

## A ORGANIZATION

The Appendices contain additional technical content and are organized as follows. In Appendix B, we present additional results for Filterpd for when the underlying distribution only has bounded  $2^{nd}$  moment and RGD – Filterpd for logistic regression and generalized linear models. In Appendix C, we present details to supplement the empirical evidence in Section 3. In Appendix D, we present supplementary experimental results for mean estimation. In Appendix E, we detail the hyperparameters used for experiments in Section 5.2. Finally, in Appendix F and G we give the proofs for the propositions in Section 2 and Appendix B.

## B SUPPLEMENTARY THEORETICAL RESULTS

### B.1 GUARANTEES FOR Filterpd UNDER BOUNDED $2^{nd}$ MOMENT

**Theorem 3.** Suppose  $\{z_i\}_{i=1}^n \sim P$ , where  $P$  has bounded  $2^{nd}$  moment and  $n$  satisfies the relation in (3). Then Algorithm 1 when instantiated for  $T^* = \lceil C \log(1/\delta) \rceil$  steps returns an estimate  $\hat{\theta}_\delta$  such that, with probability at least  $1 - 4\delta$ ,  $\delta \in (0, 0.25)$ :

$$\|\hat{\theta}_\delta - \mu\|_2 \lesssim \sqrt{\frac{\text{trace}(\Sigma) \log(p/\delta)}{n}}$$

**Remarks:** In the univariate setting, Theorem 3 shows that Filterpd achieves the optimal sub-Gaussian deviation bound. In the multivariate setting, our theoretical upper bounds are weaker than the guarantees of GMOM (Minsker, 2015); they achieve a rate of  $\sqrt{\frac{\text{trace}(\Sigma) \log(1/\delta)}{n}}$  (note the dependence on  $p$ ).

### B.2 OPTIMUM $T$ FOR HEAVY-TAILED LINEAR REGRESSION

In Theorem 2, we stated the rate of convergence of RGD – Filterpd to  $\theta^*$ , which is given by:

$$\|\hat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^* - \theta^0\|_2 + \frac{C_2 \sigma}{1 - \kappa} \left( \sqrt{\frac{\text{trace}(\Sigma_x)}{n/T}} \right) + \frac{C_2 \sigma}{1 - \kappa} \left( \sqrt{\frac{\|\Sigma_x\|_2 \log(T/\delta)}{n/T}} \right)$$

Note that setting  $T \approx \log_{1/\kappa} \left( \sigma \sqrt{\frac{\text{trace}(\Sigma_x)}{n}} + \sqrt{\frac{\|\Sigma_x\|_2 \log(1/\delta)}{n}} \right)$  suggests that upto logarithmic factors for sufficiently large number of samples, we get an error rate of  $\tilde{\mathcal{O}} \left( \sigma \left( \sqrt{\frac{\text{trace}(\Sigma_x)}{n}} + \sqrt{\frac{\|\Sigma_x\|_2 \log(1/\delta)}{n}} \right) \right)$  where  $\tilde{\mathcal{O}}(\cdot)$  hides logarithmic factors. In comparison, error rates in Prasad et al. (2020); Hsu & Sabato (2016) - which have studied this problem as well - scale as  $\tilde{\mathcal{O}} \left( \sqrt{\frac{\text{trace}(\Sigma_x) \log(1/\delta)}{n}} \right)$ . Other previous works in statistics (Fan et al., 2017; Sun et al., 2019) achieve similar rates, but under the additional assumption that the covariates are sub-Gaussian. Recently, Cherapanamjeri et al. (2020) also studied the problem of heavy-tailed linear regression, when the covariates are isotropic and have *certifiably* bounded  $8^{th}$  moments. In this setting, barring logarithmic factors, they achieve the same rate as us, but at a better worst-case sample complexity of  $p^{3/2}$ , whereas we have  $p^3$ . However, the proposed estimator in Cherapanamjeri et al. (2020) is based on a degree 8 sum-of-squares program and is not yet practical and they only focus on the when  $\Sigma_x$  is identity.

### B.3 GUARANTEES FOR RGD – Filterpd FOR GENERALIZED LINEAR MODELS

In this setting, we observe data  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , where each  $(x_i, y_i) \in \mathbb{R}^p \times \mathcal{Y}$ . We suppose that the  $(x, y)$  pairs sampled from the true distribution  $P_{\theta^*}$  are linked via a linear model such that

when conditioned on the covariates  $x$ , the response variable has the distribution:

$$P_{\theta^*}(y|x) \propto \exp\left(\frac{y\langle x, \theta^* \rangle - \Phi(\langle x, \theta^* \rangle)}{c(\sigma)}\right) \quad (7)$$

Here  $c(\sigma)$  is a fixed and known scale parameter and  $\Phi : \mathbb{R} \mapsto \mathbb{R}$  is the link function. We focus on the random design setting where the covariates  $x \in \mathbb{R}^p$ , have mean 0, and covariance  $\Sigma_x$ . We use the negative conditional log-likelihood as our loss function, i.e.

$$\bar{\mathcal{L}}(\theta; (x, y)) = -y\langle x, \theta \rangle + \Phi(\langle x, \theta \rangle) \quad (8)$$

Here we assume that the covariates have bounded 8<sup>th</sup> moment and that  $\Phi'(\cdot)$  is smooth around  $\theta^*$ . Specifically, we assume that there exist universal constants  $L_{\Phi, 2k}$ ,  $B_{2k}$  such that

$$\mathbb{E}_x \left[ |\Phi'(\langle x, \theta \rangle) - \Phi'(\langle x, \theta^* \rangle)|^{2k} \right] \leq L_{\Phi, 2k} \|\theta^* - \theta\|_2^{2k} + B_{\Phi, 2k}, \quad \text{for } k = 1, 2$$

We also assume that  $\mathbb{E}_x[|\Phi^{(t)}(\langle x, \theta^* \rangle)|^k] \leq M_{\Phi, t, k}$  for  $t \in \{1, 2, 4\}$ , where  $\Phi^{(t)}(\cdot)$  is the  $t^{\text{th}}$ -derivative of  $\Phi(\cdot)$ .

**Theorem 4.** Consider the statistical model in (7). Given  $n$  pairs of samples, where  $n$  satisfies the condition in (3), then RGD – Filterpd when initialized at  $\theta^0$  with stepsize  $\eta = 2/(\tau_u + \tau_\ell)$  and confidence  $\delta$  then, it returns iterates  $\{\hat{\theta}^t\}_{t=1}^T$  which with probability at least  $1 - \delta$  satisfy:

$$\|\hat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^* - \theta^0\|_2 + \frac{C_*}{1 - \kappa} \left( \sqrt{\frac{\text{trace}(\Sigma_x)}{n/T}} + \sqrt{\frac{\|\Sigma_x\|_2 \log(1/\delta/T)}{n/T}} \right), \quad (9)$$

where  $C_* = C_2 \left[ B_{\Phi, 4}^{\frac{1}{4}} + c(\sigma)^{1/2} M_{\Phi, 2, 2}^{\frac{1}{4}} + c(\sigma)^{\frac{3}{4}} M_{\Phi, 4, 1}^{\frac{1}{4}} \right]$  for some contraction parameter  $\kappa < 1$ .

#### B.4 GUARANTEES FOR RGD – Filterpd FOR LOGISTIC REGRESSION

In this setting, we observe data  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , where each  $(x_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$ . We suppose that the  $(x, y)$  pairs sampled from the true distribution  $P_{\theta^*}$  are linked via a linear model such that when conditioned on the covariates  $x$ , the response variable has the distribution:

$$P_{\theta^*}(y = 1|X = x) = \sigma(\langle x, \theta^* \rangle) \quad (10)$$

where  $\sigma(z) = 1/(1 + \exp(-z))$ . We seek to minimize the negative log-likelihood, given as:

$$\bar{\mathcal{L}}(\theta; (x, y)) = -\log(\sigma(\langle x, \theta \rangle)) = -y\langle x, \theta \rangle + \log(1 + \exp(\langle x, \theta \rangle)) \quad (11)$$

The Hessian of the population risk is given by

$$\nabla^2 \mathcal{R}(\theta) = \mathbb{E}_x [\sigma(\langle x, \theta \rangle)(1 - \sigma(\langle x, \theta \rangle))xx^T].$$

Note that as  $\theta$  diverges, the minimum eigenvalue of the Hessian approaches 0 and the loss is no longer strongly convex. To prevent this, we take the parameter space  $\Theta$  to be bounded i.e.  $\Theta = \{\theta : \theta \in \mathbb{R}^p, \|\theta\|_2 \leq B\}$  for some finite  $B > 0$ .

**Corollary 5.** Consider the statistical model in (10). Given  $n$  pairs of samples, where  $n$  satisfies the condition in (5), RGD – Filterpd when initialized at  $\theta^0$  with step size  $\eta = 2/(\tau_u + \tau_\ell)$  and confidence parameter  $\delta$  returns iterates  $\{\hat{\theta}^t\}_{t=1}^T$  which with probability at least  $1 - \delta$  satisfy:

$$\|\hat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^* - \theta^0\|_2 + \left( \sqrt{\frac{\text{trace}(\Sigma_x)}{n/T}} + \sqrt{\frac{\|\Sigma_x\|_2 \log(1/\delta/T)}{n/T}} \right) \quad (12)$$

for some contraction parameter  $\kappa < 1$ .

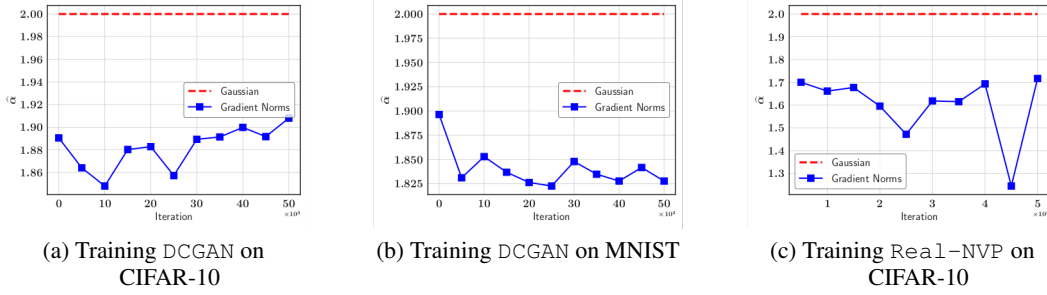


Figure 5: Variation of estimated  $\alpha$ -index across different iterations for different models and datasets. (a) shows the variation of the estimated index for the generator gradient norms of DCGAN when trained on the CIFAR-10 dataset. (b) shows the variation of the estimated index for the generator gradient norms of DCGAN when trained on the MNIST dataset. (c) shows the variation of the estimated index for complete gradient norms over iterations for Real-NVP trained on the CIFAR-10 dataset

### C SUPPLEMENTARY DETAILS REGARDING THE EMPIRICAL STUDY IN SECTION 3

First, we present the variation of the estimated  $\alpha$ -index computed using the  $\alpha$ -index estimator discussed in Section 3. To compute this  $\alpha$ -index, we sample 10000 gradients from the models. Therefore, the estimated  $\alpha$ -index at iteration  $t$  is that computed after sampling 10000 gradients at the end of the  $t^{th}$  iteration. Figure 5 showcases this variation.

While these estimated indices can be used to judge how heavy the tail of a distribution is as compared to a normal distribution, there are some drawbacks of the used estimator which have been highlighted earlier. To validate our the  $\alpha$ -index estimation for heavy-tailedness, we also use the kurtosis ratio to measure heavy-tailedness relative to a normal distribution. Given  $n$  samples  $\{X_i\}_{i=1}^n$ , the estimated kurtosis ratio is given by:

$$\hat{\kappa} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^4}{\left( \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right)^2}$$

where  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . For samples from a normal distribution, this tends to 3 as  $n \rightarrow \infty$ . A quantity greater than 3 could be indicative of heavy-tailedness.

In Figure 6, we plot the variation of  $\hat{\kappa}^{1/4}$  with iterations. We can observe that the trends more or less seem to match; the troughs and crests are attained at the same places.

### D SYNTHETIC EXPERIMENTS FOR MEAN ESTIMATION

**Setup** We generate  $x \in \mathbb{R}^p$  from an isotropic zero-mean heavy-tailed distribution, namely the multivariate Pareto distribution. For a Pareto distribution with tail-parameter  $\beta$ , the  $k^{th}$  order moments exists only if  $k < \beta$ . We fix  $\beta = 3$  in our experiments. In this setup, we experiment with different  $n, p$  and  $\delta$ . For each setting of  $(n, p, \delta)$ , cumulative metrics are reported over 2000 trials. We vary  $n$  from 100 to 500,  $p$  from 20 to 100 and  $\delta$  from 0.01 to 0.1.

**Methods** We compare Filterpd with two baselines: sample mean and GMOM (Minsker, 2015).

**Metric and Hyperparameter Tuning** For any estimator  $\hat{\theta}_{n,\delta}$ , we use  $\ell(\hat{\theta}_{n,\delta}) = \|\hat{\theta} - \mu(P)\|_2$  as our primary metric. We also measure the quantile error of the estimator, *i.e.*  $Q_\delta(\hat{\theta}_{n,\delta}) = \inf\{\alpha : \Pr(\ell(\hat{\theta}_{n,\delta}) > \alpha) \leq \delta\}$ . This can also be thought of as the length of confidence interval for a

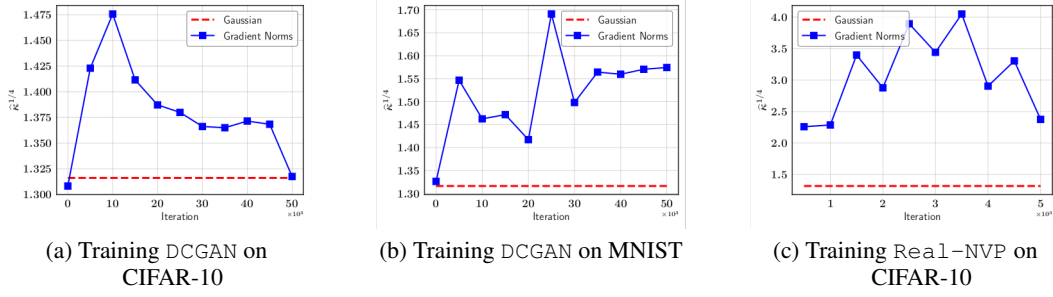


Figure 6: Variation of estimated kurtosis ratio across different iterations for different models and datasets. (a) shows the variation of the estimated ratio for the generator gradient norms of DCGAN when trained on the CIFAR-10 dataset. (b) shows the variation of the estimated ratio for the generator gradient norms of DCGAN when trained on the MNIST dataset. (c) shows the variation of the estimated ratio for complete gradient norms over iterations for Real-NVP trained on the CIFAR-10 dataset

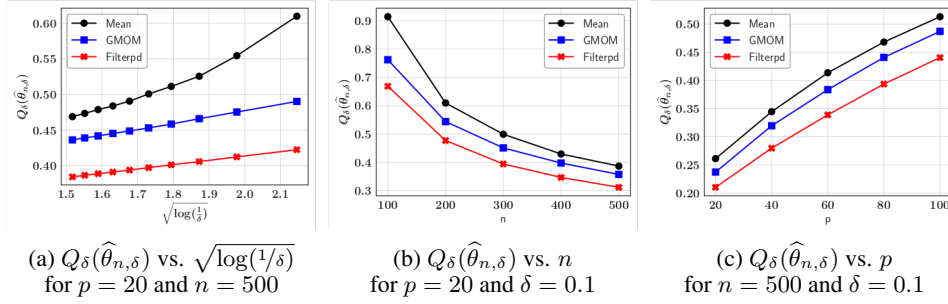


Figure 7: Results for Heavy-Tailed Mean Estimation. Smaller values for  $Q_\delta(\hat{\theta}_{n,\delta})$  are better.

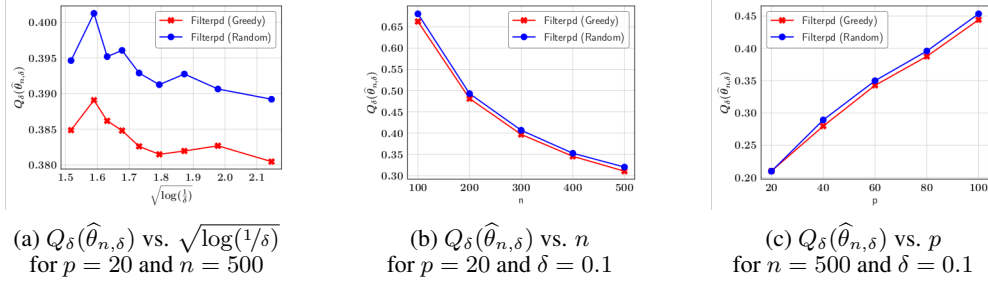


Figure 8: Results for Heavy-Tailed Mean Estimation – comparing the randomized and greedy variants of Filterpd. Smaller values for  $Q_\delta(\hat{\theta}_{n,\delta})$  are better.

confidence level of  $1 - \delta$ . For GMOM, we follow the recommendation of (Minsker, 2015) and set the number of blocks  $k$  is set to  $\lceil 3.5 \log(1/\delta) \rceil$ . We also set the number of iterations in Filterpd to  $\lceil 3.5 \log(1/\delta) \rceil$ .

**Results** Figure 7 shows that our filtering estimator clearly outperforms both baselines across several metrics. Figure 7a show that for any confidence level  $1 - \delta$ , the length of the oracle confidence interval ( $Q_\delta(\hat{\theta}_{n,\delta})$ ) for our estimator is better than all baselines. We also see better sample dependence in Figure 7b, and better dimension dependence in Figure 7c.

The version of Filterpd that we implement is the greedy version, wherein we discard the point with the maximum score, as opposed to sampling the point to discard. We conduct a comparison of both

versions in Figure 8. Note that the variations in the metric are not significant, and that the greedy variant performs better (marginally) than the randomized variant.

We also conduct a preliminary comparison to the spectral algorithm in Lei et al. (2020). Due to many tuning parameters in the fast spectral estimator proposed in Lei et al. (2020), we were unable to run a comprehensive analysis in sufficient time. However, some initial runs suggested that our estimator performed better:

$n \downarrow$ / Estimator $\rightarrow$	Filterpd	Spectral (Lei et al., 2020)
32	<b>1.17</b>	1.56
64	<b>0.81</b>	1.27
128	<b>0.58</b>	0.92

Table 3: Variation of  $\ell_2$  error  $\|\hat{\theta}_{n,\delta} - \theta^*\|_2$  with  $n$  for different algorithms

The setting considered is a 20-dimensional isotropic multivariate Pareto distribution with tail parameter 2.2.

## E HYPERPARAMETERS USED FOR THE EXPERIMENTS IN SECTION 5.2

### E.1 ADDENDUM TO SECTION 5.2.1

For our experiments, we consider the following hyperparameters settings for training:

Hyperparameter	Value
Learning rate	$2 \cdot 10^{-4}$
$\alpha$ for Streaming – Filterpd	0.75
Latent dimension	128
Discard parameter $d$ for Streaming – Filterpd	5
ADAM $(\beta_1, \beta_2)$	(0.5, 0.999)
Batch size	64
Number of points discarded in the norm removal baseline	5
Clipping parameter in Clip	10
Number of Buckets in GMOM	5

The code that we provide contains all these as defaults, and the implementation is using PyTorch 1.5.0. Further environment details are specified in the code provided.

For computing the Parzen window based log-likelihood scores, we use the original code authored by Ian Goodfellow as a part of Goodfellow et al. (2014). For computing the Inception and MODE scores, we sample 10000 images from the respective models trained on CIFAR10 and MNIST. For this, we adapt code from Shane Barratt et. al. written to complement the preprint<sup>2</sup>.

We also present a comparison of times taken by the algorithms considered in the table below.

Algorithm	Time Taken per iteration (s)
Sample Mean	0.029
Clipping	0.033
Norm Removal	0.926
GMOM	0.942
Streaming – Filterpd	1.609
Filterpd	5.624

Table 4: Time taken per iteration when training a DCGAN on the MNIST dataset.

<sup>2</sup><https://arxiv.org/abs/1801.01973>

The reason we see a vast difference between the groups Norm Removal, GMOM, Streaming – Filterpd, Filterpd and Sample Mean, Clipping is because sample mean and clipping do not require the computation of element-wise gradients, whereas the algorithms in the other group do. As specified earlier, we use PyTorch 1.5.0, and this version of PyTorch does not have the functionality to parallelize element-wise gradient computation unlike JAX, a more recent framework. However, newer versions of PyTorch have this feature available in an experimental phase, and we will update our codebase when this is available out of this phase. As expected, Filterpd takes the longest, and this is due to the multiple leading eigenvector computations. As remarked in Section 4, we see that Streaming – Filterpd is approximately  $4\times$  faster than Filterpd and due to computational costs, we do not run Filterpd.

## E.2 ADDENDUM TO SECTION 5.2.2

We implemented Streaming – Filterpd on top of the `RealNVP` implementation by Chris Chute<sup>3</sup>. The implementation that we have borrowed from has retained the same hyperparameter settings as the original implementation in Dinh et al. (2016).

## F PROOF FOR THEOREM 1

We restate the theorem for brevity:

**Theorem 1.** Suppose  $\{z_i\}_{i=1}^n \sim P$ , where  $P$  has bounded  $4^{th}$  moment and  $n$  satisfies

$$n \geq Cr^2(\Sigma) \frac{\log^2(p/\delta)}{\log(1/\delta)}, \quad r(\Sigma) \stackrel{\text{def}}{=} \frac{\text{trace}(\Sigma)}{\|\Sigma\|_2} \quad (13)$$

Then, Filterpd when instantiated for  $T^* = \lceil C \log(1/\delta) \rceil$  steps returns an estimate  $\hat{\theta}_\delta$  which satisfies with probability at least  $1 - 4\delta$ ,  $\delta \in (0, 0.25)$ :

$$\|\hat{\theta}_\delta - \mu\|_2 \lesssim \text{OPT}_{n,\Sigma,\delta}$$

*Proof Sketch.* The proof follows the steps highlighted below:

- We first show that given an arbitrary collection of points  $S$ , and information about the size of an unknown subset  $G^0 \subset S$ , then Filterpd approximates the mean of the points in  $G^0$  efficiently with high probability. This is formally stated in Lemma 1.
- We then show that given  $n$  samples  $\{x_i\}_{i=1}^n$  from a distribution  $P$  with bounded moments, there exists a *good subset* of points. This *good subset* satisfies the following properties:
  1. The size of the set is sufficiently large.
  2. The mean of the points in this set concentrates strongly around the mean of  $P$
  3. The covariance of the points is well-behaved.

We define this *good subset* via a *good point selector*  $\mathcal{O} : \mathbb{R}^p \rightarrow \{0, 1\}$  as defined below:

$$\mathcal{O}(x) = \mathbb{I} \{ \|x - \mu(P)\|_2 \leq R \}$$

and the *good subset* is the set of points  $G^0 \stackrel{\text{def}}{=} \{x_i : \mathcal{O}(x_i) = 1\}$ . We formally state the assertions in Lemma 2.

- For a specific setting of  $R$  in Lemma 2, we obtain the statement of Theorem 1.

**Lemma 1.** Let  $S$  be any arbitrary collection of points, and let  $G^0 \subset S$  be an unknown subset of size  $n_{G^0}$  such that  $8 \frac{n - n_{G^0}}{n} + 36 \frac{\log(1/\delta)}{n} < \frac{1}{4}$ . Then, when Algorithm 1 is run for  $T^* =$

<sup>3</sup><https://github.com/chrischute/real-nvp>

$\lceil 3(n - n_{G^0}) + 18 \log(1/\delta) \rceil$  steps on  $S$ , it returns an estimate  $\hat{\theta}_\delta$  such that with probability at least  $1 - \delta$ ,

$$\left\| \hat{\theta}_\delta - \frac{1}{n_{G^0}} \sum_{x_i \in G^0} x_i \right\|_2 \lesssim \|\Sigma_{G^0}\|_2^{1/2} \left( \frac{n - n_{G^0}}{n} + \frac{\log(1/\delta)}{n} \right)^{1/2},$$

where  $\Sigma_{G^0}$  is the covariance of the unknown subset of points.

**Lemma 2.** Let  $P$  be any distribution with mean  $\mu$  and covariance  $\Sigma$  and bounded  $2k$ -moments for  $k \in \{1, 2\}$ . Furthermore, define:

$$\hat{\mu}_n = \left( \sum_{i=1}^n \mathcal{O}(x_i) \right)^{-1} \left( \sum_{i=1}^n x_i \mathcal{O}(x_i) \right) \quad \hat{\Sigma}_n^\mathcal{O} = \left( \sum_{i=1}^n \mathcal{O}(x_i) \right)^{-1} \left( \sum_{i=1}^n (x_i - \hat{\mu}_n)^{\otimes 2} \mathcal{O}(x_i) \right)$$

as the sample mean and covariance of points in the good subset respectively.

For any  $\delta \in (0, 0.5)$  such that  $\left( \frac{\sqrt{\text{trace}(\Sigma)}}{R} \right)^{2k} + \frac{\log(1/\delta)}{n} < c$  with probability at least  $1 - 3\delta$ ,

$$\begin{aligned} \frac{n - n_{G^0}}{n} &\leq C_1 \frac{\log(1/\delta)}{n} + \frac{(\sqrt{\text{trace}(\Sigma)})^{2k}}{R^{2k}} \\ \|\hat{\mu}_n - \mu\|_2 &\lesssim \text{OPT}_{n, \Sigma, \delta} + \frac{R \log(1/\delta)}{n} + \|\Sigma\|_2^{1/2} \left( \frac{\sqrt{\text{trace}(\Sigma)}}{R} \right)^{2k-1} \\ \|\hat{\Sigma}_n^\mathcal{O}\|_2 &\lesssim \|\Sigma\|_2 + R \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(p/\delta)}{n}} + \frac{R^2 \log(p/\delta)}{n} \end{aligned}$$

We present the proofs of Lemmas 1 and 2 in Sections F.1 and F.2 respectively.  $\square$

*Complete Proof.* Using Chebyshev's inequality, we have that,

$$\Pr(\|x - \mu\|_2 \geq R) \leq \frac{\mathbb{E}[\|x - \mu\|_2^{2k}]}{R^{2k}}$$

Now, to see that  $\mathbb{E}[\|x - \mu\|_2^{2k}] \leq C \left( \sqrt{\text{trace}(\Sigma)} \right)^{2k}$ . The case for  $k = 1$  is clear. We now show it for  $k = 2$ . Let  $\Sigma = Q\Lambda Q^T$  and  $\{q_i\}_{i=1}^p$  be the eigenvectors of  $\Sigma$  and let  $\lambda_i = q_i^T \Sigma q_i$  be the associated eigenvalue. Then,

$$(x - \mu)^T (x - \mu) = \sum_i (q_i^T (x - \mu))^2 = \sum_i \nu_i^2, \quad (14)$$

where  $\nu_i = q_i^T (x - \mu)$ . Now,  $\|x - \mu\|_2^4 = \left( \sum_i \nu_i^2 \right)^2 = \sum_i \nu_i^4 + 2 \sum_{i \neq j} \nu_i^2 \nu_j^2$ . Now, since we assume bounded fourth moments, we get that,  $\mathbb{E}[\nu_i^4] \leq C (q_i^T \Sigma q_i)^2 = C \lambda_i^2$ . Using Cauchy-Schwarz inequality, we get that  $\mathbb{E}[\nu_i^2 \nu_j^2] \leq \sqrt{\mathbb{E}[\nu_i^4]} \sqrt{\mathbb{E}[\nu_j^4]} = C \lambda_i \lambda_j$ . Hence, we have that,

$$\mathbb{E}[\|x - \mu\|_2^4] \leq C \left( \sum_i \lambda_i^2 + 2 \sum_{i \neq j} \lambda_i \lambda_j \right) = C_4 \text{trace}(\Sigma)^2$$

Consequently, we have:

$$\Pr(\|x - \mu\|_2 \geq R) \leq \frac{\mathbb{E}[\|x - \mu\|_2^4]}{R^4} = C_4 \frac{\text{trace}(\Sigma)^2}{R^4}$$

Hence, for  $k = 1, 2$ , we have that,

$$\Pr(\|x - \mu\|_2 \geq R) \leq \frac{\left(\sqrt{\text{trace}(\Sigma)}\right)^{2k}}{R^{2k}}$$

This leads to the fact that for  $x_i \sim P$ ,  $\Pr(\mathcal{O}(x_i) = 1) \geq 1 - \alpha$ , where  $\alpha = \frac{\left(\sqrt{\text{trace}(\Sigma)}\right)^{2k}}{R^{2k}}$ . Using Bernstein's inequality, we know that with probability at least  $1 - \delta$ :

$$n_{G^0} \geq n \left(1 - C_2 \frac{\log(1/\delta)}{n}\right) \Rightarrow \frac{n - n_{G^0}}{n} \lesssim \frac{\log(1/\delta)}{n} \quad (15)$$

From Lemma 1, we have with probability at least  $1 - \delta$ :

$$\|\hat{\theta}_\delta - \hat{\mu}_n\|_2 \lesssim \|\Sigma_{G^0}\|_2^{1/2} \left(\frac{n - n_{G^0}}{n} + \frac{\log(1/\delta)}{n}\right)^{1/2} \quad (16)$$

From Lemma 2, we bound  $\|\Sigma_{G^0}\|_2^{1/2}$  as:

$$\|\Sigma_{G^0}\|_2 \leq C_1 \|\Sigma\|_2 + C_2 R \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(p/\delta)}{n_{G^0}}} + R^2 \frac{\log(p/\delta)}{n_{G^0}} \quad (17a)$$

$$\|\Sigma_{G^0}\|_2^{1/2} \leq C_1 \|\Sigma\|_2^{1/2} + C_2 \sqrt{R} \|\Sigma\|_2^{1/4} \sqrt[4]{\frac{\log(p/\delta)}{n_{G^0}}} + R \sqrt{\frac{\log(p/\delta)}{n_{G^0}}} \quad (17b)$$

Plugging the above bound and (15) in (16), with  $R = \frac{\sqrt{\text{trace}(\Sigma)}}{\left(\frac{\log(1/\delta)}{n}\right)^{1/4}}$ , we get,

$$\|\Sigma_n^{\mathcal{O}}\|_2^{1/2} \leq \underbrace{C_1 \|\Sigma\|_2^{1/2} + C_2 \text{trace}(\Sigma)^{1/4} \|\Sigma\|_2^{1/4} \frac{\left(\frac{\log(p/\delta)}{n_{G^0}}\right)^{1/4}}{\left(\frac{\log(1/\delta)}{n}\right)^{1/8}}}_{T_1} + \underbrace{\sqrt{\text{trace}(\Sigma)} \frac{\sqrt{\frac{\log(p/\delta)}{n_{G^0}}}}{\left(\frac{\log(1/\delta)}{n}\right)^{1/4}}}_{T_2} \quad (18)$$

Some algebra shows that when  $n \geq Cr^2(\Sigma) \frac{\log^2(p/\delta)}{\log(1/\delta)}$ ,  $T_1, T_2 \leq O(\|\Sigma\|_2)$ , which gives:

$$\|\hat{\theta}_\delta - \hat{\mu}_n\|_2 \lesssim \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n}} \quad (19)$$

Again invoking Lemma 2, we get with probability at  $1 - \delta$ :

$$\|\mu(P) - \hat{\mu}_n\|_2 \lesssim \underbrace{\text{OPT}_{n,\Sigma,\delta}}_{T_3} + \sqrt{\text{trace}(\Sigma)} \left(\frac{\log(1/\delta)}{n}\right)^{3/4} \quad (20)$$

Under our assumption,  $T_3 \lesssim \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n}}$ . By an application of the triangle inequality and union bound, we obtain the statement of the theorem.  $\square$



## F.1 PROOF FOR LEMMA 1

*Proof.* Our proof is split into two keys Lemmas. Firstly, in Lemma 3, we show that the with probability at least  $1 - \delta$ , when the algorithm terminates after  $T_\delta^* = \lceil 18 \log(1/\delta) + 3(n - n_{G^0}) \rceil$ , then the covariance of the remaining samples is well-behaved. Finally, in Lemma 4 we show that under our assumptions that  $8 \frac{n - n_{G^0}}{n} + 36 \frac{\log(1/\delta)}{n} < \frac{1}{4}$ , when the algorithm stops after  $T_\delta^*$  steps, the sample mean of points,  $\hat{\theta}_{S^{T_\delta^*}}$  is close to the mean of  $G^0$ . In particular, we show that

$$\|\hat{\theta}_{G^0} - \hat{\theta}_{S^{T_\delta^*}}\|_2 \leq C_1 \left( 8 \frac{n - n_{G^0}}{n} + 36 \frac{\log(1/\delta)}{n} \right)^{1/2} \|\Sigma_{G^0}\|_2^{1/2}, \quad (21)$$

which recovers the statement of the Lemma.

**Lemma 3.** *When Algorithm 1 is instantiated on  $S^0$  for  $T_\delta^* = \lceil 18 \log(1/\delta) + 3(n - n_{G^0}) \rceil$  steps, then with probability  $1 - \delta$ ,*

$$\|\Sigma_{S^{T^*}}\|_2 \leq C_2 \|\Sigma_{G^0}\|_2$$

*Proof.* At each step of Algorithm 1, we remove one sample based on the probability distribution of the scores. Let  $l = 1, 2, \dots, n$  be the steps of the algorithm. Note that the steps of the Algorithm are dependent, hence to obtain a high probability statement, we will have to use martingale style analysis. The martingale analysis in the proof mostly follows from (Xu et al., 2013; Liu et al., 2020).

Let  $\mathcal{F}^l$  be the filtration generated by the sets of events until step  $l$ . At step  $l$ , let  $S^l$  be the set of samples,  $G^l$  be the subset of  $G^0$  stil in  $S^l$ , i.e.  $\{x_i \in S^l \cap G^0\}$ . Let  $B^l = S^l \setminus G^l$  be the remaining samples. Note that  $|S^l| = n_l = n - l$ , and  $S^l, G^l, B^l \in \mathcal{F}^l$ .

Let  $\tau_i$  be some score for each point. Define  $\mathcal{E}^l$  be an event variable at step  $l$  which is True if

$$\sum_{i \in G^l} \tau_i \geq \frac{1}{(\gamma - 1)} \sum_{j \in B^l} \tau_j \quad \equiv \quad \sum_{i \in G^l} \tau_i \geq \frac{1}{\gamma} \sum_{j \in S^l} \tau_j$$

for say  $\gamma = 3$ . Intuitively, this means the event is true when the sum of the scores of the good points is larger compared to the bad points. Now, when  $\mathcal{E}^l$  is false, we sample a point  $j$  according  $\tau_j$  and remove it. Some algebra shows, that when  $\mathcal{E}^l$  is false, then with constant probability of  $2/3$ , we throw a point from  $B^l$ .

$$\Pr(\text{sample removed at Step } l \in B^l | \mathcal{F}^l) = \frac{\sum_{i \in B^l} \tau_i}{\sum_{j \in S^l} \tau_j} \geq \frac{\gamma - 1}{\gamma} = 2/3$$

Essentially, our argument shows that whenever  $\mathcal{E}^l$  is false, then we are more likely to throw a point from the bad set. This means, that in the next iteration the fraction of bad points will reduce. To argue more formally, let  $T \stackrel{\text{def}}{=} \min\{l : \mathcal{E}^l \text{ is true}\}$  be the first time that  $\mathcal{E}^l$  is True. Then, our goal is to show that  $T$  is small.

To show this, based on  $T$ , define  $Y^l$ , as

$$Y^l = \begin{cases} |B^{T-1}| + \frac{\gamma-1}{\gamma}(T-1), & \text{if } l \geq T \\ |B^l| + \frac{\gamma-1}{\gamma}l, & \text{if } l < T \end{cases}$$

Now, we show that  $\{Y^l, \mathcal{F}^l\}$  is a supermartingale, i.e.  $\mathbb{E}[Y^l | \mathcal{F}^{l-1}] \leq Y^{l-1}$ . To see this, we split it into three cases:

- **Case 1.**  $l < T$ . This means that  $\mathcal{E}^l$  is false.

$$Y^l - Y^{l-1} = |B^l| - |B^{l-1}| + \frac{\gamma - 1}{\gamma}, \quad (22)$$

Now,  $|B^l| = |B^{l-1}|$  if no bad point is thrown, and  $|B^l| = |B^{l-1}| - 1$  if the point thrown is bad. Since,  $\mathcal{E}^{l-1}$  is false, hence, we have that,

$$\mathbb{E}[Y^l - Y^{l-1} | \mathcal{F}^{l-1}] = -1(\Pr(\text{sample removed at Step } l-1 \in B^{l-1})) + \frac{\gamma - 1}{\gamma} \stackrel{(i)}{\leq} 0$$

where (i) is true because  $\mathcal{E}^{l-1}$  is false.

- **Case 2.**  $l = T$ , This follows by construction, because at  $l = T$ ,  $Y^l = Y^{l-1}$ .
- **Case 3.**  $l > T$ , This also follows by construction.

So, we have that  $Y^l, \mathcal{F}^l$  is a supermartingale. Now, we need to bound the steps  $T_\delta$  such that the probability that the algorithm doesn't stop in  $T_\delta$  steps is less than  $\delta$ , i.e.

$$\Pr\left(\bigcap_{l=1}^{T_\delta} (\mathcal{E}^l)^c\right) \leq \delta$$

Note, that,

$$\Pr\left(\bigcap_{l=1}^{T_\delta} (\mathcal{E}^l)^c\right) = \Pr(T \geq T_\delta) \stackrel{(ii)}{\leq} \Pr\left(Y^{T_\delta} \geq \frac{\gamma - 1}{\gamma} T_\delta\right) \quad (23)$$

where (ii) follows because, if  $T > T_\delta \implies Y^{T_\delta} = |B^{T_\delta}| + \frac{\gamma - 1}{\gamma} T_\delta \geq \frac{\gamma - 1}{\gamma} T_\delta$ . Now,

$$\Pr\left(Y^{T_\delta} \geq \frac{\gamma - 1}{\gamma} T_\delta\right) = \Pr\left(Y^{T_\delta} - Y^0 \geq \frac{\gamma - 1}{\gamma} T_\delta - Y_0\right)$$

Now, defining  $D^l = Y^l - Y^{l-1}$ , and let  $Z^l = D^l - \mathbb{E}[D^l | D^1, D^2, \dots, D^{l-1}]$ . Then,

$$Y^{T_\delta} - Y^0 = \sum_{l=1}^{T_\delta} D^l = \sum_{l=1}^{T_\delta} Z^l + \sum_{l=1}^{T_\delta} \mathbb{E}[D^l | D^1, D^2, \dots, D^{l-1}]$$

Since, we know that  $\{Y^l, \mathcal{F}^l\}$  is a supermartingale, hence the difference process is such that

$$\mathbb{E}[D^l | D^1, D^2, \dots, D^{l-1}] \leq 0$$

This implies that

$$Y^{T_\delta} - Y^0 \leq \sum_{l=1}^{T_\delta} Z^l \implies \Pr\left(Y^{T_\delta} - Y^0 \geq \frac{\gamma - 1}{\gamma} T_\delta - Y_0\right) \leq \Pr\left(\sum_{l=1}^{T_\delta} Z^l \geq \frac{\gamma - 1}{\gamma} T_\delta - Y_0\right)$$

Since,  $|D^l| \leq 1$ , and  $Z^l \leq 2$  are bounded, hence we can use the Azuma-Hoeffding inequality to bound the above probability. In particular,

$$\Pr\left(\sum_{l=1}^{T_\delta} Z^l \geq \frac{\gamma - 1}{\gamma} T_\delta - Y_0\right) \leq \exp\left(-\frac{\left(\frac{\gamma - 1}{\gamma} T_\delta - Y_0\right)^2}{8T_\delta}\right)$$

Now, we want a  $T_\delta$  such that,  $\exp\left(-\frac{(\frac{\gamma-1}{\gamma}T_\delta - Y_0)^2}{8T_\delta}\right) \leq \delta$ . Solving the quadratic, we need a  $T_\delta$  such that,

$$\left(\frac{\gamma-1}{\gamma}\right)^2 T_\delta^2 - \left(8\log(1/\delta) + 2Y_0 \frac{\gamma-1}{\gamma}\right) T_\delta + Y_0^2 \geq 0$$

Some algebra shows that  $T_\delta^* = \left\lceil 8\log(1/\delta) \frac{\gamma^2}{(\gamma-1)^2} + 2Y_0 \frac{\gamma}{\gamma-1} \right\rceil$  satisfies the above equation. Hence, we know that with probability at least  $1 - \delta$ , there exists at least one good event in 1 to  $T_\delta^*$  iterations. Note that  $Y^0 = n_{B^0} = n - n_{G^0}$ .

While we have established that there is at least one good event in 1 to  $T_\delta^*$  iterations, suppose  $m \in [1, T_\delta^*]$  is the first index such that  $\mathcal{E}^m$  is true. Next, we establish a series of deterministic results.

- When  $\mathcal{E}^m$  is True, then  $\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$  (See Claim 3).
- Coupling this with Claim 2, which shows that  $\|\Sigma_{G^m}\|_2 \leq 2\|\Sigma_{G^0}\|_2$ , we get that  $\|\Sigma_{S^m}\|_2 \leq 32\|\Sigma_{G^0}\|_2$ .
- Hence, we have that with probability  $1 - \delta$ , there exists a point in time  $m \in [1, T_\delta]$  such that,
$$\|\Sigma_{S^m}\|_2 \leq 32\|\Sigma_{G^0}\|_2$$
- Now, observe that  $S^{T^*} \subseteq S^m$ , i.e. the final returned set of points is a subset of the points at  $m$ . Claim 4 shows that the covariance at  $S^{T^*}$  is such that  $\|\Sigma_{S^{T^*}}\|_2 \leq \frac{n-m}{n-T^*} \|\Sigma_{S^m}\|_2 \leq C_1 \|\Sigma_{S^m}\|_2$ .

Chaining the above arguments shows that  $\|\Sigma_{T^*}\|_2 \leq C\|\Sigma_{G^0}\|_2$ .  $\square$

Next, we state and prove Lemma 4. Recall that  $\mathcal{E}^l$  is defined to be an event variable at step  $l$  which is True if

$$\sum_{i \in G^l} \tau_i \geq \frac{1}{(\gamma-1)} \sum_{j \in B^l} \tau_j \equiv \sum_{i \in G^l} \tau_i \geq \frac{1}{\gamma} \sum_{j \in S^l} \tau_j,$$

where  $S^l$  is set of samples at step  $l$ , and  $G^l = \{x_i \in S^l \cap G^0\}$  is the subset of samples from  $G^0$  which are still in  $S^l$ . Also, recall that for Algorithm 1, the sampling weights  $\tau_i$  at any step  $l$  are defined as  $\tau_i = \left(v^T(x_i - \hat{\theta}_{S^l})\right)^2$ , where  $v$  is the top unit-norm eigenvector of  $\hat{\Sigma}_{S^l}$  and  $\hat{\theta}_{S^l}$  is the sample mean of  $S^l$ . Then, in Lemma 3 we showed that with probability  $1 - \delta$ ,

$$\|\Sigma_{S^{T^*}}\|_2 \leq C_2 \|\Sigma_{G^0}\|_2.$$

**Lemma 4.** Let  $\phi = \frac{n-n_{G^0}}{n}$ . Then, under the assumption that  $8\phi + 36\frac{\log(1/\delta)}{n} < \frac{1}{4}$ , we have that for  $m = T_\delta^*$

$$\|\hat{\theta}_{G^0} - \hat{\theta}_{S^m}\|_2 \leq 10\sqrt{2} \left(8\phi + 36\frac{\log(1/\delta)}{n}\right)^{1/2} \|\Sigma_{G^0}\|_2^{1/2},$$

*Proof.* Using Lemma 6, we get that,

$$\|\hat{\theta}_{G^0} - \hat{\theta}_{S^m}\|_2 \leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}} \left(\|\Sigma_{G^0}\|_2^{1/2} + \|\Sigma_{S^m}\|_2^{1/2}\right),$$

where  $P_1$  is the equal weight discrete distribution with support on  $S^m$ , and  $P_2$  is the equal weight discrete distribution with support on  $G^0$ . Lemma 3 already controls tell us that for  $m = T_\delta^*$ ,  $\|\Sigma_{S^m}\|_2 \leq C_2 \|\Sigma_{G^0}\|_2$ . We show next that

$$TV(P_1, P_2) \leq 8\phi + 36\frac{\log(1/\delta)}{n},$$

which finishes the proof of the Lemma.

To bound the TV distance between  $P_1$  and  $P_2$ , we use triangle inequality. Let  $P_3$  be the equal weight discrete distribution with support on  $G^m$ . Let  $\tau \in [1, m] \leq T_\delta$  be the number of "good" points thrown out in  $m \leq T_\delta$  steps. For  $\gamma = 3$ , we have that,

$$T_\delta = 18 \log(1/\delta) + 3n_{B^0}$$

$$TV(P_1, P_2) \leq TV(P_1, P_3) + TV(P_3, P_2) \quad (24)$$

$$\leq \frac{n_{S^m} - n_{G^m}}{n_{S^m}} + \frac{n_{G^0} - n_{G^m}}{n_{G^0}} \quad (25)$$

$$= \frac{n - T_\delta - (n - n_{B^0} - \tau)}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} \quad (26)$$

$$= \frac{n_{B^0} + \tau - T_\delta}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} \quad (27)$$

$$\leq \frac{n_{B^0}}{n - T_\delta} + \frac{T_\delta}{n - n_{B^0}} \quad (28)$$

$$= \frac{\phi}{1 - \frac{18 \log(1/\delta)}{n} - 3\phi} + \frac{\frac{18 \log(1/\delta)}{n} + 3\phi}{1 - \phi} \quad (29)$$

where  $\phi = \frac{n_{B^0}}{n}$ . Now under the assumption that  $3\phi + \frac{18 \log(1/\delta)}{n} < \frac{1}{2}$ , the first term is less than  $2\phi$ . □

□

#### F.1.1 AUXILLARY RESULTS FOR PROOF OF LEMMA 1

**Lemma 5.** Let  $S$  be a collection of  $n$  points. And let  $G$  be a subset of  $S$  containing  $n_G$  points. Define  $\tau_i = \left(v^T(x_i - \hat{\theta}_S)\right)^2$ , where  $v$  is the top unit-norm eigenvector of  $\hat{\Sigma}_S$  and  $\hat{\theta}_S$  is the sample mean of  $S$ . Let  $\lambda = \|\Sigma_S\|_2$ . If  $\lambda > \frac{1+\psi}{\frac{n}{n_G\gamma} - \psi} \|\Sigma_G\|_2$ , then

$$\sum_{i: x_i \in G} \tau_i < \frac{1}{\gamma} \sum_{j=1}^n \tau_j,$$

where  $\psi = \left(\frac{1}{\sqrt{\frac{n}{n-n_G}-1}}\right)^2 < \frac{n}{n_G\gamma}$ .

*Proof.* Let  $\hat{\theta}_G$  be the sample mean of points in  $G$ .

$$\begin{aligned} \frac{1}{n_G} \sum_{i: x_i \in G} \tau_i &= \frac{1}{n_G} \sum_{i: x_i \in G} v^T(x_i - \hat{\theta}_S)(x_i - \hat{\theta}_S)^T v \\ &= v^T \left( \frac{1}{n_G} \sum_{i: x_i \in G} (x_i - \hat{\theta}_G)(x_i - \hat{\theta}_G)^T \right) v + \left( v^T(\hat{\theta}_G - \hat{\theta}_S) \right)^2 \\ &\leq v^T \Sigma_G v + \|\hat{\theta}_G - \hat{\theta}_S\|_2^2 \\ &\leq v^T \Sigma_G v + \underbrace{\left( \frac{1}{\sqrt{\frac{n}{n-n_G}-1}} \right)^2}_{\psi} (\|\Sigma_S\|_2 + \|\Sigma_G\|_2) \\ &\leq \|\Sigma_G\|_2 (1 + \psi) + \psi \|\Sigma_S\|_2 \end{aligned}$$

Now, if  $\|\Sigma_S\|_2 \geq \frac{1+\psi}{\frac{n}{n_G\gamma}-\psi} \|\Sigma_G\|_2$ , then we have that

$$\begin{aligned} \frac{1}{n_G} \sum_{i:x_i \in G} \tau_i &\leq \frac{n}{n_G\gamma} \|\Sigma_S\|_2 \\ &= \frac{n}{n_G\gamma} \sum_{j=1}^n (v^T(x_j - \hat{\theta}_S))^2 \\ \Rightarrow \sum_{i:x_i \in G} \tau_i &\leq \frac{1}{\gamma} \sum_{j=1}^n \tau_j \end{aligned}$$

□

**Claim 1.** Suppose  $P_1$  is the equal weight discrete distribution with support on  $S^m$ , and  $P_2$  is the equal weight discrete distribution with support on  $G^0$ . Then, when  $\phi = \frac{n_{B^0}}{n}$  is such that  $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$ ,

$$TV(P_1, P_2) \leq 8\phi + 36 \frac{\log(1/\delta)}{n}$$

*Proof.* To bound the TV distance between  $P_1$  and  $P_2$ , we use triangle inequality. Let  $P_3$  be the equal weight discrete distribution with support on  $G^m$ . Let  $\tau \in [T_\delta]$  be the number of "good" points thrown out in  $T_\delta$  steps. For  $\gamma = 3$ , we have that,

$$T_\delta = 18\log(1/\delta) + 3n_{B^0}$$

$$\begin{aligned} TV(P_1, P_2) &\leq TV(P_1, P_3) + TV(P_3, P_2) \\ &\leq \frac{n_{S^m} - n_{G^m}}{n_{S^m}} + \frac{n_{G^0} - n_{G^m}}{n_{G^0}} \\ &= \frac{n - T_\delta - (n - n_{B^0} - \tau)}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} \\ &= \frac{n_{B^0} + \tau - T_\delta}{n - T_\delta} + \frac{\tau}{n - n_{B^0}} \\ &\leq \frac{n_{B^0}}{n - T_\delta} + \frac{T_\delta}{n - n_{B^0}} \\ &= \frac{\phi}{1 - \frac{18\log(1/\delta)}{n} - 3\phi} + \frac{\frac{18\log(1/\delta)}{n} + 3\phi}{1 - \phi} \end{aligned}$$

where  $\phi = \frac{n_{B^0}}{n}$ . Now under the assumption that  $3\phi + \frac{18\log(1/\delta)}{n} < \frac{1}{2}$ , the first term is less than  $2\phi$ . □

**Lemma 6** (Kothari et al. (2018)). Given a collection of points  $S$  of size  $n$ . Let  $P_1$  and  $P_2$  be discrete empirical distributions on  $n$ . Then, we have that,

$$\|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i]\|_2 \leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}} \left( \|\hat{\Sigma}_{P_1}\|_2^{1/2} + \|\hat{\Sigma}_{P_2}\|_2^{1/2} \right) \quad (30)$$

where  $\hat{\Sigma}_{P_1}$  is the covariance matrix when  $x_i \sim P_1$ , and  $\hat{\Sigma}_{P_2}$  is the empirical covariance matrix of when  $x_i \sim P_2$

*Proof.* Consider a joint distribution (also called coupling)  $\omega^*(z, z')$  over  $S \times S$  such that its individual marginal distributions are equal to  $P_1$  and  $P_2$ ; i.e.  $\omega(z) = P_1$  and  $\omega(z') = P_2$  and

$\omega(z \neq z') = TV(P_1, P_2)$ . Then, we have that

$$\begin{aligned}
\|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i]\|_2 &= \sup_{v \in \mathcal{S}^{p-1}} |\langle v, \mathbb{E}_{w^*}[z - z'] \rangle| \\
&= \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*} [|\langle v, z - z' \rangle|] \\
&= \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*} [1(z \neq z') |\langle v, z - z' \rangle|] \\
&\leq (\mathbb{E}_{w^*} [(1(z \neq z'))^2])^{1/2} \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*} [(\langle v, z - z' \rangle)^2]^{1/2} \\
&\leq TV(P_1, P_2)^{1/2} \left( \sup_{v \in \mathcal{S}^{p-1}} \left( \mathbb{E}_{w^*} \left[ (\langle v, z - \mathbb{E}_{x_i \sim P_1}[x_i] + \mathbb{E}_{x_i \sim P_1}[x_i] \right. \right. \right. \\
&\quad \left. \left. \left. - \mathbb{E}_{x_i \sim P_2}[x_i] + \mathbb{E}_{x_i \sim P_2}[x_i] - z' \rangle)^2 \right] \right)^{1/2} \right) \\
&\leq TV(P_1, P_2)^{1/2} \left( \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}_{w^*} \left[ (\langle v, z - \mathbb{E}_{x_i \sim P_1}[x_i] \rangle)^2 \right]^{1/2} \right. \\
&\quad \left. + \|\mathbb{E}_{x_i \sim P_1}[x_i] - \mathbb{E}_{x_i \sim P_2}[x_i]\|_2 \right) \\
&\quad + TV(P_1, P_2)^{1/2} \sup_{v \in \mathcal{S}^{p-1}} \left( \mathbb{E}_{w^*} \left[ (\langle v, z - \mathbb{E}_{x_i \sim P_2}[x_i] \rangle)^2 \right]^{1/2} \right) \\
&\leq \frac{\sqrt{TV(P_1, P_2)}}{1 - \sqrt{TV(P_1, P_2)}} \left( \|\Sigma_{P_1}\|_2^{1/2} + \|\Sigma_{P_2}\|_2^{1/2} \right)
\end{aligned}$$

□

**Claim 2.** Under the assumption that  $4\phi + 18 \frac{\log(1/\delta)}{n} < \frac{1}{2}$ , we have that,

$$\|\Sigma_{G^m}\|_2 \leq 2\|\Sigma_{G^0}\|_2$$

*Proof.* We first show that  $\|\Sigma_{G^m}\|_2 \leq \frac{n_{G^0}}{n_{G^m}} \|\Sigma_{G^0}\|_2$ .

$$\Sigma_{G^0} = \frac{1}{n_{G^0}} \sum_{i \in G^0} (x_i - \hat{\theta}_{G^0})(x_i - \hat{\theta}_{G^0})^T \quad (31)$$

$$= \frac{1}{n_{G^0}} \sum_{i \in G^0} (x_i - \hat{\theta}_{G^0})(x_i - \hat{\theta}_{G^0})^T (\mathbb{I}\{x_i \in G^m\} + \mathbb{I}\{x_i \notin G^m\}) \quad (32)$$

$$= \frac{1}{n_{G^0}} \sum_{i \in G^0} (x_i - \hat{\theta}_{G^0})(x_i - \hat{\theta}_{G^0})^T (\mathbb{I}\{x_i \in G^m\}) \quad (33)$$

$$+ \underbrace{\frac{1}{n_{G^0}} \sum_{i \in G^0} (x_i - \hat{\theta}_{G^0})(x_i - \hat{\theta}_{G^0})^T (\mathbb{I}\{x_i \notin G^m\})}_{T_1} \quad (34)$$

$$= \frac{n_{G^m}}{n_{G^0}} \left( \Sigma_{G^m} + (\hat{\theta}_{G^m} - \hat{\theta}_{G^0})(\hat{\theta}_{G^m} - \hat{\theta}_{G^0})^T \right) + T_1 \quad (35)$$

Now for  $v$  being the top eigenvector of  $\Sigma_{G^m}$ , we get that,

$$\frac{n_{G^m}}{n_{G^0}} v^T \Sigma_{G^m} v + \frac{n_{G^m}}{n_{G^0}} \underbrace{(v^T (\hat{\theta}_{G^m} - \hat{\theta}_{G^0}))^2}_{\geq 0} + \underbrace{v^T T_1 v}_{\geq 0} = v^T \Sigma_{G^0} v$$

Hence, we get that,

$$\|\Sigma_{G^m}\|_2 \leq \frac{n_{G^0}}{n_{G^m}} \|\Sigma_{G^0}\|_2,$$

Now,

$$\frac{n_{G^0}}{n_{G^m}} = \frac{n - n_{B^0}}{n - n_{B^0} - \tau} \leq \frac{n - n_{B^0}}{n - n_{B^0} - T_\delta} = \frac{n - n_{B^0}}{n - 18 \log(1/\delta) - 4n_{B^0}} = \frac{1 - \phi}{1 - 18 \frac{\log(1/\delta)}{n} - 4\phi},$$

where  $\phi = \frac{n_{B^0}}{n}$ . Under our assumption, we get that,  $\frac{n_{G^0}}{n_{G^m}} < 2$ .  $\square$

**Claim 3.** Under the assumption that  $\phi = \frac{n_{B^0}}{n}$  is such that  $3\phi + \frac{18 \log(1/\delta)}{n} < \frac{1}{2}$ , and  $2\phi < 0.12$ , then when  $\mathcal{E}^m$  is True, we have that,

$$\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$$

*Proof.* Suppose  $P_1$  is the equal weight discrete distribution with support on  $S^m$  and let  $P_3$  be the equal weight discrete distribution with support on  $G^m$ . When  $\mathcal{E}^m$  is True, we know by contrapositive

of Lemma 5 that  $\|\Sigma_{S^m}\|_2 \leq \frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma} - \psi_m} \|\Sigma_{G^m}\|_2$ , where  $\psi_m = \left( \frac{\sqrt{TV(P_1, P_3)}}{1 - \sqrt{TV(P_1, P_3)}} \right)^2$ .

Note that for  $TV(P_1, P_3) = \frac{n_{S^m} - n_{G^m}}{n_{S^m}}$ . Hence,  $\frac{n_{S^m}}{n_{G^m}\gamma} = \frac{1}{\gamma(1 - TV(P_1, P_3))}$ . For  $\gamma = 3$ , the term  $\frac{1+\psi_m}{\frac{n_{S^m}}{n_{G^m}\gamma} - \psi_m}$  can be rewritten solely as a function of the  $TV(P_1, P_3)$ . In particular, it can be written as

$$f(x) = \frac{\left(1 + \left(\frac{x^{0.5}}{1 - x^{0.5}}\right)^2\right) \left(3(1 - x^{0.5})^2(1 + x^{(0.5)})\right)}{1 - x^{(0.5)} - 3x - 3x^{(1.5)}}$$

Now  $TV(P_1, P_3) = \frac{n_{S^m} - n_{G^m}}{n_{S^m}} = \frac{(n - T_\delta) - (n - n_{B^0} - \tau)}{n - T_\delta} = \frac{n_{B^0} + \tau - T_\delta}{n - T_\delta} \leq \frac{n_{B^0}}{n - T_\delta} = \frac{\phi}{1 - 18 \frac{\log(1/\delta)}{n} - 3\phi}$ .

Hence, under our assumptions,  $TV(P_1, P_3) < 0.12$ . Some algebra shows that under  $f(x)$  is monotonically increasing for  $x < 0.12$ , and in particular,  $f(0.12) < 16$ . Hence, we get that  $\|\Sigma_{S^m}\|_2 \leq 16\|\Sigma_{G^m}\|_2$ .  $\square$

**Claim 4.** Let  $S_1$  be any collection of points of size  $n_1$ . Let  $S_2 \subseteq S_1$  be a subset of size  $n_2 \leq n_1$ . Then, we have that

$$\|\Sigma_{S_2}\|_2 \leq \frac{n_1}{n_2} \|\Sigma_{S_1}\|_2$$

*Proof.* Let  $\hat{\theta}_{S_2}$  be the mean of points in  $S_2$ . Similarly, let  $\hat{\theta}_{S_1}$  be mean of points in  $S_1$ .

$$\Sigma_{S_1} = \frac{1}{n_1} \sum_{i \in S_1} (x_i - \hat{\theta}_{S_1})(x_i - \hat{\theta}_{S_1})^T \quad (36)$$

$$= \frac{1}{n_1} \sum_{i \in S_1} (x_i - \hat{\theta}_{S_1})(x_i - \hat{\theta}_{S_1})^T (\mathbb{I}\{x_i \in S_2\} + \mathbb{I}\{x_i \notin S_2\}) \quad (37)$$

$$= \frac{1}{n_1} \sum_{i \in S_1} (x_i - \hat{\theta}_{S_1})(x_i - \hat{\theta}_{S_1})^T (\mathbb{I}\{x_i \in S_2\}) \quad (38)$$

$$+ \underbrace{\frac{1}{n_1} \sum_{i \in S_1} (x_i - \hat{\theta}_{S_1})(x_i - \hat{\theta}_{S_1})^T (\mathbb{I}\{x_i \notin S_2\})}_{T_1} \quad (39)$$

$$= \frac{n_2}{n_1} \left( \Sigma_{S_2} + (\hat{\theta}_{S_2} - \hat{\theta}_{S_1})(\hat{\theta}_{S_2} - \hat{\theta}_{S_1})^T \right) + T_1 \quad (40)$$

Now for  $v$  being the top eigenvector of  $\Sigma_{S_2}$ , we get that,

$$\frac{n_2}{n_1} v^T \Sigma_{S_2} v + \frac{n_{S_2}}{n_{S_1}} \underbrace{(v^T (\hat{\theta}_{S_2} - \hat{\theta}_{S_1}))^2}_{\geq 0} + \underbrace{v^T T_1 v}_{\geq 0} = v^T \Sigma_{S_1} v$$

$\square$

## F.2 PROOF OF LEMMA 2

*Proof.* We controlled the size of  $G^0$  in the proof of Theorem 1.

**Controlling the mean of  $G^0$ .** Recall from our assumption that

$$\alpha + C_2 \frac{\log(1/\delta)}{n} < \frac{1}{2},$$

hence we have that  $|G^0| = n_{G^0} > n/2$ . Let  $\hat{\theta}_{G^0} = \hat{\mu}_n$  be the mean of the points in  $G^0$ .

1. Controlling  $\|\mu - \mathbb{E}[\hat{\theta}_{G^0}]\|_2$ . This is a deterministic statement and essentially quantifies the amount the mean can shift, when the random variable is conditioned on an event. We show this in Claim 5 which was shown in (Steinhardt, 2018; Lai et al., 2016).

**Claim 5.** [General Mean shift, (Steinhardt, 2018; Lai et al., 2016)] Suppose that a distribution  $P$  has mean  $\mu$  and covariance  $\Sigma$  and bounded  $2k$  moments. Then, for any event  $\mathcal{A}$  which occurs with probability at least  $1 - \epsilon \geq \frac{1}{2}$ ,

$$\|\mu - E[x|\mathcal{A}]\|_2 \leq 2\|\Sigma\|_2^{1/2} \epsilon^{1-\frac{1}{2k}} \quad (41)$$

Now using this Claim 5 with  $\mathcal{A}$  being the event that  $\mathcal{O}(x) = 1$ , we get that

$$\|\mu - \mathbb{E}[\hat{\theta}_{G^0}]\|_2 \leq 2\|\Sigma\|_2^{1/2} \alpha^{1-1/(2k)} \quad (42)$$

2. **Controlling  $\|\hat{\theta}_{G^0} - \mathbb{E}[\hat{\theta}_{G^0}]\|_2$ .** This term measures how quickly the samples within  $G^0$  converge to their true mean. To show this we use vector version of Bernstein's inequality. Let  $z_i \stackrel{\text{def}}{=} x_i - \mathbb{E}[\hat{\theta}_{G^0}]$  be the centered random variables. Then, we have that

$$\begin{aligned} \|z_i\|_2 &\leq \|\theta^* - \mathbb{E}[\hat{\theta}_{G^0}]\|_2 + \|x_i - \theta^*\|_2 \\ &\leq 2\|\Sigma\|_2^{1/2} \alpha^{1-1/(2k)} + R \\ &\leq 2R \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E}[\|z_i\|_2^2] &= \mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2 | x \in \mathcal{A}] \\ &= \frac{\mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2 \mathbb{I}\{x \in \mathcal{A}\}]}{P(\mathcal{A})} \\ &\leq 2\mathbb{E}[\|x - E[x|\mathcal{A}]\|_2^2] \\ &\leq 2\mathbb{E}[\|x - E[x]\|_2^2] + 2\|\theta^* - E[x|\mathcal{A}]\|_2^2 \\ &\leq 2\text{trace}(\Sigma) + 4\|\Sigma\|_2 \alpha^{2-1/(k)} \\ &\leq 4\text{trace}(\Sigma) \end{aligned}$$

Now, we first state the vector version of Bernstein's inequality.

**Lemma 7** (Foucart & Rauhut, 2013, Corollary 8.45). *Let  $Y_1, \dots, Y_M$  be independent copies of a random vector  $Y \in \mathbb{C}^p$  satisfying  $\mathbb{E}Y = 0$ . Assume  $\|Y\|_2 \leq K$  for some  $K > 0$ . Let,*

$$Z = \left\| \sum_{l=1}^M Y_l \right\|_2, \mathbb{E}[Z^2] = M\mathbb{E}[\|Y\|_2^2], \sigma^2 = \sup_{\|v\|_2 \leq 1} \mathbb{E}[|\langle v, Y \rangle|^2]$$

Then for  $t > 0$ ,

$$\Pr\left(Z \geq \sqrt{\mathbb{E}Z^2} + t\right) \leq \exp\left(-\frac{t^2/2}{M\sigma^2 + 2K\sqrt{\mathbb{E}Z^2} + tK/3}\right) \quad (43)$$



We use the above lemma, with  $Y_i = \frac{z_i}{n_{G^0}}$ . Hence, we have that,  $K = \frac{2R}{n_{G^0}}$ . Hence, we have that

$$Z = \left\| \sum_{k=1}^{n_{G^0}} Y_k \right\|_2 = \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2. \text{ Hence, we have the following,}$$

- $\mathbb{E}[Z^2] \leq n \frac{4\text{trace}(\Sigma)}{n^2} = 4 \frac{\text{trace}(\Sigma)}{n}.$
- $\sigma^2 \leq 4 \frac{\|\Sigma\|_2}{n^2}.$  To see this, for any  $v \in \mathcal{S}^{p-1},$

$$\mathbb{E}[(v^T Y)^2] = \frac{1}{n^2} \mathbb{E}[(v^T (x - \mu_A))^2 | x \in \mathcal{A}]$$

where  $\mu_A$  is the conditional mean, and  $\mathcal{A}$  is the event that  $x$  s.t.  $\|x - \mu\|_2 \leq R$ . We know that  $P(\mathcal{A}) \geq 1/2$ . Hence, we get that,

$$\begin{aligned} \mathbb{E}[(v^T Y)^2] &= \frac{1}{n^2} \frac{\mathbb{E}[(v^T (x - \mu_A))^2 \mathbb{I}\{x \in \mathcal{A}\}]}{P(\mathcal{A})} \\ &\leq \frac{2}{n^2} \mathbb{E}[(v^T (x - \mu_A))^2] \\ &= \frac{2}{n^2} (\mathbb{E}[(v^T (x - \mu))^2] + \|\mu - \mu_A\|_2^2) \\ \implies \sigma^2 &\leq \frac{2}{n^2} (\|\Sigma\|_2 + \|\Sigma\|_2 \alpha) \\ &\leq \frac{4\|\Sigma\|_2}{n^2} \end{aligned}$$

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

$$\begin{aligned} \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 &\leq C_1 \sqrt{\frac{\text{trace}(\Sigma)}{n_{G^0}}} + C_2 \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n_{G^0}}} + C_3 R^{1/2} \left( \sqrt{\frac{\text{trace}(\Sigma)}{n_{G^0}}} \right)^{1/2} \sqrt{\frac{\log(1/\delta)}{n_G^0}} \\ &\quad + C_4 R \frac{\log(1/\delta)}{n_{G^0}} \end{aligned}$$

Now, we use that  $\sqrt{ab} \leq a + b \forall a, b \geq 0$ . Hence, we get that with probability at least  $1 - \delta$

$$\|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \leq \underbrace{C_5 \sqrt{\frac{\text{trace}(\Sigma)}{n_{G^0}}} + C_2 \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n_{G^0}}}}_{T_1} + C_3 R \frac{\log(1/\delta)}{n_{G^0}}$$

Using the bound on  $\|\mathbb{E}[\widehat{\theta}_{G^0}] - \mu\|_2$  from (42), we get that,

$$\begin{aligned} \|\widehat{\theta}_{G^0} - \mu\|_2 &\leq \|\mathbb{E}[\widehat{\theta}_{G^0}] - \mu\|_2 + \|\widehat{\theta}_{G^0} - \mathbb{E}[\widehat{\theta}_{G^0}]\|_2 \\ &\leq T_1 + C_3 R \frac{\log(1/\delta)}{n_{G^0}} + 2\|\Sigma\|_2^{1/2} \left( \left( \frac{\sqrt{\text{trace}(\Sigma)}}{R} \right)^{2k} \right)^{1-1/(2k)} \\ &= T_1 + C_3 R \frac{\log(1/\delta)}{n_{G^0}} + 2\|\Sigma\|_2^{1/2} \left( \frac{(\sqrt{\text{trace}(\Sigma)})^{2k-1}}{R^{2k-1}} \right) \end{aligned}$$

Under our assumption that  $(\frac{\sqrt{\text{trace}(\Sigma)}}{R})^{2k} + \frac{\log(1/\delta)}{n} < c$ , we know that  $n_{G^0} \geq n/2$ . Hence, we get that  $T_1 \lesssim \text{OPT}_{n, \Sigma, \delta}$ .

**Controlling the covariance of points in  $G^0$ .** Let  $G^0 = \{x_i | \mathcal{O}(x_i) = 1\}$  be the empirical collection of points chosen by the oracle. Let  $n_{G^0} = |G^0|$ . Then, we study and bound the operator norm of  $\Sigma_{G^0}$ . Recall that all oracles have the form  $\mathbb{I}\{\|x_i - \mu\|_2 \leq R\}$ , i.e.,  $\forall x_i$  s.t.  $\mathcal{O}(x_i) = 1$ , we have that  $\|x_i - \mu\|_2 \leq R$ .

Let  $\Sigma_{G^0}$  be the empirical covariance matrix. Then,

$$\Sigma_{G^0} = \frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \hat{\theta}_{G^0})(x_i - \hat{\theta}_{G^0})^T,$$

where  $\hat{\theta}_{G^0}$  is the empirical mean of the points in  $G^0$ . Recentering it around the true mean  $\theta^*$  of  $P$ , we get that,

$$\Sigma_{G^0} = \frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \theta^*)(x_i - \theta^*)^T - (\hat{\theta}_{G^0} - \theta^*)(\hat{\theta}_{G^0} - \theta^*)^T$$

Hence, we have that  $\|\Sigma_{G^0}\|_2 \leq \underbrace{\left\| \frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \theta^*)(x_i - \theta^*)^T \right\|_2}_A$ . To control,  $\|A\|_2$ , we use triangle inequality,

$$\|A\|_2 \leq \underbrace{\|A - \mathbb{E}[A]\|_2}_{T_1} + \underbrace{\|\mathbb{E}[A]\|_2}_{T_2} \quad (44)$$

1. **Controlling  $T_2$ .** Note that  $\mathbb{E}[A] = \mathbb{E}[(x - \theta^*)(x - \theta^*)^T | x \in G]$ .

$$\mathbb{E}[A] = \frac{\mathbb{E}[(x - \theta^*)(x - \theta^*)^T \mathbb{I}\{x \in G^0\}]}{P(x \in G^0)} \quad (45)$$

Let  $\Pr(x \in G^0) \geq 1 - \alpha$ . Hence, for any  $v \in \mathcal{S}^{p-1}$ ,

$$v^T \mathbb{E}[A] v = \frac{\mathbb{E}[(v^T(x - \theta^*))^2 \mathbb{I}\{x \in G^0\}]}{P(x \in G^0)} \leq \frac{\|\Sigma\|_2}{1 - \alpha}$$

Under the assumption that  $\alpha < \frac{1}{2}$ , we get that,

$$\|\mathbb{E}[A]\|_2 \leq 2\|\Sigma\|_2$$

2. **Controlling  $T_1$ .** Note that  $T_1$  can be controlled using a concentration of measure argument, and in particular exploits concentration of covariance for bounded random vectors.

**Lemma 8** ((Vershynin, 2010, Theorem 5.44)). *Let  $\{y_i\}_{i=1}^n$  samples such that  $y_i \in \mathbb{R}^p$  and  $\|y_i\|_2 \leq \sqrt{m}$  and  $\mathbb{E}[yy^T] = \Sigma$ . Then, with probability at least  $1 - \delta$ ,*

$$\left\| \frac{1}{n} \sum_{i=1}^n y_i y_i^T - \Sigma \right\|_2 \leq \max \left( \|\Sigma\|_2^{1/2} \sqrt{\log(p/\delta)} \sqrt{\frac{m}{n}}, \log(p/\delta) \frac{m}{n} \right)$$

$$T_1 = \left\| \frac{1}{n_{G^0}} \sum_{i=1}^{n_{G^0}} (x_i - \theta^*)(x_i - \theta^*)^T - \mathbb{E}[A] \right\|_2 \quad (46)$$

We use Lemma 8 with  $y_i = x_i - \theta^*$ . Note that  $\sqrt{m} = R$ . This means that with probability  $1 - \delta$ ,

$$T_1 \leq C_1 R \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(p/\delta)}{n_{G^0}}} + R^2 \frac{\log(p/\delta)}{n_{G^0}}$$

Hence, we get that under the assumption that  $\alpha + \sqrt{\alpha} \sqrt{\frac{\log(1/\delta)}{n}} < \frac{1}{2}$ , we recover statement of the result.  $\square$

## G PROOF OF THEOREM 3

We restate the theorem for brevity:

**Theorem 2.** Suppose  $\{z_i\}_{i=1}^n \sim P$ , where  $P$  has bounded  $2^{\text{nd}}$  moment and  $n$  satisfies the relation in (3). Then Algorithm 1 when instantiated for  $T^* = \lceil C \log(1/\delta) \rceil$  steps returns an estimate  $\hat{\theta}_\delta$  such that, with probability at least  $1 - 4\delta$ ,  $\delta \in (0, 0.25)$ :

$$\|\hat{\theta}_\delta - \mu\|_2 \lesssim \sqrt{\frac{\text{trace}(\Sigma) \log(p/\delta)}{n}}$$

*Proof.* The proof follows a similar approach to the proof of Theorem 1, except we set a different radius parameter  $R$ .

By Chebyshev's and Bernstein's inequality, we have with probability at least  $1 - \delta$ :

$$|n_{G^0}| \geq n(1 - C \frac{\log(1/\delta)}{n}) \quad (47)$$

Hence, we have that,

$$\frac{n - n_{G^0}}{n} \lesssim \frac{\log(1/\delta)}{n} \quad (48)$$

Let  $\hat{\mu}_n$  and  $\Sigma_{G^0}$  be the empirical mean and covariance of the points in  $G^0$ .

From Lemma 1, we know that with probability at least  $1 - \delta$ ,

$$\|\hat{\theta}_\delta - \hat{\mu}_n\|_2 \lesssim \|\Sigma_{G^0}\|_2^{1/2} \left( \frac{n - n_{G^0}}{n} + \frac{\log(1/\delta)}{n} \right)^{1/2} \quad (49)$$

Using Lemma 2, we bound  $\|\Sigma_{G^0}\|_2^{1/2}$ .

$$\begin{aligned} \|\Sigma_n^\mathcal{O}\|_2 &\leq C_1 \|\Sigma\|_2 + C_2 R \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(p/\delta)}{n_{G^0}}} + R^2 \frac{\log(p/\delta)}{n_{G^0}} \\ \|\Sigma_n^\mathcal{O}\|_2^{1/2} &\leq C_1 \|\Sigma\|_2^{1/2} + C_2 R^{1/2} \|\Sigma\|_2^{1/4} \left( \frac{\log(p/\delta)}{n_{G^0}} \right)^{1/4} + R \sqrt{\frac{\log(p/\delta)}{n_{G^0}}} \end{aligned} \quad (50)$$

Plugging  $R = \frac{\sqrt{\text{trace}(\Sigma)}}{\left(\frac{\log(1/\delta)}{n}\right)^{1/2}}$ , we get,

$$\|\Sigma_n^\mathcal{O}\|_2^{1/2} \leq C_1 \|\Sigma\|_2^{1/2} + C_2 \text{trace}(\Sigma)^{1/4} \|\Sigma\|_2^{1/4} \left( \frac{\log(p/\delta)}{\log(1/\delta)} \right)^{1/4} + \frac{\sqrt{\text{trace}(\Sigma)}}{\sqrt{\frac{\log(1/\delta)}{n}}} \sqrt{\frac{\log(p/\delta)}{n}} \quad (51)$$

Plugging (48) and (51) into (49), we get that,

$$\|\hat{\theta}_\delta - \hat{\mu}_n\|_2 \lesssim \|\Sigma\|_2^{1/2} \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\text{trace}(\Sigma) \log(p/\delta)}{n}} \quad (52)$$

Using Lemma 2, and plugging  $R = \frac{\sqrt{\text{trace}(\Sigma)}}{\left(\frac{\log(1/\delta)}{n}\right)^{1/2}}$ , we get that with probability at least  $1 - \delta$ ,

$$\|\mu(P) - \hat{\mu}_n\|_2 \lesssim \text{OPT}_{n,\Sigma,\delta} + \sqrt{\frac{\text{trace}(\Sigma) \log(p/\delta)}{n}} \quad (53)$$

Combining the above equation and 52, we recover the corollary statement.  $\square$

## H PROOF OF THEOREMS FOR HEAVY-TAILED LINEAR REGRESSION AND GENERALIZED LINEAR MODELS

### H.1 COMMON PROOF TEMPLATE FOR THEOREMS 2 AND 4

We follow the template provided by Prasad et al. (2020) to prove the corollaries appearing in this section.

- In particular, given a distribution  $z \sim P$ , and a loss function  $\bar{\mathcal{L}}(\theta, z)$ , we look at the distribution of the gradients  $\nabla \bar{\mathcal{L}}(\theta^t, z)$  for any  $\theta^t$ , and in particular calculate the trace and operator norm of the covariance of gradients  $\Sigma(\bar{\mathcal{L}}(\theta^t, z))$ . We show that for linear regression and GLMs, they are of the form:

$$\text{trace}(\Sigma(\bar{\mathcal{L}}(\theta^t, z))) \leq A\|\theta^t - \theta^*\|_2^2 + B \quad (54)$$

$$\|\Sigma(\bar{\mathcal{L}}(\theta^t, z))\|_2 \leq C\|\theta^t - \theta^*\|_2^2 + D \quad (55)$$

- From Theorem 1, we know that given  $n$  samples the output of Filterpd satisfies the guarantee that with probability at least  $1 - \delta$ ,

$$\begin{aligned} & \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta^t, z)] - \text{Filterpd}(\{\nabla \bar{\mathcal{L}}(\theta^t, z_i)\}_{i=1}^n)\|_2 \\ & \leq \sqrt{\frac{\text{trace}(\Sigma(\bar{\mathcal{L}}(\theta^t, z)))}{n}} + \sqrt{\frac{\|\Sigma(\bar{\mathcal{L}}(\theta^t, z))\|_2 \log(1/\delta)}{n}} \end{aligned}$$

or equivalently from (54) and (55),

$$\begin{aligned} & \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta^t, z)] - \text{Filterpd}(\{\nabla \bar{\mathcal{L}}(\theta^t, z_i)\}_{i=1}^n)\|_2 \\ & \leq \left( \sqrt{\frac{A}{n}} + \sqrt{\frac{C}{n}} \right) \|\theta^t - \theta^*\|_2 + \left( \sqrt{\frac{B}{n}} + \sqrt{\frac{D \log(1/\delta)}{n}} \right) \end{aligned}$$

- The last step is to use the following result from Prasad et al. (2020) on the stability of gradient descent with inexact gradients.

**Lemma 9** (Prasad et al. (2020)). *For a given sample-size  $n$  and confidence parameter  $\delta \in (0, 1)$ , suppose we have a gradient estimator  $g(\theta; \{\nabla \bar{\mathcal{L}}(\theta, z_i)\}_{i=1}^n, \delta)$  such that for any fixed  $\theta \in \Theta$ , the estimator satisfies the following inequality:*

$$\|g(\theta; \{\nabla \bar{\mathcal{L}}(\theta, z_i)\}_{i=1}^n, \delta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta, z)]\|_2 \leq \alpha(n, \delta) \|\theta - \theta^*\|_2 + \beta(n, \delta). \quad (56)$$

Then Algorithm 2 initialized at  $\theta^0$  with step-size  $\eta = 2/(\tau_\ell + \tau_u)$ , returns iterates  $\{\hat{\theta}^t\}_{t=1}^T$  such that with probability at least  $1 - \delta$

$$\|\hat{\theta}^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1 - \kappa} \beta(\tilde{n}, \tilde{\delta}), \quad (57)$$

where  $\tilde{n} = n/T$ ,  $\tilde{\delta} = \delta/T$ ,  $\kappa = \sqrt{1 - \frac{2\eta\tau_\ell\tau_u}{\tau_\ell + \tau_u}} + \eta\alpha(\tilde{n}, \tilde{\delta}) < 1$  is a contraction and  $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\tilde{\mathcal{L}}(\theta, z)]$  is the minimizer of the population loss.

- Using the above we get that

$$\|\hat{\theta}^t - \theta^*\|_2 \lesssim \kappa^t \|\theta^0 - \theta^*\|_2 + \sqrt{\frac{B}{(n/T)}} + \sqrt{\frac{D \log(T/\delta)}{(Tn)}}, \quad (58)$$

as long as  $\alpha(\tilde{n}, \tilde{\delta}) < \tau_\ell$ .

- Hence, all that remains is to calculate  $(A, B, C, D)$  for linear regression and GLMs.

## H.2 PROOF OF THEOREM 2

In this section we simply focus on deriving upper bounds for the gradient distribution for Linear Regression. This result can also be found in Prasad et al. (2020), but we provide it for the sake of completeness. Recall that for linear regression we have that,  $\tilde{\mathcal{L}}(\theta, (x, y)) = \frac{1}{2}(y - x^T \theta)^2$ .

**Lemma 10** (Prasad et al. (2020)). *Consider the model in (4). Suppose the covariates  $x \in \mathbb{R}^p$  have bounded  $8^{\text{th}}$ -moments and the noise  $w$  has bounded  $4^{\text{th}}$  moments. The following statements hold true:*

$$\begin{aligned} \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)] &= \Sigma \Delta \\ \operatorname{trace}(\operatorname{Cov}(\nabla \tilde{\mathcal{L}}(\theta))) &\leq \underbrace{C_4 \operatorname{trace}(\Sigma) \|\Sigma\|_2 \|\Delta\|_2^2}_A + \underbrace{\sigma^2 \operatorname{trace}(\Sigma)}_B, \\ \|\operatorname{Cov}(\nabla \tilde{\mathcal{L}}(\theta))\|_2 &\leq \|\Delta\|_2^2 \underbrace{C_1 \|\Sigma\|_2^2}_C + \underbrace{\sigma^2 \|\Sigma\|_2}_D \\ \mathbb{E} \left[ [(\nabla \tilde{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)])^T v]^4 \right] &\leq C_2 (\operatorname{Var}[\nabla \tilde{\mathcal{L}}(\theta)^T v])^2 \end{aligned}$$

where  $\Delta = \theta - \theta^*$  and  $\mathbb{E}[xx^T] = \Sigma$ .

From the above lemma, we recover the values of  $(A, B, C, D)$  for linear regression which we simply plug into (58) to recover the statement of the corollary.

*Proof.* We start by deriving the results for  $\mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)]$ .

$$\begin{aligned} \tilde{\mathcal{L}}(\theta) &= \frac{1}{2}(y - x^T \theta)^2 = \frac{1}{2}(x^T(\Delta) - w)^2 \\ \nabla \tilde{\mathcal{L}}(\theta) &= xx^T \Delta - x.w \\ \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)] &= \Sigma \Delta. \end{aligned}$$

Next, we bound the operator norm of the covariance of the gradients  $\nabla \tilde{\mathcal{L}}(\theta)$  at any point  $\theta$ . Recall the definition of covariance below:

$$\operatorname{Cov}(\nabla \tilde{\mathcal{L}}(\theta)) = \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta) \nabla \tilde{\mathcal{L}}(\theta)^T] - \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)] \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)]^T$$

For any unit vector  $z \in \mathcal{S}^{p-1}$ , we have that,

$$\begin{aligned} z^T \operatorname{Cov}(\nabla \tilde{\mathcal{L}}(\theta)) z &= z^T \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta) \nabla \tilde{\mathcal{L}}(\theta)^T] z - (\mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta)]^T z)^2 \\ &\leq z^T \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta) \nabla \tilde{\mathcal{L}}(\theta)^T] z \\ \implies \sup_{z \in \mathcal{S}^{p-1}} z^T \operatorname{Cov}(\nabla \tilde{\mathcal{L}}(\theta)) z &\leq \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[\nabla \tilde{\mathcal{L}}(\theta) \nabla \tilde{\mathcal{L}}(\theta)^T] z \end{aligned}$$

Hence, we have that

$$\begin{aligned}
\lambda_{\max}(\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &\leq \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z \\
&= \sup_{z \in \mathcal{S}^{p-1}} z^T \mathbb{E}[(xx^T \Delta - x.w)(xx^T \Delta - x.w)^T] z \\
&= \sup_{z \in \mathcal{S}^{p-1}} z^T (\mathbb{E}[xx^T \Delta \Delta^T xx^T] + \sigma^2 \mathbb{E}[xx^T]) z \\
&\leq \sup_{z \in \mathcal{S}^{p-1}} z^T (\mathbb{E}[xx^T \Delta \Delta^T xx^T]) z + \sigma^2 \|\Sigma\|_2 \\
&\leq \sigma^2 \|\Sigma\|_2 + \|\Delta\|_2^2 \sup_{y, z \in \mathcal{S}^{p-1}} \mathbb{E}[(z^T x)^2 (y^T z)^2] \\
&\leq \sigma^2 \|\Sigma\|_2 + \|\Delta\|_2^2 \sup_{y, z \in \mathcal{S}^{p-1}} \sqrt{\mathbb{E}[(y^T x)^4]} \sqrt{\mathbb{E}[(z^T x)^4]} \\
&\leq \sigma^2 \|\Sigma\|_2 + \|\Delta\|_2^2 C_4 \|\Sigma\|_2^2
\end{aligned}$$

where the second last step follows from Cauchy-Schwarz and the last step follows from our assumption of bounded 4<sup>th</sup> moment - the constant  $C_4$  is from the definition of bounded 4<sup>th</sup> moment. Now to bound the trace of the covariance matrix,

$$\begin{aligned}
\text{Cov}(\nabla \bar{\mathcal{L}}(\theta)) &= \mathbb{E}[(xx^T - \Sigma)\Delta - xw] (xx^T - \Sigma)\Delta - xw)^T] \\
\text{trace}(\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &= \mathbb{E}[\|(xx^T - \Sigma)\Delta - xw\|_2^2] \\
&= \underbrace{\mathbb{E}[\|(xx^T - \Sigma)\Delta\|_2^2]}_{T_1} + \underbrace{\mathbb{E}[\|x\|_2^2 w^2]}_{\sigma^2 \text{trace}(\Sigma)} \\
T_1 &= \mathbb{E}[\|(xx^T - \Sigma)\Delta\|_2^2] = \Delta^T \mathbb{E}[(xx^T - \Sigma)^2] \Delta \\
&= \Delta^T \mathbb{E}[(x^T x)xx^T + \Sigma^2 - \Sigma xx^T - xx^T \Sigma] \Delta \\
&= \Delta^T \mathbb{E}[(x^T x)xx^T] \Delta - \Delta^T \Sigma^2 \Delta \\
&\leq \Delta^T \mathbb{E}[(x^T x)xx^T] \Delta \\
&\leq \|\Delta\|_2^2 \mathbb{E}[(x^T x)(x^T u)^2], \text{ where } u = \frac{\Delta}{\|\Delta\|_2} \in \mathcal{S}^{p-1} \\
&\leq \|\Delta\|_2^2 \mathbb{E}[(x^T x)^2]^{1/2} \underbrace{\mathbb{E}[(x^T u)^4]^{1/2}}_{\leq \sqrt{C_4} \|\Sigma\|_2} \\
x &\stackrel{\text{def}}{=} \sum_{i=1}^p \underbrace{(x^T q_i)}_{\nu_i} q_i, \text{ where } \{q_i\}_{i=1}^p \text{ are eigenvectors of } \Sigma \\
\mathbb{E}[(x^T x)(x^T x)] &= \mathbb{E}[(\sum_i \nu_i^2)(\sum_i \nu_i^2)] \\
&= \mathbb{E}[\sum_i \nu_i^4 + 2 \sum_{i < j} \nu_i^2 \nu_j^2] \\
\mathbb{E}[\nu_i^4] &= \mathbb{E}[(x^T q_i)^4] \leq C_4 \mathbb{E}[(x^T q_i)^2]^2 = C_4 \lambda_i^2 \\
\mathbb{E}[\nu_i^2 \nu_j^2] &\leq \sqrt{\mathbb{E}[\nu_i^4]} \sqrt{\mathbb{E}[\nu_j^4]} = C_4 \lambda_i \lambda_j \\
\mathbb{E}[(x^T x)(x^T x)] &\leq C_4 \left( \sum_i \lambda_i^2 + 2 \sum_{i < j} \lambda_i \lambda_j \right) = C_4 \text{trace}(\Sigma)^2 \\
\text{trace}(\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &\leq \sigma^2 \text{trace}(\Sigma) + C_4 \text{trace}(\Sigma) \|\Sigma\|_2 \|\Delta\|_2^2
\end{aligned}$$

We finally show that the gradients have bounded  $4^{th}$  moment under the conditions specified in the statement of the theorem. We start from the LHS:

$$\begin{aligned}
\mathbb{E} \left[ |(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v|^4 \right] &\leq \mathbb{E} \left[ |(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v|^4 \right] \\
&= \mathbb{E} \left[ |((xx^T - \Sigma)\Delta - wx)^T v|^4 \right] \\
&= \mathbb{E} \left[ |(\Delta^T x)(x^T v) - (\Sigma \Delta)^T v - wv^T x|^4 \right] \\
&\leq 8 \left[ \underbrace{8 \mathbb{E} |(\Delta^T x)(x^T v)|^4}_A + \underbrace{\mathbb{E} |(\Sigma \Delta)^T v|^4}_B + \underbrace{\mathbb{E} |w(x^T v)|^4}_C \right].
\end{aligned}$$

The last step follows from two applications of the following inequality:

**$C_r$  inequality** If  $X$  and  $Y$  are random variables such that  $\mathbb{E}|X|^r < \infty$  and  $\mathbb{E}|Y|^4 < \infty$  where  $r \geq 1$  then:

$$\mathbb{E}|X + Y|^r \leq 2^{r-1} (\mathbb{E}|X|^r + \mathbb{E}|Y|^r).$$

Now to control each term:

- **Control of  $A$ .** Using Cauchy-Schwarz and the fact that  $C_8$  is bounded for  $x$ , where  $C_8$  is the constant appearing the definition of bounded  $8^{th}$  moment, we get:

$$A \leq \sqrt{\mathbb{E}|\Delta^T x|^8} \sqrt{\mathbb{E}|x^T v|^8} \quad (59)$$

$$\lesssim \|\Delta\|_2^4 C_8 \|\Sigma\|_2^4. \quad (60)$$

- **Control of  $B$ .** Using the fact that  $|\Sigma \Delta^T v| \leq \|\Sigma\|_2 |\Delta^T v| \leq \|\Sigma\|_2 \|\Delta\|_2$ , we get:

$$B \lesssim \|\Delta\|_2^4 \|\Sigma\|_2^4$$

- **Control of  $C$ .** Since  $w$  and  $x$  are independent, and have bounded moments, we can bound  $C$  as:

$$C \lesssim C_4 \|\Sigma\|_2^2$$

Therefore the  $\mathbb{E} \left[ |(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v|^4 \right] \lesssim c + \|\Sigma\|_2^4 \|\Delta\|_2^4$ .

For the RHS:

$$\text{Var}(\nabla \bar{\mathcal{L}}(\theta)^T v)^2 = (v^T \text{Cov}(\nabla \bar{\mathcal{L}}(\theta)) v)^2 \leq \|\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2^2$$

We saw that the  $\|\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2 \lesssim c + \|\Sigma\|_2^2 \|\Delta\|_2^2$ , so both the LHS and RHS scale with  $\|\Sigma\|_2^4 \|\Delta\|_2^4$ , and this completes the proof.  $\square$

### H.3 PROOF OF THEOREM 4

In this section we simply focus on deriving upper bounds for the gradient distribution for GLMs. This result can also be found in Prasad et al. (2020), but we provide it for the sake of completeness. Recall that for generalized linear models we have that,

$$\bar{\mathcal{L}}(\theta; (x, y)) = -y \langle x, \theta \rangle + \Phi(\langle x, \theta \rangle). \quad (61)$$

**Lemma 11** (Prasad et al. (2020)). *Consider the model in (7). Under the pre-conditions of the theorem, the following statements hold true:*

$$\begin{aligned}
\text{trace}(\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &\leq \underbrace{\sqrt{C_4} \text{trace}(\Sigma) \sqrt{L_{\Phi,4}}}_{A} \|\Delta\|_2^2 \\
&\quad + \underbrace{\sqrt{C_4} \text{trace}(\Sigma) \left( \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right)}_B \\
\|\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2 &\leq \underbrace{\sqrt{c_1} \sqrt{C_4} \|\Sigma\|_2 \left( \sqrt{L_{\Phi,4}} \right)}_C \|\Delta\|_2^2 \\
&\quad + \underbrace{\sqrt{c_1} \sqrt{C_4} \|\Sigma\|_2 \left( \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right)}_D \\
\mathbb{E} \left[ [(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v]^4 \right] &\leq C_2 (\text{Var}[\nabla \bar{\mathcal{L}}(\theta)^T v])^2
\end{aligned}$$

for some universal constant  $c_1 > 0$ .

From the above lemma, we recover the values of  $(A, B, C, D)$  for GLMs which we simply plug into (58) to recover the statement of the corollary.

### H.3.1 PROOF OF LEMMA 11

*Proof.* The gradient  $\nabla \bar{\mathcal{L}}(\theta)$  and it's expectation can be written as:

$$\begin{aligned}
\nabla \bar{\mathcal{L}}(\theta) &= -y \cdot x + u(\langle x, \theta \rangle) \cdot x \\
\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] &= \mathbb{E}[x (u(x^T \theta) - u(x^T \theta^*))]
\end{aligned}$$

where  $u(t) = \Phi'(t)$ .

$$\begin{aligned}
\|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2 &= \sup_{y \in \mathbb{S}^{p-1}} y^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \\
&\leq \sup_{y \in \mathbb{S}^{p-1}} \mathbb{E}[(y^T x) (u(x^T \theta) - u(x^T \theta^*))] \\
&\leq \sup_{y \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}[(y^T x)^2]} \sqrt{\mathbb{E}[(u(x^T \theta) - u(x^T \theta^*))^2]} \\
&\leq C_1 \|\Sigma\|_2^{1/2} \sqrt{L_{\Phi,2} \|\Delta\|_2^2 + B_{\Phi,2}}
\end{aligned}$$

where the last line follows from our assumption of smoothness.

Now, to bound the maximum eigenvalue of the  $\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))$ ,

$$\begin{aligned}
\lambda_{\max}(\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &\leq \sup_{z \in \mathbb{S}^{p-1}} z^T \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] z \\
&= \sup_{z \in \mathbb{S}^{p-1}} z^T \left( \mathbb{E} \left[ x x^T (u(x^T \theta) - y)^2 \right] \right) z \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} \mathbb{E} \left[ z^T \left( x x^T (u(x^T \theta) - y)^2 \right) z \right] \\
&\leq \sup_{z \in \mathbb{S}^{p-1}} \sqrt{\mathbb{E}[(z^T x)^4]} \sqrt{\mathbb{E}[(u(x^T \theta) - y)^4]}
\end{aligned}$$



To bound  $\mathbb{E} \left[ (u(x^T \theta) - y)^4 \right]$ , we make use of the  $C_r$  inequality stated earlier.

$$\begin{aligned} \mathbb{E} \left[ (u(x^T \theta) - y)^4 \right] &\leq 8 \left( \mathbb{E} \left[ (u(x^T \theta) - u(x^T \theta^*))^4 \right] + \mathbb{E} \left[ (u(x^T \theta^*) - y)^4 \right] \right) \\ &\leq c_1 (L_{\Phi,4} \|\Delta\|_2^4 + B_{\Phi,4} + c(\sigma)^3 M_{\Phi,4,1} + 3c(\sigma)^2 M_{\Phi,2,2}) \end{aligned}$$

where the last line follows from our assumption that  $P_{\theta^*}(y|x)$  is in the exponential family, hence, the cumulants are higher order derivatives of the log-normalization function.

$$\|\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))\|_2 \leq \sqrt{c_1} \sqrt{C_4} \|\Sigma\|_2 \left( \sqrt{L_{\Phi,4}} \|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right)$$

Now, to control the trace. We have that,

$$\begin{aligned} \text{Cov}(\nabla \bar{\mathcal{L}}(\theta)) &= \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T] - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T \\ \text{trace}(\text{Cov}(\nabla \bar{\mathcal{L}}(\theta))) &= \text{trace}(\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T]) - \text{trace}(\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)] \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T) \\ &\leq \text{trace}(\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta) \nabla \bar{\mathcal{L}}(\theta)^T]) \\ &\leq \text{trace} \left( \mathbb{E} \left[ x x^T (u(x^T \theta) - y)^2 \right] \right) \\ &= \mathbb{E} \left[ \text{trace} \left( x x^T (u(x^T \theta) - y)^2 \right) \right] \\ &= \mathbb{E}[\text{trace}((x x^T) u(x^T \theta) - y)^2] \quad \because (u(x^T \theta) - y)^2 \in \mathbb{R} \\ &\leq \sqrt{\mathbb{E}[\text{trace}((x x^T))^2]} \sqrt{\mathbb{E}[(u(x^T \theta) - y)^4]} \\ &\leq \sqrt{C_4} \text{trace}(\Sigma) \left( \sqrt{L_{\Phi,4}} \|\Delta\|_2^2 + \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right) \\ &= \sqrt{C_4} \text{trace}(\Sigma) \sqrt{L_{\Phi,4}} \|\Delta\|_2^2 \\ &\quad + \sqrt{C_4} \text{trace}(\Sigma) \left( \sqrt{B_{\Phi,4}} + c(\sigma) \sqrt{3M_{\Phi,2,2}} + \sqrt{c(\sigma)^3 M_{\Phi,4,1}} \right) \end{aligned}$$

Finally, we show that the fourth moment of the gradient distribution is bounded. We have:

$$\begin{aligned} \mathbb{E} \left[ [(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v]^4 \right] &\leq \mathbb{E} \left[ |(\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)])^T v|^4 \right] \\ &\leq 8 \left[ \underbrace{\mathbb{E}[|\nabla \bar{\mathcal{L}}(\theta)|^4]}_A + \underbrace{\mathbb{E}[|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]|^4]}_B \right] \end{aligned}$$

To control each term:

- **Control of A.**

$$\begin{aligned} \mathbb{E}[|\nabla \bar{\mathcal{L}}(\theta)|^4 v^4] &= \mathbb{E}[(x^T v)^4 (u(x^T \theta) - y)^4] \\ &\leq \sqrt{\mathbb{E}[(x^T v)^8]} \sqrt{\mathbb{E}[(u(x^T \theta) - y)^8]} \\ &\leq \sqrt{C_8} \|\Sigma\|_2^2 \sqrt{\mathbb{E}[(u(x^T \theta) - u(x^T \theta^*))^8] + \mathbb{E}[(u(x^T \theta^*) - y)^8]} \\ &\leq \sqrt{C_8} \|\Sigma\|_2^2 \sqrt{L_{\Phi,8} \|\Delta\|_2^8 + B_{\Phi,8} + \sum_{t,k=2}^8 g_{t,k} M_{\Phi,t,k}} \\ &\leq \sqrt{C} \|\Sigma\|_2^2 \sqrt{L_{\Phi,8}} \|\Delta\|_2^4 + \sqrt{B_{\Phi,8}} + \sqrt{\sum_{t,k=2}^8 g_{t,k} M_{\Phi,t,k}} \end{aligned}$$

where the last step follows from the fact that the 8th central moment can be written as a polynomial involving the lower cumulants, which in turn are the derivatives of the log-normalization function.

- **Control of B.**

$$\mathbb{E}[|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T v|^4] \leq \|\mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]\|_2^4 \leq C_1 \|\Sigma\|_2^2 (L_{\Phi,2}^2 \|\Delta\|_2^2 + B_{\Phi,2}^2)$$

By assumption  $L_{\Phi,k}, B_{\Phi,k}, M_{\Phi,t,k}$  are all bounded for  $k, t \leq 8$ , which implies that there exist constants  $c_1, c_2 > 0$  such that

$$\mathbb{E} \left[ \left[ (\nabla \bar{\mathcal{L}}(\theta) - \mathbb{E}[\nabla \bar{\mathcal{L}}(\theta)]^T v \right]^4 \right] \leq c_1 \|\Sigma\|_2^2 \|\Delta\|_2^4 + c_2$$

Previously, we say that  $\|\text{Cov} \nabla \bar{\mathcal{L}}(\theta)\|_2 \leq c_3 \|\Sigma\|_2 \|\Delta\|_2^2 + c_4$ , for some universal constants  $c_3, c_4 > 0$ , hence the gradient  $\nabla \bar{\mathcal{L}}(\theta)$  has bounded fourth moments.

□