Appendix of On Differentially Private Stochastic Convex Optimization with Heavy-tailed Data

Di Wang [*1,2] **Hanshen Xiao** [*3] **Srini Devadas** [3] **Jinhui Xu** [1]

A. Omitted Proofs

Proof of Lemma 1. Before the proof, we recall the following two lemmas

Lemma 1 ((Srebro *et al.*, 2010)). If a non-negative function $f : \mathcal{W} \mapsto \mathbb{R}_+$ is β-smooth, then $\|\nabla f(w)\|_2^2 \leq 4\beta f(w)$ for all $w \in \mathcal{W}$.

subscribe

Lemma 2 ((Juditsky and Nemirovski, 2008)). Let $X_1, X_2, \cdots, X_n$ be independent copies of a zero-mean random vector X, then $\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n X_i\|_2^2 \leq \frac{1}{n}\mathbb{E}\|X\|_2^2$.

Consider $w = w^*$. Then by Assumption 1, we have $\nabla L(w^*) = \mathbb{E}[\nabla \ell(w^*, x)] = 0$. Thus, by Lemma 2 we have

$$\mathbb{E}\|\nabla \hat{L}(w^*, D)\|_2^2 \leq \frac{1}{n}\mathbb{E}[\|\nabla \ell(w^*, x)\|_2^2].$$

By Markov's inequality, we get

$$\Pr[\|\nabla \hat{L}(w^*, D)\|_2^2 \leq \frac{10}{n}\mathbb{E}[\|\nabla \ell(w^*, x)\|_2^2] \geq \frac{9}{10}.$$

Since $n \geq n_\alpha$, by the assumption we have with probability at least $\frac{5}{6}$ that $\hat{L}(w, D)$ is α strongly convex. Thus, we get

$$\begin{aligned}&\frac{\alpha}{2}\|w_D - w^*\|_2^2 \leq \\ &- \langle \nabla \hat{L}(w^*, D), w_D - w^* \rangle + \hat{L}(w_D, D) - \hat{L}(w^*, D) \\ &\leq \|\nabla \hat{L}(w^*, D)\|_2 \|w_D - w^*\|_2.\end{aligned}$$

In total, with probability at least $\frac{3}{4}$, we have

$$\|w_D - w^*\|_2 \leq \sqrt{\frac{40\mathbb{E}\|\nabla \ell(w^*, x)\|_2^2}{n\alpha^2}}.$$

□

Proof of Theorem 2. For each subsample set D_{S_i}, by the assumption we have its size $\frac{n}{m} \geq n_\alpha$. Thus, Lemma 1 holds with $n = \frac{n}{m}$. That is, (1) holds with $r = \sqrt{\frac{40m\mathbb{E}\|\nabla \ell(w^*, x)\|_2^2}{n\alpha^2}}$. Hence, by Theorem 1 we have

$$\|\mathcal{A}(D) - w^*\|_2 \leq O(\frac{\sqrt{d}r}{\epsilon}) = O(\sqrt{\frac{dm\mathbb{E}\|\nabla \ell(w^*, x)\|_2^2}{n\epsilon^2\alpha^2}}).$$

Since $L_{\mathcal{D}}(w)$ is β-smooth and $\nabla L_{\mathcal{D}}(w^*) = 0$, we have $L_{\mathcal{D}}(\mathcal{A}(D)) - L_{\mathcal{D}}(w^*) \leq \frac{\beta}{2}\|\mathcal{A}(D) - w^*\|_2^2$. Also, by Lemma 1 and the non-negative property we get

$$L_{\mathcal{D}}(\mathcal{A}(D)) - L_{\mathcal{D}}(w^*) \leq O((\frac{\beta}{\alpha})^2 \frac{dm}{n\epsilon^2} L_{\mathcal{D}}(w^*)).$$

Taking $m = \tilde{\Theta}(\frac{d^2}{\epsilon^2})$, we get the proof. □

Proof of Theorem 4. We first give the definition of zCDP in (Bun and Steinke, 2016).

Definition 1. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \mapsto \mathcal{Y}$ is ρ-zero Concentrated Differentially Private (zCDP) if for all neighboring datasets $D \sim D'$ and all $\alpha \in (1, \infty)$,

$$D_\alpha(\mathcal{A}(D)\|\mathcal{A}(D')) \leq \rho\alpha,$$

where $D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log \mathbb{E}_{X\sim P}[(\frac{P(X)}{Q(X)})^{\alpha-1}]$ denotes the Rényi divergence of order α.

We first convert (ϵ, δ)-DP to $\frac{1}{2}\tilde{\epsilon}^2$-zCDP by using the following lemma

Lemma 3 ((Bun and Steinke, 2016)). Let $M : \mathcal{X}^n \mapsto \mathcal{Y}$ be a randomized algorithm. If M is $\frac{1}{2}\epsilon^2$-zCDP, it is $(\frac{1}{2}\epsilon^2 + \epsilon \cdot \sqrt{2\log\frac{1}{\delta}}, \delta)$-DP for all $\delta > 0$.

Thus, it suffices to show that Algorithm 3 is $\frac{1}{2}\tilde{\epsilon}^2$-zCDP. We note that in each iteration and each coordinate, outputting $\nabla_{t-1,j}$ will be $\frac{1}{2}\frac{\tilde{\epsilon}^2}{dT}$-zCDP by Theorem 3. Thus by the composition property of CDP, we know that it is $\frac{1}{2}\tilde{\epsilon}^2$-zCDP. □

Proof of Lemma 2. By assumption, we know that $\mathcal{W}$ is closed and bounded, and hence it is compact. By

[*]Equal contribution [1]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY [2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia [3]CSAIL, MIT, Cambridge, MA. Correspondence to: Di Wang <dwang45@buffalo.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

(Lorentz, 1966) we know that its covering number with radius δ (will be specified later) is bounded from above as $N_\delta \leq (\frac{3\Delta}{2\delta})^d$. Denote the center of this δ-net as $\tilde{\mathcal{W}} = \{\tilde{w}_1, \tilde{w}_2, \cdots, \tilde{w}_{N_\delta}\}$.

We first fix $j \in [d]$ and consider $|\tilde{\nabla}_j(w) - \nabla_j L_D(w)|$ (we omit the subscript $t-1$). Then, we have

$$\begin{aligned}
&\mathbb{E}_{Z_j}(\tilde{\nabla}_j(w) - \nabla_j L_D(w))^2 = \\
&\mathbb{E}\big([\text{Trim}_m(D_j(w))]_{[a,b]} + \frac{1}{s} S^t_{[trim(\cdot)]_{[a,b]}}(D_j(w)) \cdot Z_j \\
&- \nabla_j L_D(w)\big)^2 \\
&\leq O(([\text{Trim}_m(D_j(w))]_{[a,b]} - \nabla_j L_D(w))^2 \\
&+ \mathbb{E}(\frac{1}{s} S^t_{[trim(\cdot)]_{[a,b]}}(D_j(w)) \cdot Z_j)^2) \\
&\leq O((\text{Trim}_m(D_j(w))] - \nabla_j L_D(w))^2 \\
&+ \mathbb{E}(\frac{1}{s} S^t_{[trim_m(\cdot)]_{[a,b]}}(D(w)) \cdot Z_j)^2), \qquad (1)
\end{aligned}$$

where $D_j(w) = \{\nabla_j \ell(w, x_i)\}_{i=1}^n$ and the last inequality is due to the property that the truncation operation reduces error.

Lemma 4. Let $a \leq \mu \leq b$ and X be a random variable. Then

$$([X]_{[a,b]} - \mu)^2 \leq (x - \mu)^2.$$

By the proof of Theorem 51 in (Bun and Steinke, 2019) and the fact that $\epsilon = \frac{\tilde{\epsilon}}{\sqrt{dT}}$, we have ($m, a, b = O(1)$)

$$\mathbb{E}_Z(\frac{1}{s} S^t_{[trim_m(\cdot)]_{[a,b]}}(D_j(w)) \cdot Z)^2 \leq O(\frac{\tau^2 dT \log n}{n\tilde{\epsilon}^2}), \qquad (2)$$

where the O-notation omits the $\log \sigma^2$ and $\log(b-a)$ factors.

Next, we bound the first term of (1). Before showing that, we first give the following estimation error on the trimming operation for sub-exponential random variables.

Lemma 5. Suppose that x_i are i.i.d υ-sub-exponential with mean μ. Then, the following holds for any $t \geq 0$,

$$\mathbb{P}\{\frac{1}{n}\sum_{i=1}^n x_i - \mu \geq t\} \leq 2\exp(-n\min\{\frac{t}{2\upsilon}, \frac{t^2}{2\upsilon^2}\}),$$

and for any $s \geq 0$,

$$\mathbb{P}[\max_{i\in[n]}\{|x_i - \mu|\} \geq s] \leq 2n\exp(-\min\{\frac{s}{2\upsilon}, \frac{s^2}{2\upsilon^2}\}),$$

and for any $m \geq 0$, under the above two events,

$$|\text{Trim}_m(\{x_i\}_{i=1}^n) - \mu| \leq \frac{nt + ms}{n - 2m}.$$

Proof of Lemma 5. Note that the first two inequalities are just the Berstein's Inequality. We only prove the last inequality.

Let $\mathcal{T} \subset [n]$ denote the set of all trimmed variables and $\mathcal{U} = [n]\backslash\mathcal{T}$. Then, we know that $\text{Trim}_m(\{x_i\}_{i=1}^n) = \frac{\sum_{i\in\mathcal{U}} x_i}{n-2m}$. Thus, we have

$$\begin{aligned}
&|\frac{\sum_{i\in\mathcal{U}} x_i}{n-2m} - \mu| = \frac{1}{n-2m}|\sum_{i\in[n]}(x_i - \mu) - \sum_{i\in\mathcal{T}}(x_i - \mu)| \\
&\leq \frac{1}{n-2m}(|\sum_{i\in[n]}(x_i - \mu)| + |\sum_{i\in\mathcal{T}}(x_i - \mu)|). \qquad (3)
\end{aligned}$$

For the second term of (3), we have $|\sum_{i\in\mathcal{T}}(x_i - \mu)| \leq m\max\{|x_i - \mu|\}$. Plugging the inequalities into (3) we get the proof. □

Now, fix any $w \in \mathcal{W}$, we know that there exists a $\tilde{w}$ which is in the δ-net, *i.e.,* $\|\tilde{w} - w\|_2 \leq \delta$. Then by using the Bernstein inequality and the sub-exponential assumption and taking the union bound, we can see that with probability at least $1 - 2dN_\delta \exp(-n\min\{\frac{t}{2\tau}, \frac{t^2}{2\tau^2}\})$, we have the following for all $j \in [d]$ and $\tilde{w} \in \tilde{\mathcal{W}}$

$$|\sum_{i=1}^n \frac{\nabla_j \ell(\tilde{w}, x_i)}{n} - \nabla_j L_{\mathcal{D}}(\tilde{w})| \leq t, \qquad (4)$$

and with probability at least $1 - 2dnN_\delta \exp(-\min\{\frac{s}{2\tau}, \frac{s^2}{2\tau^2}\})$, we get the following for all $j \in [d]$ and $\tilde{w} \in \tilde{W}$,

$$\max_{i\in[n]} |\nabla_j \ell(\tilde{w}, x_i) - \nabla_j L_{\mathcal{D}}(\tilde{w})| \leq s. \qquad (5)$$

By the β_j-smoothness of $\ell_j(\cdot, x)$ we have

$$|\sum_{i=1}^n \frac{\nabla_j \ell(\tilde{w}, x_i)}{n} - \sum_{i=1}^n \frac{\nabla_j \ell(w, x_i)}{n}| \leq \beta_j \|w - \tilde{w}\|_2 \leq \beta_j \delta, \qquad (6)$$

$$|\nabla_j L_{\mathcal{D}}(\tilde{w}) - \nabla_j L_{\mathcal{D}}(w)| \leq \beta_j \delta. \qquad (7)$$

Thus, we get

$$|\sum_{i=1}^n \frac{\nabla_j \ell(w, x_i)}{n} - \nabla_j L_{\mathcal{D}}(w)| \leq t + 2\beta_j\delta \qquad (8)$$

$$\max_{i\in[n]} |\nabla_j \ell(w, x_i) - \nabla_j L_{\mathcal{D}}(w)| \leq s + 2\beta_j\delta. \qquad (9)$$

By Lemma 5 we have for all $j \in [d]$ and $w \in \mathcal{W}$

$$|\text{Trim}_m(D_j(w)) - \nabla_j L_{\mathcal{D}}(w)| \leq \frac{nt + ms}{n - 2m} + \frac{m + n}{n - 2m} 2\beta_j\delta.$$

Combining this with (2) we have the following with probability at least $1 - 2dnN_\delta \exp(-\min\{\frac{s}{2\tau}, \frac{s^2}{2\tau^2}\}) - 2dN_\delta \exp(-n\min\{\frac{t}{2\tau}, \frac{t^2}{2\tau^2}\})$ for all $j \in [d]$ and $\tilde{w} \in \tilde{\mathcal{W}}$,

$$\mathbb{E}\|\nabla \tilde{L}(w, D) - \nabla L_{\mathcal{D}}(w)\|_2 \leq \\ \leq O(\sqrt{d}\frac{nt + ms}{n - 2m} + \hat{\beta}\delta\frac{m+n}{n-2m} + \frac{\tau d\sqrt{T\log n}}{\sqrt{n\tilde{\epsilon}}}), \quad (10)$$

where $\hat{\beta} = \sqrt{\beta_1^2 + \cdots + \beta_d^2}$. Thus, let $\delta = \frac{1}{n\hat{\beta}}, m = O(1)$,

$$t = O(\tau \max\{\frac{d}{n}\log(n\hat{\beta}\Delta), \sqrt{\frac{d}{n}\log(n\hat{\beta}\Delta)}\}),$$

$$s = O(\tau d \log(\hat{\beta} n \Delta)).$$

Then, we get the proof. □

Proof of Theorem 5. In the t-th iteration, let

$$\hat{w}^t = w^{t-1} - \eta \nabla \tilde{L}(w^{t-1}, D).$$

Then, by the property of Euclidean project we have

$$\|w^t - w^{t-1}\|_2 \leq \|\hat{w}^t - w^{t-1}\|_2.$$

Hence, we have

$$\|\hat{w}^t - w^*\|_2 \leq \|w^{t-1} - \eta\nabla\tilde{L}(w^{t-1}, D) - w^*\|_2 \\ \leq \|w^{t-1} - \eta\nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2 \\ + \eta\|\nabla\tilde{L}(w^{t-1}, D) - L_{\mathcal{D}}(w^{t-1})\|_2.$$

For the first term, by the co-coercivity of strongly convex functions (Bubeck and others, 2015), we have

$$\langle w^{t-1} - w^*, \nabla L_{\mathcal{D}}(w^{t-1})\rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|w^{t-1} - w^*\|_2^2 \\ + \frac{1}{\alpha+\beta}\|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2.$$

Thus we obtain the following by taking $\eta = \frac{1}{\beta}$

$$\|w^{t-1} - \eta\nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2^2 \leq \\ (1 - \frac{2\alpha}{\alpha+\beta})\|w^{t-1} - w^*\|_2^2 - \frac{2}{\beta(\beta+\alpha)}\|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ + \frac{1}{\beta^2}\|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ \leq (1 - \frac{2\alpha}{\alpha+\beta})\|w^{t-1} - w^*\|_2^2. \quad (11)$$

Taking the expectation w.r.t Z_{t-1} and using the inequality of $\sqrt{1-x} \leq 1 - \frac{x}{2}$ and Lemma 4, we have

$$\mathbb{E}\|\hat{w}^t - w^*\|_2 \leq (1 - \frac{\alpha}{\alpha+\beta})\mathbb{E}\|w^{t-1} - w^*\|_2 + O(\frac{\tau d\sqrt{T\log n}}{\beta\sqrt{n\tilde{\epsilon}}}). \quad (12)$$

That is,

$$\mathbb{E}\|\hat{w}^T - w^*\|_2 \leq (1 - \frac{\alpha}{\beta+\alpha})^T\Delta + O(\frac{\beta}{\alpha}\frac{\tau d\sqrt{T\log n}}{\beta\sqrt{n\tilde{\epsilon}}}).$$

Thus, taking $T = O(\frac{\beta}{\alpha}\log n)$, we have the following with probability at least $1 - \Omega(\frac{2dn\log n}{(1+n\tilde{L}\Delta)^d})$

$$\mathbb{E}\|\hat{w}^t - w^*\|_2 \leq O(\sqrt{\frac{\beta}{\alpha}}\frac{\Delta\tau d\log n}{\alpha\sqrt{n\tilde{\epsilon}}}).$$

Since $\tilde{\epsilon} = \sqrt{2\log\frac{1}{\delta} + 2\epsilon} - \sqrt{2\log\frac{1}{\delta}}$, by using the Taylor series of the function $\sqrt{x+1} - \sqrt{x}$, we have $\tilde{\epsilon} = O(\frac{\epsilon}{\sqrt{\log\frac{1}{\delta}}})$. Since $L_{\mathcal{D}}(w)$ is β-smooth we have $\mathbb{E}L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \leq \frac{\beta}{2}\mathbb{E}\|w^T - w^*\|_2^2$. Thus we get the proof. □

Proof of Theorem 7. The proof of (ϵ, δ)-DP is the same as in the proof of Theorem 3. The ℓ_2 sensitivity is $\frac{s}{n}\frac{4\sqrt{2}}{3}$.

Next, we show the upper bound. The key lemma on the uniform converge rate is the following. For convenience, we denote by

$$\hat{g}_j(w) = \frac{1}{n}\sum_{i=1}^n(\nabla_j\ell(w, x_i)\big(1 - \frac{\nabla_j^2\ell(w,x_i)}{2s^2\beta}\big) \\ -\frac{\nabla_j^3\ell(w,x_i)}{6s^2}) + \frac{1}{n}\sum_{i=1}^n C\left(\frac{\nabla_j\ell(w,x_i)}{s}, \frac{|\nabla_j\ell(w,x_i)|}{s\sqrt{\beta}}\right)$$

and $\hat{g}_j(w) = (\hat{g}_1(w), \hat{g}_2(w), \cdots, \hat{g}_d(w))$.

Lemma 6 (Lemma 8 in (Holland, 2019))**.** Under Assumptions 1 and 4, with probability at least $1 - \delta'$, the following holds for any $w \in \mathcal{W}$,

$$\|\hat{g}_j(w) - \mathbb{E}[\nabla\ell(w, x)]\|_2 \leq O(\frac{\beta d\sqrt{v\log(\frac{1}{\delta'}\Delta n)}}{\sqrt{n}}). \quad (13)$$

Thus, we have the following lemma.

Lemma 7. Under the assumptions in the previous lemma, the following holds with probability at least $1 - 2\delta'$ for any $w \in \mathcal{W}$

$$\|g_j(w) - \mathbb{E}[\nabla\ell(w, x)]\|_2 \leq O(\frac{\beta d\sqrt{vT\log(\frac{1}{\delta'}\Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}}). \quad (14)$$

The remaining proof is almost the same as the proof of Theorem 5 by using Lemma 7. We omit it here for convenience. □

Proof of Theorem 8. Let $\hat{w}^t$ denote the same notation as in the proof of Theorem 5. Then, we have

$$\begin{aligned}&\|\hat{w}^t - w^*\|_2 \le \|w^{t-1} - \eta g^{t-1}(w^{t-1}) - w^*\|_2 \\ &\le \|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2 \\ &+ \eta \|g^{t-1}(w^{t-1}) - L_{\mathcal{D}}(w^{t-1})\|_2,\end{aligned}$$

and

$$\begin{aligned}&\|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2^2 \le \|w^{t-1} - w^*\|_2^2 \\ &- 2\eta \langle \nabla L_{\mathcal{D}}(w^{t-1}), w^{t-1} - w^* \rangle + \eta^2 \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ &\le \|w^{t-1} - w^*\|_2^2 - 2\eta \frac{1}{\beta} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 + \eta^2 \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ &\le \|w^{t-1} - w^*\|_2^2.\end{aligned}$$

Thus by Lemma 7 we have with probability at least $1 - 2\delta'$

$$\|\hat{w}^t - w^*\|_2 \le \|w^{t-1} - w^*\|_2 + O(\frac{d\sqrt{vT\log(\frac{1}{\delta'}\Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}}). \tag{15}$$

Hence, when $O(\frac{dT\sqrt{vT\log(\frac{1}{\delta'}\Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}}) \le \|w^0 - w^*\|_2$, we have $\hat{w}^t \in \mathcal{W}$ for all $t = \{1, \cdots, T\}$ with probability at least $1 - 2\delta' T$. This means that $\hat{w}^t = w^t$ for all $t \in [T]$. Hence, we proceed to study the algorithm without projection. Let $D_t = \|w^0 - w^*\|_2 + O(\frac{dt\sqrt{vT\log(\frac{1}{\delta'}\Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}})$ for $t = \{0, 1, \cdots, T\}$. By the smoothness of $L_{\mathcal{D}}(\cdot)$ we have

$$\begin{aligned}&L_{\mathcal{D}}(w^t) \le L_{\mathcal{D}}(w^{t-1}) + \langle \nabla L_{\mathcal{D}}(w^{t-1}), w^t - w^{t-1} \rangle \\ &+ \frac{\beta}{2}\|w^t - w^{t-1}\|_2^2 \\ &= L_{\mathcal{D}}(w^{t-1}) + \eta \langle \nabla L_{\mathcal{D}}(w^{t-1}), -g^{t-1}(w^{t-1}) + \nabla L_{\mathcal{D}}(w^{t-1}) \\ &- \nabla L_{\mathcal{D}}(w^{t-1}) \rangle + \eta^2 \frac{\beta}{2} \|g^{t-1}(w^{t-1}) - \nabla L_{\mathcal{D}}(w^{t-1}) \\ &+ \nabla L_{\mathcal{D}}(w^{t-1})\|_2^2.\end{aligned}$$

Since $\eta = \frac{1}{\beta}$, by simple calculation we have

$$\begin{aligned}L_{\mathcal{D}}(w^t) \le L_{\mathcal{D}}(w^{t-1}) - \frac{1}{2\beta}\|\nabla L_(w^{t-1})\|^2 \\ + O(\frac{\beta d^2 v T \log(\frac{1}{\delta'}\Delta n)}{n\tilde{\epsilon}}).\end{aligned} \tag{16}$$

Next we show the following lemma

Lemma 8. Assume that events (14) hold for all $t = \{1, \cdots, T\}$. Then there exists at least one $t \in \{1, \cdots, T\}$ such that

$$L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \le 16 D_0 \chi,$$

where $\chi = O(\frac{\beta d\sqrt{vT\log(\frac{1}{\delta'}\Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}})$.

Proof. We note that $D_t \le 2D_0$ for all $t = 0, \cdots, T$. Thus we have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \le \|\nabla L_{\mathcal{D}}(w)\|_2 \|w - w^*\|_2,$$

which implies that

$$\|\nabla L_{\mathcal{D}}(w)\|_2 \ge \frac{L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*)}{\|w - w^*\|_2}.$$

Suppose that there exists $t \in \{1, 2, \cdots, T\}$ such that $\|\nabla L_{\mathcal{D}}(w^t)\|_2 < \sqrt{2}\chi$. Then, we have $L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \le \|\nabla L_{\mathcal{D}}(w^t)\|_2 \|w^t - w^*\|_2 \le 2\sqrt{2} D_0 \chi$.

Otherwise suppose that for all $\{1, 2, \cdots, T\}$, $\|\nabla L_{\mathcal{D}}(w^t) \ge \sqrt{2}\chi$. Then, we have the following for all $t \le T$,

$$\begin{aligned}&L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \le L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*) \\ &- \frac{1}{4\beta}\|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ &\le L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*) - \frac{1}{4\beta D_{t-1}^2}(L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)).\end{aligned}$$

Multiplying both side by $[(L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*))(L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*))]^{-1}$ we get

$$\begin{aligned}\frac{1}{L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)} &\ge \frac{1}{L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)} \\ &+ \frac{1}{4\beta D_{t-1}^2} \frac{L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)}{L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)} \\ &\ge \frac{1}{L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)} + \frac{1}{16\beta D_0^2},\end{aligned}$$

where the last inequality is due to the facts that $D_t \le 2D_0$ and $L_{\mathcal{D}}(w^{t-1}) \ge L_{\mathcal{D}}(w^t)$.

Hence, we have

$$\frac{1}{L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*)} \ge \frac{T}{16\beta D_0^2} \ge \frac{1}{16 D_0 \chi} \tag{17}$$

using the fact that $T = \frac{\beta D_0}{\chi}$, that is, $T = \tilde{O}\left(\frac{\|w^0 - w^*\|_2 \sqrt{n}\sqrt{\tilde{\epsilon}}}{d}\right)^{\frac{2}{3}}$. Thus $\chi = \tilde{O}(\Delta \frac{d^{\frac{2}{3}}}{(n\tilde{\epsilon})^{\frac{1}{3}}})$. □

Next we show that

$$L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \le 16 D_0 \chi + \frac{1}{2\beta}\chi^2. \tag{18}$$

Let $t = t_0$ be the first time that $L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \le 16 D_0 \chi$. We show that for any $t \ge t_0$, $L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \le 16 D_0 \chi + \frac{1}{2\beta}\chi^2$. If not, let t_1 be the first time that $L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) > 16 D_0 \chi + \frac{1}{2\beta}\chi^2$. Then, we must have $L_{\mathcal{D}}(w^{t_1}) > L_{\mathcal{D}}(w^{t_1 - 1})$. By (16) we have

$$\begin{aligned}&L_{\mathcal{D}}(w^{t_1 - 1}) - L_{\mathcal{D}}(w^*) \ge \\ &L_{\mathcal{D}}(w^{t_1}) - L_{\mathcal{D}}(w^*) - \frac{1}{2\beta}\chi^2 > 16 D_0 \chi.\end{aligned}$$

Thus, we have

$$\|\nabla L_{\mathcal{D}}(w^{t_1-1})\|_2 \geq \frac{L_{\mathcal{D}}(w^{t_1-1}) - L_{\mathcal{D}}(w^*)}{\|w^{t_1-1} - w^*\|_2} \geq 8\chi.$$

By (16) we have $L_{\mathcal{D}}(w^{t_1}) \leq L_{\mathcal{D}}(w^{t_1-1})$ which is a contradiction. □

B. Explicit Form of $C(a, b)$ in (10)

We first define the following notations:

$$V_- := \frac{\sqrt{2}-a}{b}, V_+ = \frac{\sqrt{2}+a}{b} \tag{19}$$

$$F_- := \Phi(-V_-), F_+ := \Phi(-V_+) \tag{20}$$

$$E_- := \exp(-\frac{V_-^2}{2}), E_+ := \exp(-\frac{V_+^2}{2}), \tag{21}$$

where Φ denotes the CDF of the standard Gaussian distribution. Then

$$C(a, b) = T_1 + T_2 + \cdots + T_5, \tag{22}$$

where

$$T_1 := \frac{2\sqrt{2}}{3}(F_- - F_+) \tag{23}$$

$$T_2 := -(a - \frac{a^3}{6})(F_- + F_+) \tag{24}$$

$$T_3 := \frac{b}{\sqrt{2\pi}}(1 - \frac{a^2}{2})(E_+ - E_-) \tag{25}$$

$$T_4 := \frac{ab^2}{2}\left(F_+ + F_- + \frac{1}{\sqrt{2\pi}}(V_+E_+ + V_-E_-)\right) \tag{26}$$

$$T_5 := \frac{b^3}{6\sqrt{2\pi}}\left((2 + V_-^2)E_- - (2 + V_+^2)E_+\right). \tag{27}$$

C. Full description of experiments

For the synthetic data generation, we select the parameters $(\mu = 1, \sigma = 1)$ and $(\mu = 0.2, \sigma = 0.2)$ for the Lognormal and Loglogistic noises underlying, respectively. The step size of Algorithm 3 is set to 0.01 where $m = 0.05n$. As for algorithm 4, $v = 5$, failure probability $\delta' = 0.01$ and the step size is set to 0.1. For the stochastic Algorithm 4, the step size is selected as $\frac{1}{\sqrt{t}}$, where t is the iteration number. Accordingly, $\bar{w}^T = \frac{\sum_{t=1}^{T} w^t}{T}$. Corresponding to Fig. 1 and 2, we present the results which also mark the difference between the best and the worst performances as follows.

To measure the impact from dimension on performances, we fix $n = 10^5$ and test d varying from 10 to 50 through stochastic Algorithm 4 and RGD under the same setup as above. To test the impact from the size of the dataset, we fix $d = 20$ and test n varying from 2×10^4 to 10^5.

References

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

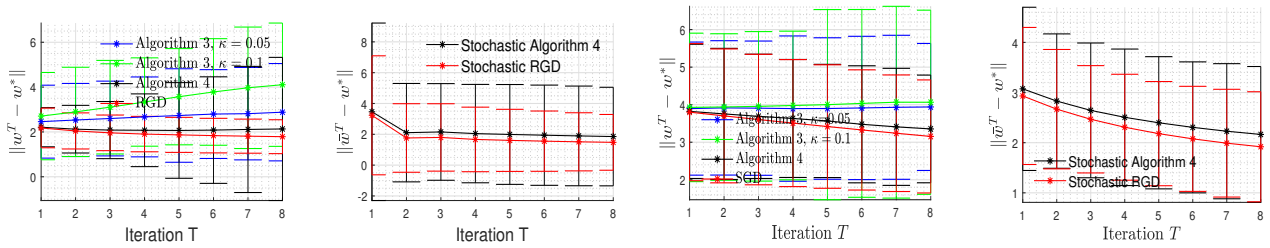
Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *arXiv preprint arXiv:1906.02830*, 2019.

Matthew J Holland. Robust descent using smoothed multiplicative noise. In *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 703–711, 2019.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

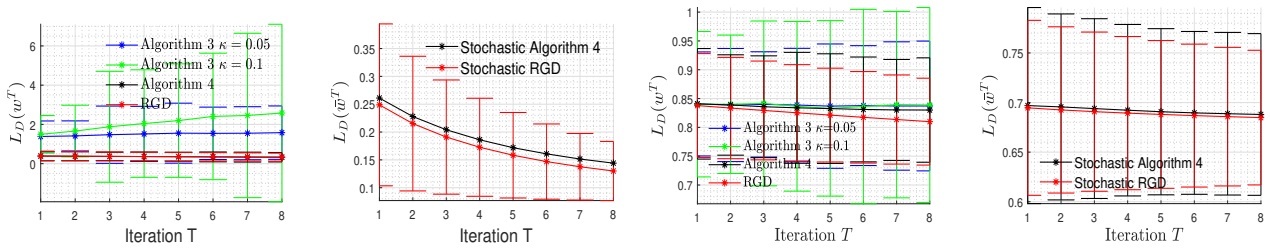
GG Lorentz. Metric entropy and approximation. *Bulletin of the American Mathematical Society*, 72(6):903–937, 1966.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.



(a) $\epsilon = 1$ (b) $\epsilon = 0.5$ (c) $\epsilon = 1$ (d) $\epsilon = 0.5$

Figure 1. Experiments on synthetic datasets. Figures (a) and (b) are for ridge regressions over synthetic data with Lognormal noises. Figures (c) and (d) are for logistic regressions over synthetic data with Loglogistic noises.



(a) $\epsilon = 1$ (b) $\epsilon = 0.5$ (c) $\epsilon = 1$ (d) $\epsilon = 0.5$

Figure 2. Experiments on UCI Adult dataset. Figures (a) and (b) are for ridge regressions. Figures (c) and (d) are for logistic regressions.