Improved Rates for Differentially Private Stochastic Convex Optimization with Heavy-Tailed Data

Gautam Kamath, University of Waterloo
Xingtu Liu, University of Waterloo
Huanyu Zhang, Meta

Table of contents

Contents

Problem formulation

Stochastic convex optimization (SCO)

A fundamental optimization problem in machine learning.

- n : sample size.

Stochastic convex optimization (SCO)

A fundamental optimization problem in machine learning.

- n : sample size.
- $\mathcal{D}$: an unknown distribution over $\mathcal{X} \subseteq \mathbb{R}^d$.

Stochastic convex optimization (SCO)

A fundamental optimization problem in machine learning.

- n : sample size.
- $\mathcal{D}$: an unknown distribution over $\mathcal{X} \subseteq \mathbb{R}^d$.
- $\mathcal{W}$: parameter space, a subset of $\mathbb{R}^d$.

Stochastic convex optimization (SCO)

A fundamental optimization problem in machine learning.

- n : sample size.
- $\mathcal{D}$: an unknown distribution over $\mathcal{X} \subseteq \mathbb{R}^d$.
- $\mathcal{W}$: parameter space, a subset of $\mathbb{R}^d$.
- $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_+$, loss function.

Stochastic convex optimization (SCO)

A fundamental optimization problem in machine learning.

- n : sample size.
- $\mathcal{D}$: an unknown distribution over $\mathcal{X} \subseteq \mathbb{R}^d$.
- $\mathcal{W}$: parameter space, a subset of $\mathbb{R}^d$.
- $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_+$, loss function.
- $L_{\mathcal{D}}(w) : \mathbb{E}_{x \sim \mathcal{D}} [\ell(w, x)]$, population risk.

Stochastic convex optimization (SCO)

A fundamental optimization problem in machine learning.

- n : sample size.
- $\mathcal{D}$: an unknown distribution over $\mathcal{X} \subseteq \mathbb{R}^d$.
- $\mathcal{W}$: parameter space, a subset of $\mathbb{R}^d$.
- $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}_+$, loss function.
- $L_{\mathcal{D}}(w) : \mathbb{E}_{x \sim \mathcal{D}} [\ell(w, x)]$, population risk.

Goal: given n i.i.d. samples from an unknown distribution $\mathcal{D}$, find $\hat{w}$ to minimize

$$L_{\mathcal{D}}(\hat{w}) - \min_{w^* \in \mathcal{W}} L_{\mathcal{D}}(w^*).$$

Privacy

Data may contain **sensitive** information.

(a) Navigation

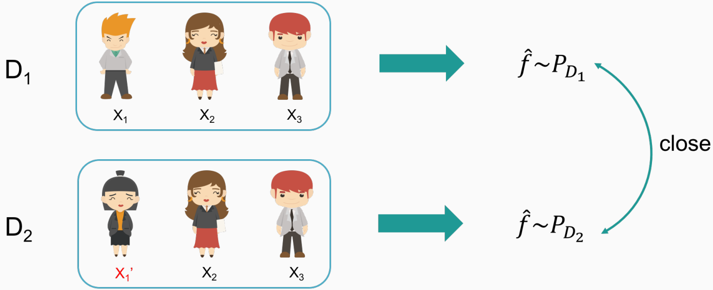
(b) Medical data

We want to protect the privacy while learning from samples.

Differential privacy (DP) [Dwork et al., 2006]

$\hat{f}$ is (ε, δ)-DP for any D_1 and D_2, with $d_{Ham}(D_1, D_2) \leq 1$, for all measurable S,

$$\forall S, \ \ \Pr\left(\hat{f}(D_1) \in S\right) \leq e^{\varepsilon} \cdot \Pr\left(\hat{f}(D_2) \in S\right) + \delta.$$



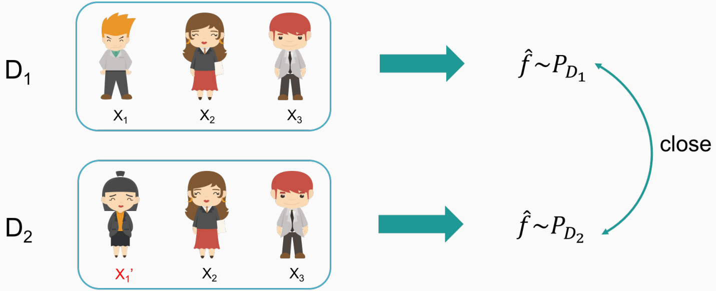
Pure DP: $\delta = 0$; **approximate DP:** $\delta \neq 0$

Concentrated Differential Privacy [Bun and Steinke, 2016]

$\hat{f}$ is ε^2-CDP if for any D_1 and D_2, with $d_{Ham}(D_1, D_2) \leq 1$,

$$\forall \alpha \in (1, \infty), D_\alpha\left(\hat{f}(D_1), \hat{f}(D_2)\right) \leq \varepsilon^2 \alpha,$$

where $D_\alpha\left(\hat{f}(D_1), \hat{f}(D_2)\right)$ is the α-Rényi divergence.



ε^2-CDP lies between $(O(\varepsilon), 0)$-DP and $(O(\varepsilon), \delta)$-DP.

Gradients can be unbounded!

- ℓ is usually assumed to be Lipschitz, i.e., $\|\nabla \ell(w, x)\|_2$ is bounded for $\forall w, x$ [Bassily et al., 2014, Bassily et al., 2019].

Gradients can be unbounded!

- ℓ is usually assumed to be Lipschitz, i.e., $\|\nabla \ell(w, x)\|_2$ is bounded for $\forall w, x$ [Bassily et al., 2014, Bassily et al., 2019].
- Convenient for analysis, but unrealistic in practice.

Gradients can be unbounded!

- ℓ is usually assumed to be Lipschitz, i.e., $\|\nabla \ell(w, x)\|_2$ is bounded for $\forall w, x$ [Bassily et al., 2014, Bassily et al., 2019].
- Convenient for analysis, but unrealistic in practice.
- Following [Wang et al., 2020] and [Holland, 2019], we assume heavy-tailed gradients:

Gradients can be unbounded!

- ℓ is usually assumed to be Lipschitz, i.e., $\|\nabla \ell(w, x)\|_2$ is bounded for $\forall w, x$ [Bassily et al., 2014, Bassily et al., 2019].
- Convenient for analysis, but unrealistic in practice.
- Following [Wang et al., 2020] and [Holland, 2019], we assume heavy-tailed gradients:

Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. We assume for every $w \in W$,

$$\mathbb{E}_{x \in \mathcal{D}} \left[\left| \langle \nabla \ell(w, x), e_j \rangle \right|^k \right] \leq 1, \forall j \in [d],$$

where e_j is the j-th standard basis vector.

Gradients can be unbounded!

Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. We assume for every $w \in W$,

$$\mathbb{E}_{x \sim \mathcal{D}} \left[|\langle \nabla \ell(w, x), e_j \rangle|^k \right] \leq 1, \forall j \in [d],$$

where e_j is the j-th standard basis vector.

- The k-th moment of each dimension is bounded.

Gradients can be unbounded!

Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. We assume for every $w \in W$,

$$\mathbb{E}_{x \sim \mathcal{D}} \left[|\langle \nabla \ell(w, x), e_j \rangle|^k \right] \leq 1, \forall j \in [d],$$

where e_j is the j-th standard basis vector.

- The k-th moment of each dimension is bounded.
- Stronger assumption when k increases.

Gradients can be unbounded!

Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. We assume for every $w \in W$,

$$\mathbb{E}_{x \sim \mathcal{D}} \left[|\langle \nabla \ell(w, x), e_j \rangle|^k \right] \leq 1, \forall j \in [d],$$

where e_j is the j-th standard basis vector.

- The k-th moment of each dimension is bounded.
- Stronger assumption when k increases.

- $k = 2$ throughout this talk.

CDP SCO with heavy-tailed gradients

Goal: given n i.i.d. samples from an unknown distribution $\mathcal{D}$, and the gradient distribution satisfying the heavy-tailed assumption, we want to design a ε^2-CDP algorithm w^{priv} that minimizes

$$L_{\mathcal{D}}(w^{priv}) - \min_{w^* \in \mathcal{W}} L_{\mathcal{D}}(w^*).$$

Results

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].
- Previous work: $\widetilde{O}\left(\frac{d}{(\varepsilon^2 n)^{\frac{1}{3}}}\right)$ [Wang et al., 2020].

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].
- Previous work: $\widetilde{O}\left(\frac{d}{(\varepsilon^2 n)^{\frac{1}{3}}}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{\sqrt{\varepsilon n}}\right)$, $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].
- Previous work: $\widetilde{O}\left(\frac{d}{(\varepsilon^2 n)^{\frac{1}{3}}}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{\sqrt{\varepsilon n}}\right)$, $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.

Strongly convex setting:

- Non-private: $\Theta\left(\frac{d}{n}\right)$ [Holland, 2019].

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].
- Previous work: $\widetilde{O}\left(\frac{d}{(\varepsilon^2 n)^{\frac{1}{3}}}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{\sqrt{\varepsilon n}}\right)$, $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.

Strongly convex setting:

- Non-private: $\Theta\left(\frac{d}{n}\right)$ [Holland, 2019].
- Previous work: $O\left(\frac{d^3}{\varepsilon^2 n}\right)$ [Wang et al., 2020].

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].
- Previous work: $\widetilde{O}\left(\frac{d}{(\varepsilon^2 n)^{\frac{1}{3}}}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{\sqrt{\varepsilon n}}\right)$, $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.

Strongly convex setting:

- Non-private: $\Theta\left(\frac{d}{n}\right)$ [Holland, 2019].
- Previous work: $O\left(\frac{d^3}{\varepsilon^2 n}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{n} + \frac{d^{\frac{3}{2}}}{\varepsilon n}\right)$, $\Omega\left(\frac{d}{n} + \frac{d^{\frac{3}{2}}}{\varepsilon n}\right)$.

ε^2-CDP SCO with heavy-tailed gradients

Convex setting:

- Non-private: $\Theta\left(\sqrt{\frac{d}{n}}\right)$ [Holland, 2019].
- Previous work: $\widetilde{O}\left(\frac{d}{(\varepsilon^2 n)^{\frac{1}{3}}}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{\sqrt{\varepsilon n}}\right)$, $\Omega\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.

Strongly convex setting:

- Non-private: $\Theta\left(\frac{d}{n}\right)$ [Holland, 2019].
- Previous work: $O\left(\frac{d^3}{\varepsilon^2 n}\right)$ [Wang et al., 2020].
- This work: $\widetilde{O}\left(\frac{d}{n} + \frac{d^{\frac{3}{2}}}{\varepsilon n}\right)$, $\Omega\left(\frac{d}{n} + \frac{d^{\frac{3}{2}}}{\varepsilon n}\right)$.
- Tight up to a logarithmic factor.

Our techniques

CDP mean estimation

- Given n i.i.d. samples from an unknown heavy-tailed distribution $\mathcal{D}$ over $\mathbb{R}^d$, privately estimate the distribution mean under ℓ_2 distance.

CDP mean estimation

- Given n i.i.d. samples from an unknown heavy-tailed distribution $\mathcal{D}$ over $\mathbb{R}^d$, privately estimate the distribution mean under ℓ_2 distance.
- Heavy-tailed assumption: $\mathbb{E}_{X \sim \mathcal{D}} \left[|\langle X, e_j \rangle|^2 \right] \leq 1, \forall j \in [d]$.

CDP mean estimation

- Given n i.i.d. samples from an unknown heavy-tailed distribution $\mathcal{D}$ over $\mathbb{R}^d$, privately estimate the distribution mean under ℓ_2 distance.
- Heavy-tailed assumption: $\mathbb{E}_{X\sim\mathcal{D}}\left[|\langle X, e_j\rangle|^2\right] \leq 1, \forall j \in [d]$.
- ε^2-CDP result: $\Theta\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.

CDP mean estimation

- Given n i.i.d. samples from an unknown heavy-tailed distribution $\mathcal{D}$ over $\mathbb{R}^d$, privately estimate the distribution mean under ℓ_2 distance.
- Heavy-tailed assumption: $\mathbb{E}_{X\sim\mathcal{D}}\left[|\langle X, e_j\rangle|^2\right] \le 1, \forall j \in [d]$.
- ε^2-CDP result: $\Theta\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.
- ε-DP result: $\Theta\left(\sqrt{\frac{d}{n}} + \frac{d}{\sqrt{\varepsilon n}}\right)$.

CDP mean estimation

- Given n i.i.d. samples from an unknown heavy-tailed distribution $\mathcal{D}$ over $\mathbb{R}^d$, privately estimate the distribution mean under ℓ_2 distance.
- Heavy-tailed assumption: $\mathbb{E}_{X\sim\mathcal{D}}\left[|\langle X, e_j\rangle|^2\right] \leq 1, \forall j \in [d]$.
- ε^2-CDP result: $\Theta\left(\sqrt{\frac{d}{n}} + \frac{d^{\frac{3}{4}}}{\sqrt{\varepsilon n}}\right)$.
- ε-DP result: $\Theta\left(\sqrt{\frac{d}{n}} + \frac{d}{\sqrt{\varepsilon n}}\right)$.
- A new separation between pure DP and approximate DP!

CDP mean estimation (upper bound)

- We generalize the idea and analysis from [Kamath et al., 2020], which has a slightly different assumption.

CDP mean estimation (upper bound)

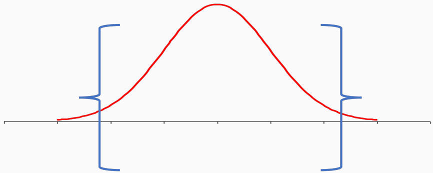
- We generalize the idea and analysis from [Kamath et al., 2020], which has a slightly different assumption.
- Intuitively, the algorithm outputs $\sum_x \text{clip}(x) + N(0, \sigma^2 \mathbb{I}_d)$.

CDP mean estimation (upper bound)

- We generalize the idea and analysis from [Kamath et al., 2020], which has a slightly different assumption.
- Intuitively, the algorithm outputs $\sum_x \mathsf{clip}(x) + N(0, \sigma^2 \mathbb{I}_d)$.
- Clipping decides both bias and variance (σ^2).

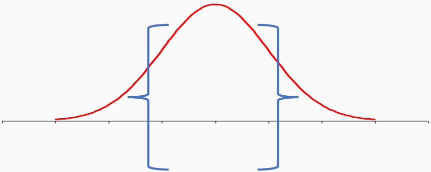
CDP mean estimation (upper bound)

- We generalize the idea and analysis from [Kamath et al., 2020], which has a slightly different assumption.
- Intuitively, the algorithm outputs $\sum_x \text{clip}(x) + N(0, \sigma^2 \mathbb{I}_d)$.
- Clipping decides both bias and variance (σ^2).
- Larger clipping range leads to less bias and higher variance.



CDP mean estimation (upper bound)

- We generalize the idea and analysis from [Kamath et al., 2020], which has a slightly different assumption.
- Intuitively, the algorithm outputs $\sum_x \text{clip}(x) + N(0, \sigma^2 \mathbb{I}_d)$.
- Clipping decides both bias and variance (σ^2).
- Larger clipping range leads to less bias and higher variance.
- Smaller clipping range leads to larger bias and less variance.



CDP mean estimation (upper bound)

- We generalize the idea and analysis from [Kamath et al., 2020], which has a slightly different assumption.
- Intuitively, the algorithm outputs $\sum_x \mathrm{clip}(x) + N(0, \sigma^2 \mathbb{I}_d)$.
- Clipping decides both bias and variance (σ^2).
- Larger clipping range leads to less bias and higher variance.
- Smaller clipping range leads to larger bias and less variance.
- Wisely select the clipping range to balance the bias and variance!

CDP SCO upper bound (convex)

Our algorithm is an adaption of full gradient descent.

- Initialize $w^0 \in \mathcal{W}$.

CDP SCO upper bound (convex)

Our algorithm is an adaption of full gradient descent.

- Initialize $w^0 \in \mathcal{W}$.
- Let $G_t = \mathrm{MeanOracle}(\{\nabla \ell(w^{t-1}, x_i)\}_{i \in [n]})$.

CDP SCO upper bound (convex)

Our algorithm is an adaption of full gradient descent.

- Initialize $w^0 \in \mathcal{W}$.
- Let $G_t = \mathrm{MeanOracle}(\{\nabla \ell(w^{t-1}, x_i)\}_{i \in [n]})$.
- $w^t = \mathrm{Proj}_{\mathcal{W}}(w^{t-1} - \eta_{t-1} G_t)$

CDP SCO upper bound (convex)

Our algorithm is an adaption of full gradient descent.

- Initialize $w^0 \in \mathcal{W}$.
- Let $G_t = \text{MeanOracle}(\{\nabla \ell(w^{t-1}, x_i)\}_{i \in [n]})$.
- $w^t = \text{Proj}_{\mathcal{W}}(w^{t-1} - \eta_{t-1} G_t)$

CDP SCO upper bound (convex)

Our algorithm is an adaption of full gradient descent.

- Initialize $w^0 \in \mathcal{W}$.
- Let $G_t = \text{MeanOracle}(\{\nabla \ell(w^{t-1}, x_i)\}_{i \in [n]})$.
- $w^t = \text{Proj}_{\mathcal{W}}(w^{t-1} - \eta_{t-1} G_t)$

Theorem (privacy)

Our CDP SCO algorithm satisfies ε^2-CDP suppose the mean estimation oracle satisfies ε^2/T-CDP.

CDP SCO upper bound (convex)

Our algorithm is an adaption of full gradient descent.

- Initialize $w^0 \in \mathcal{W}$.
- Let $G_t = \text{MeanOracle}(\{\nabla \ell(w^{t-1}, x_i)\}_{i \in [n]})$.
- $w^t = \text{Proj}_{\mathcal{W}}(w^{t-1} - \eta_{t-1} G_t)$

Theorem (privacy)

Our CDP SCO algorithm satisfies ε^2-CDP suppose the mean estimation oracle satisfies ε^2/T-CDP.

Proof: DP post-processing and composition.

CDP SCO upper bound (convex)

Theorem (utility, informal)

Suppose the mean estimation oracle guarantees that the bias is smaller than B and the variance is smaller than G^2, the algorithm outputs w^{priv} such that

$$\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)\right] \leq O\left(\frac{1}{\sqrt{T}} + \frac{G^2}{\sqrt{T}} + B\right).$$

- Setting $G = B$ leads to the optimal performance for one single round.

CDP SCO upper bound (convex)

Theorem (utility, informal)

Suppose the mean estimation oracle guarantees that the bias is smaller than B and the variance is smaller than G^2, the algorithm outputs w^{priv} such that

$$\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)\right] \leq O\left(\frac{1}{\sqrt{T}} + \frac{G^2}{\sqrt{T}} + B\right).$$

- Setting $G = B$ leads to the optimal performance for one single round.
- However, it is sub-optimal for CDP SCO.

CDP SCO upper bound (convex)

Theorem (utility, informal)

Suppose the mean estimation oracle guarantees that the bias is smaller than B and the variance is smaller than G^2, the algorithm outputs w^{priv} such that

$$\mathbb{E}\left[L_{\mathcal{D}}(w^{priv}) - L_{\mathcal{D}}(w^*)\right] \leq O\left(\frac{1}{\sqrt{T}} + \frac{G^2}{\sqrt{T}} + B\right).$$

- Setting $G = B$ leads to the optimal performance for one single round.
- However, it is sub-optimal for CDP SCO.
- Instead we set $B = \frac{G^2}{\sqrt{T}}$ to balance the second and third terms.

CDP SCO lower bound

- Following a similar argument in [Bassily et al., 2014], we reduce CDP mean estimation to CDP SCO.

CDP SCO lower bound

- Following a similar argument in [Bassily et al., 2014], we reduce CDP mean estimation to CDP SCO.
- We propose CDP Fano's inequality, generalizing the results in [Acharya et al., 2021] and [Bun and Steinke, 2016].

CDP SCO lower bound

- Following a similar argument in [Bassily et al., 2014], we reduce CDP mean estimation to CDP SCO.
- We propose CDP Fano's inequality, generalizing the results in [Acharya et al., 2021] and [Bun and Steinke, 2016].

Theorem (ε^2-CDP Fano's inequality)

Let $\mathcal{V} = \{p_1, ..., p_M\}$ be a set of distributions, θ be a parameter of interest, and ℓ be a loss function. Suppose for all $i \neq j$, it satisfies (a) $\ell(\theta(p_i), \theta(p_j)) \geq r$, (b) $d_{TV}(p_i, p_j) \leq \alpha$. Then for any ε^2-CDP estimator $\hat{\theta}$,

$$\frac{1}{M} \sum_{i \in [M]} \mathbb{E}\left[\ell\Big(\hat{\theta}(X), \theta(p_i)\Big)\right] \geq \frac{r}{2}\left(1 - \frac{\varepsilon^2\big(n^2\alpha^2 + n\alpha(1-\alpha)\big) + \log 2}{\log M}\right).$$

The End

Paper ID: 4892

Details in paper online:

https://arxiv.org/abs/2106.01336

Acharya, J., Sun, Z., and Zhang, H. (2021).
Differentially private assouad, fano, and le cam.
In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 48–78. JMLR, Inc.

Bassily, R., Feldman, V., Talwar, K., and Thakurta, A. G. (2019).
Private stochastic convex optimization with optimal rates.
In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 11282–11291. Curran Associates, Inc.

Bassily, R., Smith, A., and Thakurta, A. (2014).
Private empirical risk minimization: Efficient algorithms and tight error bounds.

In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 464–473, Washington, DC, USA. IEEE Computer Society.

Bun, M. and Steinke, T. (2016).
Concentrated differential privacy: Simplifications, extensions, and lower bounds.
In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pages 635–658, Berlin, Heidelberg. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006).
Calibrating noise to sensitivity in private data analysis.
In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg. Springer.

Holland, M. J. (2019).

Robust descent using smoothed multiplicative noise.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR.

Kamath, G., Singhal, V., and Ullman, J. (2020).
Private mean estimation of heavy-tailed distributions.
In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, pages 2204–2235.

Wang, D., Xiao, H., Devadas, S., and Xu, J. (2020).
On differentially private stochastic convex optimization with heavy-tailed data.
In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 10081–10091. JMLR, Inc.