A. Setting Constants in Section 3

In this section, we describe how to appropriately set the universal constants $c_1, \ldots, c_4 \geq 1$ in Section 3. These constants are set in the following order: c_1, c_3, c_4, c_2. In this order, each c_i only depends on the constants set before it, and there is only a lower bound requirement on the value of each c_i so we can set c_i to a sufficiently large constant.

The constant c_1 appears in Condition (3). and is related to the constants involved in the concentration inequalities required to establish this condition. With the right sample complexity, Condition (3) holds with high probability for $\delta = c_1 \epsilon \ln(1/\epsilon)$.

For the remaining three constants, recall that by assumption $r = \|\mu_w - \mu^\star\|_2 \geq c_2 \epsilon \sqrt{\ln(1/\epsilon)} \geq \epsilon \sqrt{\ln(1/\epsilon)}$.

Next we choose c_3 such that $c_3 \geq 5c_1$. This is to guarantee that, in the proof of Lemma 3.5, we have $2c_1 \ln(1/\epsilon) + 2\sqrt{2c_1 \ln(1/\epsilon)} \cdot r \leq c_3 \cdot \frac{r^2}{\epsilon^2}$.

The constant c_4 appears in the proof of Lemma 3.4. There are two inequalities related to c_4. We need $c_4 \geq 50c_1$ so that $\frac{c_4 r^2}{\epsilon} - c_1 \epsilon \ln(1/\epsilon) \geq 0.98 \cdot c_4 \cdot \frac{r^2}{\epsilon}$, and we require $c_4 \geq \max(100, 6c_3)$ so that $\frac{0.49 \cdot c_4 \cdot r^2}{\epsilon^2} - \frac{1.4 \cdot \sqrt{c_4} \cdot r^2}{\epsilon} > 2c_3 \cdot \frac{r^2}{\epsilon^2}$.

Finally, we set the value of c_2, which appears in our final guarantee: we show that any stationary point w of $f(w)$ satisfies $\|\mu_w - \mu^\star\|_2 \leq c_2 \epsilon \sqrt{\ln(1/\epsilon)}$. The constant c_2 only depends on c_4. At the beginning of the proof of Lemma 3.4, we need that if $\|\mu_w - \mu^\star\|_2 \geq c_2 \epsilon \sqrt{\ln(1/\epsilon)}$, then $\|\Sigma_w\|_2 \geq 1 + c_4 \cdot \frac{r^2}{\epsilon}$. By Lemma 2.1 from (Diakonikolas et al., 2016), we know that this is possible if we set c_2 to be sufficiently large.

B. Missing Proofs from Section 4

In this section, we prove Theorem 3.2 and Lemma 4.2 from Section 4. These two statements play an important role in showing that projected sub-gradient descent efficiently finds an approximate stationary point w, and that w is a good solution to our robust mean estimation task.

We briefly recall our notation. We use $X \in \mathbb{R}^{d \times N}$ to denote the sample matrix, $\Sigma_w = \left(X \operatorname{diag}(w) X^\top - X w w^\top X^\top\right)$, $F(w,u) = u^\top \Sigma_w u$, $f(w) = \max_u F(w,u) = \|\Sigma_w\|_2$, and $\Delta_{N,\epsilon} = \left\{w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1-\epsilon)N} \forall i\right\}$.

Note that we can assume without loss of generality that no input samples have very large ℓ_2-norm. This is because we can perform a standard preprocessing step that centers the input samples at the coordinate-wise median, which does not affect our mean estimation task. We can then throw away all samples that are $\Omega(\sqrt{d \log d})$ far from the coordinate-wise median. With high probability, the coordinate-wise median of all good samples are $O(\sqrt{d \log d})$ far from the true mean. Assuming this happens, then no good samples are thrown away and the remaining samples satisfy $\max_i \|X_i\|_2 = O(\sqrt{d \log d})$. Consequently, we have $\|\mu_w\|_2 = O(\sqrt{d \log d})$ for any $w \in \Delta_{N,\epsilon}$.

In Lemma B.1, we show that the function $F(w,u) = u^\top \Sigma_w u$ is Lipschitz and smooth with respect to w.

Lemma B.1. *The function $F(w,u)$ is L-Lipschitz and β-smooth for $L = \widetilde{O}(\sqrt{N}d)$ and $\beta = \widetilde{O}(Nd)$. That is,*

$$|F(w,u) - F(\widetilde{w},u)| \leq L \|\widetilde{w} - w\|_2 \quad \textit{for all } w, \widetilde{w}, \in \Delta_{N,2\epsilon} \textit{ and all unit vectors } u \in \mathbb{R}^d$$
$$\|\nabla_w F(w,u) - \nabla_w F(\widetilde{w},u)\|_2 \leq \beta \|\widetilde{w} - w\|_2 \quad \textit{for all } w, \widetilde{w}, \in \Delta_{N,2\epsilon} \textit{ and all unit vectors } u \in \mathbb{R}^d .$$

Proof. We use the ℓ_2-norm of the gradient to bound L from above. We have

$$\begin{aligned}
\|\nabla_w F(w,u)\|_2 &= \left\|X^\top u \odot X^\top u - 2(u^\top X w) X^\top u\right\|_2 \\
&\leq \sqrt{N} \max_i (X_i^\top u)^2 + 2 \left\|u^\top X\right\|_\infty \|w\|_1 \|X\|_2 \|u\|_2 \\
&\leq \sqrt{N} \max_i \|X_i\|_2^2 + 2 \max_i \|X_i\|_2 \|X\|_2 \ .
\end{aligned}$$

To bound from above the smoothness parameter, we have

$$\|\nabla_w F(w,u) - \nabla_w F(\widetilde{w},u)\|_2 = 2 \left|u^\top X (w - \widetilde{w})\right| \left\|X^\top u\right\|_2 \leq 2 \|X\|_2^2 \|w - \widetilde{w}\|_2 \ .$$

We conclude the proof by observing that, after the preprocessing step, we have $\max_i \|X_i\|_2 = O(\sqrt{d \log d})$ and consequently $\|X\|_2 = O(\sqrt{Nd \log d})$. Therefore, $L = O(\sqrt{N} d \log d)$ and $\beta = O(Nd \log d)$. □

Recall that the Moreau envelope $f_\beta(w)$ is defined as

$$f_\beta(w) = \min_{\widetilde{w}} \mathcal{I}_{\mathcal{K}}(\widetilde{w}) + F(\widetilde{w}) + \beta \left\|\widetilde{w} - w\right\|_2^2 = \min_{\widetilde{w} \in \mathcal{K}} f(\widetilde{w}) + \beta \left\|\widetilde{w} - w\right\|_2^2 \ ,$$

where $\mathcal{I}_{\mathcal{K}}(\cdot)$ is the support function of $\mathcal{K}$.

We restate Theorem 3.2 before proving it.

Theorem 3.2. *Consider the spectral norm loss $f(w) = \|\Sigma_w\|_2$ with $f_\beta(w)$ denoting the corresponding Moreau envelope function per Definition 2.4 with $\beta = 2\|X\|_2^2$. Then, for any $w \in \Delta_{N,2\epsilon}$ obeying*

$$\left\|\nabla f_\beta(w)\right\|_2 = O(\log(1/\epsilon)),$$

we have $\left\|\mu_w - \mu^\star\right\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.

Proof. Let $\delta = \frac{c_3 c_2^2 \ln(1/\epsilon)}{\sqrt{2}}$, where c_2 and c_3 are the positive universal constants from Lemma 3.4. We show that any $w \in \Delta_{N,2\epsilon}$ obeying $\left\|\nabla f_\beta(w)\right\|_2 \le \delta$ must satisfy that $\left\|\mu_w - \mu^\star\right\|_2 \le O(\epsilon\sqrt{\log(1/\epsilon)})$.

The condition $\left\|\nabla f_\beta(w)\right\|_2 \le \delta$ implies that there exists a vector $\widehat{w}$ such that (see, e.g., Rockafellar (2015)):

$$\left\|\widehat{w} - w\right\|_2 = \frac{\delta}{2\beta} \quad \text{and} \quad \min_{g \in \partial f(\widehat{w}) + \partial \mathcal{I}_{\mathcal{K}}(\widehat{w})} \|g\|_2 \le \delta \ .$$

We first show that $\widehat{w}$ is a good solution.

It is well known that the subdifferential of the support function is the normal cone, which is in turn the polar of the tangent cone. That is,

$$\partial \mathcal{I}_{\mathcal{K}}(\widehat{w}) = \mathcal{N}_{\mathcal{K}}(\widehat{w}) = (\mathcal{C}_{\mathcal{K}}(\widehat{w}))^\circ \ .$$

Thus, there exists a vector $g = \nu + v$ with $\|g\|_2 \le \delta$ such that $\nu \in \partial f(\widehat{w})$ and $v \in (\mathcal{C}_{\mathcal{K}}(\widehat{w}))^\circ$. Now consider any unit vector $u \in \mathcal{C}_{\mathcal{K}}(\widehat{w})$:

$$-\delta \le u^\top g = u^\top \nu + u^\top v \le u^\top \nu \ ,$$

where the last step follows from the definition of the polar set. In other words, there exists a vector $\nu \in \partial f(\widehat{w})$ such that

$$-\nu^\top u \le \delta \quad \text{for all unit vectors } \ u \in \mathcal{C}_{\mathcal{K}}(\widehat{w}) \ . \tag{5}$$

Suppose $\left\|\mu_{\widehat{w}} - \mu^\star\right\|_2 \ge c_2\epsilon\sqrt{\ln(1/\epsilon)}$. Then for the $v \in \partial f(\widehat{w})$ in question, we can use Lemmas 3.4 and 3.5 to find two coordinates i and j such that

$$\widehat{w}_i > 0, \ \widehat{w}_j < \frac{1}{(1-2\epsilon)N}, \ \text{ and } \ \nu_i - \nu_j > c_3 \frac{\left\|\mu_{\widehat{w}} - \mu^\star\right\|_2^2}{\epsilon^2} \ge c_3 c_2^2 \ln(1/\epsilon) = \sqrt{2}\delta \ .$$

However, this contradicts Condition (5), because for the unit vector $u = \frac{1}{\sqrt{2}}(e_j - e_i)$, where e_i is the i-th basis vector, we have $u \in \mathcal{C}_{\Delta_{N,2\epsilon}}(\widehat{w})$ but

$$-\nu^\top u = \frac{\nu_i - \nu_j}{\sqrt{2}} > \delta \ .$$

Therefore, $\widehat{w}$ must satisfy $\left\|\mu_{\widehat{w}} - \mu^\star\right\|_2 < c_2\epsilon\sqrt{\ln(1/\epsilon)}$.

We conclude the proof by noticing that w is very close to $\widehat{w}$, so if $\widehat{w}$ is a good solution, then w must also be a good solution:

$$\begin{aligned}
\left\|\mu_w - \mu^\star\right\|_2 &\le \left\|\mu_w - \mu_{\widehat{w}}\right\|_2 + \left\|\mu_{\widehat{w}} - \mu^\star\right\|_2 \\
&\le \|X\|_2 \left\|w - \widehat{w}\right\|_2 + c_2\epsilon\sqrt{\ln(1/\epsilon)} \\
&= O(\beta^{-1/2}\delta + \epsilon\sqrt{\log(1/\epsilon)}) = O(\epsilon\sqrt{\log(1/\epsilon)}) \ .
\end{aligned}$$

In the last two steps, we used the fact that $\left\|\widehat{w} - w\right\|_2 = \frac{\delta}{2\beta}$ and $\beta = 2\left\|X\right\|_2^2$ (see Lemma B.1). This completes the proof of Theorem 3.2. □

We restate Lemma 4.2 before proving it. We note that the proof of Lemma 4.2 is directly inspired by the proof of Theorem 2.1 in (Davis & Drusvyatskiy, 2018).

Lemma 4.2. *Let $\mathcal{K}$ be a closed convex set. Let $F(w,u)$ be a function which is L-Lipschitz and β-smooth with respect to w. Consider the following optimization problem $\min_{w\in\mathcal{K}}\max_{\|u\|_2=1} F(w,u)$.*

Starting from any initial point $w_0\in\mathcal{K}$, we run iterative updates of the form:

$$\begin{aligned}&\text{Find } u_\tau \text{ with } F(w_\tau,u_\tau)\geq(1-\epsilon')\max_u F(w_\tau,u_\tau);\\ &w_{\tau+1}=\mathcal{P}_{\mathcal{K}}(w_\tau-\eta\nabla_w F(w_\tau,u_\tau)\ ,\end{aligned}$$

for T iterations with step size $\eta=\frac{\gamma}{\sqrt{T}}$. Then, we have

$$\begin{aligned}&\min_{0\leq\tau<T}\|\nabla f_\beta(w_\tau)\|_2^2\\ &\quad\leq\frac{2}{\sqrt{T}}\left(\frac{f_\beta(w_0)-\min_w f(w)}{\gamma}+\gamma\beta L^2\right)+4\beta\epsilon'\ ,\end{aligned}$$

where $f_\beta(w)$ is the Moreau envelope, as in Definition 2.4.

Proof. Note that since f is β-smooth with respect to w and u_τ is an approximate maximizer for w_τ, for any $\widetilde{w}\in\mathcal{K}$, we have that

$$\begin{aligned}f(\widetilde{w})\geq F(\widetilde{w},u_\tau)&\geq F(w_\tau,u_\tau)+(\nabla_w F(w_\tau,u_\tau))^\top(\widetilde{w}-w_\tau)-\frac{\beta}{2}\|\widetilde{w}-w_\tau\|_2^2\\ &\geq f(w_\tau)-\epsilon'+(\nabla_w F(w_\tau,u_\tau))^\top(\widetilde{w}-w_\tau)-\frac{\beta}{2}\|\widetilde{w}-w_\tau\|_2^2\ .\end{aligned}\tag{6}$$

To continue, define the proximal function

$$\mathrm{prox}_{f_\beta}(w)=\arg\min_{\widetilde{w}\in\mathcal{K}}\left(f(\widetilde{w})+\beta\|\widetilde{w}-w\|_2\right)\ ,$$

and let $\widehat{w}_\tau=\mathrm{prox}_{f_\beta}(w_\tau)$.

Now we have

$$\begin{aligned}f_\beta(w_{\tau+1})&\leq f(\widehat{w}_\tau)+\beta\|\widehat{w}_\tau-w_{\tau+1}\|_2\\ &=f(\widehat{w}_\tau)+\beta\|\widehat{w}_\tau-\Pi_{\mathcal{K}}(w_\tau-\eta\nabla_w F(w_\tau,u_\tau))\|_2\\ &\leq f(\widehat{w}_\tau)+\beta\|\widehat{w}_\tau-w_\tau+\eta\nabla_w F(w_\tau,u_\tau)\|_2 && \text{(convexity of }\mathcal{K})\\ &=f(\widehat{w}_\tau)+\beta\|\widehat{w}_\tau-w_\tau\|_2^2+2\eta\beta(\nabla_w F(w_\tau,u_\tau))^\top(\widehat{w}_\tau-w_\tau)+\eta^2\beta\|\nabla_w F(w_\tau,u_\tau)\|_2^2\\ &=f_\beta(w_\tau)+2\eta\beta(\nabla_w F(w_\tau,u_\tau))^\top(\widehat{w}_\tau-w_\tau)+\eta^2\beta\|\nabla_w F(w_\tau,u_\tau)\|_2^2 && (\widehat{w}_\tau=\mathrm{prox}_{f_\beta}(w_\tau))\\ &\leq f_\beta(w_\tau)+2\eta\beta(\nabla_w F(w_\tau,u_\tau))^\top(\widehat{w}_\tau-w_\tau)+\eta^2\beta L^2 && (F(w,u)\text{ is }L\text{-Lipschitz in }w)\\ &\leq f_\beta(w_\tau)+2\eta\beta\left(f(\widehat{w}_\tau)-f(w_\tau)+\epsilon'+\frac{\beta}{2}\|\widehat{w}_\tau-w_\tau\|_2^2\right)+\eta^2\beta L^2\ . && \text{(by Inequality (6))}\end{aligned}$$

Summing the above over τ, we obtain

$$f_\beta(w_T)\leq f_\beta(w_0)+2\eta\beta\sum_{\tau=0}^{T-1}\left(f(\widehat{w}_\tau)-f(w_\tau)+\frac{\beta}{2}\|\widehat{w}_\tau-w_\tau\|_2^2\right)+\eta^2\beta L^2T+2\eta\beta T\epsilon'\ .$$

Dividing by $2\eta\beta T$, we get

$$\begin{aligned}\frac{1}{T}\sum_{\tau=0}^{T-1}\left(f(w_\tau)-f(\widehat{w}_\tau)-\frac{\beta}{2}\|\widehat{w}_\tau-w_\tau\|_2^2\right)&\leq\frac{f_\beta(w_0)-f_\beta(w_T)}{2\eta\beta T}+\frac{\eta L^2}{2}+\epsilon'\\ &\leq\frac{f_\beta(w_0)-\min_w f(w)}{2\eta\beta T}+\frac{\eta L^2}{2}+\epsilon'\ .\end{aligned}$$

Observe that the function $w \to f(w) + \beta \|w - w_\tau\|_2^2$ is β-strongly convex, therefore

$$\begin{aligned}
&f(w_\tau) - f(\widehat{w}_\tau) - \frac{\beta}{2}\|\widehat{w}_\tau - w_\tau\|_2^2 \\
&= \left(f(w_\tau) + \beta\|w_\tau - w_\tau\|_2^2\right) - \left(f(\widehat{w}_\tau) + \beta\|w_\tau - \widehat{w}_\tau\|_2^2\right) + \frac{\beta}{2}\|w_\tau - \widehat{w}_\tau\|_2^2 \\
&\geq \frac{\beta}{2}\|\widehat{w}_\tau - w_\tau\|_2^2 + \frac{\beta}{2}\|\widehat{w}_\tau - w_\tau\|_2^2 \qquad \text{(strong convexity)} \\
&= \beta\|\widehat{w}_\tau - w_\tau\|_2^2 = \frac{1}{4\beta}\|\nabla f_\beta(w_\tau)\|_2^2 \ .
\end{aligned}$$

In the above, we used the fact that for a β-strongly convex function $h(w) = \mathcal{I}_{\mathcal{K}}(w) + f(w) + \beta\|w - w_\tau\|_2^2$, we have $g(w_\tau) - g(\widehat{w}_\tau) \geq \frac{\beta}{2}\|w_\tau - \widehat{w}_\tau\|_2^2$.

Combining the two inequalities above, we arrive at

$$\frac{1}{T}\sum_{\tau=0}^{T-1}\|\nabla f_\beta(w_\tau)\|_2^2 \leq 2\frac{f_\beta(w_0) - \min_w f(w)}{\eta T} + 2\eta\beta L^2 + 4\beta\epsilon' \ .$$

Finally, setting the step size $\eta = \frac{\gamma}{\sqrt{T}}$, we conclude that

$$\min_{0\leq\tau<T}\|\nabla f_\beta(w_\tau)\|_2^2 \leq \frac{2}{\sqrt{T}}\left(\frac{f_\beta(w_0) - \min_w f(w)}{\gamma} + \gamma\beta L^2\right) + 4\beta\epsilon' \ .$$

This completes the proof of Lemma 4.2. □

C. Minimizing Softmax of Spectral Norm

In this section, we analyze our alternate non-convex formulation that replaces the spectral norm with a softmax. Note that when the largest eigenvalue of Σ_w is not unique, the spectral norm of Σ_w may not be differentiable with respect to w. Instead of considering sub-gradients, we can minimize the softmax of the eigenvalues of Σ_w, which is a smoothed version of spectral norm that is differentiable everywhere.

Formally, we minimize the following non-convex objective function:

$$f(w) = \mathrm{smax}_\rho(\Sigma_w) = \frac{1}{\rho}\ln\mathrm{tr}(\exp(\rho\Sigma_w)) \qquad \text{for} \qquad \rho = \frac{\ln d}{\epsilon} \ , \tag{7}$$

where $X \in \mathbb{R}^{d\times N}$ is the sample matrix, and $\Sigma_w = X\,\mathrm{diag}(w)X^\top - Xww^\top X^\top$ is the weighted empirical covariance matrix.

The structure of this section is as follows: In Section C.1, we start by recording some useful properties of the softmax objective. In Section C.2, we prove our key structural result for this section (Theorem C.5), establishing that any approximate stationary point w of $f(w)$ provides a good estimate μ_w of the true mean $\mu^\star$. In Section C.3, we present our algorithmic result (Theorem 1.4), which states that we can efficiently find an approximate stationary point of $f(w)$ via projected gradient descent.

C.1. Basic Properties of Softmax

Lemma C.1 (Duality of softmax)**.** *For any $Z \in \mathbb{R}^{n\times n}$ and $\rho > 0$, let $\mathrm{smax}_\rho(Z) := \frac{1}{\rho}\ln\mathrm{tr}(\exp(\rho Z))$. We have the following identity*

$$\mathrm{smax}_\rho(Z) = \max_{Y\in\Delta_{n\times n}}\left(Y \bullet Z - \frac{1}{\rho}Y\bullet\log Y\right) \ .$$

Proof. Fix $Z \in \mathbb{R}^{n\times n}$. Let $f(Y) = Y\bullet Z - \frac{1}{\rho}Y\bullet\log Y$. Using the KKT conditions, we know that when $f(Y)$ is maximized, we have $\frac{\partial f}{\partial Y} = \lambda I$, for some $\lambda \in \mathbb{R}$. Combining this with $\frac{\partial f}{\partial Y} = Z - \frac{1}{\rho}(\log Y + I)$, it follows that $f(Y)$ is maximized at

$$Y^\star = \exp(\rho Z - (\rho\lambda + 1)I) = \frac{\exp(\rho Z)}{\mathrm{tr}(\exp(\rho Z))} \ ,$$

where the second equality holds because $Y^\star \in \Delta_{n\times n}$. One can substitute $Y^\star$ into the definition of $f(Y)$ and verify that $f(Y^\star) = \mathrm{smax}_\rho(Z)$. □

Corollary C.2 (Softmax and max)**.** *For any PSD matrix $Z \in \mathbb{R}^{n\times n}$ and $\rho > 0$, we have that $\lambda_{\max}(Z) \le \mathrm{smax}_\rho(Z) \le \lambda_{\max}(Z) + \frac{\ln n}{\rho}$. Moreover, for $Y = \frac{\exp(\rho Z)}{\mathrm{tr}(\exp(\rho Z))}$, we have that $Y \bullet Z \ge \mathrm{smax}_\rho(Z) - \frac{\ln n}{\rho}$.*

Proof. Observe that

$$\mathrm{smax}_\rho(Z) = \frac{1}{\rho}\ln \mathrm{tr}(\exp(\rho Z)) \ge \frac{1}{\rho}\ln \lambda_{\max}(\exp(\rho Z)) = \lambda_{\max}(Z)\ ,$$

and

$$\mathrm{smax}_\rho(Z) = \frac{1}{\rho}\ln \mathrm{tr}(\exp(\rho Z)) \le \frac{1}{\rho}\ln(n\cdot \lambda_{\max}(\exp(\rho Z))) = \lambda_{\max}(Z) + \frac{\ln n}{\rho}\ .$$

For the second claim, by Lemma C.1, we know that $\mathrm{smax}_\rho(Z) = Y \bullet Z - \frac{1}{\rho} Y \bullet \log Y$. The claim then follows from the fact that $Y \bullet \log Y \ge -\ln n$ for all $Y \in \Delta_{n\times n}$. □

When working with the matrix exponentials in our softmax objective function f, the following chain rule formula will be useful to compute the Hessian of f (see, e.g., (Wilcox, 1967)).

Lemma C.3 (Derivative of matrix exponential)**.** *For a symmetric matrix function $X(t)$ that depends on a scalar t, we have that*

$$\frac{d}{dt}\exp(X(t)) = \int_0^1 \exp(\alpha X(t))\frac{dX(t)}{dt}\exp((1-\alpha)X(t))d\alpha\ .$$

C.2. Structural Result: Any Approximate Stationary Point Suffices

The gradient of our softmax objective function is

$$\nabla f(w) = \mathrm{diag}(X^\top Y X) - 2X^\top Y X w\ , \quad \text{where} \quad Y = \frac{\exp(\rho\Sigma_w)}{\mathrm{tr}(\exp(\rho\Sigma_w))}\ . \tag{8}$$

Notice that $Y \in \Delta_{N\times N}$ is a convex combination of directions. That is, we can write $Y = \sum_{k=1}^d \lambda_k u_k u_k^\top$, where $u_k \in \mathbb{R}^d$ and $\sum_k \lambda_k = 1$. The gradient $\nabla f(w)$ is the same as the gradient of w for the one-dimensional problem, where the input samples are $(X_i^\top Y^{1/2})_{i=1}^N$. Equivalently, $\nabla f(w)$ tries to move w towards minimizing the average variance

$$\sum_k \lambda_k \left(\sum_i w_i (X_i^\top u_k)^2 - \left(\sum_i w_i (X_i^\top u_k)\right)^2\right)$$

of the projections of X along the directions $\{u_k\}$.

The intuition is as follows: The goal is to show that $\lambda_{\max}(\Sigma_w)$ is small at any stationary point w of $\mathrm{smax}_\rho(\Sigma_w)$. Now fix some $w \in \Delta_{N,2\epsilon}$, where $\lambda_{\max}(\Sigma_w)$ is large. Then $\mathrm{smax}_\rho(\Sigma_w)$ must be large. By the duality of softmax, there is a combination of directions Y such that: (1) the one-dimensional samples $(X_i^\top Y^{1/2})_{i=1}^N$ weighted by w have large variance, and (2) the derivative of $\mathrm{smax}_\rho(\Sigma_w)$ is the same as the derivative for minimizing variance on this one-dimensional instance. We proceed by examining this one-dimensional instance, which is easier to analyze. We show that w cannot be a stationary point, because we can always reduce the variance by increasing the weight on one of the good samples and reducing the weight on one of the bad samples.

Formally, we use the following notion of approximate stationarity for our constrained non-convex minimization problem.

Definition C.4. *Fix a convex set $\mathcal{K}$. For $\delta > 0$, we say $x \in \mathcal{K}$ is a δ-stationary point of f if the following condition holds: For any unit vector u where $x + \alpha u \in \mathcal{K}$ for some $\alpha > 0$, we have $u^\top \nabla f(x) \ge -\delta$.*

Our main structural result in this section is the following theorem.

Theorem C.5 (Any stationary point of $f(w)$ is a good solution)**.** *Let S be an ϵ-corrupted set of $N = \widetilde{\Omega}(d/\epsilon^2)$ samples drawn from a d-dimensional Gaussian $\mathcal{N}(\mu^\star, I)$ with unknown mean $\mu^\star$. Suppose S satisfies Condition (3) and Lemma 2.1.*

Let $f(w)$ be the softmax objective as defined in Equation (7). Let $\delta = c\ln(1/\epsilon)$ for some universal constant c. For any $w \in \Delta_{N,2\epsilon}$ that is a δ-stationary point of $f(w)$, we have $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.

Theorem C.5 follows directly from Lemmas C.6, C.7, and C.8.

For the rest of this subsection, we assume the input samples satisfy Condition (3) and Lemma 2.1, and we fix an approximate stationary point $w \in \Delta_{N,2\epsilon}$ of the softmax objective. We establish the following bimodal sub-gradient property which holds at all (approximate) stationary points.

Lemma C.6 (Bimodal sub-gradient property at stationary points)**.** *Fix $w \in \Delta_{N,2\epsilon}$. Let $S_- = \{i : w_i > 0\}$ and $S_+ = \{i : w_i < \frac{1}{(1-2\epsilon)N}\}$ denote the set of coordinates of w that can decrease and increase respectively. If w is a δ-stationary point of $f(w)$, then $\nabla f(w)_i \le \nabla f(w)_j + \sqrt{2}\delta$ for all $i \in S_-$ and $j \in S_+$.*

Proof. Suppose there is some $i \in S_-$ and $j \in S_+$ such that $\nabla f(w)_i > \nabla f(w)_j + \sqrt{2}\delta$.

Consider the unit vector $u = \frac{1}{\sqrt{2}}(e_j - e_i)$, where e_i is the i-th basis vector. We have $w + \alpha u \in \Delta_{N,2\epsilon}$ for $\alpha = \min(w_i, \frac{1}{(1-2\epsilon)N} - w_j) > 0$, but

$$u^\top \nabla f(x) = \frac{\nabla f(w)_j - \nabla f(w)_i}{\sqrt{2}} < -\delta\ ,$$

which violates the assumption that w is a δ-approximate stationary point (Definition C.4). □

At a high level, we prove Theorem C.5 by showing that if μ_w is far from $\mu^\star$, then w violates Lemma C.6. More specifically, if μ_w is far from $\mu^\star$, then there exists a bad sample with index $j \in S_-$ whose gradient is large (Lemma C.7). Meanwhile, the concentration bound in Condition (3) guarantees that there exists a good sample with index $i \in S_+$ whose gradient is small (Lemma C.8).

We frequently use the partial derivative of $f(w)$ with respect to w_i in our analysis:

$$\begin{aligned}\nabla f(w)_i &= X_i^\top Y X_i - 2X_i^\top Y \mu_w \\ &= (X_i - \mu^\star)^\top Y (X_i - \mu^\star) - 2(X_i - \mu^\star)^\top Y(\mu_w - \mu^\star) \\ &\quad + {\mu^\star}^\top Y(\mu^\star - 2\mu_w)\ .\end{aligned}$$

Notice that the last term in $\nabla f(w)_i$ is the same for all i. Since our goal is to identify $i \in S_-$ and $j \in S_+$ such that $\nabla f(w)_i > \nabla f(w)_j$, we can focus on the first two terms.

We have the following lemmas:

Lemma C.7. *Fix $w \in \Delta_{N,2\epsilon}$ and assume that Condition (3) and Lemma 2.1 hold. Let c_2 and c_3 be universal constants. Let $r = \|\mu_w - \mu^\star\|_2$ and suppose $r \ge c_2\epsilon\sqrt{\ln(1/\epsilon)}$. Then, there exists $i \in (B \cap S_-)$ such that*

$$\nabla f(w)_i - {\mu^\star}^\top Y(\mu^\star - 2\mu_w) > 2c_3 \cdot \frac{r^2}{\epsilon^2}\ .$$

Lemma C.8. *Consider the same setting as in Lemma C.7. There exists $j \in (G \cap S_+)$ such that*

$$\nabla f(w)_j - {\mu^\star}^\top Y(\mu^\star - 2\mu_w) \le c_3 \cdot \frac{r^2}{\epsilon^2}\ .$$

We defer the proofs of Lemmas C.7 and C.8 to Section C.2.1, and we first use them to prove Theorem C.5.

Proof of Theorem C.5. Suppose that w is a bad solution where $\|\mu_w - \mu^\star\|_2 \ge c_2\epsilon\sqrt{\ln(1/\epsilon)}$. Since we assume Condition (3) and Lemma 2.1 both hold on the input samples, we can use Lemmas C.7 and C.8 to find two coordinates $i \in S_-$ and $j \in S_+$, such that the bimodal sub-gradient property in Lemma C.6 does not hold at w. Therefore, w is not a δ-approximate stationary point for some $\delta = \sqrt{2}c_3\frac{\|\mu_w - \mu^\star\|_2^2}{\epsilon^2} \ge \sqrt{2}c_3c_2^2\ln(1/\epsilon)$, that is, we can set $c = \sqrt{2}c_3c_2^2$. □

C.2.1. Proofs of Lemmas C.7 and C.8

In this section, we prove Lemmas C.7 and C.8.

The proofs of these lemmas are conceptually similar to the proofs of related lemmas (Lemmas 3.4 and 3.5) in Section 3. We include their proofs here to make this section self-contained. The main difference is that we switch to the softmax objective, and consequently, we need to work with multiple directions simultaneously. That is, we consider the projections using Y instead of the projections along the maximum eigenvector of Σ_w.

Lemma C.7 states that when μ_w is far from $\mu^\star$, there exists an index $i \in (B \cap S_-)$ such that the gradient $\nabla f(w)_i$ is relatively large.

Recall that the gradient $\nabla f(w)$ in Equation (8) is the same as the gradient of the variance (weighted by w) of the one-dimensional samples $\left(X_i^\top Y^{1/2}\right)_{i=1}^N$. For this one-dimensional problem, a sample far from the (projected) true mean must have large gradient. Our objective is to find such a sample for which we can decrease its weight. More specifically, since w is assumed to be a bad solution, and the softmax objective is close to the spectral norm of Σ_w, the weighted empirical variance of the projected samples is very large. Because the good samples cannot have this much variance, most of the variance comes from the bad samples. We prove that among these bad samples that contribute a lot to the variance, one of them must be very far from the (projected) true mean and hence has a large gradient, which satisfies Lemma C.7.

We use $c_1, \ldots, c_4$ to denote universal positive constants that are independent of N, d, and ϵ. These constants can be set in a way that is similar to that in Section 3 (see Appendix A). The universal constant c in Theorem C.5 can be set as $c = \sqrt{2}c_3c_2^2$ after we set c_2 and c_3.

Proof of Lemma C.7. We first show that $\Sigma_w \bullet Y$ is relatively large. By Lemma 2.1, we know that if $\|\mu_w - \mu^\star\|_2 \geq r$ and $r \geq c_2\epsilon\sqrt{\ln(1/\epsilon)}$, then

$$\lambda_{\max}(\Sigma_w) \geq 1 + c_4 \cdot \frac{r^2}{\epsilon} \,.$$

By Corollary C.2, for $Y = \frac{\exp(\rho\Sigma_w)}{\mathrm{tr}(\exp(\rho\Sigma_w))}$ and $\rho = \frac{\ln d}{\epsilon}$, we have

$$\Sigma_w \bullet Y \geq \mathrm{smax}_\rho(\Sigma_w) - \epsilon \geq \lambda_{\max}(\Sigma_w) - \epsilon \geq 1 - \epsilon + \frac{c_4 r^2}{\epsilon} \,.$$

Recall that $\Sigma_w = \sum_{i=1}^N w_i(X_i - \mu_w)(X_i - \mu_w^\top)$. If we replace μ_w with $\mu^\star$, we have

$$\sum_{i=1}^N w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top \succeq \Sigma_w \,,$$

and therefore,

$$\left(\sum_{i=1}^N w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \geq \Sigma_w \bullet Y \geq 1 - \epsilon + \frac{c_4 r^2}{\epsilon} \,.$$

Next we show that most of the variance is due to bad samples. By Condition (3),

$$\left(\sum_{i\in G} w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \leq 1 + c_1 \cdot \epsilon\ln(1/\epsilon) \,.$$

Consequently,

$$\left(\sum_{i\in B} w_i(X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \geq \frac{c_4 r^2}{\epsilon} - \epsilon - c_1\epsilon\ln(1/\epsilon) \geq 0.98 \cdot c_4 \cdot \frac{r^2}{\epsilon} \,.$$

The last step is because $r \geq c_2 \cdot \epsilon\sqrt{\ln(1/\epsilon)}$ and we can choose c_2 and c_4 to be sufficiently large.

At this point, we know that when $r = \|\mu_w - \mu^\star\|_2$ is large, most of the variance is due to the bad samples. However, the total weight w_B on the bad samples is at most $\epsilon N \cdot \frac{1}{(1-2\epsilon)N} \leq 2\epsilon$. Therefore, there must be some $i \in B$ with $w_i > 0$ and

$$\left((X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \bullet Y \geq \frac{0.98 \cdot c_4 \cdot r^2 \cdot \epsilon^{-1}}{w_B} \geq 0.49 \cdot c_4 \cdot \frac{r^2}{\epsilon^2} \,.$$

By definition, $i \in B \cap S_-$. It remains to show that $\nabla f(w)_i$ is large.

$$\begin{aligned}\nabla f(w)_i - {\mu^\star}^\top Y(\mu^\star - 2\mu_w) &= \left((X_i-\mu^\star)(X_i-\mu^\star)^\top\right) \bullet Y - 2\left((X_i-\mu^\star)(\mu_w-\mu^\star)^\top\right) \bullet Y \\ &\geq \left\|Y^{1/2}(X_i-\mu^\star)\right\|_2^2 - 2\left\|Y^{1/2}(X_i-\mu^\star)\right\|_2 \cdot \left\|Y^{1/2}\right\|_2 \cdot \|\mu_w-\mu^\star\|_2 \\ &\geq \frac{0.49\cdot c_4 \cdot r^2}{\epsilon^2} - 2\cdot\frac{0.7\cdot\sqrt{c_4}\cdot r}{\epsilon}\cdot 1\cdot r \\ &> 2c_3\cdot\frac{r^2}{\epsilon^2}\ .\end{aligned}$$

The first inequality is because $Y \in \Delta_{d\times d}$. The last step uses the fact that c_4 can be sufficiently large. This completes the proof of Lemma C.7. □

Lemma C.8 states that there exists an index $j \in (G \cap S_+)$ such that the gradient $\nabla f(w)_j$ is relatively small. Similar to the proof of Lemma C.7, for the projected one-dimensional instance, a sample close to the (projected) true mean should have small gradient. Our goal is to find such a sample for which we can increase its weight. Recall that S^+ contains the samples whose weight can be increased. We first prove that there are at least ϵN good samples in S^+. Among these ϵN good samples, the concentration bounds imply that there must exist some X_j that is close to the (projected) true mean. The derivative $\nabla f(w)_j$ satisfies Lemma C.8.

Proof of Lemma C.8. Recall that S^+ contains every coordinate i where $w_i < \frac{1}{(1-2\epsilon)N}$. Since at most $(1-2\epsilon)N$ samples can have the maximum weight $\frac{1}{(1-2\epsilon)N}$, we know that $|S^+| \geq 2\epsilon N$. Combining this with $|G| = (1-\epsilon)N$, we know that $|G \cap S^+| \geq \epsilon N$.

Fix a subset $G^+ \subseteq (G \cap S^+)$ of size $|G^+| = \epsilon N$. We first show that, on average, samples in G^+ do not contribute much to the variance.

Let w' be the uniform weight vector on G, i.e., $w'_i = \frac{1}{(1-\epsilon)N}$ for all $i \in G$ and $w'_i = 0$ otherwise. Since $w' \in \Delta_{N,2\epsilon}$, by Condition (3), we have that

$$\left\|\sum_{i\in G}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top - I\right\|_2 \leq c_1\cdot\epsilon\ln(1/\epsilon)\ .$$

Let w'' be the uniform weight vector on $S \setminus G^+ = (G\setminus G^+)\cup B$, i.e., $w''_i = \frac{1}{(1-\epsilon)N}$ for all $i \in ((G\setminus G^+)\cup B)$ and $w''_i = 0$ otherwise. Since $w'' \in \Delta_{N,2\epsilon}$, again by Condition (3), we have that

$$\left\|\sum_{i\in G\setminus G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top - I\right\|_2 \leq c_1\cdot\epsilon\ln(1/\epsilon)\ .$$

Combining the previous two concentration bounds, we obtain that

$$\begin{aligned}\left\|\sum_{i\in G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top\right\|_2 &\leq \left\|\sum_{i\in G}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top - I\right\|_2 \\ &\quad + \left\|\sum_{i\in G\setminus G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top - I\right\|_2 \leq 2c_1\cdot\epsilon\ln(1/\epsilon)\ .\end{aligned}$$

As a result, because $Y \in \Delta_{d\times d}$, it follows that

$$\left(\sum_{i\in G^+}\frac{1}{|G|}(X_i-\mu^\star)(X_i-\mu^\star)^\top\right)\bullet Y \leq 2c_1\cdot\epsilon\ln(1/\epsilon)\ .$$

Now we know that, on average, samples in G^+ do not contribute much to the variance. We continue to show that one of these samples satisfies the lemma.

Let $j = \arg\min_{i \in G^+} \left(Y \bullet (X_i - \mu^\star)(X_i - \mu^\star)^\top\right)$. We have that

$$\left((X_j - \mu^\star)(X_j - \mu^\star)^\top\right) \bullet Y \leq \frac{|G|}{|G^+|} \cdot 2c_1 \cdot \epsilon \ln(1/\epsilon) \leq 2c_1 \ln(1/\epsilon) \; .$$

Finally, because $(X_j - \mu^\star)^\top Y (X_j - \mu^\star) \leq 2c_1 \ln(1/\epsilon)$, we can bound $\nabla f(w)_j$ from above as follows:

$$\begin{aligned}
\nabla f(w)_j - \mu^{\star\top} Y (\mu^\star - 2\mu_w) &= \left((X_j - \mu^\star)(X_j - \mu^\star)^\top\right) \bullet Y - 2\left((X_j - \mu^\star)(\mu_w - \mu^\star)^\top\right) \bullet Y \\
&\leq \left\|Y^{1/2}(X_j - \mu^\star)\right\|_2^2 + 2\left\|Y^{1/2}(X_j - \mu^\star)\right\|_2 \cdot \left\|Y^{1/2}\right\|_2 \cdot \|\mu_w - \mu^\star\|_2 \\
&\leq 2c_1 \ln(1/\epsilon) + 2\sqrt{2c_1 \ln(1/\epsilon)} \cdot 1 \cdot r \\
&\leq \frac{c_3}{2} \cdot \frac{r^2}{\epsilon^2} + \frac{c_3}{2} \cdot \frac{r}{\epsilon} \cdot r \leq c_3 \cdot \frac{r^2}{\epsilon^2} \; .
\end{aligned}$$

The last step uses that c_3 is sufficiently large, as well as the fact that $\ln(1/\epsilon) \leq \frac{r^2}{\epsilon^2}$, because $r \geq c_2 \epsilon \sqrt{\ln(1/\epsilon)}$. This completes the proof of Lemma C.8. □

C.3. Convergence Rate of Minimizing Softmax

Algorithm 2 Robust Mean Estimation via Projected Gradient Descent on the Softmax Objective

Input: ϵ-corrupted set of N samples $\{X_i\}_{i=1}^N$ on $\mathbb{R}^d$ satisfying Condition (3), and $\epsilon < \epsilon_0$.
Output: $w \in \mathbb{R}^N$ with $\|\mu_w - \mu^\star\|_2 \leq O(\epsilon\sqrt{\log(1/\epsilon)})$.
Let $\rho = \ln d/\epsilon$.
Let $\beta = \widetilde{O}(Nd^2/\epsilon)$ be the smoothness parameter of the softmax objective $f(w) = \mathrm{smax}_\rho(\Sigma_w)$.
Let w_0 be an arbitrary weight vector in $\Delta_{N,2\epsilon}$.
Let $T = \widetilde{O}(Nd^3/\epsilon)$ and $\eta = 1/\beta$.
for $\tau = 0$ **to** $T - 1$ **do**
  $w_{\tau+1} = \mathcal{P}_{\Delta_{N,2\epsilon}}\left(w_\tau - \eta \nabla f(w)\right)$, where $\mathcal{P}_{\mathcal{K}}(\cdot)$ is the ℓ_2-projection operator onto $\mathcal{K}$.
end for
return $w_{\tau^\star}$ where $\tau^\star = \arg\min_{0 \leq \tau < T} \|w_{\tau+1} - w_\tau\|_2$.

In this section, we prove our algorithmic result for the softmax objective (Theorem 1.4). We show that the projected gradient descent algorithm (Algorithm 2) on f can efficiently find an approximate stationary point w, and that w is a good solution to our robust mean estimation task.

We first restate Theorem 1.4 (correctness and iteration count of Algorithm 2).

Theorem 1.4. *Let S be an ϵ-corrupted set of $N = \widetilde{\Omega}(d/\epsilon^2)$ samples drawn from a d-dimensional Gaussian $\mathcal{N}(\mu^\star, I)$ with unknown mean $\mu^\star$. Suppose S satisfies Condition (3) and Lemma 2.1.*

Let $f(w)$ be the softmax objective as defined in Equation (7). After $\widetilde{O}(Nd^3/\epsilon)$ iterations, projected gradient descent on $f(w)$ outputs a point w such that $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)})$.

Theorem 1.4 follows immediately from Lemmas C.9, C.10, and C.11.

Lemma C.9 analyzes the convergence rate of (nonconvex) projected gradient descent. The number of iterations in Lemma C.9 depends on the range and smoothness of the objective function. Lemmas C.10 and C.11 upper bounds these two parameters for our softmax objective.

We note that Lemma C.9 appears to be folklore in the optimization literature, see, e.g., (Beck, 2017). For the sake of completeness, we provide a self-contained proof in the following subsection.

Lemma C.9. *Fix a (possibly non-convex) function f and a convex set $\mathcal{K}$. Suppose f is β-smooth on $\mathcal{K}$ and $0 \leq f(x) \leq B$ for all $x \in \mathcal{K}$. If we run projected gradient descent with step size $\eta = \frac{1}{\beta}$ starting from an arbitrary $x_0 \in \mathcal{K}$:*

$$x_{\tau+1} = \Pi_{\mathcal{K}}\left(x_\tau - \eta \nabla f(x_\tau)\right) \; ,$$

where $\Pi_{\mathcal{K}}$ is the projection onto $\mathcal{K}$, we can compute a δ-stationary point of f in $O(\frac{\beta \cdot B}{\delta^2})$ iterations.

Recall that the softmax objective is $f(w) = \mathrm{smax}_\rho(\Sigma_w) = \frac{1}{\rho}\ln \mathrm{tr}(\exp(\rho\Sigma_w))$ with $\rho = \frac{\ln d}{\epsilon}$. A differentiable function f is β-smooth on $\mathcal{K}$ if $\|\nabla f(x) - \nabla f(y)\|_2 \le \beta \|x - y\|_2$ for all $x, y \in \mathcal{K}$.

Lemma C.10 (Smoothness of f)**.** *The softmax objective f is β-smooth on $\Delta_{N,2\epsilon}$ for $\beta = \widetilde{O}(Nd^2/\epsilon)$.*

Lemma C.11 (Range of f)**.** *The softmax objective f satisfies that $0 \le f(w) \le \widetilde{O}(d)$ for all $w \in \Delta_{N,2\epsilon}$.*

We defer the proofs of Lemmas C.9, C.10, and C.11 to the next subsections and first use them to prove Theorem 1.4.

Proof of Theorem 1.4. We first prove the correctness of Algorithm 2. Let c be the universal constant in Theorem C.5 and let $\delta = c\ln(1/\epsilon)$. We run Algorithm 2 to obtain a δ-stationary point w. Since we assume the input samples satisfy Condition (3) and Lemma 2.1, Theorem C.5 states that w is a good solution with $\|\mu_w - \mu^\star\|_2 = O(\epsilon\sqrt{\ln(1/\epsilon)})$.

We now analyze the number of iterations T. By Lemma C.9, it is sufficient to set $T = O(\frac{\beta \cdot B}{\delta^2})$, as in Algorithm 2. Substituting the upper bounds on β and B from Lemmas C.10 and C.11, and our choice of δ, we get

$$T = O(\beta \cdot B \cdot \delta^{-2}) = \widetilde{O}(Nd^2/\epsilon) \cdot \widetilde{O}(d) \cdot O(\log^{-2}(1/\epsilon)) = \widetilde{O}(Nd^3/\epsilon)\ ,$$

as claimed. □

C.4. Proof of Lemma C.9

In this section, we prove Lemma C.9.

Lemma C.9 analyzes the convergence rate of projected gradient descent, when we use it to minimize a smooth non-convex function with constraints. Lemma C.9 follows directly from Lemmas C.12 and C.13.

Lemma C.12 defines a "truncated gradient" mapping g and relates the progress in the τ-th iteration with $\|g(x_\tau)\|_2^2$. Because we cannot keep decreasing $f(x)$, we know that after many iterations, there exists some τ such that $\|g(x_\tau)\|_2$ is very small. Lemma C.13 shows that if $\|g(x_\tau)\|_2$ is very small, that is, if projected gradient descent moves very little between x_τ and $x_{\tau+1}$, then $x_{\tau+1}$ is an approximate stationary point.

Lemma C.12. *Fix a convex set $\mathcal{K}$. Suppose f is β-smooth on $\mathcal{K}$ and $0 \le f(x) \le B$ for all $x \in \mathcal{K}$. Suppose we run projected gradient descent with step size $\eta = \frac{1}{\beta}$ starting from an arbitrary $x_0 \in \mathcal{K}$, i.e.,*

$$x_{\tau+1} = \Pi_{\mathcal{K}}\left(x_\tau - \eta\nabla f(x_\tau)\right)\ ,$$

where $\Pi_{\mathcal{K}}$ is the ℓ_2-projection onto $\mathcal{K}$. Then we have that

$$\min_{0\le\tau<T} \frac{1}{\eta}\left\|\Pi_{\mathcal{K}}\left(x_\tau - \eta\nabla f(x_\tau)\right) - x_\tau\right\|_2 \le \sqrt{\frac{2\beta B}{T}}\ .$$

Proof. Define the mapping

$$g(x) = \frac{x - \Pi_{\mathcal{K}}(x - \eta\nabla f(x))}{\eta}\ .$$

Let $y_{\tau+1} = x_\tau - \eta\nabla f(x_\tau)$. Notice that $x_{\tau+1} = \Pi_{\mathcal{K}}(y_{\tau+1}) = x_s - \eta g(x_\tau)$.

By the convexity of $\mathcal{K}$, we have

$$(x_{\tau+1} - x_\tau)^\top (x_{\tau+1} - y_{\tau+1}) \le 0\ ,$$

which is equivalent to

$$\nabla f(x_\tau)^\top (x_{\tau+1} - x_\tau) \le g(x_\tau)^\top (x_{\tau+1} - x_\tau)\ .$$

Using the quadratic upper bound combined with the above inequality, we have

$$\begin{aligned} f(x_{\tau+1}) &\le f(x_\tau) + \nabla f(x_\tau)^\top (x_{\tau+1} - x_\tau) + \frac{\beta}{2} \|x_{\tau+1} - x_\tau\|_2^2 \\ &\le f(x_\tau) + g(x_\tau)^\top (x_{\tau+1} - x_\tau) + \frac{\beta}{2} \|x_{\tau+1} - x_\tau\|_2^2 \\ &= f(x_\tau) - \eta \|g(x_\tau)\|_2^2 + \frac{\eta^2 \beta}{2} \|g(x_\tau)\|_2^2 \\ &= f(x_\tau) - \frac{1}{2\beta} \|g(x_\tau)\|_2^2 \ . \end{aligned}$$

Therefore, after T iterations, we have

$$\min_{0 \le \tau < T} \|g(x_\tau)\|_2^2 \le \frac{1}{T} \sum_{\tau=0}^{T-1} \|g(x_\tau)\|_2^2 \le \frac{2\beta}{T} \left(f(x_0) - f(x_T)\right) \le \frac{2\beta B}{T} \ . \qquad \square$$

Lemma C.13. *Consider the same setting as in Lemma C.12. Define the tangent cone of $\mathcal{K}$ at a point $x \in \mathcal{K}$ as $\mathcal{C}_\mathcal{K}(x) = \mathrm{cone}(\mathcal{K} - \{x\})$. If for some τ we have*

$$\|\Pi_\mathcal{K} \left(x_\tau - \eta \nabla f(x_\tau)\right) - x_\tau\|_2 \le \frac{\delta}{2} \ ,$$

then for all unit vector $u \in \mathcal{C}_\mathcal{K}(x)$,

$$\nabla f(x_{\tau+1})^\top u \le \delta \ .$$

Proof. By the convexity of $\mathcal{K}$, we know that for any $z \in \mathcal{K}$,

$$(y_{\tau+1} - x_{\tau+1})^\top (z - x_{\tau+1}) \le 0 \ .$$

Consequently, for any $u \in \mathcal{C}_\mathcal{K}(x_{\tau+1})$, we have

$$(y_{\tau+1} - x_{\tau+1})^\top u \le 0 \ ,$$

which is equivalent to

$$-\nabla f(x_\tau)^\top u \le -g(x_\tau)^\top u \ .$$

Using the fact that u is a unit vector together with the above inequality, we get

$$\begin{aligned} -\nabla f(x_{\tau+1})^\top u &\le -\nabla f(x_{\tau+1})^\top u + \nabla f(x_\tau)^\top u - g(x_\tau)^\top u \\ &\le \|f(x_{\tau+1}) - \nabla f(x_\tau)\|_2 + \|g(x_\tau)\|_2 \\ &\le \beta \|x_{\tau+1} - x_\tau\|_2 + \|g(x_\tau)\|_2 \\ &= 2 \|g(x_\tau)\|_2 \le \delta \ . \end{aligned}$$

$\square$

Proof of Lemma C.9. As in Algorithm 2, we run projected gradient descent, track the value of $\|g(x_\tau)\|_2$ in each iteration, and return the x_τ that has the minimum $\|g(x_\tau)\|_2$. Combining Lemmas C.12 and C.13, if we want a δ-stationary point, we should set T such that $\sqrt{2\beta B/T} \le \delta/2$, i.e., $T \ge 8\beta B \delta^{-2} = O(\beta B \delta^{-2})$. $\square$

C.5. Proofs of Lemmas C.10 and C.11

In this subsection, we bound from above the smoothness and maximum value of the softmax objective.

For these two lemmas, we can assume without loss of generality that no input samples have very large ℓ_2-norm. This is because we can perform a standard preprocessing step that centers the input samples at the coordinate-wise median, which does not affect our mean estimation task. We then throw away all samples that are $\Omega(\sqrt{d \log d})$ far from the coordinate-wise median. With high probability, the coordinate-wise median and all good samples are $O(\sqrt{d \log d})$ far from the true mean. Assuming this happens, then no good samples are thrown away and all remaining samples satisfies $\max_i \|X_i\|_2 = O(\sqrt{d \log d})$. Consequently, we have $\|\mu_w\|_2 = O(\sqrt{d \log d})$ for any $w \in \Delta_{N,\epsilon}$.

Proof of Lemma C.10. We proceed to bound from above the spectral norm of the Hessian of f. Recall that $X \in \mathbb{R}^{d\times N}$ and the partial derivative of f with respect to w_i is

$$\nabla f(w)_i = X_i^\top Y X_i - 2X_i^\top Y \mu_w = \left(X_i X_i^\top - X_i\mu_w^\top - \mu_w X_i^\top\right) \bullet Y,$$

where $Y = \frac{\exp(\rho\Sigma_w)}{\operatorname{tr}\exp(\rho\Sigma_w)}$ is a PSD matrix. Observe that $Y \succeq 0$, $\operatorname{tr}(Y) = 1$, and Y depends on w.

We can compute the (i,j)-th entry in the Hessian matrix of f, as follows

$$\nabla^2 f(w)_{i,j} = \frac{df(w)_i}{dw_j} = \left(X_i X_i^\top - X_i\mu_w^\top - \mu_w X_i^\top\right) \bullet \frac{dY}{dw_j} - \left(X_i X_j^\top + X_j X_i^\top\right) \bullet Y \ .$$

By the chain rule, we have

$$\begin{aligned}\frac{dY}{dw_j} &= \frac{1}{\operatorname{tr}(\exp(\rho\Sigma_w))^2}\left[\frac{d\exp(\rho\Sigma_w)}{dw_j}\operatorname{tr}(\exp(\rho\Sigma_w)) - \frac{d\operatorname{tr}(\exp(\rho\Sigma_w))}{dw_j}\exp(\rho\Sigma_w)\right] \\ &= \frac{1}{\operatorname{tr}(\exp(\rho\Sigma_w))}\left[\frac{d\exp(\rho\Sigma_w)}{dw_j} - \frac{d\operatorname{tr}(\exp(\rho\Sigma_w))}{dw_j}\cdot Y\right] \ .\end{aligned}$$

Using Lemma C.3 to compute the derivative of matrix exponential, we have

$$\begin{aligned}\frac{dY}{dw_j} &= \frac{1}{\operatorname{tr}(\exp(\rho\Sigma_w))}\left[\frac{d\exp(\rho\Sigma_w)}{dw_j} - \frac{d\operatorname{tr}(\exp(\rho\Sigma_w))}{dw_j}\ Y\right] \\ &= \frac{1}{\operatorname{tr}\exp(\rho\Sigma_w)}\left[\int_{\alpha=0}^1 \exp(\alpha\rho\Sigma_w)\frac{d(\rho\Sigma_w)}{dw_j}\exp((1-\alpha)\rho\Sigma_w)d\alpha - \left(\frac{d(\rho\Sigma_w)}{dw_j} \bullet \exp(\rho\Sigma_w)\right) Y\right] \\ &= \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \exp(\alpha\Sigma_w)\frac{d\Sigma_w}{dw_j}\exp((1-\alpha)\rho\Sigma_w)d\alpha - \rho\left(\frac{d\Sigma_w}{dw_j} \bullet Y\right) Y \ .\end{aligned}$$

Since $\frac{d\Sigma_w}{dw_j} = X_j X_j^\top - X_j\mu_w^\top - \mu_w X_j^\top$, putting it all together, we have,

$$\begin{aligned}\nabla^2 f(w)_{i,j} = &-\left(X_i^\top Y(X_i - 2\mu_w)\right)\left(X_j^\top Y(X_j - 2\mu_w)\right) - 2X_i^\top Y X_j \\ &+ \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\cdot \\ &\int_{\alpha=0}^1 \operatorname{tr}\left(\left(X_i X_i^\top - X_i\mu_w^\top - \mu_w X_i^\top\right)\exp(\alpha\rho\Sigma_w)\left(X_j X_j^\top - X_j\mu_w^\top - \mu_w X_j^\top\right)\exp((1-\alpha)\rho\Sigma_w)\right)d\alpha \ .\end{aligned}$$

Let $R = \max(\|\mu_w\|_2, \max_i \|X_i\|_2)$. From the preprocessing step, we know that $R = \widetilde{O}(d^{1/2})$. Using this fact, we obtain

$$\left|\nabla^2 f(w)_{i,j}\right| \le 9R^4 + 2R^2 + 9\rho R^4 = \widetilde{O}(\rho d^2) \ .$$

This is because the first term can be bounded from above by

$$\begin{aligned}-\left(X_i^\top Y(X_i - 2\mu_w)\right)\left(X_j^\top Y(X_j - 2\mu_w)\right) &\le \|X_i\|_2 \|Y\|_2 \|X_i - 2\mu_w\|_2 \|X_j\|_2 \|Y\|_2 \|X_j - 2\mu_w\|_2 \\ &\le 9R^4 \ .\end{aligned}$$

Similarly, the second term is at most $2R^2$. The third term can be split into 9 terms of the form

$$\begin{aligned}&\frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \operatorname{tr}\left(\left(X_i X_i^\top\right)\exp(\alpha\rho\Sigma_w)\left(X_j X_j^\top\right)\exp((1-\alpha)\rho\Sigma_w)\right)d\alpha \\ &= \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \left(X_i^\top\exp(\alpha\rho\Sigma_w)X_j\right)\left(X_j^\top\exp((1-\alpha)\rho\Sigma_w)X_i\right)d\alpha \\ &\le \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\int_{\alpha=0}^1 \|X_i\|_2\|\exp(\alpha\rho\Sigma_w)\|_2\|X_j\|_2\|X_j\|_2\|\exp((1-\alpha)\rho\Sigma_w)\|_2\|X_i\|_2 d\alpha \\ &= \frac{\rho}{\operatorname{tr}\exp(\rho\Sigma_w)}\cdot R^4\cdot\|\exp(\rho\Sigma_w)\|_2 \le \rho R^4 \ .\end{aligned}$$

To conclude the proof, we bound from above the smoothness parameter by the spectral norm of the Hessian matrix. For any $w \in \Delta_{N,2\epsilon}$,

$$\left\|\nabla^2 f(w)\right\|_2 \leq N \cdot \max_{ij} \left|\nabla^2 f(w)_{ij}\right| \leq O(N\rho d^2) = \widetilde{O}(Nd^2/\epsilon) \ ,$$

where the last step uses that $\rho = \ln d/\epsilon$. □

Proof of Lemma C.11. Fix any $w \in \Delta_{N,2\epsilon}$. By Corollary C.2 and our choice of $\rho = \frac{\ln d}{\epsilon}$, we have

$$f(w) = \mathrm{smax}_\rho(\Sigma_w) \leq \lambda_{\max}(\Sigma_w) + \epsilon.$$

Therefore, it is sufficient to bound from above $\lambda_{\max}(\Sigma_w)$ by $O(d \log d)$.

The preprocessing step guarantees that all samples have ℓ_2-norm at most $\widetilde{O}(d^{1/2})$, consequently, the weighted empirical mean μ_w has ℓ_2-norm is at most $\widetilde{O}(d^{1/2})$ as well. Consequently,

$$\begin{aligned}\|\Sigma_w\|_2 &= \left\|\sum_{i=1}^N w_i (X_i - \mu_w)(X_i - \mu_w)^\top\right\|_2 \\ &\leq \sum_{i=1}^N w_i \left\|(X_i - \mu_w)(X_i - \mu_w)^\top\right\|_2 \leq \max_{i\in[N]} \|X_i - \mu_w\|_2^2 \leq \widetilde{O}(d) \ .\end{aligned}$$

The proof is now complete. □