Appendix

A. Convergence of SGD and AdaGrad with biased gradients estimates

For the sake of our analysis, we find it helpful to first study the convergence of SGD and AdaGrad when the stochastic estimates of the subgradients may be biased and noisy (Algorithms 4 and 5.)

Algorithm 4 Biased SGD

Require: Dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, convex set $\mathcal{X}$, mini-batch size b, number of iterations T.
1: Choose arbitrary initial point $x^0 \in \mathcal{X}$;
2: **for** $k = 0;\ k \leq T-1;\ k = k+1$ **do**
3: Sample a batch $\mathcal{D}_k := \{z_i^k\}_{i=1}^b$ from $\mathcal{S}$ uniformly with replacement;
4: Set $g^k := \frac{1}{b}\sum_{i=1}^b g^{k,i}$ where $g^{k,i} \in \partial F(x^k; z_i^k)$;
5: Set $\tilde{g}^k$ be the biased estimate of g^k;
6: Set $\hat{g}^k := \tilde{g}^k + \xi^k$ where ξ^k is a zero-mean random variable, independent from previous information;
7: $x^{k+1} := \text{proj}_{\mathcal{X}}(x^k - \alpha_k \hat{g}^k)$;
8: **end for**
9: **Return:** $\overline{x}^T := \frac{1}{T}\sum_{k=1}^T x^k$.

Algorithm 5 Biased Adagrad

Require: Dataset $\mathcal{S} = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, convex set $\mathcal{X}$, mini-batch size b, number of iterations T.
1: Choose arbitrary initial point $x^0 \in \mathcal{X}$;
2: **for** $k = 0;\ k \leq T-1;\ k = k+1$ **do**
3: Sample a batch $\mathcal{D}_k := \{z_i^k\}_{i=1}^b$ from $\mathcal{S}$ uniformly with replacement;
4: Set $\tilde{g}^k$ be the biased estimate of g^k;
5: Set $\hat{g}^k := \tilde{g}^k + \xi^k$ where ξ^k is a zero-mean random variable, independent from previous information;
6: Set $H_k = \text{diag}\left(\sum_{i=1}^k \hat{g}^i \hat{g}^{i^T}\right)^{\frac{1}{2}} / \text{diam}_\infty(\mathcal{X})$;
7: $x^{k+1} = \text{proj}_{\mathcal{X}}(x^k - H_k^{-1}\hat{g}^k)$ where the projection is with respect to $\|\cdot\|_{H_k}$;
8: **end for**
9: **Return:** $\overline{x}^T := \frac{1}{T}\sum_{k=1}^T x^k$.

Also, let

$$\mathsf{bias}_{\|\cdot\|}(\tilde{g}^k) = \mathbb{E}_{\mathcal{D}_k}\left[\left\|\tilde{g}^k - g^k\right\|\right]$$

be the bias of $\tilde{g}_k$ with respect to a general norm $\|\cdot\|$. The next two theorems characterize the convergence of these two algorithms using this term.

Theorem 6. *Consider the biased SGD method (Algorithm 4) with a non-increasing sequence of stepsizes $\{\alpha_k\}_{k=0}^{T-1}$. Then for any $x^\star \in \text{argmin}_{\mathcal{X}} f$, we have*

$$\mathbb{E}[f(\overline{x}^T) - f(x^\star)] \leq \frac{\text{diam}_2(\mathcal{X})^2}{2T\alpha_{T-1}} + \frac{1}{2T}\sum_{k=0}^{T-1}\mathbb{E}[\alpha_k \left\|\hat{g}^k\right\|_2^2] + \frac{\text{diam}_{\|\cdot\|_*}(\mathcal{X})}{T}\sum_{k=0}^{T-1}\mathsf{bias}_{\|\cdot\|}(\tilde{g}^k).$$

Proof We first consider the progress of a single step of the gradient-projected stochastic gradient method. We have

$$\begin{aligned}\frac{1}{2}\|x^{k+1} - x^\star\|_2^2 &\leq \frac{1}{2}\|x^k - x^\star\|_2^2 - \alpha_k\langle\hat{g}^k, x^k - x^\star\rangle + \frac{\alpha_k^2}{2}\|\hat{g}^k\|_2^2 \\ &= \frac{1}{2}\|x^k - x^\star\|_2^2 - \alpha_k\langle f'(x^k), x^k - x^\star\rangle + \alpha_k E_k + \frac{\alpha_k^2}{2}\|\hat{g}^k\|_2^2,\end{aligned}$$

where the error random variable E_k is given by

$$E_k := \langle f'(x^k) - g^k, x^k - x^\star\rangle + \langle g^k - \tilde{g}^k, x^k - x^\star\rangle + \langle\tilde{g}^k - \hat{g}^k, x^k - x^\star\rangle.$$

Using that $\langle f'(x^k), x^k - x^\star \rangle \leq f(x^k) - f(x^\star)$ then yields

$$f(x^k) - f(x^\star) \leq \frac{1}{2\alpha_k}\left(\|x^k - x^\star\|_2^2 - \|x^{k+1} - x^\star\|_2^2\right) + \frac{\alpha_k}{2}\|\hat{g}^k\|_2^2 + E_k.$$

Summing for $k = 0, \ldots, T-1$, by rearranging the terms and using that the stepsizes are non-increasing, we obtain

$$\sum_{k=1}^{T}[f(x^k) - f(x^\star)] \leq \frac{\mathrm{diam}(\mathcal{X})^2}{2\alpha_{T-1}} + \sum_{k=0}^{T-1}\frac{\alpha_k}{2}\|\hat{g}^k\|_2^2 + \sum_{k=0}^{T-1} E_k. \tag{12}$$

Taking expectations from both sides, we have

$$\begin{aligned}\mathbb{E}[E_k] &= \mathbb{E}\left[\langle f'(x^k) - g^k, x^k - x^\star\rangle\right] + \mathbb{E}\left[\langle g^k - \tilde{g}^k, x^k - x^\star\rangle\right] + \mathbb{E}\left[\langle \tilde{g}^k - \hat{g}^k, x^k - x^\star\rangle\right] \\ &= \mathbb{E}\left[\langle g^k - \tilde{g}^k, x^k - x^\star\rangle\right] \\ &\leq \mathsf{bias}_{\|\cdot\|}(\tilde{g}^k) \cdot \mathrm{diam}_{\|\cdot\|_*}(\mathcal{X}),\end{aligned}$$

where the second equality comes from the fact that the two other expectations are zero and the last inequality follows from the Holder's inequality. □

Remark This result holds in the case that α_k's are adaptive and depend on observed gradients.

Next theorem states the convergence of biased Adagrad (Algorithm 5).

Theorem 7. *Consider the biased Adagrad method (Algorithm 5). Then for any $x^\star \in \mathrm{argmin}_{\mathcal{X}} f$, we have*

$$\mathbb{E}[f(\overline{x}^T) - f(x^\star)] \leq \frac{\mathrm{diam}_\infty(\mathcal{X})}{T}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=0}^{T-1}(\hat{g}_j^k)^2}\right] + \frac{\mathrm{diam}_{\|\cdot\|_*}(\mathcal{X})}{T}\sum_{k=0}^{T-1}\mathsf{bias}_{\|\cdot\|}(\tilde{g}^k).$$

Proof Recall that x^{k+1} is the projection of $x^k - H_k^{-1}\hat{g}^k$ into $\mathcal{X}$ with respect to $\|.\|_{\mathcal{H}_k}$. Hence, since $x^\star \in \mathcal{X}$ and projections are non-expansive, we have

$$\left\|x^{k+1} - x^\star\right\|_{\mathcal{H}_k}^2 \leq \left\|x^k - H_k^{-1}\hat{g}^k - x^\star\right\|_{\mathcal{H}_k}^2. \tag{13}$$

Now, expanding the right hand side yields

$$\begin{aligned}\frac{1}{2}\left\|x^{k+1} - x^\star\right\|_{\mathcal{H}_k}^2 &\leq \frac{1}{2}\left\|x^k - x^\star\right\|_{\mathcal{H}_k}^2 - \langle \hat{g}^k, x^k - x^\star\rangle + \frac{1}{2}\left\|\hat{g}^k\right\|_{\mathcal{H}_k^{-1}}^2 \\ &= \frac{1}{2}\left\|x^k - x^\star\right\|_{\mathcal{H}_k}^2 - \langle g^k, x^k - x^\star\rangle + \langle g^k - \hat{g}^k, x^k - x^\star\rangle + \frac{1}{2}\left\|\hat{g}^k\right\|_{\mathcal{H}_k^{-1}}^2.\end{aligned}$$

Taking expectation and using that $\mathbb{E}\left[\langle g^k, x^k - x^\star\rangle\right] \geq \mathbb{E}\left[f(x^k) - f(x^\star)\right]$ from convexity along with the fact that $\mathbb{E}\left[\langle g^k - \hat{g}^k, x^k - x^\star\rangle\right] = \mathbb{E}\left[\langle g^k - \tilde{g}^k, x^k - x^\star\rangle\right]$, we have

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\left[\left\|x^{k+1} - x^\star\right\|_{\mathcal{H}_k}^2\right] \\ &\quad\leq \mathbb{E}\left[\frac{1}{2}\left\|x^k - x^\star\right\|_{\mathcal{H}_k}^2 - (f(x^k) - f(x^\star)) + \frac{1}{2}\left\|\hat{g}^k\right\|_{\mathcal{H}_k^{-1}}^2\right] + \mathbb{E}\left[\langle g^k - \tilde{g}^k, x^k - x^\star\rangle\right].\end{aligned}$$

Thus, using Holder's inequality, we have

$$f(x^k) - f(x^\star) \leq \frac{1}{2}\mathbb{E}\left[\left\|x^k - x^\star\right\|_{\mathcal{H}_k}^2 - \left\|x^{k+1} - x^\star\right\|_{\mathcal{H}_k}^2 + \left\|\hat{g}^k\right\|_{\mathcal{H}_k^{-1}}^2\right] + \mathsf{bias}_{\|\cdot\|}(\tilde{g}^k) \cdot \mathrm{diam}_{\|\cdot\|_*}(\mathcal{X}).$$

Now the claim follows using standard techniques for Adagrad (as for example Corollary 4.3.8 in (Duchi, 2018)). □

B. Proofs of Section 3

B.1. Proof of Lemma 3.1

The proof mainly follows from Theorem 1 in (Abadi et al., 2016) where the authors provide a tight privacy bound for mini-batch SGD with bounded gradient using the Moments Accountant technique. Here we do not have the bounded gradient assumption. However, recall that we have

$$\hat{g}^k = \frac{1}{b}\sum_{i=1}^{b} \tilde{g}^{k,i} + \frac{\sqrt{\log(1/\delta)}}{b\varepsilon}\xi^k, \quad \tilde{g}^{k,i} = \pi_{A_k}(g^{k,i}),$$

where $\left\|\tilde{g}^{k,i}\right\|_{A_k} \leq 1$ and $\xi^k \overset{\text{iid}}{\sim} \mathsf{N}(0, A_k^{-1})$. Note that for any Borel-measurable set $O \subset \mathbb{R}^d$, $A_k^{1/2}O$ is also Borel-measurable, and furthermore, we have

$$\mathbb{P}\left(\hat{g}^k \in O\right) = \mathbb{P}\left(A_k^{1/2}\hat{g}^k \in A_k^{1/2}O\right) = \mathbb{P}\left(\frac{1}{b}\sum_{i=1}^{b} A_k^{1/2}\tilde{g}^{k,i} + \frac{\sqrt{\log(1/\delta)}}{b\varepsilon}A_k^{1/2}\xi^k \in A_k^{1/2}O\right),$$

where, now, $\left\|A_k^{1/2}\tilde{g}^{k,i}\right\|_2 \leq 1$ and $A_k^{1/2}\xi^k \overset{\text{iid}}{\sim} \mathsf{N}(0, I_d)$ and we can use Theorem 1 in (Abadi et al., 2016).

B.2. The proof deferred from Example 1

Note that $\nabla F(x; z) = \nabla g(x) + Z$, and hence we could take $G(Z, C) = \sup_{x\in\mathcal{X}} \|\nabla g(x)\|_C + \|Z\|_C$. As a result, by Minkowski inequality, we have

$$\mathbb{E}\left[G(Z,C)^p\right]^{1/p} \leq \sup_{x\in\mathcal{X}} \|\nabla g(x)\|_C + \mathbb{E}\left[\|Z\|_C^p\right]^{1/p} \leq \mu + \sup_{x\in\mathcal{X}} \mathbb{E}\left[\|Z\|_C^p\right]^{1/p}. \tag{14}$$

Now note that $C^{1/2}Z$ is $(C_{11}\sigma_1^2, \ldots, C_{dd}\sigma_d^2)$ sub-Gaussian. Also, we also know that if X is σ^2 sub-gaussian, then $\mathbb{E}[|X|^p]^{1/p} \leq O(\sigma\sqrt{p})$, which implies the desired result.

B.3. Intermediate Results

Before discussing the proofs of Theorems 1 and 2, we need to state a few intermediate results which will be used in our analysis.

First, recall the definition of $\mathsf{bias}_{\|\cdot\|}(\tilde{g}^k)$ from Section A:

$$\mathsf{bias}_{\|\cdot\|}(\tilde{g}^k) = \mathbb{E}_{\mathcal{D}_k}\left[\left\|\tilde{g}^k - g^k\right\|\right]$$

Here, we first bound the bias term. To do so, we use the following lemma:

Lemma B.1 (Lemma 3, (Barber & Duchi, 2014))**.** *Consider the ellipsoid projection operator π_D. Then, for any random vector X with $\mathbb{E}[\|X\|_C^p]^{1/p} \leq G$, we have*

$$\mathbb{E}_X[\|\pi_D(X) - X\|_C] \leq \frac{G^p}{(p-1)B^{p-1}}.$$

We will find this lemma useful in our proofs. Another useful lemma that we will use it is the following:

Lemma B.2. *Let $a_1, a_2, \ldots$ be an arbitrary sequence in $\mathbb{R}$. Let $a_{1:k} = (a_1, \ldots, a_i) \in \mathbb{R}^i$. Then*

$$\sum_{k=1}^{n} \frac{a_k^2}{\|a_{1:k}\|_2} \leq 2\|a_{1:n}\|_2.$$

Proof We proceed by induction. The base case that $n = 1$ is immediate. Now, let us assume the result holds through

index $n-1$, and we wish to prove it for index n. The concavity of $\sqrt{\cdot}$ guarantees that $\sqrt{b+a} \leq \sqrt{b} + \frac{1}{2\sqrt{b}}a$, and so

$$\begin{aligned}\sum_{k=1}^{n} \frac{a_k^2}{\|a_{1:k}\|_2} &= \sum_{k=1}^{n-1} \frac{a_k^2}{\|a_{1:k}\|_2} + \frac{a_n^2}{\|a_{1:n}\|_2} \\ &\leq 2\|a_{1:n-1}\|_2 + \frac{a_n^2}{\|a_{1:n}\|_2} = 2\sqrt{\|a_{1:n}\|_2^2 - a_n^2} + \frac{a_n^2}{\|a_{1:n}\|_2} \\ &\leq 2\|a_{1:n}\|_2 ,\end{aligned}$$

where the first inequality follows from the inductive hypothesis and the second one uses the concavity of $\sqrt{\cdot}$. □

B.4. Proof of Theorem 1

We first state a more general version of the theorem here:

Theorem 8. *Let $\mathcal{S}$ be a dataset with n points sampled from distribution P. Let C also be a diagonal and positive definite matrix. Consider running Algorithm 1 with $T = cn^2/b^2$, $A_k = C/B^2$ where $B > 0$ is a positive real number and c is given by Lemma 3.1. Then, with probability $1 - 1/n$, we have*

$$\begin{aligned}\mathbb{E}[f(\overline{x}^T; \mathcal{S}) - \min_{x\in\mathcal{X}} f(x;\mathcal{S})] \leq \mathcal{O}(1) \Bigg(&\frac{\mathrm{diam}_2(\mathcal{X})}{T}\sqrt{\sum_{k=1}^{T}\mathbb{E}[\|g^k\|_2^2]} \\ &+ \frac{\mathrm{diam}_2(\mathcal{X}) B\sqrt{\mathrm{tr}(C^{-1})\log(1/\delta)}}{n\varepsilon} + \frac{\mathrm{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})\,(2G_{2p}(C))^p}{(p-1)B^{p-1}}\Bigg),\end{aligned}$$

where the expectation is taken over the internal randomness of the algorithm.

Proof Let $x^\star \in \mathrm{argmin}_{x\in\mathcal{X}} f(x;\mathcal{S})$. Also, for simplicity, we suppress the dependence of f on $\mathcal{S}$ throughout the proof. First, by Lemma 3.2, we know that with probability at least $1 - 1/n$, we have

$$\hat{G}_p(\mathcal{S}; C) \leq 2G_{2p}(C),$$

We consider the setting that this bound holds. Now, note that by Theorem 6 we have

$$\mathbb{E}[f(\overline{x}^T) - f(x^\star)] \leq \frac{\mathrm{diam}_2(\mathcal{X})^2}{2T\alpha_{T-1}} + \frac{1}{2T}\sum_{k=0}^{T-1}\mathbb{E}[\alpha_k \left\|\hat{g}^k\right\|_2^2] + \frac{\mathrm{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})}{T}\sum_{k=0}^{T-1}\mathsf{bias}_{\|\cdot\|_C}(\tilde{g}^k). \tag{15}$$

Using Lemma B.1, we immediately obtain the following bound

$$\mathsf{bias}_{\|\cdot\|_C}(\tilde{g}^k) = \mathbb{E}\left[\left\|\tilde{g}^k - g^k\right\|_C\right] \leq \frac{\hat{G}_p(\mathcal{S};C)^p}{(p-1)B^{p-1}} \leq \frac{(2G_{2p}(C))^p}{(p-1)B^{p-1}} \tag{16}$$

Plugging (16) into (15), we obtain

$$\mathbb{E}[f(\overline{x}^T) - f(x^\star)] \leq \frac{\mathrm{diam}_2(\mathcal{X})^2}{2T\alpha_{T-1}} + \frac{1}{2T}\sum_{k=0}^{T-1}\mathbb{E}[\alpha_k \left\|\hat{g}^k\right\|_2^2] + \frac{\mathrm{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})\,(2G_{2p}(C))^p}{(p-1)B^{p-1}}. \tag{17}$$

Next, we substitute the value of α_k and use Lemma B.2 to obtain

$$\sum_{k=0}^{T-1}\mathbb{E}[\alpha_k \left\|\hat{g}^k\right\|_2^2] \leq 2\mathrm{diam}_2(\mathcal{X})\sqrt{\sum_{k=1}^{T}\mathbb{E}[\|\hat{g}^k\|_2^2]},$$

and by replacing it in (17), we obtain

$$\mathbb{E}[f(\overline{x}^T) - f(x^\star)] \leq \frac{3\mathrm{diam}_2(\mathcal{X})}{2T}\sqrt{\sum_{k=1}^{T}\mathbb{E}[\|\hat{g}^k\|_2^2]} + \frac{\mathrm{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})\,(2G_{2p}(C))^p}{(p-1)B^{p-1}}. \tag{18}$$

Finally, note that

$$\begin{aligned}\sqrt{\sum_{k=1}^{T}\mathbb{E}[\|\hat{g}^k\|_2^2]} &= \sqrt{\sum_{k=1}^{T}\mathbb{E}[\|\tilde{g}^k\|_2^2] + \frac{\log(1/\delta)}{b^2\epsilon^2}\sum_{k=0}^{T-1}\operatorname{tr}(A_k^{-1})} \\ &= \sqrt{\sum_{k=1}^{T}\mathbb{E}[\|\tilde{g}^k\|_2^2] + T\frac{B^2\log(1/\delta)\operatorname{tr}(C^{-1})}{b^2\epsilon^2}} \\ &\leq \sqrt{2}\left(\sqrt{\sum_{k=1}^{T}\mathbb{E}[\|\tilde{g}^k\|_2^2]} + \frac{B\sqrt{\log(1/\delta)\operatorname{tr}(C^{-1})}}{b\epsilon}\sqrt{T}\right), \end{aligned} \tag{19}$$

where the last inequality follows from the fact that $\sqrt{x+y} \leq \sqrt{2}\left(\sqrt{x}+\sqrt{y}\right)$ for nonnegative real numbers x and y. Plugging (19) into (18) completes the proof. □

B.5. Proof of Theorem 2

We first state the more general version of theorem:

Theorem 9. *Let $\mathcal{S}$ be a dataset with n points sampled from distribution P. Let C also be a diagonal and positive definite matrix. Consider running Algorithm 1 with $T = cn^2/b^2$ $A_k = C/B^2$ where $B > 0$ is a positive real number and c is given by Lemma 3.1. Then, with probability $1 - 1/n$, we have*

$$\begin{aligned}\mathbb{E}[f(\overline{x}^T;\mathcal{S}) - \min_{x\in\mathcal{X}} f(x;\mathcal{S})] \leq \mathcal{O}(1)\Bigg(&\frac{\operatorname{diam}_\infty(\mathcal{X})}{T}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(g_j^k)^2}\right] \\ &+\frac{\operatorname{diam}_\infty(\mathcal{X})B\sqrt{\log(1/\delta)}(\sum_{j=1}^{d}C_{jj}^{-\frac{1}{2}})}{n\varepsilon} + \frac{\operatorname{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})(2G_{2p}(C))^p}{(p-1)B^{p-1}}\Bigg),\end{aligned}$$

where the expectation is taken over the internal randomness of the algorithm.

Proof Similar to the proof of Theorem 1, we choose $x^\star \in \operatorname{argmin}_{x\in\mathcal{X}} f(x;\mathcal{S})$. We suppress the dependence of f on $\mathcal{S}$ throughout this proof as well. Again, we focus on the case that the bound

$$\hat{G}_p(\mathcal{S};C) \leq 2G_{2p}(C),$$

which we know its probability is at least $1 - 1/n$.

Using Theorem 7, we have

$$\mathbb{E}[f(\overline{x}^T) - f(x^\star)] \leq \frac{\operatorname{diam}_\infty(\mathcal{X})}{T}\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=0}^{T-1}(\hat{g}_j^k)^2}\right] + \frac{\operatorname{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})}{T}\sum_{k=0}^{T-1}\mathsf{bias}_{\|\cdot\|_C}(\tilde{g}^k).$$

Similar to the proof of Theorem 1, and by using Lemma B.1, we could bound the second term with

$$\frac{\operatorname{diam}_{\|\cdot\|_{C^{-1}}}(\mathcal{X})(2G_{2p}(C))^p}{(p-1)B^{p-1}}.$$

Now, it just suffices to bound the first term. Note that

$$\begin{aligned}
\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(\hat{g}_j^k)^2}\right] &= \sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(\tilde{g}_j^k+\xi_j^k)^2}\right] \\
&\le \sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}2\left((\tilde{g}_j^k)^2+(\xi_j^k)^2\right)}\right] \\
&\le 2\sum_{j=1}^{d}\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(\tilde{g}_j^k)^2}\right]+\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(\xi_j^k)^2}\right]\right) \qquad (20)\\
&\le 2\sum_{j=1}^{d}\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(\tilde{g}_j^k)^2}\right]+\sqrt{\mathbb{E}\left[\sum_{k=1}^{T}(\xi_j^k)^2\right]}\right) \qquad (21)\\
&\le 2\sum_{j=1}^{d}\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}(\tilde{g}_j^k)^2}\right]+2B\sqrt{T\log(1/\delta)}\frac{\sum_{j=1}^{d}C_{jj}^{-1/2}}{b\epsilon}, \qquad (22)
\end{aligned}$$

where (20) is obtained by using $\sqrt{x+y}\le\sqrt{2}\left(\sqrt{x}+\sqrt{y}\right)$ with $x=\sum_{k=1}^{T}(\tilde{g}_j^k)^2$ and $y=\sum_{k=1}^{T}(\xi_j^k)^2$, and (21) follows from $\mathbb{E}[X]\le\sqrt{\mathbb{E}[X^2]}$ with $X=\sqrt{\sum_{k=1}^{T}(\xi_j^k)^2}$.

□

C. Proof of Theorem 3

We begin with the following lemma, which upper bounds the bias from truncation.

Lemma C.1. *Let Z be a random vector satisfying Definition 4.1. Let $\sigma_j^2=\mathbb{E}[z_j^2]$ and $\Delta\ge 4r\sigma_j\log r$. Then we have*

$$|\mathbb{E}[\min(z_j^2,\Delta^2)]-\mathbb{E}[z_j^2]|\le\sigma_j^2/8.$$

Proof Let $\sigma_j^2=\mathbb{E}[z_j^2]$. To upper bound the bias, we need to upper bound $P(z_j^2\ge t\Delta^2)$. We have that z_j is $r^2\sigma_j^2$-sub-Gaussian therefore

$$P(z_j^2\ge tr^2\sigma_j^2)\le 2e^{-t}.$$

Thus, if $Y=|\min(z_j^2,\Delta^2)-z_j^2|$ then $P(Y\ge tr^2\sigma_j^2)\le 2e^{-t}$ hence

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\infty P(Y\ge t)dt \\
&= \int_0^\infty P(z_j^2\ge\Delta^2+t)dt \\
&\le \int_0^\infty 2e^{-(\Delta^2+t)/r^2\sigma_j^2}dt \\
&\le 2r^2\sigma_j^2e^{-\Delta^2/r^2\sigma_j^2}\le\sigma_j^2/8,
\end{aligned}$$

where the last inequality follows since $\Delta=4r\sigma_j\log r$. □

The following lemma demonstrates that the random variable $Y_i=\min(z_{i,j}^2,\Delta^2)$ quickly concentrates around its mean.

Lemma C.2. *Let Z be a random vector satisfying Definition 4.1. Then with probability at least $1-\beta$,*

$$\left|\frac{1}{n}\sum_{i=1}^{n}\min(z_{i,j}^2,\Delta^2)-\mathbb{E}[\min(z_j^2,\Delta^2)]\right|\le\frac{2r^2\sigma_j^2\sqrt{\log(2/\beta)}}{\sqrt{n}}.$$

Proof Let $Y_i = \min(z_{i,j}^2, \Delta^2)$. Since z_j is $r^2\sigma_j^2$-sub-Gaussian, we get that z_j^2 is $r^4\sigma_j^4$-sub-exponential, meaning that $\mathbb{E}[(z_j^2)^k]^{1/k} \le \mathcal{O}(k)r^2\sigma_j^2$ for all $k \ge 1$. Thus Y_i is also $r^4\sigma_j^4$-sub-exponential, and using Bernstein's inequality (Vershynin, 2019, Theorem 2.8.1), we obtain

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}[Y_i]\right| \ge t\right) \le 2\exp\left(-\min\left\{\frac{nt^2}{2r^4\sigma_j^4}, \frac{nt}{r^2\sigma_j^2}\right\}\right).$$

Setting $t = r^2\sigma_j^2\frac{2\sqrt{\log(2/\beta)}}{\sqrt{n}}$ yields the result. □

Given Lemmas C.1 and C.2, we are now ready to finish the proof of Theorem 3.

Proof of Theorem 3 First, privacy follows immediately, as each iteration t is $(\varepsilon/T, \delta/T)$-DP (using standard properties of the Gaussian mechanism (Dwork & Roth, 2014)), so basic composition implies that the final output is (ε, δ)-DP. We now proceed to prove the claim about utility. Let ρ_t^2 be the truncation value at iterate t, i.e., $\rho_t = 4r\log r/2^{t-1}$. First, note that Lemma C.2 implies that with probability $1-\beta/2$ for every $j \in [d]$

$$\left|\frac{1}{n}\sum_{i=1}^n \min(z_{i,j}^2, \rho_t^2) - \mathbb{E}[\min(z_j^2, \rho_t^2)]\right| \le \frac{2r^2\sigma_j^2\sqrt{\log(8d/\beta)}}{\sqrt{n}} \le \sigma_j^2/10,$$

and similar arguments show that

$$\left|\sigma_j^2 - \frac{1}{n}\sum_{i=1}^n z_{i,j}^2\right| \le \frac{2r^2\sigma_j^2\sqrt{\log(8d/\beta)}}{\sqrt{n}} \le \sigma_j^2/10,$$

where the last inequality follows since $n \ge 400r^4\log(8d/\beta)$. Moreover, for σ_j such that $\rho_t \ge 4r\sigma_j\log r$, Lemma C.1 implies that

$$|\mathbb{E}[\min(z_j^2, \rho_t^2) - \sigma_j^2| \le \sigma_j^2/8.$$

Let us now prove that if $\sigma_j = 2^{-k}$ then its value will be set at most at iterate $t = k$. Indeed at iterate $t = k$ we have $\rho_t = 4r2^{-k}\log r \ge 4r\sigma_j\log r$ hence we have that using the triangle inequality and standard concentration resutls for Gaussian distributions that with probability $1-\beta/2$

$$|\hat\sigma_{k,j}^2 - \sigma_j^2| \le \sigma_j^2/5 + \frac{16r^2T\sqrt{d}\log^2 r\log(T/\delta)\log(4d/\beta)}{2^{2k}n\varepsilon} \le \sigma_j^2/4,$$

where the last inequality follows since $n\varepsilon \ge 1000r^2T\sqrt{d}\log^2 r\log(T/\delta)\log(4d/\beta)$. Thus, in this case we get that $\hat\sigma_{k,j}^2 \ge \sigma_j^2/2 \ge 2^{-k-1}$ hence the value of coordinate j will best set at most at iterate k hence $\hat\sigma_j \ge \sigma_j/2$.

On the other hand, we now assume that $\sigma_j = 2^{-k}$ and show that the value of $\hat\sigma_j$ cannot be set before the iterate $t = k-3$ and hence $\hat\sigma_j \le 2^{-k+3} \le 8\sigma_j$. The above arguments show that at iterate t we have $\hat\sigma_{t,j}^2 \le 3/2\sigma_j^2 + \frac{1}{10\cdot 2^{2k}} \le 2^{-2k+1} + \frac{1}{10\cdot 2^{2k}} \le 2^{-2k+2}$ hence the first part of the claim follows.

To prove the second part, first note that z_j is $r\sigma_j$-sub-Gaussian, hence using Theorem 2, it is enough to show that $G_{2p}(\hat C) \le O(G_{2p}(C))$ and that $\sum_{j=1}^d \hat C_j^{-1/2} \le O(1)\cdot\sum_{j=1}^d C_j^{-1/2}$ where $C = (r\sigma_j)^{-4/3}$ is the optimal choice of C as in the bound (6). The first condition immediately follows from the definition of G_{2p} since $\hat C_j \le C_j$ for all $j \in [d]$. The latter condition follows immediately since $\frac{1}{2}\max(\sigma_j, 1/d^2) \le \hat\sigma_j$, implying

$$\sum_{j=1}^d \hat C_j^{-1/2} \le O(r^{-2/3})\sum_{j=1}^d \hat\sigma_j^{-2/3} \le O(r^{-2/3})\sum_{j=1}^d \sigma_j^{-2/3} + 1/d \le O(r^{-2/3})\sum_{j=1}^d \sigma_j^{-2/3}.$$

D. Proofs of Section 5 (Lower bounds)

D.1. Proof of Proposition 1

We begin with the following lemma which gives a lower bound for the sign estimation problem when $\sigma_j = \sigma$ for all $j \in [d]$. Asi et al. (2021) use similar result to prove lower bounds for private optimization over ℓ_1-bounded domains. For completeness, we give a proof in Section D.2.

Lemma D.1. *Let M be (ε, δ)-DP and $\mathcal{S} = (z_1, \ldots, z_n)$ where $z_i \in \mathcal{Z} = \{-\sigma, \sigma\}^d$. Then*

$$\sup_{\mathcal{S} \in \mathcal{Z}^n} \mathbb{E}\left[\sum_{j=1}^d |\bar{z}_j| 1\{\mathrm{sign}(M_j(\mathcal{S})) \neq \mathrm{sign}(\bar{z}_j)\}\right] \geq \min\left(\sigma d, \frac{\sigma d^{3/2}}{n\varepsilon \log d}\right).$$

We are now ready to complete the proof of Proposition 1 using bucketing-based techniques. First, we assume without loss of generality that $\sigma_j \leq 1$ for all $1 \leq j \leq d$ (otherwise we can divide by $\max_{1 \leq j \leq d} \sigma_j$). Now, we define buckets of coordinates $B_0, \ldots, B_K$ such that

$$B_i = \{j : 2^{-i-1} \leq \sigma_j \leq 2^{-i}\}.$$

For $i = K$, we set $B_K = \{j : \sigma_j \leq 2^{-K}\}$. We let $\sigma_{\max}(B_i) = \max_{j \in B_i} \sigma_j$ denote the maximal value of σ_j inside B_i. Similarly, we define $\sigma_{\min}(B_i) = \min_{j \in B_i} \sigma_j$. Focusing now on the i'th bucket, since $\sigma_j \geq \sigma_{\min}(B_i)$ for all $j \in B_i$, Lemma D.1 now implies (as $d \log^2 d \leq (n\varepsilon)^2$) the lower bound

$$\sup_{\mathcal{S} \in \mathcal{Z}^n} \mathbb{E}\left[\sum_{j \in B_i} |\bar{z}_j| 1\{\mathrm{sign}(M_j(\mathcal{S})) \neq \mathrm{sign}(\bar{z}_j)\}\right] \geq \frac{\sigma_{\min}(B_i)|B_i|^{3/2}}{n\varepsilon \log d}.$$

Therefore this implies that

$$\sup_{\mathcal{S} \in \mathcal{Z}^n} \mathbb{E}\left[\sum_{j=1}^d |\bar{z}_j| 1\{\mathrm{sign}(M_j(\mathcal{S})) \neq \mathrm{sign}(\bar{z}_j)\}\right] \geq \max_{0 \leq i \leq K} \frac{\sigma_{\min}(B_i)|B_i|^{3/2}}{n\varepsilon \log d}.$$

To finish the proof of the theorem, it is now enough to prove that

$$\sum_{j=1}^d \sigma_j^{2/3} \leq O(1)\ \log d \max_{0 \leq i \leq K} \sigma_{\min}(B_i)^{2/3}|B_i|.$$

We now have

$$\begin{aligned}
\sum_{j=1}^d \sigma_j^{2/3} &\leq \sum_{i=0}^K |B_i| \sigma_{\max}(B_i)^{2/3} \\
&\leq K \max_{0 \leq i \leq K-1} |B_i| \sigma_{\max}(B_i)^{3/2} \\
&\leq 4K \max_{0 \leq i \leq K-1} |B_i| \sigma_{\min}(B_i)^{3/2},
\end{aligned}$$

where the second inequality follows since the maximum cannot be achieved for $i = K$ given our choice of $K = 10 \log d$, and the last inequality follows since $\sigma_{\max}(B_i) \leq 2\sigma_{\min}(B_i)$ for all $i \leq K - 1$. This proves the claim.

D.2. Proof of Lemma D.1

Instead of proving lower bounds on the error of private mechanisms, it is more convenient for this result to prove lower bounds on the sample complexity required to achieve a certain error. Given a mechanism M and data $\mathcal{S} \in \mathcal{Z}^n$, define the error of the mechanism to be:

$$\mathsf{Err}(M, \mathcal{S}) = \mathbb{E}\left[\sum_{j=1}^d |\bar{z}_j| 1\{\mathrm{sign}(M_j(\mathcal{S})) \neq \mathrm{sign}(\bar{z}_j)\}\right].$$

The error of a mechanism for datasets of size n is $\mathsf{Err}(M, n) = \sup_{\mathcal{S} \in \mathcal{Z}^n} \mathsf{Err}(M, \mathcal{S})$.

We let $n^\star(\alpha, \varepsilon)$ denote the minimal n such that there is an (ε, δ)-DP (with $\delta = n^{-\omega(1)}$) mechansim M such that $\mathsf{Err}(M, n^\star(\alpha, \varepsilon)) \leq \alpha$. We prove the following lower bound on the sample complexity.

Proposition 3. *If* $\|z\|_\infty \le 1$ *then*

$$n^\star(\alpha, \varepsilon) \ge \Omega(1) \cdot \frac{d^{3/2}}{\alpha\varepsilon \log d}.$$

To prove this result, we first state the following lower bound for constant α and ε which follows from Theorem 3.2 in (Talwar et al., 2015).

Lemma D.2 (Talwar et al. (2015), Theorem 3.2)**.** *Under the above setting,*

$$n^\star(\alpha = d/4, \varepsilon = 0.1) \ge \Omega(1) \cdot \frac{\sqrt{d}}{\log d}.$$

We now prove a lower bound on the sample complexity for small values of α and ε which implies Proposition 3.

Lemma D.3. *Let* $\varepsilon_0 \le 0.1$*. For* $\alpha \le \alpha_0/2$ *and* $\varepsilon \le \varepsilon_0/2$*,*

$$n^\star(\alpha, \varepsilon) \ge \frac{\alpha_0\varepsilon_0}{\alpha\varepsilon} n^\star(\alpha_0, \varepsilon_0).$$

Proof Assume there exists an (ε, δ)-DP mechanism M such that $\mathsf{Err}(M, n) \le \alpha$. Then we now show that there is M' that is $(\varepsilon_0, \frac{2\varepsilon_0}{\varepsilon}\delta)$-DP with $n' = \Theta(\frac{\alpha\varepsilon}{\alpha_0\varepsilon_0} n)$ such that $\mathsf{Err}(M', n') \le \alpha_0$. This proves the claim. Let us now show how to define M' given M. Let $k = \lfloor \log(1+\varepsilon_0)/\varepsilon \rfloor$. For $\mathcal{S}' \in \mathcal{Z}^{n'}$, we define $\mathcal{S}$ to have k copies of $\mathcal{S}'$ and $(n - kn')/2$ users which have $z_i = (\sigma, \ldots, \sigma)$ and $(n - kn')/2$ users which have $z_i = (-\sigma, \ldots, -\sigma)$. Then we simply define $M'(\mathcal{S}') = M(\mathcal{S})$. Notice that now we have

$$\bar{z} = \frac{kn'}{n} \bar{z'}.$$

Therefore for a given $\mathcal{S}'$ we have that:

$$\mathsf{Err}(M', \mathcal{S}') = \frac{n}{kn'} \mathsf{Err}(M, \mathcal{S}) \le \frac{n\alpha}{kn'}$$

Thus if $n' \ge \frac{2n\alpha}{k\alpha_0}$ then

$$\mathsf{Err}(M', \mathcal{S}') \le \alpha_0.$$

Thus it remains to argue for the privacy of M'. By group privacy, M' is $(k\varepsilon, \frac{e^{k\varepsilon}-1}{e^\varepsilon - 1}\delta)$-DP, hence our choice of k implies that $k\varepsilon \le \varepsilon_0$ and $\frac{e^{k\varepsilon}-1}{e^\varepsilon-1}\delta \le \frac{2\varepsilon_0}{\varepsilon}\delta$. □

D.3. Proof of Theorem 5

We assume without loss of generality that $\sigma_j \le 1$ for all $1 \le j \le d$ (otherwise we can divide by $\max_{1 \le j \le d} \sigma_j$). We follow the bucketing-based technique we had in the proof of Proposition 1. We define buckets of coordinates $B_0, \ldots, B_K$ such that

$$B_i = \{j : 2^{-i-1} \le \sigma_j \le 2^{-i}\}.$$

For $i = K$, we set $B_K = \{j : \sigma_j \le 2^{-K}\}$. We let $\sigma_{\max}(B_i) = \max_{j \in B_i} \sigma_j$ denote the maximal value of σ_j inside B_i. Similarly, we define $\sigma_{\min}(B_i) = \min_{j \in B_i} \sigma_j$. Focusing now on the i'th bucket, since $\sigma_j \ge \sigma_{\min}(B_i)$ for all $j \in B_i$, Proposition 2 now implies the lower bound

$$\sup_{\mathcal{S} \in \mathcal{Z}^n} \mathbb{E}\left[f(M(\mathcal{S}); \mathcal{S}) - f(x^\star_{\mathcal{S}}; \mathcal{S})\right] \ge \min\left(\sigma_{\min}(B_i)\sqrt{|B_i|}, \frac{|B_i|\sigma_{\min}(B_i)}{n\varepsilon}\right).$$

Since $d \le (n\varepsilon)^2$, taking the maximum over buckets, we get that the error of any mechanism is lower bounded by:

$$\sup_{\mathcal{S} \in \mathcal{Z}^n} \mathbb{E}\left[f(M(\mathcal{S}); \mathcal{S}) - f(x^\star_{\mathcal{S}}; \mathcal{S})\right] \ge \max_{0 \le i \le K} \frac{|B_i|\sigma_{\min}(B_i)}{n\varepsilon}.$$

To finish the proof, we only need to show now that

$$\frac{\sum_{j=1}^{d} \sigma_j}{\log d} \leq O(1) \max_{0 \leq i \leq K} |B_i| \sigma_{\min}(B_i).$$

Indeed, we have that

$$\begin{aligned}
\sum_{j=1}^{d} \sigma_j &\leq \sum_{i=0}^{K} |B_i| \sigma_{\max}(B_i) \\
&\leq K \max_{0 \leq i \leq K-1} |B_i| \sigma_{\max}(B_i) \\
&\leq 2K \max_{0 \leq i \leq K-1} |B_i| \sigma_{\min}(B_i),
\end{aligned}$$

where the second inequality follows since the maximum cannot be achieved for $i = K$ given our choice of $K = 10 \log d$, and the last inequality follows since $\sigma_{\max}(B_i) \leq 2\sigma_{\min}(B_i)$ for all $i \leq K - 1$. The claim follows.