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \text{f}(\{u^T x_i\}_{i=1}^n, \epsilon, \frac{\delta}{5^p})|$, where f is our univariate estimator. Let $\theta^* = \mathbb{E}[x]$ be the true mean. Then, we can write the ℓ_2 error in its variational form.

$$\|\widehat{\theta}_\delta - \theta^*\|_2 = \sup_{u \in \mathcal{S}^{p-1}} |u^T(\widehat{\theta}_\delta - \theta^*)| \tag{35}$$

Suppose $\{y_i\}$ is a $\frac{1}{2}$-cover of the net, so there exist a y_j such that $u = y_j + v$, where $\|v\|_2 \le \epsilon$.

$$\begin{aligned}
\|\widehat{\theta}_\delta - \theta^*\|_2 &\le \sup_{u \in \mathcal{S}^{p-1}} |y_j^T(\widehat{\theta}_\delta - \theta^*)| + |v^T(\widehat{\theta}_\delta - \theta^*)| \\
&\le \sup_{y_j \in \mathcal{N}^{\frac{1}{2}}(\mathcal{S}^{p-1})} |y_j^T(\widehat{\theta}_\delta - \theta^*)| + \|v\|_2 \|\widehat{\theta}_\delta - \theta^*\|_2 \\
&\le 2 \sup_{y_j \in \mathcal{N}^{\frac{1}{2}}(\mathcal{S}^{p-1})} |y_j^T(\widehat{\theta}_\delta - \theta^*)|
\end{aligned}$$

$$\begin{aligned}
\|\widehat{\theta}_\delta - \theta^*\|_2 &\le 2 \sup_{u \in \mathcal{N}^{1/2}} |u^T(\widehat{\theta} - \theta^*)| && (36) \\
&\le 2 \left[\sup_{u \in \mathcal{N}^{1/2}} |u^T\widehat{\theta} - f(u^T P_n, \epsilon; \tilde{\delta})| + \sup_{u \in \mathcal{N}^{1/2}} |u^T\theta^* - f(u^T P_n, \epsilon; \tilde{\delta})| \right] && (37) \\
&\le 4 \sup_{u \in \mathcal{N}^{1/2}} |u^T\theta^* - f(u^T P_n, \epsilon; \tilde{\delta})| && (38)
\end{aligned}$$

For a fixed u, the distribution $u^T P$ has mean $u^T\theta^*$, where θ^* is the mean of the multivariate distribution P. Hence, we get that, for a confidence level $\tilde{\delta}$, when the univariate estimator is applied to the projection of the data long u, it returns a real number such that, with probability at least $1 - \tilde{\delta}$

$$|f(u^T P_n; \epsilon; \tilde{\delta}) - u^T\theta^*| \le C_1 \omega_f(\epsilon, u^T P, \tilde{\delta})$$

Taking a union bound over the elements of the cover, and using the fact that $|\mathcal{N}^{1/2}(\mathcal{S}^{p-1})| \le 5^p$ (Wainwright, 2019), we substitute $\tilde{\delta} = \delta/(5^p)$ and recover the statement of the Lemma.

□

A.5 Proof of Lemma 6

Proof. Let $\widehat{\theta}_\delta = \underset{\theta \in \Theta_s}{\text{argmin}} \sup_{u \in \mathcal{N}_{2s}^{1/2}(\mathcal{S}^{p-1})} |u^T\theta - \text{f}(\{u^T x_i\}_{i=1}^n, \epsilon, \frac{\delta}{(6ep/s)^s})|$, where $f(\cdot)$ is our univariate estimator. Observe that since $\widehat{\theta}_\delta$ and the true mean θ^* are both s-sparse. Hence, the error vector $\widehat{\theta} - \theta^*$ is atmost $2s$-sparse. Then, we can write the ℓ_2 error in its variational form,

$$\|\widehat{\theta}_\delta - \theta^*\|_2 = \sup_{u \in \mathcal{S}^{p-1} \cap \mathcal{B}_{2s}} |u^T(\widehat{\theta}_\delta - \theta^*)|, \tag{39}$$

where $\mathcal{S}^{p-1} \cap \mathcal{B}_{2s}$ is the set of unit vectors which are $2s$-sparse. The remaining of the proof follows along the lines of proof of Lemma 5, coupled with the fact that the cardinality of the half-cover of an $2s$-sparse ball, *i.e.* $\left|\mathcal{N}^{\frac{1}{2}}(\mathcal{S}^{p-1})\right| \le (\frac{6ep}{s})^s$ (Vershynin, 2009).

□

A.6 Proof of Lemma 7

Let $\widehat{\Theta}_{\mathrm{f}} = \mathrm{argmin}_{\Theta \in \mathcal{F}} \sup_{u \in \mathcal{N}^{1/4}(\mathcal{S}^{p-1})} |u^T \Theta u - f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p})|$, where f is a univariate estimator, and z_i are the pseudo-samples obtained by $z_i = (x_{i+n/2} - x_i)/\sqrt{2}$. We begin by first using one-step discretization,

$$\begin{aligned}\|\widehat{\Theta}_{\mathrm{f}} - \Sigma(P)\|_2 &= \sup_{u \in \mathcal{S}^{p-1}} |u^T(\widehat{\Theta}_{\mathrm{f}} - \Sigma(P))u| \\ &\leq \frac{1}{1-2\gamma} \sup_{y \in \mathcal{N}^{\gamma}} |y^T(\widehat{\Theta}_{\mathrm{IM}} - \Sigma(P))y|,\end{aligned}$$

where $\mathcal{N}^{\gamma}$ is the γ-cover of the unit sphere. We set $\gamma = 1/4$.

$$\begin{aligned}\|\widehat{\Theta}_{\mathrm{f}} - \Sigma(P)\|_2 &\leq 2 \sup_{u \in \mathcal{N}^{1/4}} |u^T(\widehat{\Theta}_{\mathrm{f}} - \Sigma(P))u| &(40)\\ &\leq 2\left[\sup_{u \in \mathcal{N}^{1/4}} |u^T \widehat{\Theta}_{\mathrm{IM}} u - f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p}))| + \sup_{u \in \mathcal{N}^{1/4}} |u^T \Sigma(P) u - f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p})|\right] &(41)\\ &\leq 4 \sup_{u \in \mathcal{N}^{1/2}} |u^T \Sigma(P) u - f(u^T \mathcal{X}_n, \epsilon; \tilde{\delta})| &(42)\end{aligned}$$

For a fixed u, for the clean samples in z_i, $(u^T z_i)^2$ has mean $u^T \Sigma(P) u$, and variance $C_4(u^T\Sigma(P)u)^2$. Note that the scalar random variables $(u^T z_i)^2$ have bounded k moments, whenever x_i has bounded $2k$-moments. Hence, for a fixed u, we get that with probability at least $1-\delta$,

$$|f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{9^p}) - u^T \Sigma(P) u| \lesssim \omega_f(2\epsilon, u^T P^{\otimes 2}, \tilde{\delta})$$

Taking a union bound over the elements of the cover, and using the fact that $|\mathcal{N}^{1/4}(\mathcal{S}^{p-1})| \leq 9^p$ (Wainwright, 2019), we substitute $\tilde{\delta} = \delta/(9^p)$ and recover the statement of the Lemma.

A.7 Proof of Lemma 8

Let $\widehat{\Theta}_{\mathrm{f,s}} = \mathrm{argmin}_{\Theta \in \mathcal{F}_s} \sup_{u \in \mathcal{N}_{2s}^{1/4}(\mathcal{S}^{p-1})} |u^T \Theta u - f(\{(u^T z_i)^2\}_{i=1}^n, 2\epsilon, \frac{\delta}{(9ep/s)^s})|$, where f is a univariate estimator, and z_i are the pseudo-samples obtained by $z_i = (x_{i+n/2} - x_i)/\sqrt{2}$.

Observe that since $\widehat{\Theta}_{\mathrm{f,s}}$ and the true covariance $\Sigma(P)$ are both in $\mathcal{F}_s$. Hence, the difference matrix $\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P)$ has atmost $2s$ non-zero off diagonal elements. Hence, we can write that $\|\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P)\|_2 = \sup_{u \in \mathcal{B}_{2s} \cap \mathcal{S}^{p-1}} |u^T(\widehat{\Theta}_{\mathrm{IM}}^{(s)} - \Sigma(P))u|$, where $\mathcal{B}_{2s} \cap \mathcal{S}^{p-1}$ is the set of unit vectors which are atmost $2s$-sparse. Using the one-step discretization, we get that,

$$\|\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P)\|_2 \leq 2 \sup_{u \in \mathcal{N}^{1/4}(\mathcal{B}_{2s} \cap \mathcal{S}^{p-1})} |u^T(\widehat{\Theta}_{\mathrm{f,s}} - \Sigma(P))u|$$

The remainder of the proof follows from the proof of Lemma 7 coupled with the fact that the cardinality of the 1/4-cover of an $2s$-sparse ball $\left|\mathcal{N}^{1/4}(\mathcal{S}^{p-1})\right| \leq (\frac{9ep}{s})^s$ (Vershynin, 2009).

A.8 Proof of Corollary 5

Proof. From Corollary 4, we know that the with probability at least $1-\delta$ sparse covariance estimator satisfies,

$$\underbrace{\|\widehat{\Theta}_{\mathrm{IM,s}} - \Sigma(P)\|_2 \lesssim \|\Sigma(P)\|_2 \epsilon^{1-1/k} + \|\Sigma(P)\|_2 \sqrt{\frac{s \log p}{n}} + \|\Sigma(P)\|_2 \sqrt{\frac{\log 1/\delta}{n}}}_{T1}$$

Let $\widehat{\Theta}_{\mathrm{IM,s}} - \Sigma(P) = \Delta$, then, we have that $\|\Delta\|_2 \leq T1$. Using Weyl's Inequality, we know that,

$$|\lambda_{r+1}(\widehat{\Theta}_{\mathrm{IM,s}}) - \lambda_{r+1}(\Sigma(P))| \leq \|\Delta\|_2$$

We know that $\lambda_{r+1}(\Sigma(P)) = 1$. Hence, we have that $\lambda_{r+1}(\widehat{\Theta}_{\text{IM,s}}) \in 1 \pm T1$. We also know that $\lambda_r(\Sigma(P)) = 1 + \Lambda_r$. Hence, we can now lower bound the eigengap, *i.e.*

$$|\lambda_r(\Sigma) - \lambda_{r+1}(\widehat{\Theta}_{\text{IM,s}})| \geq \Lambda_r - T1$$

Under the assumption that $T1 < \frac{1}{2}\Lambda_r$, and using Davis-Kahan Theorem (Davis and Kahan, 1970), we get that,

$$\|VV^T - \hat{V}\hat{V}^T\|_F \leq \frac{\|\Theta_\delta - \Sigma\|_2}{\Lambda_r - T1} \leq C\frac{T1}{\Lambda_r}$$

□

A.9 Proof of Lemma 4

Note that the proof of this follows from Lemma 6 (Altschuler et al., 2018), but we provide it for completeness. Let F be a CDF and let $Q_{L,F}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$ and $Q_{R,F}(p) = \inf\{x \in \mathbb{R} : F(x) > p\}$ be the left and right quantile functions. Let

$$R_F(t) \geq \max\{Q_{R,F}(\frac{1}{2} + t) - m, m - Q_{L,F}(\frac{1}{2} - t)\},$$

where m is the median. Then, given n-samples from the mixture model, let $\hat{m}(\{x_i\}_{i=1}^n)$ be the empirical median. Then, we have that with probability at least $1 - \delta$,

$$|\hat{m} - m| \leq R(\frac{\epsilon}{2(1-\epsilon)} + \sqrt{\frac{2\log(2/\delta)}{n}}).$$

To see this, for each sample x_i define an indicator variable $L_i \in \{0, 1\}$.

$$L_i = \mathbb{I}\left\{x_i \sim Q, or(x_i \sim P \text{ and } x_i \geq Q_{R,F}(\frac{1}{2(1-\epsilon)} + a))\right\},$$

for $a = \frac{\sqrt{\log(2/\delta)}}{(1-\epsilon)\sqrt{n}}$. Note that

$$\Pr(L_i = 1) \leq \epsilon + (1-\epsilon)(1 - (a + \frac{1}{2(1-\epsilon)}))$$
$$\equiv \frac{1}{2} - (1-\epsilon)a$$
$$\hat{m} \geq Q_{R,F}(\frac{1}{2(1-\epsilon)} + a) \implies \sum_i L_i \geq n/2$$

Hence, we have that,

$$\Pr(\hat{m} > Q_{R,F}(\frac{1}{2(1-\epsilon)} + a)) \leq \Pr(\sum_i L_i \geq n/2) \leq \exp(-2n(1-\epsilon)^2 a^2) = \frac{\delta}{2}$$

The other side is also symmetric. Hence, we have that with probability at least $1 - \delta$,

$$|\hat{m} - m| \leq R(\frac{\epsilon}{2(1-\epsilon)} + a),$$

where $a = \frac{1}{(1-\epsilon)}\sqrt{\frac{\log(2/\delta)}{n}}$. Note that under our assumption that $P \in \mathcal{P}_{\text{sym}}^{t_0,\kappa}$, we have that $R(t) \leq \kappa t$ for all $t \leq t_0$. Hence, as long as the contamination level ϵ, and confidence level δ are such that,

$$\frac{\epsilon}{2(1-\epsilon)} + \frac{1}{(1-\epsilon)}\sqrt{\frac{\log(2/\delta)}{n}} \leq t_0,$$

we have that with probability at least $1 - \delta$,

$$|\hat{m} - m| \lesssim \kappa\epsilon + \kappa\sqrt{\frac{\log(2/\delta)}{n}}$$