Optimal Rates for Pure Differentially Private Stochastic Convex Optimization with Heavy Tails

Andrew Lowy
CISPA Helmholtz Center for Information Security
lowy@cispa.de

Abstract

We study stochastic convex optimization (SCO) with heavy-tailed gradients under pure ε-differential privacy (DP). Instead of assuming a bound on the worst-case Lipschitz parameter of the loss, we assume only a bounded k-th moment. This assumption allows for unbounded, heavy-tailed stochastic gradient distributions, and can yield sharper excess risk bounds. Prior work characterized the minimax optimal rate for ρ-zero-concentrated DP SCO up to logarithmic factors in this setting, but the pure ε-DP case has remained open. We characterize the minimax optimal excess-risk rate for pure ε-DP heavy-tailed SCO up to logarithmic factors. Our algorithm achieves this rate in polynomial time with high probability. Moreover, it runs in deterministic polynomial time when the worst-case Lipschitz parameter is polynomially bounded. For important structured problem classes — including hinge/ReLU-type and absolute-value losses on Euclidean balls, ellipsoids, and polytopes — we achieve deterministic polynomial time even when the worst-case Lipschitz parameter is infinite. Our approach is based on a novel framework for privately optimizing *Lipschitz extensions* of the empirical loss. We complement our upper bound with a nearly matching high-probability lower bound.

1 Introduction

Stochastic convex optimization (SCO) is a fundamental problem in machine learning. Given i.i.d. data $Z = (z_1, \dots, z_n)$ drawn from an unknown distribution P, the goal is to solve

$$\min_{w \in \mathcal{W}} \{F(w) := \mathbb{E}_{z \sim P}[f(w, z)]\}, \tag{1}$$

where $\mathcal{W} \subset \mathbb{R}^d$ is a convex compact parameter domain of ℓ_2-diameter D and $f(\cdot, z)$ is a convex loss function. The quality of a solution w of (1) is measured by its *excess risk* $F(w) - \min_{w' \in \mathcal{W}} F(w')$.

In many applications, the data used to train models contains sensitive information. *Differential privacy* (DP) provides a rigorous framework for protecting individual data contributions while enabling statistical learning [DMNS06]. The strongest standard notion of DP is pure ε-DP.

A large body of work has studied DP SCO under the *uniform Lipschitz* assumption[1]:

$$\sup_{w \in \mathcal{W}, z \in \mathcal{Z}} \|\nabla f(w, z)\| \leq L. \tag{2}$$

Under this assumption, optimal excess risk bounds scale with the worst-case Lipschitz parameter L [BFTT19, FKT20, AFKT21, BGM23, LLA24]. In many applications, however, L can be extremely large or infinite, making such guarantees overly pessimistic or vacuous. For example, in linear

[1]Throughout, $\nabla f(w, z)$ denotes the gradient with respect to w when $f(\cdot, z)$ is differentiable at w, and otherwise denotes an arbitrary subgradient in $\partial f(\cdot, z)(w)$.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

regression with squared loss, the gradient norm scales with the squared feature norm, so if the feature distribution is unbounded then L is unbounded as well.

To address these issues, a recent line of work studies *heavy-tailed DP SCO* under weaker moment assumptions on the gradients [WXDX20, ADF+21, TWZW21, KLZ22, HNXW22, WX22, ALT24, LR25]. Instead of requiring uniformly bounded gradients, one assumes *bounded gradient moments*:

$$\mathbb{E}_{z\sim P}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w,z)\|^j\right] \le G_j^j, \qquad j \in \{2,k\}, \tag{3}$$

for some $k \ge 2$. Write $\mathcal{P}(G_2, G_k)$ for the distributions satisfying these two bounds. For nonsmooth f, the supremum in (3) is taken also over subgradient norms. (3) allows unbounded heavy-tailed gradient distributions while controlling the average behavior, and captures a broad class of realistic learning problems. Note $G_2 \le G_k \le L$ for all k and often $G_k \ll L$: e.g., for linear regression, G_k scales with the $2k$-th moment of the feature data. Thus, excess risk bounds that scale with G_k instead of L are often sharper. For ρ-zCDP [BS16] heavy-tailed SCO, the optimal excess risk rate is known up to logarithmic factors [ALT24, LR25].

The pure-DP gap. Despite this progress, the case of *pure* ε-differential privacy remains poorly understood. The work of [BD14] proved an in-expectation pure DP lower bound, but no direct algorithm achieving this rate was previously known. Existing heavy-tailed DP SCO algorithms rely on noisy clipped-gradient methods, which are suboptimal under pure ε-DP. This raises the following:

> **Question 1.** What is the minimax optimal excess risk for heavy-tailed stochastic convex optimization under pure ε-differential privacy?

Contribution 1: Optimal excess risk for pure-DP heavy-tailed SCO (up to logarithms). We determine the minimax optimal pure ε-DP excess risk over $\mathcal{P}(G_2, G_k)$, obtaining with probability $\ge 1-\delta$

$$O\Bigg(D\min\Bigg\{G_2,\ G_2\sqrt{\frac{\log(1/\delta)}{n}} + \min\left\{G_2\sqrt{\frac{d\log(1/\delta)}{n\varepsilon}},\ G_k\left(\frac{d\log(1/\delta)}{n\varepsilon}\right)^{1-1/k}\right\}\Bigg\}\Bigg). \tag{4}$$

This bound is tight up to a $O(\log(1/\delta)^{1-1/k})$ factor: we prove a nearly matching high-probability lower bound that is sharper than the in-expectation lower bound of [BD14].

> **Question 2.** Can the minimax optimal excess risk for pure-DP heavy-tailed SCO be achieved by a computationally efficient algorithm?

Contribution 2: Polynomial-time algorithms for pure-DP heavy-tailed SCO. Our main result is that (4) can be achieved in polynomial time with high probability; if the worst-case Lipschitz parameter is finite and polynomially bounded, the runtime is polynomial with probability 1. This is the first polynomial-time ε-DP algorithm achieving the high-probability excess-risk guarantee (4). A black-box purification route from zCDP is looser and more cumbersome than our direct pure-DP approach; see Appendix B.

For certain structured subclasses — including hinge/ReLU-type and absolute-value losses on Euclidean balls, ellipsoids, and polytopes — we prove a stronger guarantee: deterministic polynomial time, even when the worst-case Lipschitz parameter is infinite.

Contribution 3: A new framework for privately optimizing Lipschitz extensions. To obtain our upper bounds, we move away from clipped-gradient methods and instead use the *Lipschitz extension*

$$f_C(w,z) := \inf_{y\in\mathcal{W}}\big[f(y,z) + C\|w-y\|\big]. \tag{5}$$

This reduces heavy-tailed regularized ERM to Lipschitz regularized ERM. To optimize the resulting objective efficiently under pure DP, we develop a novel jointly convex reformulation together with adaptive (inexact) projected subgradient methods with deterministic accuracy guarantees.

Table 1: Heavy-tailed DP SCO: summary of results. Upper and lower bounds match up to logarithms.

Result	Privacy	Runtime	Setting
zCDP upper/lower bounds [ALT24, LR25]	ρ-zCDP	polytime w.p. 1	general
Our upper bound (Thm. 3.1)	pure ε-DP	polytime w.h.p.	general
Our structured-subclass upper bound (Corollary E.7)	pure ε-DP	polytime w.p. 1	structured
Our lower bound (Theorem 4.1)	pure ε-DP	—	general

1.1 Challenges and Techniques

We build on the population-level localization framework of [ALT24], reducing heavy-tailed SCO to regularized ERM. The key question is how to solve heavy-tailed ERM efficiently under pure ε-DP.

Challenge 1: Clipped gradients are insufficient. Prior heavy-tailed DP SCO algorithms rely on noisy clipped gradient methods. These are optimal for zCDP, but suboptimal under pure DP because the more forgiving zCDP composition is unavailable under pure DP. Our algorithm abandons gradient clipping and instead uses the *Lipschitz extension* (5).

Challenge 2: Optimizing the Lipschitz extension under pure DP. The Lipschitz extension (5) is defined by an inner optimization problem. Exact evaluation of $f_C(w, z)$ from fixed finite local oracle information is impossible (see Theorem C.1), and certified approximation is nontrivial because no Lipschitz bound for the inner problem is available. For pure DP this is especially problematic, since the required sensitivity control cannot fail even with small probability. We address this via a *jointly convex reformulation* and *adaptive inexact projected subgradient methods*, which provide certified approximate minimizers without requiring prior knowledge of the Lipschitz parameter.

Challenge 3: The bias of the Lipschitz extension is too large on the original domain. Over the full domain $\mathcal{W}$, the Lipschitz-extension bias is too large. We therefore use *localized Lipschitz extensions* over the current set U, so the bias scales with $\mathrm{diam}(U)$, and *recursively shrink* U around the regularized ERM minimizer via output perturbation. Geometric allocation of privacy and failure probabilities avoids extra log-log factors. Combining these techniques yields our main algorithm.

Lower bound techniques. We construct two different hard instances: one for the private error term and one for the non-private error. To prove the private term, we combine the packing technique of [BD14] with a reduction from quantile estimation to decoding. The non-private term is proved via a bounded two-point construction together with the high-probability testing framework of [MVS24].

1.2 Preliminaries

Let $\|\cdot\|$ denote the ℓ_2 norm. $\mathbb{B}(w_0, r)$ denotes ℓ_2-ball of radius r around w_0. For a function $h : \mathcal{W} \to \mathbb{R}$, a vector $g \in \mathbb{R}^d$ is a *subgradient* of h at $w \in \mathcal{W}$ if $h(u) \geq h(w) + \langle g, u - w\rangle \ \forall u \in \mathcal{W}$. We write $\partial h(w)$ for the set of all subgradients of h at w. When h is differentiable at w, $\partial h(w) = \{\nabla h(w)\}$, the gradient of h at w. For $\lambda \geq 0$, we say that h is λ-*strongly convex* if for every $w, u \in \mathcal{W}$ and every $g \in \partial h(w)$, $h(u) \geq h(w) + \langle g, u - w\rangle + \frac{\lambda}{2}\|u - w\|^2$. If $\lambda = 0$, we say h is *convex*. For a closed convex set $K \subseteq \mathbb{R}^d$, the Euclidean *projection* of $y \in \mathbb{R}^d$ onto K is $\Pi_K(y) := \arg\min_{u \in K} \|u - y\|$. Denote $h^* := \min_{w \in \mathcal{W}} h(w)$. Throughout, $c_{(\cdot)}$ denotes an absolute constant. We use $O(\cdot)$ and $\lesssim$ to hide absolute constants, and $\tilde{O}(\cdot)$ to additionally hide logarithmic factors.

Throughout the paper, we assume the following:

Assumption 1.1.
1. *The loss function $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is such that $f(\cdot, z)$ is convex for every $z \in \mathcal{Z}$, and* (3) *holds for some constant $k \geq 2$ where $0 < G_2 \leq G_k$ are publicly known (c.f. Remark B.1).*
2. *The domain $\mathcal{W} \subset \mathbb{R}^d$ is closed and convex with ℓ_2-diameter D, and is projection-friendly: for every $y \in \mathbb{R}^d$, the Euclidean projection $\Pi_{\mathcal{W}}(y)$ can be computed in polynomial time.*
3. *Given $z \in \mathcal{Z}$, $w \in \mathcal{W}$, one can compute $f(w, z)$ and $g \in \partial f(w, z)$ in polynomial time.*

By monotonicity of moments, $G_2 \leq G_k \leq \sup_{w,z} \|\nabla f(w, z)\|$ for all $k \geq 2$.

Algorithm 1: POP–LOCALIZE$(Z, \varepsilon, \mathcal{A}_{\mathrm{ERM}}, \delta)$ [ALT24]

Input: Data $Z \in \mathcal{Z}^n$, privacy ε, private regularized ERM solver $\mathcal{A}_{\mathrm{ERM}}$, error prob. δ.
Output: $\widehat{w} \in \mathcal{W}$.

Choose $J = \Theta(\log(\log n/\delta))$
Set number of phases $T = \lfloor \log_2(n/(2J)) \rfloor$
Initialize $\overline{w}_1 \in \mathcal{W}$ arbitrarily
Choose base regularization $\lambda_1 > 0$
Partition Z into disjoint $Z^1, \dots, Z^T$ with $|Z^t| = n_t := \lfloor n/2^t \rfloor$, discarding unused samples
for $t = 1$ **to** T **do**
  Set $\lambda_t := 32^{t-1}\lambda_1$.
  Partition Z^t into J disjoint blocks $Z^{(t,1)}, \dots, Z^{(t,J)}$ of size $|Z^{(t,j)}| = m_t := \lfloor n_t/J \rfloor$.
  for $j = 1$ **to** J **do**
    Compute $\widehat{w}_{t,j} \leftarrow \mathcal{A}_{\mathrm{ERM}}(Z^{(t,j)}, \varepsilon, \lambda_t, \overline{w}_t)$.
  end
  Aggregate $\{\widehat{w}_{t,j}\}_{j=1}^{J}$ via geometric aggregation to obtain $\overline{w}_{t+1}$
end
return $\overline{w}_{T+1}$

Examples. Let $z = (x, y)$ and $\mathcal{W} \subseteq \mathbb{B}(0, B)$, so $D \le 2B$. Logistic and hinge losses with $y \in \{-1, 1\}$ have sample Lipschitz parameter at most $\|x\|$, allowing $G_r = (\mathbb{E}\|x\|^r)^{1/r}$ for $r \in \{2, k\}$. For squared loss $f(w, z) = \frac{1}{2}(\langle w, x\rangle - y)^2$, one may take $G_r = (\mathbb{E}[\|x\|^r(|y| + B\|x\|)^r])^{1/r}$. Thus, G_k can be finite even when the data is unbounded and the worst-case Lipschitz parameter is infinite.

Differential Privacy. Differential privacy ensures that no attacker can infer much more about any individual's data than they could have inferred had that person's data not been used.

Definition 1.2 (Pure Differential Privacy [DMNS06])**.** A randomized algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{O}$ is *ε-differentially private* (DP) if for every pair of neighboring datasets $Z, Z' \in \mathcal{Z}^n$ differing in one entry, and every measurable set $S \subseteq \mathcal{O}$, $\Pr(\mathcal{A}(Z) \in S) \le e^{\varepsilon} \Pr(\mathcal{A}(Z') \in S)$.

Zero-concentrated differential privacy (zCDP) is weaker than ε-DP; see Appendix A.

2 Algorithmic Building Blocks

This section develops the key ingredients that will be used by our main algorithm.

2.1 Reduction from SCO to ERM

We use the population-localization framework of [ALT24], which reduces heavy-tailed SCO to a sequence of private regularized ERM problems; see Algorithm 1.

Guarantee for Algorithm 1. For a sample $Z = (z_1, \dots, z_m) \in \mathcal{Z}^m$, a center $w_0 \in \mathcal{W}$, and a regularization parameter $\lambda > 0$, define the regularized empirical objective

$$\widehat{F}_{\lambda,Z}^{(w_0)}(w) := \frac{1}{m}\sum_{i=1}^{m} f(w, z_i) + \frac{\lambda}{2}\|w - w_0\|^2, \qquad \widehat{w}_\lambda(Z; w_0) := \operatorname*{arg\,min}_{w \in \mathcal{W}} \widehat{F}_{\lambda,Z}^{(w_0)}(w).$$

By the following theorem, it suffices to design an ε-DP regularized ERM solver $\mathcal{A}_{\mathrm{ERM}}$ such that, on every instance $(Z, \varepsilon, \lambda, w_0)$ with $|Z| = m$, with probability at least 0.6 its output satisfies

$$\left\|\mathcal{A}_{\mathrm{ERM}}(Z, \varepsilon, \lambda, w_0) - \widehat{w}_\lambda(Z; w_0)\right\| \le \frac{c_{\mathrm{erm}}}{\lambda}\left(G_k\Big(\frac{d}{m\varepsilon}\Big)^{1-\frac{1}{k}} + \frac{G_2}{\sqrt{m}}\right). \tag{6}$$

Theorem 2.1 (Regularized ERM implies SCO [ALT24])**.** *Fix $\delta \in (0, 1/2)$. Suppose $\mathcal{A}_{\mathrm{ERM}}$ is an ε-DP algorithm such that for every center $w_0 \in \mathcal{W}$ and every $\lambda > 0$, its output satisfies* (6) *with*

probability at least 0.6 *over* $\mathcal{A}_{\mathrm{ERM}}$ *and* $Z \sim P^m$*. Then, Algorithm 1 is* ε*-DP and there exists a choice of parameters such that, with probability at least* $1-\delta$*,*

$$F(\widehat{w}) - F^* \lesssim G_k D \left(\frac{d\log(1/\delta)}{n\varepsilon}\right)^{1-\frac{1}{k}} + G_2 D\sqrt{\frac{\log(1/\delta)}{n}}.$$

Moreover, if one call to $\mathcal{A}_{\mathrm{ERM}}$ *on a dataset of size* m *takes time* $\mathrm{Time}(\mathcal{A}_{\mathrm{ERM}}, m, d, \varepsilon, \lambda)$*, then the total runtime of Algorithm 1 is bounded by* $J\sum_{t=1}^{T}\mathrm{Time}(\mathcal{A}_{\mathrm{ERM}}, m_t, d, \varepsilon, \lambda_t)$*, plus the cost of geometric aggregation.*

Therefore, the rest of the paper focuses on designing ε-DP regularized ERM solvers satisfying (6).

2.2 Lipschitz Extension

The Lipschitz extension (5) transforms any convex $f(\cdot, z)$ into a convex C-Lipschitz function.

Lemma 2.2 ([HUL13])**.** *Let* $f(\cdot, z)$ *be convex on* $\mathcal{W}$*. Then:*

1. $f_C(\cdot, z)$ *is convex on* $\mathcal{W}$*;*
2. $f_C(\cdot, z)$ *is* C*-Lipschitz on* $\mathcal{W}$*;*
3. $f_C(w, z) \le f(w, z)$ *for all* $w \in \mathcal{W}$*.*

This suggests reducing heavy-tailed regularized ERM to regularized Lipschitz ERM, provided we can control the error introduced by replacing f with f_C. For $Z = (z_1, \dots, z_m)$, $w_0 \in \mathcal{W}$, $\lambda > 0$, define the *regularized empirical* C*-Lipschitz-extension* $\widehat{F}^{(w_0)}_{C,\lambda,Z}(w) := \frac{1}{m}\sum_{i=1}^{m} f_C(w, z_i) + \frac{\lambda}{2}\|w - w_0\|^2$.

Why the global extension is insufficient. Output perturbation gives the optimal distance bound for C-Lipschitz, λ-strongly convex ε-DP ERM. Running it on $\widehat{F}^{(w_0)}_{C,\lambda,Z}$ yields w_{DP} such that w.p. ≥ 0.6

$$\|w_{\mathrm{DP}} - \widehat{w}_\lambda(Z; w_0)\| \lesssim \frac{Cd}{\lambda m\varepsilon} + \sqrt{\frac{DG_k^k}{\lambda C^{k-1}}}.$$

The first term is the DP optimization error and the second is the bias from replacing f by f_C; see Appendix D.1. Even after optimizing over C, this is too large to imply (6) unless the diameter D is small. Our solution is to replace the global diameter D by the diameter of a localized search set.

2.3 Shrinking the Lipschitz Extension Bias by Localization

To reduce the Lipschitz extension bias, we use Lipschitz extensions localized to the current search set. For any closed convex set $U \subseteq \mathcal{W}$, define the U*-localized* C*-Lipschitz extension* by

$$f_{C,U}(w, z) := \inf_{y \in U}\big[f(y, z) + C\|w - y\|\big], \qquad w \in U. \tag{7}$$

When $U = \mathcal{W}$, this is the global Lipschitz extension (5). As in Lemma 2.2, $f_{C,U}(\cdot, z)$ is convex and C-Lipschitz on U, and $f_{C,U}(w, z) \le f(w, z)$ for all $w \in U$. For $Z = (z_1, \dots, z_m)$, define

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w) := \frac{1}{m}\sum_{i=1}^{m} f_{C,U}(w, z_i) + \frac{\lambda}{2}\|w - w_0\|^2, \qquad \widehat{w}_{C,U,\lambda}(Z; w_0) := \operatorname*{arg\,min}_{w \in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w),$$

for $w \in U$. The key is that the bias of the localized extension is proportional to $\mathrm{diam}(U)$, not D.

Lemma 2.3 (Uniform bias bound for localized Lipschitz extensions)**.** *With probability at least* $1-\beta$*,*

$$\|\widehat{w}_{C,U,\lambda}(Z; w_0) - \widehat{w}_\lambda(Z; w_0)\| \le \sqrt{\frac{2RG_k^k}{\lambda\beta(k-1)C^{k-1}}}$$

simultaneously for every closed convex $U \subseteq \mathcal{W}$ *with* $\mathrm{diam}(U) \le R$ *and* $\widehat{w}_\lambda(Z; w_0) \in U$*.*

The uniformity over U is important: in Algorithm 3, the sets U are chosen adaptively from previous private outputs, but Lemma 2.3 still applies simultaneously to all candidate sets of diameter $\le R$.

Algorithm 2: LOCALIZED-OUTPUTPERT$(Z, \varepsilon, \lambda, w_0, U, R, C, \beta)$

Input: Data $Z = (z_1, \dots, z_m)$, privacy ε, regularization λ, center w_0, convex set $U \subseteq \mathcal{W}$, deterministic diameter bound $R \geq \operatorname{diam}(U)$, Lipschitz parameter C, failure probability β

Output: A center $w_{\mathrm{loc}} \in \mathcal{W}$ and localized domain $U^+ \subseteq \mathcal{W}$

1 Set $\ell = \log(1/\beta)$ and $\alpha = \frac{C^2}{100\lambda m^2} \min\left\{1, \frac{d^2\ell^2}{\varepsilon^2}\right\}$.

2 Compute any point $\tilde{w} \in U$ satisfying $\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\tilde{w}) - \min_{w \in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w) \leq \alpha$.

3 Sample isotropic Laplace noise b with density proportional to $\exp\left(-\frac{\varepsilon\lambda m}{10C}\|b\|_2\right)$.

4 Set $w_{\mathrm{loc}} := \Pi_{\mathcal{W}}(\tilde{w} + b)$ and $R^+ := 100\left[\sqrt{\frac{RG_k^k}{\lambda\beta C^{k-1}}} + \frac{Cd\ell}{\lambda m\varepsilon}\right]$.

5 Set $U^+ := \mathcal{W} \cap \mathbb{B}(w_{\mathrm{loc}}, R^+)$

6 **return** (w_{loc}, U^+).

Localizing the domain via output perturbation. Given a current set U with diameter bound R and failure probability β, LOCALIZED-OUTPUTPERT (Algorithm 2) privately minimizes $\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w)$ over U, adds isotropic Laplace noise with scale $Cd/(\lambda m\varepsilon)$, and returns a ball around the noisy point with radius $R^+ \lesssim \sqrt{\frac{RG_k^k}{\beta\lambda C^{k-1}}} + \frac{Cd\log(1/\beta)}{\lambda m\varepsilon}$.

Lemma D.3 shows that Algorithm 2 is ε-DP and if $\widehat{w}_\lambda(Z; w_0) \in U$, then with probability $\geq 1 - 2\beta$, U^+ also contains $\widehat{w}_\lambda(Z; w_0)$. Choosing C optimally in line 4 of Algorithm 2 yields

$$R^+ \lesssim R_\star(R/R_\star)^{1/(k+1)}, \qquad R_\star := (G_k/\lambda)(d/(m\varepsilon))^{1-1/k}.$$

Thus one localization step contracts the radius toward $R_\star$ from R to R^+, but not all the way to $R_\star$. To achieve (6), it suffices to privately obtain the center of a ball of radius $O(R_\star)$ that contains $\widehat{w}_\lambda(Z; w_0)$ with constant probability. We accomplish this by applying Algorithm 2 recursively.

Recursive localization. Algorithm 3 applies LOCALIZED-OUTPUTPERT recursively. Writing R_j for a deterministic upper bound on $\operatorname{diam}(U_j)$, step j constructs a private ball U_{j+1} around $w_{\mathrm{loc},j+1}$ that still contains $\widehat{w}_\lambda(Z; w_0)$ with high probability, while updating

$$R_{j+1} \lesssim R_\star\left(\frac{R_j}{R_\star}\right)^{1/(k+1)}.$$

Equivalently, if $R_j = R_\star M_j$, then $M_{j+1} \lesssim M_j^{1/(k+1)}$. Hence after $S \approx \log_{k+1}\log(eD/R_\star)$ rounds the radius is $O(R_\star)$. The geometric choice of ε_j, β_j avoids extra log-log factors.

Proposition 2.4. *Let $Z \sim P^m$, fix $w_0 \in \mathcal{W}$, and $\lambda > 0$. Algorithm 3 is ε-DP and*

$$\Pr\left(\|w_{\mathrm{loc},S} - \widehat{w}_\lambda(Z; w_0)\| \leq c_2\frac{G_k}{\lambda}\left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}}\right) \geq 0.7.$$

Remark 2.5 (**Summary of the reduction**). Proposition 2.4 gives the distance guarantee (6) required by Theorem 2.1, yielding the optimal statistical rate. Thus, to prove our main result, it remains to implement line 2 of Algorithm 2 *efficiently*: compute an approximate minimizer of $\widehat{F}^{(w_0)}_{C,U,\lambda,Z}$ over U.

2.4 Efficient Optimization of the Regularized Lipschitz Extension

We now give a primitive for optimizing the empirical regularized Lipschitz-extension.

Joint convex reformulation. Fix a dataset $Z = (z_1, \dots, z_m)$, a center $w_0 \in \mathcal{W}$, a regularization parameter $\lambda > 0$, a Lipschitz parameter $C > 0$, and a closed convex set $U \subseteq \mathcal{W}$. Define

$$\Phi_{Z,U}(w, y_1, \dots, y_m) := \frac{1}{m}\sum_{i=1}^m\left[f(y_i, z_i) + C\|w - y_i\|\right] + \frac{\lambda}{2}\|w - w_0\|^2, \qquad (w, y_1, \dots, y_m) \in U^{m+1}. \tag{8}$$

Algorithm 3: RECURSIVE-LOCALIZED-OUTPUTPERT

Input: Dataset $Z = (z_1, \dots, z_m)$, privacy ε, regularization λ, center w_0
Output: A private center $w_{\mathrm{loc},S} \in \mathcal{W}$

Set $R_\star := \frac{G_k}{\lambda}\left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}}$
if $D \le R_\star$ **then**
  Choose any fixed $w_{\mathrm{loc},S} \in \mathcal{W}$ and **return** $w_{\mathrm{loc},S}$
end
Set $a := 1/(k+1)$ and $S := \left\lceil \log_{k+1} \log\left(\frac{eD}{R_\star}\right)\right\rceil$
Initialize $U_0 := \mathcal{W}$, $R_0 := D$
for $j = 0$ **to** $S-1$ **do**
  Set $\theta_j := \frac{(1-a)a^{S-1-j}}{1-a^S}$, $\varepsilon_j := \frac{\varepsilon}{2}\theta_j$, $\beta_j := \beta_0\theta_j$, where $\beta_0 = 1/100$, and $\ell_j := \log(1/\beta_j)$
  Set $C_j := G_k^{k/(k+1)}(\lambda R_j)^{1/(k+1)}\left(\frac{m\varepsilon_j}{d\ell_j}\right)^{2/(k+1)}\beta_j^{-1/(k+1)}$
  Run $(w_{\mathrm{loc},j+1}, U_{j+1}) \leftarrow$ LOCALIZED-OUTPUTPERT$(Z, \varepsilon_j, \lambda, w_0, U_j, R_j, C_j, \beta_j)$
  Set $R_{j+1} := \min\left\{D,\ 1000\left(\frac{G_k}{\lambda}\right)^{k/(k+1)} R_j^{1/(k+1)}\left(\frac{d\ell_j}{m\varepsilon_j}\right)^{(k-1)/(k+1)}\beta_j^{-1/(k+1)}\right\}$
end
return $w_{\mathrm{loc},S}$

This reformulation avoids evaluating each $f_{C,U}(w, z_i)$: all inner minimizations defining the Lipschitz extensions are folded into one convex program over $(w, y_1, \dots, y_m)$. Lemma 2.6 reduces minimization of the localized Lipschitz-extension objective $\widehat{F}^{(w_0)}_{C,U,\lambda,Z}$ over U to minimization of $\Phi_{Z,U}$ over U^{m+1}.

Lemma 2.6 (Localized joint convex reformulation)**.** *The function $\Phi_{Z,U}$ is convex on U^{m+1}, and*

$$\min_{(w,y_1,\dots,y_m)\in U^{m+1}} \Phi_{Z,U}(w, y_1, \dots, y_m) = \min_{w\in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w).$$

Moreover, if $u_\alpha = (w_\alpha, y_{1,\alpha}, \dots, y_{m,\alpha}) \in U^{m+1}$ satisfies

$$\Phi_{Z,U}(u_\alpha) - \min_{u\in U^{m+1}} \Phi_{Z,U}(u) \le \alpha,$$

then

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w_\alpha) - \min_{w\in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w) \le \alpha.$$

Certified adaptive projected subgradient method solver. We do not know the Lipschitz constant of Φ, since $f(\cdot, z)$ may have unbounded worst-case Lipschitz parameter. Standard first-order methods therefore do not directly yield a certified α-approximate minimizer. We instead use the adaptive projected subgradient method in Algorithm 4, which finds an α-minimizer without requiring knowledge of the Lipschitz constant: it adaptively chooses the step size and uses a stopping criterion based on the norms of observed subgradients. This algorithm terminates in finite time with probability 1 and in polynomial time with high probability because (3) implies that the Lipschitz parameter of Φ is finite with probability 1 and polynomially bounded with high probability.

In the recursive localization procedure, the current set always has the form $U_j = \mathcal{W} \cap \mathbb{B}(c_j, r_j)$. Even if $\mathcal{W}$ is projection-friendly, such intersections need not admit an immediate exact projection oracle. The next lemma shows that this is not an obstacle: projection onto U_j can be implemented to arbitrary accuracy using only logarithmically many calls to the projection oracle for $\mathcal{W}$.

Lemma 2.7 (Efficient ξ-inexact projection onto $\mathcal{W} \cap \mathbb{B}(c, r)$)**.** *Let $\mathcal{W} \subseteq \mathbb{R}^d$ satisfy Assumption 1.1, let $c \in \mathcal{W}$, and $r > 0$. Define $U := \mathcal{W} \cap \mathbb{B}(c, r)$. Then, for every $y \in \mathbb{R}^d$ and every $\xi > 0$, one can compute in polynomial time a point $\widetilde{\Pi}^\xi_U(y) \in U$ satisfying*

$$\left\|\widetilde{\Pi}^\xi_U(y) - \Pi_U(y)\right\|_2 \le \xi.$$

Algorithm 4: ADAPTIVE-INEXACT-PROJSUBGRAD(K, Φ, α, D)

Input: Closed convex set $K \subseteq \mathbb{R}^q$ with $\operatorname{diam}(K) \le D$, convex Φ, target accuracy $\alpha > 0$
Output: A point $u_\alpha \in K$

Choose any $x_1 \in K$
for $t = 1, 2, \dots$ **do**
  Query the first-order oracle at x_t and obtain a subgradient $g_t \in \partial\Phi(x_t)$
  if $g_t = 0$ **then**
    return x_t
  else
    Set $S_t := \sum_{s=1}^{t} \|g_s\|_2^2, \eta_t := \frac{D}{\sqrt{S_t}}, \xi_t := \min\{D, \frac{\alpha\,\eta_t}{6D}\}$
    Compute an ξ_t-inexact projection $x_{t+1} := \widetilde{\Pi}_K^{\xi_t}(x_t - \eta_t g_t)$
    if $3D\sqrt{S_t} \le \alpha t$ **then**
      return $\overline{x}_t := \frac{1}{t}\sum_{s=1}^{t} x_s$
    end
  end
end

The proof exploits the KKT conditions for projection onto $\mathcal{W} \cap \mathbb{B}(w_0, r)$ to reduce the problem to a one-dimensional search over the Lagrange multiplier. We use bisection method to construct an ξ-inexact projector using only logarithmically many calls to the projection oracle for $\mathcal{W}$.

Next, Proposition 2.8 shows that adaptive projected subgradient descent with inexact projection oracle still returns a certified α-minimizer in finite time whenever the realized Lipschitz constant is finite. Thus, the method can be applied over U_j, and over U_j^{m+1} by projecting each coordinate.

Proposition 2.8 (Adaptive Inexact Projected Subgradient Method)**.** *Let $K \subseteq \mathbb{R}^q$ be a nonempty closed convex set equipped with an ξ-inexact projection oracle for every $\xi > 0$, and suppose $\operatorname{diam}(K) \le D$. Let $\Phi : K \to \mathbb{R}$ be convex and L-Lipschitz on K for some finite but unknown $L > 0$. Suppose we are given an exact subgradient oracle for Φ. Fix $\alpha > 0$. Then, Algorithm 4 halts after at most $T_\alpha := \left\lceil \left(\frac{3DL}{\alpha}\right)^2 \right\rceil$ iterations, and its output $u_\alpha \in K$ satisfies $\Phi(u_\alpha) - \min_{u \in K} \Phi(u) \le \alpha$. Hence, if each subgradient query and ξ_t-inexact projection can be performed in polynomial time, then the total running time is polynomial in q, D, L, and $1/\alpha$.*

Proposition 2.8 extends standard adaptive projected subgradient method guarantees (c.f. [DHS11, SS$^+$12]) to inexact projections by charging projection error into the descent estimate. It ensures that we can find an α-minimizer of $\Phi_{Z,U}$ — which, by Lemma 2.6, yields an α-minimizer of $\widehat{F}_{C,U,\lambda,Z}^{(w_0)}$ over U — without knowing a bound on the realized Lipschitz constant. This is essential for pure-DP output perturbation.

Application to localized regularized Lipschitz extensions. We use Lemma 2.6 together with Algorithm 4 to implement the approximate minimization step inside Algorithm 2. Namely, whenever the current localized set is $U = \mathcal{W} \cap \mathbb{B}(c, r)$, we minimize the jointly convex objective $\Phi_{Z,U}$ over U^{m+1} to the accuracy α required by Algorithm 2. Projection onto U is implemented using Lemma 2.7, with w_0 replaced by the current center c. Projection onto the product set U^{m+1} is implemented coordinatewise. Thus Algorithm 3 can be implemented efficiently.

3 Optimal Pure-DP Heavy-Tailed SCO

We now combine the ingredients from Sections 2.1, 2.3 and 2.4. Our main regularized ERM solver is RECURSIVE-LOCALIZED-OUTPUTPERT (Algorithm 3), with each call to LOCALIZED-OUTPUTPERT implemented efficiently by minimizing the localized joint convex objective Φ_{Z,U_j} over U_j^{m+1} using Algorithm 4 with diameter bound $\sqrt{m+1}R_j$ and the inexact projection oracle from Lemma 2.7; see Algorithm 5 for pseudocode. Plugging the efficient ERM solver Algorithm 5 into line 10 of Algorithm 1 gives our main algorithm: **POP-RECURSIVE-LOCALIZED-OUTPUTPERT**.

Public parameter selection. Choose $r \in \{2, k\}$ minimizing $G_r \left(\frac{d \log(1/\delta)}{n\varepsilon}\right)^{1-1/r}$ and use moment order r and bound G_r throughout the algorithm. Return a fixed $w_0 \in \mathcal{W}$ instead if $G_2 \leq G_2\sqrt{\frac{\log(1/\delta)}{n}} + G_r \left(\frac{d \log(1/\delta)}{n\varepsilon}\right)^{1-1/r}$.

Theorem 3.1 (Main theorem)**.** *Let $\delta, \rho \in (0, 1/5)$ and $\varepsilon \in (0, 1]$.* **POP-RECURSIVE-LOCALIZED-OUTPUTPERT***, with the public parameter selection above, is ε-differentially private and with probability at least $1 - \delta$ its output satisfies*

$$F(\widehat{w}) - F^* \leq cD \min\Bigg\{ G_2,\ G_2\sqrt{\frac{\log(1/\delta)}{n}} + \min\left\{ G_2\sqrt{\frac{d \log(1/\delta)}{n\varepsilon}},\ G_k \left(\frac{d \log(1/\delta)}{n\varepsilon}\right)^{1-1/k} \right\} \Bigg\}.$$

Moreover, its runtime is bounded with probability at least $1 - \rho$ by a polynomial in the problem parameters. If $\sup_{z \in \mathcal{Z}} \max_{w \in \mathcal{W}} \|\nabla f(w, z)\|$ is finite and polynomially bounded in these parameters, its runtime is polynomial with probability 1.

The excess risk bound in Theorem 3.1 is optimal up to a factor of $O((\log(1/\delta))^{1-1/k})$. The runtime randomness comes only from the sample Z: by (3), the realized Lipschitz constant of $\Phi_{Z,U}$ is finite almost surely and polynomially bounded with high probability. By a union bound, the excess risk and runtime guarantees both hold simultaneously with probability $\geq 1 - \delta - \rho$. If the worst-case Lipschitz parameter is itself polynomially bounded, the runtime is polynomial with probability 1.

Theorem 3.1 follows from Theorem 2.1 and Proposition 3.2, which gives (6) efficiently under ε-DP:

Proposition 3.2 (Regularized ERM solver)**.** *Fix $w_0 \in \mathcal{W}$, $\lambda > 0$, and $\rho \in (0, 1/5)$. Algorithm 5 is ε-differentially private. If $Z \sim P^m$, then its output w_{DP} satisfies*

$$\Pr\left(\|w_{\mathrm{DP}} - \widehat{w}_\lambda(Z; w_0)\| \leq c_{\mathrm{erm}} \frac{1}{\lambda} \left(G_k \left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}} + \frac{G_2}{\sqrt{m}} \right) \right) \geq 0.7.$$

Moreover, with probability at least $1 - \rho$, the runtime is bounded by a polynomial in $m,\ d,\ D,\ \lambda,\ \frac{1}{\lambda},\ \frac{1}{\varepsilon},\ G_k,\ \frac{1}{\rho}$. Further, if $\sup_{z \in \mathcal{Z}} \max_{w \in \mathcal{W}} \|\nabla f(w, z)\|$ is finite and polynomially bounded in the problem parameters, then the runtime is polynomial with probability 1.

Proof overview. Privacy and the distance guarantee follow from Proposition 2.4. Each approximate minimization step in Algorithm 2 is carried out by the localized joint convex reformulation and the adaptive inexact projected subgradient method. By Lemma 2.6, Lemma 2.7 and Proposition 2.8, the resulting implementation has finite runtime whenever the realized Lipschitz parameter of $\Phi_{Z,U}$ is finite, and polynomial runtime whenever this parameter is polynomially bounded. Under (3), these events hold with probability 1 and with high probability, respectively.

Oracle complexity. For the selected moment order $r \in \{2, k\}$, with probability at least $1 - \rho$, the full SCO algorithm uses

$$\widetilde{O}\left(n^{20} \left(1 + \frac{\lambda_1 D}{G_r}\right)^4 \rho^{-2/r} (d/\varepsilon)^{4/r} \right) \tag{9}$$

sample-subgradient evaluations and calls to $\Pi_{\mathcal{W}}$, where λ_1 is the initial regularization in Algorithm 1. This is a conservative upper bound; Appendix E.2 gives its derivation and a more precise bound that retains the dependence on each phase's parameters. The main cost is solving the joint convex program: each iteration uses m sample subgradients and $\widetilde{O}(m)$ projections, and the iteration bound scales as α^{-2} for optimization accuracy α.

3.1 Deterministic polynomial time for structured subclasses

For the full heavy-tailed function class, Theorem 3.1 yields optimal risk up to logarithmic factors in polynomial time with high probability, and with probability 1 if the worst-case Lipschitz parameter is polynomially bounded. Appendix E.4 describes a different sufficient condition under which the

same excess risk guarantee can also be obtained in deterministic polynomial time, even with infinite worst-case Lipschitz parameter: *efficient approximation of localized Lipschitz-extension subgradients*. Under this condition, the approximate minimization steps in Algorithm 2 can be implemented directly by projected subgradient descent on the localized Lipschitz-extension objective, whose Lipschitz constant is deterministically bounded by C.

Examples from Machine Learning. Several important subclasses of convex losses admit arbitrarily accurate subgradient approximation of localized Lipschitz extensions in polynomial time under mild assumptions on $\mathcal{W}$. For example: polyhedral losses on compact domains admitting an explicit second-order-cone programming representation with a strict interior point. This covers hinge/ReLU-type and absolute-value losses on Euclidean balls, ellipsoids, and polytopes. Thus, for such problems, our algorithm achieves excess risk (4) in deterministic polynomial time.

4 High Probability Lower Bound

We provide a novel high-probability excess risk lower bound, which is tighter than the expected excess risk lower bound derived from [BD14] by logarithmic factors.

Theorem 4.1 (SCO lower bound)**.** *Let $\delta \in (0, 1/5)$, $\varepsilon \in (0, 1]$, and $0 < G_2 \le G_k$. For every ε-DP $\mathcal{A}$, there exists a problem instance with $P \in \mathcal{P}(G_2, G_k)$ satisfying Assumption 1.1 such that*

$$\Pr_{Z\sim P^n}\left(F(\mathcal{A}(Z)) - F^* \ge cD\min\left\{G_2,\ G_2\sqrt{\frac{\log(1/\delta)}{n}}\right.\right.$$
$$\left.\left. + \min\left\{G_2\sqrt{\frac{d+\log(1/\delta)}{n\varepsilon}},\ G_k\left(\frac{d+\log(1/\delta)}{n\varepsilon}\right)^{1-1/k}\right\}\right\}\right) \ge \delta.$$

A fixed output has excess risk at most DG_2. Thus the upper and lower bounds match up to the gap between $d\log(1/\delta)$ and $d + \log(1/\delta)$, for every $0 < G_2 \le G_k$.

Proof idea. The non-private term follows from the high-probability testing bound of [MVS24] applied to a bounded two-point family. For the private term, we extend the packing argument of [BD14] to a high-probability bound using an estimation-to-decoding reduction. Both hard families satisfy the prescribed moment bounds. A standard reduction from mean estimation to linear SCO completes the proof; see Appendix F.

Logarithmic gap. The factor $d\log(1/\delta)$ in our upper bound comes from the population-localization reduction of [ALT24]; a similar gap remains in the zCDP setting. We do not know whether the dependence can be improved to $d + \log(1/\delta)$.

5 Conclusion

We determined the optimal rate for ε-DP heavy-tailed SCO up to a logarithmic factor. Our algorithm achieves this rate in polynomial time with high probability. If the worst-case Lipschitz parameter is polynomially bounded, then runtime is polynomial with probability 1. We also identified a condition that permits deterministic polynomial time even with infinite Lipschitz parameter: efficient approximation of the Lipschitz extension subgradient, which holds for some important ML problems.

Three natural directions remain: (1) Can the $d\log(1/\delta)$ term be improved to $d + \log(1/\delta)$ in the excess risk bound? (2) Is optimal ε-DP risk in deterministic polynomial time possible for arbitrary losses? (3) Can our algorithm be practically implemented for ML model training?

Acknowledgments

We thank Adam Smith for helpful early feedback on the proposed problem and Yu-Xiang Wang for pointing us to [LWMW25].

This work received no third-party funding or support. The author has no competing financial interests to declare.

References

[ADF+21] Hilal Asi, John Duchi, Alireza Fallah, Omid Javidbakht, and Kunal Talwar. Private adaptive gradient methods for convex optimization. In *International Conference on Machine Learning*, pages 383–392. PMLR, 2021.

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *ICML*, 2021.

[ALD21] Hilal Asi, Daniel Lévy, and John C Duchi. Adapting to function difficulty and growth conditions in private optimization. *Advances in Neural Information Processing Systems*, 34:19069–19081, 2021.

[ALT24] Hilal Asi, Daogao Liu, and Kevin Tian. Private stochastic convex optimization with heavy tails: Near-optimality from simple reductions. *Advances in Neural Information Processing Systems*, 37:59174–59215, 2024.

[BD14] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

[BFTT19] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[BGM23] Raef Bassily, Cristóbal Guzmán, and Michael Menart. Differentially private algorithms for the stochastic saddle point problem with optimal rates for the strong gap. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2482–2508. PMLR, 2023.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[DHS11] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[GLZW25] Changyu Gao, Andrew Lowy, Xingyu Zhou, and Stephen J Wright. Optimal rates for robust stochastic convex optimization. In *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, pages 9–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2025.

[GTU23] Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Universality of langevin diffusion for private optimization, with applications to sampling from rashomon sets. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1730–1773. PMLR, 2023.

[HNXW22] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 227–236, 2022.

[HUL13] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I & II*, volume 305. Springer science & business media, 2013.

[KLZ22] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10633–10660. PMLR, 2022.

[LL25] Andrew Lowy and Daogao Liu. Differentially private bilevel optimization: Efficient algorithms with near-optimal rates. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[LLA24] Andrew Lowy, Daogao Liu, and Hilal Asi. Faster algorithms for user-level private stochastic convex optimization. *Advances in Neural Information Processing Systems*, 37:96071–96100, 2024.

[LR21] Andrew Lowy and Meisam Razaviyayn. Output perturbation for differentially private convex optimization: Faster and more general. *arXiv preprint arXiv:2102.04704*, 2021.

[LR25] Andrew Lowy and Meisam Razaviyayn. Private stochastic optimization with large worst-case lipschitz parameter. *Journal of Privacy and Confidentiality*, 15:1, 2025.

[LWMW25] Yingyu Lin, Erchi Wang, Yi-An Ma, and Yu-Xiang Wang. Purifying approximate differential privacy with randomized post-processing. *arXiv preprint arXiv:2503.21071*, 2025.

[MVS24] Tianyi Ma, Kabir A Verchand, and Richard J Samworth. High-probability minimax lower bounds. *arXiv preprint arXiv:2406.13447*, 2024.

[SS$^+$12] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[TWZW21] Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits. *arXiv preprint arXiv:2106.02575*, 2021.

[WX22] Di Wang and Jinhui Xu. Differentially private ℓ_1-norm linear regression with heavy-tailed data. *arXiv preprint arXiv:2201.03204*, 2022.

[WXDX20] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

Appendix

A Zero-Concentrated Differential Privacy

We recall the standard definition of zero-concentrated differential privacy (zCDP).

Definition A.1 (zCDP)**.** A randomized algorithm $\mathcal{A}$ is ρ-zCDP if for every pair of neighboring datasets Z, Z' and every $\alpha > 1$,

$$D_\alpha(\mathcal{A}(Z)\|\mathcal{A}(Z')) \le \rho\alpha,$$

where D_α denotes the order-α Rényi divergence.

The following standard facts show that zCDP lies between pure and approximate DP.

Lemma A.2 ([BS16])**.** *1. If $\mathcal{A}$ is pure ε-DP, then $\mathcal{A}$ is $(\varepsilon^2/2)$-zCDP.*

2. If $\mathcal{A}$ is ρ-zCDP, then for every $\delta \in (0,1)$, $\mathcal{A}$ is

$$\left(\rho + 2\sqrt{\rho\log(1/\delta)},\ \delta\right)\text{-DP}.$$

B Additional Discussion of Related Work

This appendix provides additional context on related work. A concise discussion appears in the introduction; here we elaborate on the most closely related lines.

Private uniformly Lipschitz SCO. Differentially private stochastic convex optimization has been studied extensively over the past decade under uniform Lipschitz assumptions on the loss function, with sharp excess-risk guarantees now known in many settings [BFTT19, FKT20, AFKT21, BGM23, LLA24, LL25]. In particular, optimal rates for pure ε-DP SCO under the uniform Lipschitz assumption are known up to logarithmic factors [ALD21, GTU23, BST14]. This literature assumes a finite worst-case Lipschitz parameter L, and the resulting upper and lower bounds scale with L. In contrast, the present paper focuses on the heavy-tailed regime, where L may be arbitrarily large and only a finite k-th moment bound is assumed; our bounds scale with G_k rather than L.

Heavy-tailed private SCO. A recent line of work studies differentially private SCO under moment assumptions on the sample-wise Lipschitz parameters or gradients rather than a uniform Lipschitz bound [TWZW21, ADF+21, KLZ22, HNXW22, WX22, ALT24, LR25]. These works show that one can obtain substantially sharper guarantees in heavy-tailed settings by replacing worst-case dependence on L with dependence on the moment parameter G_k. In particular, the heavy-tailed regime is well understood under relaxed privacy notions such as ρ-zCDP, with efficient clipped-gradient-based algorithms achieving the optimal rates up to logarithmic factors.

The pure-DP gap. The pure ε-DP heavy-tailed case has remained open on the algorithmic side. Prior to this work, the main known lower-bound ingredient was the pure-DP mean-estimation lower bound of [BD14], which implies via the standard reduction from stochastic convex optimization to mean estimation an in-expectation / constant-probability SCO lower bound of the form

$$\Omega\left(G_k D\left(\frac{d}{n\varepsilon}\right)^{1-\frac{1}{k}} + \frac{G_2 D}{\sqrt{n}}\right).$$

However, no matching pure-DP upper bound for heavy-tailed SCO was previously known. Existing heavy-tailed algorithmic approaches were based on clipped noisy gradients and were tailored to zCDP or approximate DP, where privacy accounting under composition is significantly more forgiving. Our paper closes this algorithmic gap by giving the first direct pure ε-DP algorithms matching the minimax rate up to logarithmic factors.

In addition, we prove sharper high-probability lower bounds. For the non-private term, we use a bounded two-point construction together with the high-probability testing framework of [MVS24]. For the private term, we build on the pure-DP packing ideas underlying [BD14], together with a direct reduction from quantile estimation to decoding on a packing. This yields a high-probability private lower bound with explicit dependence on the failure probability parameter.

Relation to clipped-gradient methods. Clipping-based methods remain central in private optimization, including in heavy-tailed settings. Our main contribution is not merely an optimal statistical upper bound in polynomial time, but a different algorithmic route: we replace clipped-gradient optimization by private optimization of a Lipschitz extension of the empirical loss.

Relation to black-box purification from zCDP. One possible black-box route to pure DP is to start from the ρ-zCDP heavy-tailed SCO algorithm of [ALT24], convert its guarantee to approximate DP using the standard zCDP-to-approximate-DP implication, and then apply the purification framework of [LWMW25]. However, this route does not recover our main result. The most direct composition yields a pure-DP excess risk upper bound that is *looser* than our upper bound in Theorem 3.1 by $\mathrm{poly}(\log(nd/\varepsilon))$ factors. Moreover, this *indirect* approach is algorithmically more cumbersome: it requires combining the population-level localization framework of [ALT24] with the purification procedure of [LWMW25], thereby passing through multiple intermediate privacy notions. In contrast, our algorithm is designed directly for pure DP and achieves a high-probability excess-risk guarantee without passing through an intermediate zCDP-to-approximate-DP conversion or an additional purification step, and eliminating spurious polylogarithmic factors.

Lipschitz extensions in optimization and privacy. Lipschitz extensions are classical objects in convex analysis [HUL13] and have appeared in several privacy-related contexts. In our setting, the challenge is not only to use the extension as an analytical device, but to optimize it efficiently under pure DP. This requires handling the fact that exact evaluation of the Lipschitz extension cannot be certified from a fixed finite collection of local oracle queries in general (Appendix C) and that pure-DP output perturbation requires deterministic optimization accuracy guarantees. Our jointly convex reformulation and adaptive certified projected subgradient framework are designed to address exactly this obstacle.

Output perturbation and localization. Output perturbation for strongly convex empirical risk minimization goes back at least to [CMS11, BST14, LR21], and localization ideas play an important role in modern private optimization. Our use of output perturbation is structurally related to these earlier works, but serves a different purpose: we use recursive output-perturbation localization steps to shrink the domain so that the localized Lipschitz-extension bias becomes small enough. The final private center is then used as the phasewise regularized ERM output.

Summary of positioning. In summary, prior work resolved the heavy-tailed SCO problem under zCDP and established a pure-DP lower bound, but did not provide a matching pure-DP upper bound. This paper is, to the best of our knowledge, the first to obtain a direct polynomial-time pure-DP algorithm for heavy-tailed SCO with a nearly-tight high-probability excess-risk guarantee. In particular, unlike the black-box purification route from zCDP to approximate DP and then purifying, our guarantee does not incur the additional conversion and polylogarithmic overheads discussed above. Moreover, we establish novel high-probability lower bounds by combining pure-DP packing ideas with high-probability testing and decoding arguments.

Remark B.1 (On publicly known moment bounds). We assume that G_2, G_k are publicly known, as in prior work on heavy-tailed private optimization. Using larger valid moment bounds preserves the utility guarantee with those bounds substituted; underestimating either can invalidate it. Underestimating k is safe but may worsen the rate; overestimating it can invalidate the utility guarantee. Privately adapting to unknown moment bounds remains open.

C Finite Local Information Does Not Determine the Lipschitz Extension

The Lipschitz extension

$$f_C(w) = \inf_{y \in \mathcal{W}} \{f(y) + C\|w - y\|\}$$

is a global object. The following oracle lower-bound example shows that, even for convex C^∞ functions, a fixed finite set of local queries does not determine the exact value of $f_C(w)$. This motivates our use of certified approximate optimization of a joint convex reformulation rather than exact evaluation of the extension.

Theorem C.1 (Finite nonadaptive local information does not determine the Lipschitz extension)**.** *Fix $k \geq 2$, $0 < C < G_k$, and an integer $T \geq 1$. Let $q_1, \ldots, q_T \in B_d := \{y \in \mathbb{R}^d : \|y\| \leq 1\}$ be any*

fixed query points, where $d \geq T+1$. Then there exist two convex C^∞ functions $f_0, f_1 : B_d \to \mathbb{R}$ such that

$$\max_{y \in B_d} \|\nabla f_i(y)\| \leq G_k \qquad \textit{for } i \in \{0,1\},$$

their complete local information agrees at every query point,

$$\nabla^m f_0(q_t) = \nabla^m f_1(q_t) \qquad \textit{for every } t \in [T] \textit{ and every } m \geq 0,$$

where $m = 0$ denotes equality of function values, but their Lipschitz-extension values at the interior point $w = 0$ are different:

$$(f_0)_C(0) \neq (f_1)_C(0), \qquad (f_i)_C(0) := \inf_{y \in B_d} \big\{ f_i(y) + C\|y\| \big\}.$$

Consequently, exact evaluation of $f_C(0)$ cannot be determined, in general, from any fixed finite collection of local oracle queries.

Proof. **Step 1: A smooth convex function flat at the origin.** Define

$$\eta(s) := \begin{cases} e^{-1/s^2}, & s \neq 0, \\ 0, & s = 0. \end{cases}$$

It is standard that $\eta \in C^\infty(\mathbb{R})$, $\eta(s) \geq 0$, and

$$\eta^{(m)}(0) = 0 \qquad \text{for all } m \geq 0.$$

Now set

$$\varphi(s) := \int_0^s \int_0^u \eta(r)\, dr\, du.$$

Then $\varphi \in C^\infty(\mathbb{R})$, $\varphi''(s) = \eta(s) \geq 0$, so φ is convex, and

$$\varphi^{(m)}(0) = 0 \qquad \text{for all } m \geq 0.$$

Moreover, $\varphi(s) > 0$ for every $s \neq 0$, and

$$\varphi(s) = o(s) \qquad \text{as } s \to 0.$$

Let

$$M := \sup_{|s| \leq 1} |\varphi'(s)| < \infty.$$

Step 2: Define two one-dimensional convex profiles with identical jets at 0. Choose any $a \in (C, G_k)$. Since $a < G_k$, we may choose $0 < \beta_0 < \beta_1$ such that

$$a + \beta_1 M \leq G_k.$$

For $i \in \{0,1\}$, define

$$g_i(t) := -at + \beta_i \varphi(t), \qquad t \in [-1,1].$$

Each g_i is convex and C^∞. Since every derivative of φ vanishes at 0,

$$g_0^{(m)}(0) = g_1^{(m)}(0) \qquad \text{for all } m \geq 0.$$

Step 3: Lift to high dimension along a direction hidden from the fixed queries. Since $d \geq T+1$, there exists a unit vector $v \in \mathbb{R}^d$ orthogonal to all query points:

$$\langle v, q_t \rangle = 0 \qquad \text{for } t = 1, \ldots, T.$$

Define

$$f_i(y) := g_i(\langle v, y \rangle), \qquad y \in B_d, \quad i \in \{0,1\}.$$

Since g_i is convex and $y \mapsto \langle v, y \rangle$ is linear, each f_i is convex and C^∞. Moreover,

$$\nabla f_i(y) = g_i'(\langle v, y \rangle) v = \big(-a + \beta_i \varphi'(\langle v, y \rangle) \big) v,$$

so for every $y \in B_d$,

$$\|\nabla f_i(y)\| \le a + \beta_i \sup_{|s|\le 1} |\varphi'(s)| \le a + \beta_1 M \le G_k.$$

At each query point q_t, we have $\langle v, q_t\rangle = 0$. Therefore, for $m = 0$,

$$f_0(q_t) = g_0(0) = g_1(0) = f_1(q_t).$$

For every integer $m \ge 1$,

$$\nabla^m f_i(q_t) = g_i^{(m)}(0)\, v^{\otimes m}.$$

Since $g_0^{(m)}(0) = g_1^{(m)}(0)$ for all $m \ge 1$, the complete local information of f_0 and f_1 agrees at every q_t.

Step 4: The extension values at 0 differ. Write any $y \in B_d$ as

$$y = tv + u, \qquad \langle u, v\rangle = 0, \qquad t^2 + \|u\|^2 \le 1.$$

Then

$$f_i(y) = g_i(t), \qquad \|y\| = \sqrt{t^2 + \|u\|^2} \ge |t|,$$

with equality when $u = 0$. Hence

$$(f_i)_C(0) = \inf_{\|y\|\le 1}\{f_i(y) + C\|y\|\} = \inf_{t\in[-1,1]}\{g_i(t) + C|t|\}.$$

Define

$$h_i(t) := g_i(t) + C|t|.$$

For $t < 0$,

$$h_i(t) = -at + \beta_i\varphi(t) + C|t| = (a + C)|t| + \beta_i\varphi(t) > 0,$$

and $h_i(0) = 0$. For $t > 0$,

$$h_i(t) = -(a - C)t + \beta_i\varphi(t).$$

Since $a > C$ and $\varphi(t) = o(t)$ as $t \downarrow 0$, there exists $t_i > 0$ such that $h_i(t_i) < 0$. Thus every minimizer of h_i over $[-1, 1]$ lies in $(0, 1]$.

Finally, for every $t > 0$,

$$h_1(t) = h_0(t) + (\beta_1 - \beta_0)\varphi(t) > h_0(t).$$

Since both minima are attained at positive points, it follows that

$$(f_1)_C(0) = \min_{t\in[-1,1]} h_1(t) > \min_{t\in[-1,1]} h_0(t) = (f_0)_C(0).$$

This proves the theorem. □

D Deferred Proofs for Section 2

D.1 Proofs for Section 2.2

For a dataset $Z = (z_1, \dots, z_m)$, define

$$\widehat{F}_Z(w) := \frac{1}{m}\sum_{i=1}^m f(w, z_i), \qquad \widehat{F}_{C,Z}(w) := \frac{1}{m}\sum_{i=1}^m f_C(w, z_i).$$

Lemma D.1 (Bias of the Empirical Lipschitz Extension)**.** *Assume $f(\cdot, z)$ is convex on a domain $\mathcal{W}$ of diameter D. Denote $a_+ := \max(a, 0)$. Then for every dataset $Z = (z_1, \dots, z_m)$ and every $w \in \mathcal{W}$,*

$$\widehat{F}_Z(w) - \widehat{F}_{C,Z}(w) \le \frac{D}{m}\sum_{i=1}^m (A(z_i) - C)_+, \qquad A(z) := \sup_{u\in\mathcal{W}} \sup_{g\in\partial f(\cdot,z)(u)} \|g\|.$$

Consequently, under the moment condition (3)*,*

$$\mathbb{E}_Z\left[\max_{w\in\mathcal{W}}\big(\widehat{F}_Z(w) - \widehat{F}_{C,Z}(w)\big)\right] \le \frac{DG_k^k}{(k-1)C^{k-1}}.$$

Proof. Fix z and define

$$A(z) := \sup_{u \in \mathcal{W}} \sup_{g \in \partial f(\cdot, z)(u)} \|g\|.$$

Since $f(\cdot, z)$ is convex on the convex set $\mathcal{W}$, it is $A(z)$-Lipschitz on $\mathcal{W}$. Hence for every $w, y \in \mathcal{W}$,

$$|f(w, z) - f(y, z)| \le A(z)\|w - y\|.$$

By definition of the Lipschitz extension,

$$f_C(w, z) = \inf_{y \in \mathcal{W}} \big[f(y, z) + C\|w - y\| \big],$$

so

$$\begin{aligned} f(w, z) - f_C(w, z) &= \sup_{y \in \mathcal{W}} \big(f(w, z) - f(y, z) - C\|w - y\| \big) \\ &\le \sup_{y \in \mathcal{W}} (A(z) - C)\|w - y\| \\ &\le D(A(z) - C)_+. \end{aligned}$$

Averaging over $z_1, \dots, z_m$ gives the first claim.

Taking expectation over Z and using i.i.d. sampling,

$$\mathbb{E}_Z \left[\max_{w \in \mathcal{W}} \big(\widehat{F}_Z(w) - \widehat{F}_{C,Z}(w) \big) \right] \le D\, \mathbb{E}_{z \sim P}[(A(z) - C)_+].$$

Using the layer-cake representation and Markov's inequality,

$$\begin{aligned} \mathbb{E}[(A(z) - C)_+] &= \int_C^\infty \mathbb{P}(A(z) \ge s)\, ds \\ &\le \int_C^\infty \frac{\mathbb{E}[A(z)^k]}{s^k}\, ds = \frac{G_k^k}{(k-1)C^{k-1}}. \end{aligned}$$

□

D.2 Details and proofs for Section 2.3

For a dataset $Z = (z_1, \dots, z_m)$, define, for $w \in U$,

$$\widehat{F}_Z(w) := \frac{1}{m} \sum_{i=1}^m f(w, z_i), \qquad \widehat{F}_{C,U,Z}(w) := \frac{1}{m} \sum_{i=1}^m f_{C,U}(w, z_i),$$

and

$$\widehat{F}_{\lambda,Z}^{(w_0)}(w) := \widehat{F}_Z(w) + \frac{\lambda}{2}\|w - w_0\|^2, \qquad \widehat{F}_{C,U,\lambda,Z}^{(w_0)}(w) := \widehat{F}_{C,U,Z}(w) + \frac{\lambda}{2}\|w - w_0\|^2.$$

Finally, define

$$\widehat{w}_\lambda(Z; w_0) := \arg\min_{w \in \mathcal{W}} \widehat{F}_{\lambda,Z}^{(w_0)}(w), \qquad \widehat{w}_{C,U,\lambda}(Z; w_0) := \arg\min_{w \in U} \widehat{F}_{C,U,\lambda,Z}^{(w_0)}(w).$$

Lemma D.2 (Precise Statement of Lemma 2.3)**.** *Let $R > 0$, and define*

$$A(z) := \sup_{u \in \mathcal{W}} \sup_{g \in \partial f(\cdot, z)(u)} \|g\|.$$

Then, for every dataset $Z = (z_1, \dots, z_m)$, simultaneously for every closed convex set $U \subseteq \mathcal{W}$ with $\operatorname{diam}(U) \le R$,

$$\max_{w \in U} \big(\widehat{F}_Z(w) - \widehat{F}_{C,U,Z}(w) \big) \le R \cdot \frac{1}{m} \sum_{i=1}^m (A(z_i) - C)_+.$$

Consequently,

$$\mathbb{E}_Z \left[\sup_{\substack{U \subseteq \mathcal{W} \text{ closed,convex} \\ \operatorname{diam}(U) \le R}} \max_{w \in U} \big(\widehat{F}_Z(w) - \widehat{F}_{C,U,Z}(w) \big) \right] \le \frac{R G_k^k}{(k-1)C^{k-1}}.$$

and for every $\beta \in (0,1)$, with probability at least $1-\beta$, simultaneously for every closed convex $U \subseteq \mathcal{W}$ with $\operatorname{diam}(U) \le R$

$$\max_{w \in U} \big(\widehat{F}_Z(w) - \widehat{F}_{C,U,Z}(w)\big) \le \frac{RG_k^k}{\beta(k-1)C^{k-1}}.$$

Moreover, with probability at least $1-\beta$, the following holds simultaneously for every closed convex $U \subseteq \mathcal{W}$ with $\operatorname{diam}(U) \le R$ and $\widehat{w}_\lambda(Z; w_0) \in U$:

$$\|\widehat{w}_{C,U,\lambda}(Z; w_0) - \widehat{w}_\lambda(Z; w_0)\| \le \sqrt{\frac{2RG_k^k}{\lambda\beta(k-1)C^{k-1}}}.$$

Proof. Fix z and a closed convex set $U \subseteq \mathcal{W}$. For every $w, y \in U$, Lipschitz continuity implies

$$f(w,z) - f(y,z) \le A(z)\|w - y\|.$$

Thus, for every $w \in U$,

$$\begin{aligned} f(w,z) - f_{C,U}(w,z) &= \sup_{y \in U} \big(f(w,z) - f(y,z) - C\|w-y\|\big) \\ &\le \sup_{y \in U} (A(z) - C)\|w - y\| \\ &\le \operatorname{diam}(U)(A(z) - C)_+. \end{aligned}$$

Therefore, if $\operatorname{diam}(U) \le R$, then

$$\max_{w \in U} \big(\widehat{F}_Z(w) - \widehat{F}_{C,U,Z}(w)\big) \le R \cdot \frac{1}{m} \sum_{i=1}^m (A(z_i) - C)_+.$$

This deterministic inequality holds simultaneously for all such U.

Taking expectation and using

$$\mathbb{E}[(A(z) - C)_+] \le \int_C^\infty \frac{\mathbb{E}[A(z)^k]}{s^k}\, ds \le \frac{G_k^k}{(k-1)C^{k-1}}$$

proves the expectation bound. The high-probability bound follows from Markov's inequality applied to $\frac{1}{m}\sum_{i=1}^m (A(z_i) - C)_+$.

It remains to prove the distance bound. The distance statement is a consequence of the preceding high-probability event. Indeed, on that event, fix any closed convex $U \subseteq \mathcal{W}$ with $\operatorname{diam}(U) \le R$ and $\widehat{w}_\lambda(Z; w_0) \in U$, and let

$$B := \frac{RG_k^k}{\beta(k-1)C^{k-1}}.$$

On the high-probability event above, suppose $\widehat{w}_\lambda(Z; w_0) \in U$. Since $f_{C,U} \le f$ on U, we have

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\widehat{w}_\lambda) \le \widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_\lambda).$$

By optimality of $\widehat{w}_{C,U,\lambda}$ over U,

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\widehat{w}_{C,U,\lambda}) \le \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\widehat{w}_\lambda).$$

Finally, by the localized bias bound applied at $\widehat{w}_{C,U,\lambda} \in U$,

$$\widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_{C,U,\lambda}) \le \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\widehat{w}_{C,U,\lambda}) + B.$$

Combining the last three displays gives

$$\widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_{C,U,\lambda}) - \widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_\lambda) \le B.$$

Since $\widehat{F}^{(w_0)}_{\lambda,Z}$ is λ-strongly convex and minimized at $\widehat{w}_\lambda$,

$$\frac{\lambda}{2}\|\widehat{w}_{C,U,\lambda} - \widehat{w}_\lambda\|^2 \le B,$$

which proves the claimed distance bound. □

Localizing the domain via output perturbation. Given a current set U, Algorithm 2 privately computes an approximate minimizer of the U-localized Lipschitz-extension objective, adds Laplace noise, and returns a ball around the noisy point. The radius is chosen so that, if U contains the regularized ERM minimizer $\widehat{w}_\lambda(Z; w_0)$, the returned set also contains it with high probability.

Lemma D.3 (Guarantees of Algorithm 2)**.** *Let $Z \sim P^m$, $w_0 \in \mathcal{W}$, and $\lambda > 0$. For every fixed closed convex set $U \subseteq \mathcal{W}$ and deterministic $R \geq \operatorname{diam}(U)$, Algorithm 2 is ε-DP. Moreover, the utility statement remains valid when U is chosen adaptively from previous private outputs, provided $R \geq \operatorname{diam}(U)$ is the deterministic diameter bound used by the algorithm: if $\widehat{w}_\lambda(Z; w_0) \in U$, then with probability at least $1 - 2\beta$,*

$$\widehat{w}_\lambda(Z; w_0) \in U^+.$$

Proof. Let

$$\widehat{w}_{C,U,\lambda}(Z; w_0) := \arg\min_{w \in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w).$$

Privacy. For fixed U, each $f_{C,U}(\cdot, z)$ is C-Lipschitz on U. Hence exact minimizers of neighboring λ-strongly convex empirical objectives have ℓ_2-sensitivity at most $2C/(\lambda m)$. Since $\tilde{w}$ is an α-approximate minimizer and $\widehat{F}^{(w_0)}_{C,U,\lambda,Z}$ is λ-strongly convex,

$$\|\tilde{w} - \widehat{w}_{C,U,\lambda}(Z; w_0)\| \leq \sqrt{\frac{2\alpha}{\lambda}} = \frac{C}{\sqrt{50}\lambda m} \min\left\{1, \frac{d\ell}{\varepsilon}\right\} \leq \frac{C}{\sqrt{50}\lambda m}.$$

Therefore the ℓ_2-sensitivity of the deterministic map $Z \mapsto \tilde{w}(Z)$ is at most

$$\frac{2C}{\lambda m} + 2 \cdot \frac{C}{\sqrt{50}\lambda m} \leq \frac{3C}{\lambda m}.$$

The isotropic Laplace mechanism with density proportional to

$$\exp\left(-\frac{\varepsilon \lambda m}{10C}\|b\|\right)$$

is therefore ε-DP. Projection onto $\mathcal{W}$ and the construction of U^+ are post-processing. For adaptively chosen U, this privacy statement applies conditionally on the previous private outputs, and the full recursive use follows by adaptive composition.

Uniform localized-bias event. Define

$$A(z) := \sup_{u \in \mathcal{W}} \sup_{g \in \partial f(\cdot, z)(u)} \|g\|$$

and the event

$$E_{\text{bias}} := \left\{ \frac{1}{m} \sum_{i=1}^{m} (A(z_i) - C)_+ \leq \frac{G_k^k}{\beta(k-1)C^{k-1}} \right\}.$$

By the tail-integral bound and Markov's inequality, $\Pr(E_{\text{bias}}) \geq 1 - \beta$. On E_{bias}, by Lemma 2.3, simultaneously for every closed convex $V \subseteq \mathcal{W}$ with $\operatorname{diam}(V) \leq R$,

$$\max_{w \in V} \big(\widehat{F}_Z(w) - \widehat{F}_{C,V,Z}(w)\big) \leq \frac{RG_k^k}{\beta(k-1)C^{k-1}}.$$

This uniformity is what allows the utility proof to apply to adaptively chosen sets U.

Distance from the localized-extension minimizer to the original ERM minimizer. Assume $\widehat{w}_\lambda(Z; w_0) \in U$. On E_{bias}, using $\operatorname{diam}(U) \leq R$,

$$\max_{w \in U} \big(\widehat{F}_Z(w) - \widehat{F}_{C,U,Z}(w)\big) \leq \frac{RG_k^k}{\beta(k-1)C^{k-1}}.$$

For every $w \in U$,

$$\widehat{F}^{(w_0)}_{\lambda,Z}(w) \leq \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w) + \max_{u \in U} \big(\widehat{F}_Z(u) - \widehat{F}_{C,U,Z}(u)\big).$$

Moreover, since $\widehat{w}_\lambda \in U$ and $f_{C,U} \le f$ on U,

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\widehat{w}_{C,U,\lambda}) \le \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(\widehat{w}_\lambda) \le \widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_\lambda).$$

Combining the last two displays gives

$$\widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_{C,U,\lambda}) - \widehat{F}^{(w_0)}_{\lambda,Z}(\widehat{w}_\lambda) \le \frac{RG_k^k}{\beta(k-1)C^{k-1}}.$$

Since $\widehat{F}^{(w_0)}_{\lambda,Z}$ is λ-strongly convex,

$$\|\widehat{w}_{C,U,\lambda} - \widehat{w}_\lambda\| \le \sqrt{\frac{2RG_k^k}{\lambda\beta(k-1)C^{k-1}}}$$

on E_{bias}.

Noise tail and conclusion. Let

$$s := \frac{10C}{\varepsilon\lambda m}, \qquad \ell := \log(1/\beta).$$

For isotropic Laplace noise with density proportional to $\exp(-\|b\|/s)$, the radial variable $\|b\|/s$ has $\mathrm{Gamma}(d,1)$ distribution. Hence by a Chernoff bound

$$\Pr\left(\|b\| \le 12\frac{Cd\ell}{\lambda m\varepsilon}\right) \ge 1-\beta.$$

Call this event E_{noise}.

On $E_{\mathrm{bias}} \cap E_{\mathrm{noise}}$, using nonexpansiveness of projection onto $\mathcal{W}$ and $\widehat{w}_\lambda \in \mathcal{W}$,

$$\begin{aligned}
\|w_{\mathrm{loc}} - \widehat{w}_\lambda\| &= \|\Pi_{\mathcal{W}}(\tilde{w}+b) - \Pi_{\mathcal{W}}(\widehat{w}_\lambda)\| \\
&\le \|\tilde{w} + b - \widehat{w}_\lambda\| \\
&\le \|\tilde{w} - \widehat{w}_{C,U,\lambda}\| + \|\widehat{w}_{C,U,\lambda} - \widehat{w}_\lambda\| + \|b\| \\
&\le \sqrt{\frac{2RG_k^k}{\lambda\beta(k-1)C^{k-1}}} + \frac{20Cd\ell}{\lambda m\varepsilon} \le r^+
\end{aligned}$$

Here we used the choice of α, which gives

$$\|\tilde{w} - \widehat{w}_{C,U,\lambda}\| \le \frac{C}{\sqrt{50}\lambda m}\min\left\{1, \frac{d\ell}{\varepsilon}\right\} \le \frac{Cd\ell}{\sqrt{50}\lambda m\varepsilon}.$$

Therefore

$$\widehat{w}_\lambda(Z; w_0) \in \mathcal{W} \cap \mathbb{B}(w_{\mathrm{loc}}, R^+) = U^+,$$

on $E_{\mathrm{bias}} \cap E_{\mathrm{noise}}$.

Finally,

$$\Pr(E_{\mathrm{bias}} \cap E_{\mathrm{noise}}) \ge 1 - 2\beta,$$

which proves the lemma. □

Proof of Proposition 2.4. **Privacy:** Privacy follows from basic adaptive composition:

$$\sum_{j=0}^{S-1} \varepsilon_j = \frac{\varepsilon}{2}\sum_{j=0}^{S-1}\theta_j = \frac{\varepsilon}{2} \le \varepsilon.$$

Thus the recursive localization algorithm is ε-DP.

Distance Bound: For each round j, define the bias-good event

$$B_j := \left\{ \frac{1}{m} \sum_{i=1}^{m} (A(z_i) - C_j)_+ \le \frac{G_k^k}{\beta_j (k-1) C_j^{k-1}} \right\}, \qquad A(z) := \sup_{u \in \mathcal{W}} \sup_{g \in \partial f(\cdot, z)(u)} \|g\|.$$

By Markov's inequality and Lemma 2.3, $\Pr(B_j) \ge 1 - \beta_j$. Crucially, B_j controls the localized-extension bias simultaneously for every closed convex $U \subseteq \mathcal{W}$ with $\operatorname{diam}(U) \le R_j$, so it applies to the adaptively chosen set U_j.

Let

$$N_j := \left\{ \|b_j\| \le c_{\mathrm{Lap}} \frac{C_j d \ell_j}{\lambda m \varepsilon_j} \right\}$$

denote the corresponding Laplace-noise good event in round j, where b_j is the isotropic Laplace noise that is added in the j-th call of Algorithm 2 and c_{Lap} is a sufficiently large absolute constant that we will absorb into the final constant. Then $\Pr(N_j) \ge 1 - \beta_j$. On the event

$$\mathcal{G} := \bigcap_{j=0}^{S-1} (B_j \cap N_j),$$

an induction using Lemma D.3 gives

$$\widehat{w}_\lambda(Z; w_0) \in U_j \qquad \text{for all } j = 0, \dots, S.$$

Moreover,

$$\Pr(\mathcal{G}) \ge 1 - 2 \sum_{j=0}^{S-1} \beta_j = 1 - 2\beta_0 \ge 0.98.$$

It remains to bound the final radius. Set

$$a := \frac{1}{k+1}, \qquad b := \frac{k-1}{k+1}, \qquad R_\star := \frac{G_k}{\lambda} \left(\frac{d}{m\varepsilon} \right)^{1 - \frac{1}{k}}.$$

If $R_\star \ge D$, then

$$\|w_{\mathrm{loc},S} - \widehat{w}_\lambda(Z; w_0)\| \le D \le R_\star,$$

so the claim holds. Hence assume $R_\star < D$.

By the radius update and the definitions of ε_j, β_j,

$$R_{j+1} \le c \left(\frac{G_k}{\lambda} \right)^{k/(k+1)} R_j^a \left(\frac{d \ell_j}{m \varepsilon_j} \right)^b \beta_j^{-a}.$$

Dividing by $R_\star$, and using

$$\varepsilon_j = \frac{\varepsilon}{2} \theta_j, \qquad \beta_j = \beta_0 \theta_j,$$

we obtain

$$\frac{R_{j+1}}{R_\star} \le C_k \left(\frac{R_j}{R_\star} \right)^a \theta_j^{-(a+b)} \ell_j^b,$$

where C_k is an absolute constant depending only on k and β_0. Here $a + b = k/(k+1)$. Indeed, the identity follows from

$$R_\star^{1-a} = \left(\frac{G_k}{\lambda} \right)^{k/(k+1)} \left(\frac{d}{m\varepsilon} \right)^{(k-1)/(k+1)}.$$

Unrolling the recurrence gives

$$\frac{R_S}{R_\star} \le C_k^{1 + a + \cdots + a^{S-1}} \left(\frac{R_0}{R_\star} \right)^{a^S} \prod_{j=0}^{S-1} \left(\theta_j^{-(a+b)} \ell_j^b \right)^{a^{S-1-j}}.$$

Since $a < 1$, the prefactor $C_k^{1+a+\cdots+a^{S-1}}$ is bounded by a constant depending only on k, and we absorb it into C_k. Thus

$$\frac{R_S}{R_\star} \le C_k \left(\frac{R_0}{R_\star}\right)^{a^S} \prod_{j=0}^{S-1} \left(\theta_j^{-(a+b)} \ell_j^b\right)^{a^{S-1-j}}.$$

It remains to bound the product. Write $q = S - 1 - j$. By the choice

$$\theta_j = \frac{(1-a)a^{S-1-j}}{1-a^S},$$

we have

$$\theta_j \asymp_k a^q, \qquad \ell_j = \log(1/\beta_j) = \log(1/(\beta_0\theta_j)) \lesssim_k 1 + q.$$

Therefore

$$\log \prod_{j=0}^{S-1} \left(\theta_j^{-(a+b)} \ell_j^b\right)^{a^{S-1-j}} = \sum_{q=0}^{S-1} a^q \left[(a+b)\log(1/\theta_{S-1-q}) + b \log \ell_{S-1-q}\right],$$

by changing variables $q = S - 1 - j$. Using $\theta_{S-1-q} \asymp_k a^q$ and $\ell_{S-1-q} \lesssim_k 1 + q$, this is at most

$$C_k + C_k \sum_{q=0}^{\infty} a^q q + C_k \sum_{q=0}^{\infty} a^q \log(1+q) \le C_k$$

after enlarging C_k, because $a \in (0,1)$. Hence the product is bounded by a constant depending only on k and β_0. Consequently,

$$R_S \le C_k R_\star \left(\frac{R_0}{R_\star}\right)^{a^S}.$$

Since $R_0 = D$ and

$$S = \left\lceil \log_{k+1} \log\left(\frac{eD}{R_\star}\right) \right\rceil,$$

we have

$$\left(\frac{D}{R_\star}\right)^{a^S} \le e.$$

Hence

$$R_S \le C_k R_\star$$

after enlarging the constant C_k.

On $\mathcal{G}$, the final center $w_{\mathrm{loc},S}$ satisfies

$$\|w_{\mathrm{loc},S} - \widehat{w}_\lambda(Z; w_0)\| \le R_S \le C_k \frac{G_k}{\lambda} \left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}}.$$

This proves the proposition. □

D.3 Proofs for Section 2.4

Lemma D.4 (Re-statement of Lemma 2.6)**.** *The function $\Phi_{Z,U}$ is convex on U^{m+1}, and*

$$\min_{(w,y_1,\ldots,y_m)\in U^{m+1}} \Phi_{Z,U}(w, y_1, \ldots, y_m) = \min_{w\in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w).$$

Moreover, if

$$u_\alpha = (w_\alpha, y_{1,\alpha}, \ldots, y_{m,\alpha}) \in U^{m+1}$$

satisfies

$$\Phi_{Z,U}(u_\alpha) - \min_{u\in U^{m+1}} \Phi_{Z,U}(u) \le \alpha,$$

then

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w_\alpha) - \min_{w\in U} \widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w) \le \alpha.$$

Proof. Convexity is immediate since each $f(\cdot, z_i)$ is convex and $(w, y_i) \mapsto \|w - y_i\|$ is jointly convex. For fixed $w \in U$, the variables $y_1, \ldots, y_m$ separate, and

$$\min_{y_i \in U} \big[f(y_i, z_i) + C\|w - y_i\| \big] = f_{C,U}(w, z_i).$$

Substituting this identity into (8) proves the equality of optima. The final claim follows from

$$\widehat{F}^{(w_0)}_{C,U,\lambda,Z}(w_\alpha) = \min_{y_1,\ldots,y_m \in U} \Phi_{Z,U}(w_\alpha, y_1, \ldots, y_m) \le \Phi_{Z,U}(u_\alpha). \qquad \square$$

Proposition D.5 (Precise version of Proposition 2.8)**.** *Let $K \subseteq \mathbb{R}^q$ be a nonempty compact convex set equipped with an ξ-inexact projection oracle for every $\xi > 0$, and suppose $\operatorname{diam}(K) \le D$. Let $\Phi : K \to \mathbb{R}$ be convex and L-Lipschitz on K for some finite but unknown $L > 0$. Suppose we are given an exact subgradient oracle: for every $x \in K$, the oracle returns a vector $g(x) \in \mathbb{R}^q$ satisfying*

$$\Phi(y) \ge \Phi(x) + \langle g(x), y - x \rangle \qquad \textit{for all } y \in K.$$

Fix $\alpha > 0$. Then Algorithm 4 halts after at most

$$T_\alpha := \left\lceil \left(\frac{3DL}{\alpha} \right)^2 \right\rceil$$

iterations, and its output $u_\alpha \in K$ satisfies

$$\Phi(u_\alpha) - \min_{u \in K} \Phi(u) \le \alpha.$$

In particular, the algorithm uses at most T_α calls to the subgradient oracle and at most T_α calls to the inexact projection oracle. Hence, if each subgradient query and each ξ-inexact projection can be performed in polynomial time, then the total running time is polynomial in q, D, L, and $1/\alpha$.

Proof. Let $x^\star \in \arg\min_{x \in K} \Phi(x)$, which exists since K is compact and Φ is continuous.

If the algorithm encounters $g_t = 0$, then x_t is an exact minimizer and the claim is immediate. Hence assume $g_t \neq 0$ before termination.

Fix $t \ge 1$, and let

$$y_t := x_t - \eta_t g_t, \qquad p_t := \Pi_K(y_t), \qquad e_t := 2D\xi_t + \xi_t^2.$$

Since $\|x_{t+1} - p_t\| \le \xi_t$ and $\operatorname{diam}(K) \le D$,

$$\|x_{t+1} - x^\star\|^2 \le \|p_t - x^\star\|^2 + 2D\xi_t + \xi_t^2 = \|p_t - x^\star\|^2 + e_t.$$

Projection is nonexpansive, so

$$\|p_t - x^\star\| \le \|y_t - x^\star\|.$$

Therefore

$$\|x_{t+1} - x^\star\|^2 \le \|x_t - x^\star\|^2 - 2\eta_t \langle g_t, x_t - x^\star \rangle + \eta_t^2 \|g_t\|^2 + e_t.$$

By the subgradient inequality,

$$\Phi(x_t) - \Phi(x^\star) \le \langle g_t, x_t - x^\star \rangle.$$

Hence

$$2\eta_t(\Phi(x_t) - \Phi(x^\star)) \le \|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2 + \eta_t^2 \|g_t\|^2 + e_t.$$

Summing from $t = 1$ to T and using the standard AdaGrad telescoping argument yields

$$\sum_{t=1}^{T} (\Phi(x_t) - \Phi(x^\star)) \le \frac{D^2}{2\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \|g_t\|^2 + \frac{1}{2} \sum_{t=1}^{T} \frac{e_t}{\eta_t}.$$

Since $\eta_T = D/\sqrt{S_T}$,

$$\frac{D^2}{2\eta_T} = \frac{D}{2} \sqrt{S_T}.$$

Also,

$$\sum_{t=1}^{T} \eta_t \|g_t\|^2 = D \sum_{t=1}^{T} \frac{\|g_t\|^2}{\sqrt{S_t}} \le 2D\sqrt{S_T}.$$

Finally, because $\xi_t = \min\{D, \alpha\eta_t/(6D)\}$, we have

$$e_t = 2D\xi_t + \xi_t^2 \le 3D\xi_t \le \frac{\alpha\eta_t}{2},$$

so

$$\frac{1}{2}\sum_{t=1}^{T} \frac{e_t}{\eta_t} \le \frac{\alpha T}{4}.$$

Combining the bounds,

$$\sum_{t=1}^{T} (\Phi(x_t) - \Phi(x^\star)) \le \frac{3D}{2}\sqrt{S_T} + \frac{\alpha T}{4}.$$

By convexity,

$$\Phi(\overline{x}_T) - \Phi(x^\star) \le \frac{1}{T}\sum_{t=1}^{T} (\Phi(x_t) - \Phi(x^\star)) \le \frac{3D}{2T}\sqrt{S_T} + \frac{\alpha}{4}.$$

Thus whenever $3D\sqrt{S_T} \le \alpha T$, the returned point $\overline{x}_T$ satisfies

$$\Phi(\overline{x}_T) - \Phi(x^\star) < \alpha.$$

To show finite termination, since $\|g_t\| \le L$,

$$S_T \le TL^2.$$

Hence $3D\sqrt{S_T} \le 3DL\sqrt{T}$, so the stopping condition is guaranteed once

$$3DL\sqrt{T} \le \alpha T,$$

i.e. once

$$T \ge \left(\frac{3DL}{\alpha}\right)^2.$$

This proves the claim. □

Lemma D.6 (Precise version of Lemma 2.7)**.** *Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a nonempty compact convex set, let $c \in \mathcal{W}$, and let $r > 0$. Define*

$$U := \mathcal{W} \cap B(c, r) = \{w \in \mathcal{W} : \|w - c\|_2 \le r\}.$$

Assume that Euclidean projection onto $\mathcal{W}$ can be computed exactly in time $\mathrm{T}_{\mathrm{proj}}(d)$. Then, for every $y \in \mathbb{R}^d$ and every $\xi > 0$, one can compute a point $\widetilde{\Pi}_U^\xi(y) \in U$ satisfying

$$\left\|\widetilde{\Pi}_U^\xi(y) - \Pi_U(y)\right\|_2 \le \xi$$

using

$$O\left(\log\left(1 + \frac{\|y - c\|_2^2}{r\xi}\right)\right)$$

calls to the projection oracle for $\mathcal{W}$, plus polynomial-time arithmetic in d. In particular, if projection onto $\mathcal{W}$ is polynomial-time computable, then so is ξ-inexact projection onto U.

Proof. Fix $y \in \mathbb{R}^d$ and $\xi > 0$, and let

$$x^\star := \Pi_U(y).$$

Since U is nonempty, compact, and convex, $x^\star$ is well-defined and unique.

If $y = c$, then $c \in U$ and $x^\star = c$, so we can return c. Hence assume $R := \|y - c\|_2 > 0$.

Step 1: Multiplier representation. If $\Pi_{\mathcal{W}}(y) \in U$, then $\Pi_U(y) = \Pi_{\mathcal{W}}(y)$, so we are done. Thus assume

$$\|\Pi_{\mathcal{W}}(y) - c\|_2 > r.$$

Consider the projection problem

$$\min_{x \in \mathcal{W}} \frac{1}{2}\|x - y\|_2^2 \qquad \text{s.t.} \qquad \frac{1}{2}\big(\|x - c\|_2^2 - r^2\big) \le 0.$$

The feasible set is U. Since $c \in \mathcal{W}$ and $r > 0$, the set $\mathcal{W} \cap B(c, r)$ has nonempty relative interior in $\mathrm{aff}(\mathcal{W})$: indeed, because $\mathrm{ri}(\mathcal{W})$ is dense in $\mathcal{W}$, there exists $u \in \mathrm{ri}(\mathcal{W})$ with $\|u - c\|_2 < r$. Therefore Slater's condition holds relative to $\mathrm{aff}(\mathcal{W})$, and the KKT conditions are necessary and sufficient.

Hence there exists a multiplier $\lambda^\star \ge 0$ such that

$$0 \in x^\star - y + N_{\mathcal{W}}(x^\star) + \lambda^\star(x^\star - c),$$

with complementary slackness

$$\lambda^\star\big(\|x^\star - c\|_2^2 - r^2\big) = 0.$$

Since $\Pi_{\mathcal{W}}(y) \notin U$, the ball constraint is active at $x^\star$, so $\lambda^\star > 0$ and $\|x^\star - c\|_2 = r$.

The stationarity condition is exactly the optimality condition for minimizing over $\mathcal{W}$ the strongly convex function

$$x \mapsto \frac{1}{2}\|x - y\|_2^2 + \frac{\lambda^\star}{2}\|x - c\|_2^2.$$

Completing the square gives

$$x^\star = \Pi_{\mathcal{W}}\left(\frac{y + \lambda^\star c}{1 + \lambda^\star}\right).$$

For $\lambda \ge 0$, define

$$z_\lambda := \frac{y + \lambda c}{1 + \lambda}, \qquad x_\lambda := \Pi_{\mathcal{W}}(z_\lambda), \qquad g(\lambda) := \|x_\lambda - c\|_2^2 - r^2.$$

The preceding argument shows that $x^\star = x_{\lambda^\star}$ for some $\lambda^\star > 0$ satisfying $g(\lambda^\star) = 0$.

Step 2: Monotonicity and bracketing. We claim that g is continuous and nonincreasing. Continuity follows from continuity of z_λ and nonexpansiveness of projection onto $\mathcal{W}$. For monotonicity, define

$$\psi(\lambda) := \min_{x \in \mathcal{W}} \left\{ \frac{1}{2}\|x - y\|_2^2 + \frac{\lambda}{2}\big(\|x - c\|_2^2 - r^2\big) \right\}.$$

This is the pointwise infimum of affine functions of λ, hence ψ is concave. By Danskin's theorem,

$$\psi'(\lambda) = \frac{1}{2}\big(\|x_\lambda - c\|_2^2 - r^2\big) = \frac{1}{2}g(\lambda)$$

at every point of differentiability, and the same conclusion holds in the subgradient sense. Since the derivative/subgradient of a concave function is nonincreasing, g is nonincreasing.

By assumption,

$$g(0) = \|\Pi_{\mathcal{W}}(y) - c\|_2^2 - r^2 > 0.$$

Let

$$\overline{\lambda} := \frac{R}{r}.$$

Because $c \in \mathcal{W}$, nonexpansiveness of projection gives

$$\|x_{\overline{\lambda}} - c\|_2 = \|\Pi_{\mathcal{W}}(z_{\overline{\lambda}}) - \Pi_{\mathcal{W}}(c)\|_2 \le \|z_{\overline{\lambda}} - c\|_2 = \frac{R}{1 + \overline{\lambda}} = \frac{Rr}{R + r} < r.$$

Thus $g(\overline{\lambda}) < 0$. Hence $\lambda^\star \in (0, \overline{\lambda})$.

Algorithm 5: EFFICIENT-RECURSIVE-LOCALIZED-OUTPUTPERT$(Z, \varepsilon, \lambda, w_0)$

Input: Dataset $Z = (z_1, \dots, z_m)$, privacy parameter ε, regularization parameter λ, center $w_0 \in \mathcal{W}$

Output: A private point $w_{\mathrm{DP}} \in \mathcal{W}$

Run Algorithm 3 on $(Z, \varepsilon, \lambda, w_0)$, implementing each call to LOCALIZED-OUTPUTPERT as follows: for the current localized set U_j, minimize the localized joint convex objective Φ_{Z,U_j} over U_j^{m+1} to the required deterministic accuracy using Algorithm 4 with diameter bound $\sqrt{m+1}R_j$ and the inexact projection oracle from Lemma 2.7

Let $w_{\mathrm{loc},S}$ be the center returned by Algorithm 3

return $w_{\mathrm{DP}} := w_{\mathrm{loc},S}$

Step 3: Bisection. Perform bisection on $[0, \overline{\lambda}]$, using the sign of $g(\lambda)$. After N iterations, we obtain

$$0 \le \lambda_- \le \lambda^\star \le \lambda_+ \le \overline{\lambda}$$

such that

$$g(\lambda_-) \ge 0, \qquad g(\lambda_+) \le 0, \qquad \lambda_+ - \lambda_- \le \frac{\overline{\lambda}}{2^N}.$$

Since $g(\lambda_+) \le 0$, the point x_{λ_+} belongs to U. We return

$$\widetilde{\Pi}_U^\xi(y) := x_{\lambda_+}.$$

Step 4: Error bound. Projection onto $\mathcal{W}$ is nonexpansive, so for any $\lambda, \mu \ge 0$,

$$\|x_\lambda - x_\mu\|_2 \le \|z_\lambda - z_\mu\|_2 = \frac{|\lambda - \mu|}{(1+\lambda)(1+\mu)}\|y - c\|_2 \le R|\lambda - \mu|.$$

Therefore

$$\|\widetilde{\Pi}_U^\xi(y) - x^\star\|_2 = \|x_{\lambda_+} - x_{\lambda^\star}\|_2 \le R(\lambda_+ - \lambda_-) \le R\frac{\overline{\lambda}}{2^N} = \frac{R^2}{r\,2^N}.$$

Thus it suffices to take

$$N = \left\lceil \log_2\left(1 + \frac{\|y - c\|_2^2}{r\xi}\right)\right\rceil.$$

This proves the claimed oracle complexity and runtime. □

E Deferred Proofs for Section 3

For completeness, we provide formal pseudocode of our efficient ERM solver in Algorithm 5.

E.1 Proof of Proposition 3.2

Proposition E.1 (Re-statement of Proposition 3.2)**.** *Fix $m \in \mathbb{N}$, $w_0 \in \mathcal{W}$, $\lambda > 0$, and $\rho \in (0, 1/5)$. Then, Algorithm 5 is ε-differentially private. If $Z \sim P^m$, then its output w_{DP} satisfies*

$$\Pr\left(\|w_{\mathrm{DP}} - \widehat{w}_\lambda(Z; w_0)\| \le c_{\mathrm{erm}}\frac{1}{\lambda}\left(G_k\left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}} + \frac{G_2}{\sqrt{m}}\right)\right) \ge 0.7.$$

Moreover, with probability at least $1 - \rho$, the runtime is bounded by a polynomial in

$$m,\ d,\ D,\ \lambda,\ \frac{1}{\lambda},\ \frac{1}{\varepsilon},\ G_k,\ \frac{1}{\rho}.$$

Further, if $\sup_{z\in\mathcal{Z}} \max_{w\in\mathcal{W}} \|\nabla f(w, z)\|$ is finite and polynomially bounded in the problem parameters, then the runtime is polynomial with probability 1.

Proof. Privacy and the distance guarantee follow directly from Proposition 2.4, since Algorithm 5 outputs the final center $w_{\mathrm{loc},S}$ of Algorithm 3. In particular, with probability at least 0.7,

$$\|w_{\mathrm{DP}} - \widehat{w}_\lambda(Z; w_0)\| \le c_2 \frac{G_k}{\lambda} \left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}}.$$

This implies the displayed bound after increasing c_{erm}, since the additional $G_2/\sqrt{m}$ term is nonnegative.

It remains to justify the runtime. Each call to Algorithm 2 requires computing an α-approximate minimizer of

$$\widehat{F}^{(w_0)}_{C_j,U_j,\lambda,Z}(w)$$

over the current localized set U_j. By Lemma 2.6, this is achieved by minimizing the jointly convex objective Φ_{Z,U_j} over U_j^{m+1}. Since $U_j = \mathcal{W} \cap \mathbb{B}(w_{\mathrm{loc},j}, r_j)$, Lemma 2.7 provides a polynomial-time inexact projection oracle for U_j, and hence for the product set U_j^{m+1}.

Let

$$A_Z := \max_{1\le i\le m} \sup_{w\in\mathcal{W}} \|\nabla f(w, z_i)\|.$$

By the k-th moment assumption, $A_Z < \infty$ with probability 1. For any round j, every subgradient of the localized joint objective

$$\Phi_{Z,U_j}(w, y_1, \dots, y_m) = \frac{1}{m}\sum_{i=1}^m \big[f(y_i, z_i) + C_j\|w - y_i\|\big] + \frac{\lambda}{2}\|w - w_0\|^2$$

has norm at most an absolute constant times

$$L_{Z,j} := A_Z + C_j + \lambda D.$$

Indeed, a subgradient has components

$$g^{(w)} = \frac{C_j}{m}\sum_{i=1}^m s_i + \lambda(w - w_0), \qquad g^{(y_i)} = \frac{1}{m}(g_i - C_j s_i),$$

where $s_i \in \partial\|w - y_i\|$, $\|s_i\| \le 1$, and $g_i \in \partial f(y_i, z_i)$, $\|g_i\| \le A_Z$. Thus

$$\|g^{(w)}\| \le C_j + \lambda D, \qquad \left(\sum_{i=1}^m \|g^{(y_i)}\|^2\right)^{1/2} \le A_Z + C_j,$$

up to absolute constants.

The product domain U_j^{m+1} has diameter at most

$$\sqrt{m+1}\,\operatorname{diam}(U_j) \le \sqrt{m+1}\,D.$$

Therefore, by Proposition 2.8, each adaptive projected-subgradient invocation terminates in finite time with probability 1.

For the high-probability polynomial runtime bound, Markov's inequality and a union bound give

$$\Pr\left(A_Z \le G_k\left(\frac{m}{\rho}\right)^{1/k}\right) \ge 1 - \rho.$$

On this event, all realized Lipschitz parameters $L_{Z,j}$ are polynomially bounded in

$$m,\ d,\ D,\ \lambda,\ \frac{1}{\lambda},\ \frac{1}{\varepsilon},\ G_k,\ \frac{1}{\rho},$$

because the parameters $S, R_j, C_j, \varepsilon_j, \beta_j$ in Algorithm 3 are explicit functions of these quantities and $S = O(\log\log(eD/R_\star))$. Hence the iteration complexity from Proposition 2.8, together with the projection oracle from Lemma 2.7, gives the claimed polynomial runtime bound with probability at least $1 - \rho$.

Finally, if

$$L_* := \sup_{z\in\mathcal{Z}} \max_{w\in\mathcal{W}} \|\nabla f(w, z)\|$$

is finite and polynomially bounded in the problem parameters, then $A_Z \le L_*$ deterministically. The same runtime bound therefore holds with probability 1. □

E.2 Explicit oracle complexity

We bound the number of sample-subgradient evaluations and calls to $\Pi_{\mathcal{W}}$ in Algorithm 5. Throughout, $k \geq 2$ is fixed.

One regularized ERM call. Let $Z = (z_1, \ldots, z_m)$ and $A_Z = \max_{i \leq m} \sup_{w \in \mathcal{W}} \|\nabla f(w, z_i)\|$. In round j, the product domain has diameter at most $\sqrt{m+1}R_j$ and the joint objective has subgradient norm at most $c(A_Z + C_j + \lambda D)$. Proposition 2.8 therefore bounds the number of iterations by

$$\begin{aligned} T_j &\leq 1 + c\frac{(m+1)R_j^2}{\alpha_j^2}\left(A_Z^2 + C_j^2 + \lambda^2 D^2\right), \\ \alpha_j &= \frac{C_j^2}{100\lambda m^2}\min\left\{1, \frac{d^2\ell_j^2}{\varepsilon_j^2}\right\}. \end{aligned} \tag{10}$$

Each iteration evaluates m sample subgradients and projects $m+1$ coordinates onto U_j. For product-space accuracy ξ, each coordinate projection needs accuracy $\xi/\sqrt{m+1}$. By Lemma D.6, these projections use $\widetilde{O}(m)$ calls to $\Pi_{\mathcal{W}}$ per iteration. Including the projection in the output-perturbation step, each oracle count is

$$\widetilde{O}\left(mS + m^2\sum_{j=0}^{S-1}\frac{R_j^2}{\alpha_j^2}\left[A_Z^2 + C_j^2 + \lambda^2 D^2\right]\right). \tag{11}$$

Markov's inequality and a union bound give $\Pr(A_Z \leq G_k(m/\rho)^{1/k}) \geq 1 - \rho$, yielding a high-probability bound by substitution in (11).

The full SCO algorithm. Phase t makes J ERM calls, each with m_t samples and regularization λ_t. Add a subscript t to the inner parameters, so that, for example, $R_{t,j}$ is the radius bound in round j of phase t. A union bound over all n samples gives

$$\Pr\left(\max_{i \leq n}\sup_{w \in \mathcal{W}}\|\nabla f(w, z_i)\| \leq G_k(n/\rho)^{1/k}\right) \geq 1 - \rho.$$

On this event, summing (11) over the ERM calls bounds each oracle count by

$$\widetilde{O}\left(J\sum_{t=1}^{T}\left\{m_t S_t + m_t^2\sum_{j=0}^{S_t-1}\frac{R_{t,j}^2}{\alpha_{t,j}^2}\left[G_k^2(n/\rho)^{2/k} + C_{t,j}^2 + \lambda_t^2 D^2\right]\right\}\right). \tag{12}$$

Geometric aggregation makes no calls to either oracle.

Assume $0 < \varepsilon \leq 1$ and $G_k, D > 0$. Then the minimum in (10) equals 1. Phases with $R_{\star,t} \geq D$ make no optimization calls. In every other phase, the radius recurrence gives

$$R_{\star,t} = \frac{G_k}{\lambda_t}\left(\frac{d}{m_t\varepsilon}\right)^{1-1/k} \leq R_{t,j} \leq D.$$

Substituting this lower bound and $\varepsilon_{t,j} = (\varepsilon/2)\theta_{t,j}$, $\beta_{t,j} = \beta_0\theta_{t,j}$ into the formula for $C_{t,j}$ gives

$$\frac{C_{t,j}}{G_k} \gtrsim_k \left(\frac{m_t\varepsilon}{d}\right)^{1/k}\theta_{t,j}^{1/(k+1)}\ell_{t,j}^{-2/(k+1)}.$$

The schedules satisfy

$$\max_{j < S_t}\theta_{t,j}^{-1} = O_k\left(\log\frac{eD}{R_{\star,t}}\right), \qquad S_t = O_k\left(1 + \log\log\frac{eD}{R_{\star,t}}\right).$$

Absorbing these logarithmic factors and the projection costs into $\widetilde{O}$, and substituting $\alpha_{t,j}$ and $R_{t,j} \leq D$ in (12), we obtain

$$\begin{aligned} \widetilde{O}\Bigg(J\sum_{t=1}^{T}\Bigg[m_t + m_t^6\left(\frac{\lambda_t D}{G_k}\right)^2\Bigg\{\left(\frac{d}{m_t\varepsilon}\right)^{2/k} \\ + \left(\frac{d}{m_t\varepsilon}\right)^{4/k}\left[\left(\frac{n}{\rho}\right)^{2/k} + \left(\frac{\lambda_t D}{G_k}\right)^2\right]\Bigg\}\Bigg]\Bigg). \end{aligned} \tag{13}$$

Since $\lambda_t = 32^{t-1}\lambda_1$ and $1 \le m_t \le n/2^t$,

$$\frac{\lambda_t D}{G_k} \le \left(\frac{n}{m_t}\right)^5 \frac{\lambda_1 D}{G_k}.$$

Using this inequality in (13) and absorbing TJ into $\widetilde{O}$ gives

$$\begin{aligned}\widetilde{O}\Bigg(n + n^{10}\left(\frac{\lambda_1 D}{G_k}\right)^2 \left[\left(\frac{d}{\varepsilon}\right)^{2/k} + \left(\frac{n}{\rho}\right)^{2/k}\left(\frac{d}{\varepsilon}\right)^{4/k}\right] \\ + n^{20}\left(\frac{\lambda_1 D}{G_k}\right)^4 \left(\frac{d}{\varepsilon}\right)^{4/k}\Bigg).\end{aligned} \tag{14}$$

This is at most

$$\widetilde{O}\left(n^{20}\left(1 + \frac{\lambda_1 D}{G_k}\right)^4 \rho^{-2/k}\left(\frac{d}{\varepsilon}\right)^{4/k}\right).$$

Applying this bound with the selected r, G_r in place of k, G_k proves (9). The fixed-output branch uses no oracle calls.

The factor m_t^6 comes from the squared product-domain diameter (m_t), the cost of an iteration (m_t), and the factor m_t^4 in $\alpha_{t,j}^{-2}$. The n^{20} term comes from bounding the λ_t^4 contribution using $\lambda_t/\lambda_1 \le (n/m_t)^5$.

E.3 Proof of Theorem 3.1

Theorem E.2 (Re-statement of Theorem 3.1)**.** *Let $\delta, \rho \in (0, 1/5)$ and $\varepsilon \in (0, 1]$.* **POP-RECURSIVE-LOCALIZED-OUTPUTPERT***, with the public parameter selection in Section 3, is ε-differentially private and with probability at least $1 - \delta$ its output satisfies*

$$\begin{aligned}F(\widehat{w}) - F^* \le cD \min\Bigg\{G_2,\ G_2\sqrt{\frac{\log(1/\delta)}{n}} \\ + \min\left\{G_2\sqrt{\frac{d\log(1/\delta)}{n\varepsilon}},\ G_k\left(\frac{d\log(1/\delta)}{n\varepsilon}\right)^{1-1/k}\right\}\Bigg\}.\end{aligned}$$

Moreover, its runtime is bounded with probability at least $1 - \rho$ by a polynomial in the problem parameters. If $\sup_{z\in\mathcal{Z}} \max_{w\in\mathcal{W}} \|\nabla f(w, z)\|$ is finite and polynomially bounded in these parameters, its runtime is polynomial with probability 1.

Proof. Instantiate line 10 of Algorithm 1 with $\mathcal{A}_{\mathrm{ERM}}$ = EFFICIENT-RECURSIVE-LOCALIZED-OUTPUTPERT from Proposition 3.2. By Proposition 3.2, this phasewise primitive is ε-DP and satisfies

$$\Pr\left(\|\mathcal{A}_{\mathrm{ERM}}(Z, \varepsilon, \lambda, w_0) - \widehat{w}_\lambda(Z; w_0)\| \le c_{\mathrm{erm}}\frac{1}{\lambda}\left(G_k\left(\frac{d}{m\varepsilon}\right)^{1-\frac{1}{k}} + \frac{G_2}{\sqrt{m}}\right)\right) \ge 0.7.$$

Applying Theorem 2.1 yields an ε-DP algorithm whose output $\widehat{w}$ satisfies

$$F(\widehat{w}) - F^* \le c_3 G_k D\left(\frac{d\log(1/\delta)}{n\varepsilon}\right)^{1-\frac{1}{k}} + c_4 G_2 D\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$. Applying the same argument with moment order 2 gives

$$c_3 D G_2\sqrt{\frac{d\log(1/\delta)}{n\varepsilon}} + c_4 D G_2\sqrt{\frac{\log(1/\delta)}{n}}.$$

For a fixed $w_0 \in \mathcal{W}$, convexity and Cauchy–Schwarz give

$$F(w_0) - F^* \le D\mathbb{E}\left[\sup_{w\in\mathcal{W}} \|\nabla f(w, z)\|\right] \le DG_2.$$

Both moment bounds hold on $\mathcal{P}(G_2, G_k)$, so the public selection in Section 3 gives the minimum of these three bounds, up to an absolute constant. Only the selected algorithm is run, with privacy budget ε and failure probability δ.

For runtime, Algorithm 1 performs

$$T \lesssim \log(n)$$

phases and

$$J \lesssim \log(\log n/\delta)$$

repetitions per phase. By Proposition 3.2, each regularized ERM call terminates in finite time with probability 1, so the full algorithm also terminates with probability 1.

Now fix any $\rho \in (0, 1/5)$. Allocate runtime failure probability

$$\rho_{t,j} := \frac{\rho}{2TJ}$$

to each ERM call in phase t, repetition j. By Proposition 3.2, with probability at least $1 - \rho_{t,j}$, the runtime of that call is bounded by a polynomial in

$$m_t,\ d,\ D,\ \lambda_t,\ \frac{1}{\varepsilon},\ G_k,\ \frac{1}{\rho_{t,j}}.$$

A union bound over all TJ calls shows that with probability at least $1 - \rho/2$, all phasewise ERM calls satisfy their polynomial runtime bounds simultaneously. Since $m_t \le n$, $T = O(\log n)$, $J = O(\log(\log n/\delta))$, and the regularization schedule λ_t is explicit and geometric, the total runtime is bounded with probability at least $1 - \rho$ by a polynomial in

$$n,\ d,\ D,\ \frac{1}{\varepsilon},\ G_k,\ \log\frac{n}{\delta},\ \frac{1}{\rho}.$$

Here we use that, in the instantiation of Theorem 2.1, the base regularization parameter λ_1 is chosen explicitly as a polynomial function of the problem parameters, so the dependence on $1/\lambda_t$ is polynomially bounded.

The order-2 algorithm has the same runtime guarantee with G_2 in place of G_k.

Finally, if $\sup_{z\in\mathcal{Z}} \max_{w\in\mathcal{W}} \|\nabla f(w, z)\|$ is polynomially bounded, each ERM call runs in deterministic polynomial time by Proposition 3.2. Since there are only finitely many such calls, the entire algorithm runs in deterministic polynomial time. □

E.4 Deterministic polynomial time under deterministic approximate localized-extension oracles

The main algorithm of Theorem 3.1 is polynomial time with high probability in the general heavy-tailed model, because the realized Lipschitz constant of the joint reformulation is finite with probability 1 and polynomially bounded with high probability. In this subsection we isolate a structural condition under which the same statistical guarantee can be obtained in deterministic polynomial time, even when the worst-case Lipschitz parameter

$$\sup_{z\in\mathcal{Z}} \sup_{w\in\mathcal{W}} \|\nabla f(w, z)\|$$

is infinite.

The key point is that the joint reformulation is only needed because we do not know how to access the Lipschitz extension directly. If, for the relevant localized sets $U = \mathcal{W} \cap \mathbb{B}(c, r)$, we can compute deterministic approximate subgradients of the localized Lipschitz extension $f_{C,U}$, then each optimization subroutine in Algorithm 2 can be implemented directly by projected subgradient descent on the localized Lipschitz-extension objective. Since $f_{C,U}$ is C-Lipschitz by construction, the resulting iteration complexity is deterministically polynomial.

Definition E.3 (Deterministic approximate subgradient oracle for localized Lipschitz extensions)**.** Fix $C > 0$. We say that the loss class admits a deterministic polynomial-time approximate subgradient oracle for localized C-Lipschitz extensions if the following holds.

For every $z \in \mathcal{Z}$, every set

$$U = \mathcal{W} \cap \mathbb{B}(c, r) \qquad (c \in \mathcal{W},\ r > 0),$$

every $w \in U$, and every accuracy parameter $B > 0$, there is an algorithm running in time polynomial in the problem parameters and $1/B$ that returns a vector

$$\widetilde{g}_{C,U,B}(w,z) \in \mathbb{R}^d$$

satisfying

$$f_{C,U}(u,z) \geq f_{C,U}(w,z) + \langle \widetilde{g}_{C,U,B}(w,z), u-w\rangle - B \qquad \forall u \in U,$$

and

$$\|\widetilde{g}_{C,U,B}(w,z)\|_2 \leq 2C.$$

Lemma E.4 (Projected subgradient method with biased subgradients and inexact projections)**.** *Let $K \subseteq \mathbb{R}^q$ be a nonempty compact convex set with $\operatorname{diam}(K) \leq D_K$, and let $\Phi : K \to \mathbb{R}$ be convex. Suppose that at each query point $x_t \in K$, an oracle returns $\tilde{g}_t \in \mathbb{R}^q$ satisfying*

$$\Phi(u) \geq \Phi(x_t) + \langle \tilde{g}_t, u - x_t\rangle - B \qquad \forall u \in K,$$

and $\|\tilde{g}_t\|_2 \leq L$. Consider iterates satisfying

$$\|x_{t+1} - \Pi_K(x_t - \eta\tilde{g}_t)\|_2 \leq \xi, \qquad t = 1,\dots,T,$$

and let $\overline{x}_T := T^{-1}\sum_{t=1}^T x_t$. Then

$$\Phi(\overline{x}_T) - \min_{x\in K}\Phi(x) \leq \frac{D_K^2}{2\eta T} + \frac{\eta L^2}{2} + B + \frac{D_K\xi}{\eta} + \frac{\xi^2}{2\eta}.$$

In particular, if

$$\eta = \frac{D_K}{L\sqrt{T}}, \qquad T \geq \left(\frac{4D_KL}{\alpha}\right)^2, \qquad B \leq \frac{\alpha}{4}, \qquad \xi \leq \min\left\{D_K, \frac{\alpha\eta}{6D_K}\right\},$$

then $\Phi(\overline{x}_T) - \min_K \Phi \leq \alpha$.

Proof. Fix $x^\star \in \arg\min_{x\in K}\Phi(x)$, and let

$$p_t := \Pi_K(x_t - \eta\tilde{g}_t).$$

Since $x_{t+1}, p_t, x^\star \in K$ and $\operatorname{diam}(K) \leq D_K$,

$$\|x_{t+1} - x^\star\|^2 \leq \|p_t - x^\star\|^2 + 2D_K\xi + \xi^2.$$

By nonexpansiveness of projection,

$$\|p_t - x^\star\|^2 \leq \|x_t - \eta\tilde{g}_t - x^\star\|^2 = \|x_t - x^\star\|^2 - 2\eta\langle\tilde{g}_t, x_t - x^\star\rangle + \eta^2\|\tilde{g}_t\|^2.$$

Combining and rearranging gives

$$\langle\tilde{g}_t, x_t - x^\star\rangle \leq \frac{\|x_t - x^\star\|^2 - \|x_{t+1} - x^\star\|^2}{2\eta} + \frac{\eta}{2}\|\tilde{g}_t\|^2 + \frac{D_K\xi}{\eta} + \frac{\xi^2}{2\eta}.$$

Using the biased subgradient inequality with $u = x^\star$,

$$\Phi(x_t) - \Phi(x^\star) \leq \langle\tilde{g}_t, x_t - x^\star\rangle + B.$$

Summing over $t = 1,\dots,T$, dividing by T, using $\|x_1 - x^\star\| \leq D_K$, $\|\tilde{g}_t\| \leq L$, and convexity of Φ, we obtain

$$\Phi(\overline{x}_T) - \Phi(x^\star) \leq \frac{D_K^2}{2\eta T} + \frac{\eta L^2}{2} + B + \frac{D_K\xi}{\eta} + \frac{\xi^2}{2\eta}.$$

The final claim follows by substituting the stated choices. □

Theorem E.5 (Deterministic polynomial time from localized-extension oracles)**.** *Grant Assumption 1.1. Suppose, in addition, that the loss class admits a deterministic polynomial-time approximate subgradient oracle for localized C-Lipschitz extensions in the sense of Definition E.3. Then there exists a pure ε-DP algorithm such that, for every $\delta \in (0, 1/5)$, its output $\widehat{w}$ satisfies*

$$F(\widehat{w}) - F^* \leq cD\min\Bigg\{G_2,\, G_2\sqrt{\frac{\log(1/\delta)}{n}} + \min\left\{G_2\sqrt{\frac{d\log(1/\delta)}{n\varepsilon}},\, G_k\left(\frac{d\log(1/\delta)}{n\varepsilon}\right)^{1-1/k}\right\}\Bigg\},$$

with probability at least $1-\delta$. Moreover, the algorithm runs in deterministic polynomial time.

Proof. We use the same outer population-localization framework as in Theorem 3.1. It suffices to implement the phasewise regularized ERM primitive Algorithm 5 in deterministic polynomial time.

Fix one regularized ERM instance $(Z, \varepsilon, \lambda, w_0)$ of sample size m. The recursive localization proof of Proposition 2.4 is unchanged. The only issue is how to implement the approximate minimization step inside each call to Algorithm 2.

At round j, the current set has the form

$$U_j = \mathcal{W} \cap \mathbb{B}(c_j, r_j),$$

and the algorithm needs a deterministic α_j-approximate minimizer of

$$\widehat{F}^{(w_0)}_{C_j,U_j,\lambda,Z}(w) = \frac{1}{m}\sum_{i=1}^{m} f_{C_j,U_j}(w, z_i) + \frac{\lambda}{2}\|w - w_0\|^2$$

over U_j, where

$$\alpha_j = \frac{C_j^2}{100\lambda m^2} \min\left\{1, \frac{d^2\ell_j^2}{\varepsilon_j^2}\right\}.$$

By Definition E.3, for each z_i and $w \in U_j$ we can compute a B-approximate subgradient $\widetilde{g}_i(w)$ of $f_{C_j,U_j}(\cdot, z_i)$ over U_j, with $\|\widetilde{g}_i(w)\| \le 2C_j$. Therefore

$$\widetilde{g}(w) := \frac{1}{m}\sum_{i=1}^{m} \widetilde{g}_i(w) + \lambda(w - w_0)$$

is an B-approximate subgradient of $\widehat{F}^{(w_0)}_{C_j,U_j,\lambda,Z}$, and

$$\|\widetilde{g}(w)\| \le 2C_j + \lambda D.$$

Thus the objective is accessible through deterministically bounded biased subgradients.

Projection onto $U_j = \mathcal{W} \cap \mathbb{B}(c_j, r_j)$ is implemented to arbitrary accuracy using Lemma 2.7. Applying Lemma E.4 with

$$L = 2C_j + \lambda D, \qquad B \le \alpha_j/4,$$

and with projection accuracy chosen as in the lemma, returns a point $\tilde{w}_j \in U_j$ satisfying

$$\widehat{F}^{(w_0)}_{C_j,U_j,\lambda,Z}(\tilde{w}_j) - \min_{w \in U_j} \widehat{F}^{(w_0)}_{C_j,U_j,\lambda,Z}(w) \le \alpha_j$$

in deterministic polynomial time. The iteration complexity is polynomial because $C_j, \lambda, D, 1/\alpha_j$, and the projection accuracies are explicit polynomially bounded quantities in the phasewise parameters.

Thus every approximate minimization step in Algorithm 3 can be implemented deterministically in polynomial time. The privacy and distance guarantee of Proposition 2.4 are unchanged, because the optimization accuracy is deterministic. Hence the phasewise ERM primitive satisfies

$$\Pr\left(\|w_{\mathrm{DP}} - \widehat{w}_\lambda(Z; w_0)\| \le c_{\mathrm{erm}}\frac{1}{\lambda}\left[G_k\left(\frac{d}{m\varepsilon}\right)^{1-1/k} + \frac{G_2}{\sqrt{m}}\right]\right) \ge 0.7$$

and runs in deterministic polynomial time.

Plugging this primitive into Algorithm 1 and applying Theorem 2.1 gives the order-k bound. Using order 2 and the public selection in Section 3 gives the stated minimum. Since the outer algorithm makes only finitely many phasewise calls with polylogarithmic overhead, the full runtime is deterministic polynomial time. □

Proposition E.6 (Polyhedral losses on compact explicitly SOCP-representable localized domains)**.** *Assume that $\mathcal{W} \subseteq \mathbb{R}^d$ admits an explicit compact second-order-cone representation with a strict interior point. Suppose moreover that for every $z \in \mathcal{Z}$,*

$$f(w, z) = \max_{j \in [M(z)]} \{a_j(z)^\top w + b_j(z)\}$$

is polyhedral. Then the hypothesis of Theorem E.5 holds for every localized set $U = \mathcal{W} \cap \mathbb{B}(c, r)$ that admits a strict feasible point in the corresponding SOCP representation. In particular, this holds for the localized sets used by the algorithm after arbitrarily small harmless radius enlargements.

Proof. Fix $C > 0$, $z \in \mathcal{Z}$, a localized set $U = \mathcal{W} \cap \mathbb{B}(c, r)$, a point $w \in U$, and $B > 0$. The localized Lipschitz extension is

$$f_{C,U}(w, z) = \inf_{y \in U} \left[\max_{j \in [M(z)]} \{a_j(z)^\top y + b_j(z)\} + C\|w - y\|_2 \right].$$

Introduce variables x, y, u, s, t, where u denotes the auxiliary variable in the SOCP representation of $\mathcal{W}$. Consider the conic program

$$\begin{aligned} \min_{x,y,u,s,t} \quad & s + Ct \\ \text{s.t.} \quad & x + y = w, \\ & \|x\|_2 \le t, \\ & s \ge a_j(z)^\top y + b_j(z) \qquad \forall j \in [M(z)], \\ & y \in \mathcal{W}, \\ & \|y - c\|_2 \le r. \end{aligned}$$

Using the assumed SOCP representation of $\mathcal{W}$, this is an explicit SOCP, and its optimal value is exactly $f_{C,U}(w, z)$.

By the strict-feasibility assumption, Slater's condition holds, so strong duality holds. Let ν be any dual-feasible point, and let g denote the dual multiplier associated with the equality constraint $x + y = w$. Since w appears only in this equality constraint, the dual objective has the form

$$D_w(\nu) = \langle g, w \rangle + \beta(\nu),$$

where $\beta(\nu)$ is independent of w. Therefore, for every $u' \in U$,

$$D_{u'}(\nu) = D_w(\nu) + \langle g, u' - w \rangle.$$

The SOC constraint $\|x\|_2 \le t$ has dual variable (p, q) in the self-dual second-order cone. Stationarity with respect to x and t gives

$$g + p = 0, \qquad q = C,$$

and dual feasibility gives $\|p\| \le q$. Hence

$$\|g\| \le C.$$

By deterministic interior-point methods for SOCP, for any $B > 0$ one can compute in polynomial time in the problem parameters and $1/B$ a dual-feasible point ν_B with

$$D_w(\nu_B) \ge f_{C,U}(w, z) - B.$$

Let g_B be its multiplier for $x + y = w$. For every $u' \in U$, weak duality gives

$$f_{C,U}(u', z) \ge D_{u'}(\nu_B) = D_w(\nu_B) + \langle g_B, u' - w \rangle \ge f_{C,U}(w, z) + \langle g_B, u' - w \rangle - B.$$

Moreover, $\|g_B\| \le C$. Thus g_B is a B-approximate subgradient of $f_{C,U}(\cdot, z)$ at w over U, with norm at most C. This proves the required oracle condition. □

Corollary E.7 (Concrete practical examples)**.** *Under the assumptions of Proposition E.6, the deterministic polynomial-time guarantee of Theorem E.5 applies to:*

1. *affine losses $f(w, z) = a(z)^\top w + b(z)$;*
2. *hinge / ReLU-type losses $f(w, z) = \max\{0, a(z)^\top w + b(z)\}$;*
3. *absolute-value losses $f(w, z) = |a(z)^\top w + b(z)|$.*

In particular, the resulting pure ε-DP heavy-tailed SCO algorithm runs in deterministic polynomial time for these losses on compact Euclidean balls, ellipsoids, and polytopes, subject to the mild strict-feasibility condition for the localized SOCPs.

Proof. Affine losses are polyhedral with one affine piece. Hinge/ReLU-type losses are polyhedral with two pieces:

$$\max\{0, a(z)^\top w + b(z)\}.$$

Absolute-value losses are polyhedral because

$$|a(z)^\top w + b(z)| = \max\{a(z)^\top w + b(z), -a(z)^\top w - b(z)\}.$$

Euclidean balls, ellipsoids, and polytopes admit compact SOCP representations. The claim follows from Proposition E.6 and Theorem E.5. □

F High-probability lower bounds

We record here high-probability excess risk lower bounds for heavy-tailed DP SCO and mean estimation. These lower bounds are sharper by $\mathrm{poly}(\log(1/\delta))$ factors than the corresponding in-expectation lower bounds of [BD14] and nearly match the upper bound in Theorem 3.1.

The main goal of this section is to prove Theorem 4.1.

Risk notation. For the linear losses below, the moment class is

$$\mathcal{P}(G_2, G_k) = \left\{P : \mathbb{E}_P\|z\|_2^2 \le G_2^2,\ \mathbb{E}_P\|z\|_2^k \le G_k^k\right\}.$$

Let $\mathcal{P}$ be any class of distributions on $\mathbb{R}^d$. Given an ε-DP mean estimator $\widehat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ and $\zeta \in (0,1)$, define its $(1-\zeta)$-*quantile squared-error risk* over $\mathcal{P}$ by

$$R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}; \mathcal{P}) := \sup_{P\in\mathcal{P}} \inf\Big\{r \ge 0 : \Pr_{Z\sim P^n}\big(\|\widehat{\mu}(Z) - \mu_P\|_2^2 > r\big) \le \zeta\Big\},$$

where $Z = (z_1, \dots, z_n)$ and $\mu_P := \mathbb{E}_P[z]$.

Let

$$\mathcal{W} := B_2(0, D/2) = \{w \in \mathbb{R}^d : \|w\|_2 \le D/2\},$$

so that $\mathrm{diam}(\mathcal{W}) = D$. For a class $\mathcal{P}$ of distributions on $\mathbb{R}^d$, define the associated linear losses

$$f(w, z) := \langle z, w\rangle, \qquad F_P(w) := \mathbb{E}_{z\sim P}[f(w,z)] = \langle \mu_P, w\rangle.$$

Given an ε-DP algorithm $\mathcal{A} : (\mathbb{R}^d)^n \to \mathcal{W}$ for SCO, define its $(1-\zeta)$-*quantile excess risk* over $\mathcal{P}$ by

$$R^{\mathrm{sco}}_{n,\varepsilon,\zeta}(\mathcal{A}; \mathcal{P}) := \sup_{P\in\mathcal{P}} \inf\Big\{r \ge 0 : \Pr_{Z\sim P^n}\big(F_P(\mathcal{A}(Z)) - \inf_{w\in\mathcal{W}} F_P(w) > r\big) \le \zeta\Big\}.$$

Lower-bound strategy. The non-private term and the private term are witnessed by different hard distribution classes. This is sufficient for a minimax lower bound: since the risk is a supremum over instances, a lower bound from one subclass and a lower bound from another imply a lower bound by the maximum of the two obstructions, and hence by their sum up to constants. Both hard families below belong to $\mathcal{P}(G_2, G_k)$. This separation is conceptually useful: the non-private term is a two-point phenomenon and is most cleanly handled via the quantile version of Le Cam's method from [MVS24], whereas the private term is a packing phenomenon and is most cleanly handled by combining the pure-DP packing lower bound of [BD14] with a direct reduction from estimation to decoding.

Hard distribution classes. We will use two different distribution classes.

For the non-private term in our lower bound, define

$$\mathcal{P}^{\mathrm{bdd}}(G_2) := \{P \text{ supported on } \{\pm G_2 e_1\}\}.$$

Every $P \in \mathcal{P}^{\mathrm{bdd}}(G_2)$ satisfies

$$\mathbb{E}_P\|z\|_2^2 = G_2^2 \qquad \text{and} \qquad \mathbb{E}_P\|z\|_2^k = G_2^k \le G_k^k$$

whenever $G_2 \le G_k$.

For the private term, let $V \subseteq \mathbb{S}^{d-1}$ be a finite packing of the unit sphere, and for parameters $p \in (0,1]$ and $a > 0$ define

$$P_\nu := (1-p)\delta_0 + p\,\delta_{a\nu}, \qquad \nu \in V.$$

Choose

$$a = \min\{G_2 p^{-1/2}, G_k p^{-1/k}\}, \qquad \mathcal{P}_k^{\mathrm{pack}}(G_2, G_k; V, p) = \{P_\nu : \nu \in V\}.$$

Then

$$\mathbb{E}_{P_\nu}\|z\|_2^2 = pa^2 \le G_2^2, \qquad \mathbb{E}_{P_\nu}\|z\|_2^k = pa^k \le G_k^k,$$

so $\mathcal{P}_k^{\mathrm{pack}}(G_2, G_k; V, p) \subseteq \mathcal{P}(G_2, G_k)$.

Reduction from SCO to mean estimation. We begin with the standard reduction from linear SCO lower bounds to mean-estimation lower bounds.

Lemma F.1 (Linear SCO reduces to mean estimation)**.** *Let $\mathcal{P}$ be any family of distributions on $\mathbb{R}^d$ such that $\|\mu_P\|_2 = m$ for every $P \in \mathcal{P}$. For any algorithm $\mathcal{A} : (\mathbb{R}^d)^n \to \mathcal{W}$, define*

$$\widehat{\mu}_{\mathcal{A}}(Z) := \begin{cases} -m\,\mathcal{A}(Z)/\|\mathcal{A}(Z)\|_2, & \mathcal{A}(Z) \neq 0, \\ me_1, & \mathcal{A}(Z) = 0. \end{cases}$$

Then for every $P \in \mathcal{P}$,

$$\|\widehat{\mu}_{\mathcal{A}}(Z) - \mu_P\|_2^2 \le \frac{8m}{D}\Big(F_P(\mathcal{A}(Z)) - \inf_{w\in\mathcal{W}} F_P(w)\Big) \qquad \textit{almost surely.}$$

Consequently,

$$R^{\mathrm{sco}}_{n,\varepsilon,\zeta}(\mathcal{A};\mathcal{P}) \ge \frac{D}{8m}\, R^{\mathrm{mean}}_{n,\varepsilon,\zeta}\big(\widehat{\mu}_{\mathcal{A}};\mathcal{P}\big).$$

Proof. Fix $P \in \mathcal{P}$ and write $\mu_P = mu$ with $u \in \mathbb{S}^{d-1}$. The minimizer of $F_P(w) = \langle \mu_P, w\rangle$ over $\mathcal{W} = B_2(0, D/2)$ is $w^\star = -(D/2)u$, so

$$\inf_{w\in\mathcal{W}} F_P(w) = -\frac{Dm}{2}.$$

If $\mathcal{A}(Z) \neq 0$, then

$$\begin{aligned} F_P(\mathcal{A}(Z)) - \inf_{w\in\mathcal{W}} F_P(w) &= \langle mu, \mathcal{A}(Z)\rangle + \frac{Dm}{2} \\ &\ge \frac{Dm}{4}\left(1 - \left\langle u, -\frac{\mathcal{A}(Z)}{\|\mathcal{A}(Z)\|_2}\right\rangle\right) \\ &= \frac{D}{8m}\|\widehat{\mu}_{\mathcal{A}}(Z) - \mu_P\|_2^2. \end{aligned}$$

To verify the inequality, write $v = \mathcal{A}(Z)/\|\mathcal{A}(Z)\|$. If $\langle u, v\rangle \ge 0$, the excess risk is at least $Dm/2$. If $\langle u, v\rangle < 0$, then $\|\mathcal{A}(Z)\| \le D/2$ gives the lower bound $Dm(1 + \langle u, v\rangle)/2$. If $\mathcal{A}(Z) = 0$, then

$$F_P(\mathcal{A}(Z)) - \inf_{w\in\mathcal{W}} F_P(w) = \frac{Dm}{2},$$

while $\|\widehat{\mu}_{\mathcal{A}}(Z) - \mu_P\|_2^2 \le 4m^2$. Combining the two cases gives the pointwise inequality. The final claim follows directly from the definition of $R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}_{\mathcal{A}};\mathcal{P})$. □

Mean estimation lower bound.

Proposition F.2 (Non-private high-probability term)**.** *There exists a universal constant $c > 0$ such that for all $k \ge 2$, $G_2 > 0$, $G_k \ge G_2$, $n \in \mathbb{N}$, and $\zeta \in (0, 1/4]$, every estimator $\widehat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ satisfies*

$$R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}; \mathcal{P}^{\mathrm{bdd}}(G_2)) \ge c\, G_2^2 \min\left\{\frac{\log(1/\zeta)}{n},\, 1\right\}.$$

Proof. For $\rho \in [0, 1/2]$, define $P_+, P_- \in \mathcal{P}^{\mathrm{bdd}}(G_2)$ by

$$P_+(z = G_2 e_1) = \frac{1+\rho}{2}, \qquad P_-(z = G_2 e_1) = \frac{1-\rho}{2}.$$

Then

$$\mu_{P_+} = \rho G_2 e_1, \qquad \mu_{P_-} = -\rho G_2 e_1,$$

so

$$\|\mu_{P_+} - \mu_{P_-}\|_2 = 2\rho G_2.$$

A direct calculation shows that

$$\mathrm{KL}(P_+, P_-) = \frac{1+\rho}{2}\log\frac{1+\rho}{1-\rho} + \frac{1-\rho}{2}\log\frac{1-\rho}{1+\rho} \le 4\rho^2$$

for all $\rho \in [0, 1/2]$. Hence

$$\mathrm{KL}(P_+^{\otimes n}, P_-^{\otimes n}) \le 4n\rho^2.$$

Choose

$$\rho := c_0 \min\left\{\sqrt{\frac{\log(1/\zeta)}{n}},\, 1\right\}$$

with $c_0 > 0$ small enough that

$$\mathrm{KL}(P_+^{\otimes n}, P_-^{\otimes n}) \le \log \frac{1}{4\zeta(1-\zeta)}.$$

Corollary 6 of [MVS24] then implies

$$R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}; \{P_+, P_-\}) \ge c\,\rho^2 G_2^2.$$

Since $\{P_+, P_-\} \subseteq \mathcal{P}^{\mathrm{bdd}}(G_2)$, this yields

$$R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}; \mathcal{P}^{\mathrm{bdd}}(G_2)) \ge c\,G_2^2 \min\left\{\frac{\log(1/\zeta)}{n},\, 1\right\}.$$

□

For the private term, we use a direct decoder reduction.

Lemma F.3 (Quantile estimation implies decoding on a packing)**.** *Let $\{\theta_\nu : \nu \in V\} \subset \mathbb{R}^d$ be such that*

$$\|\theta_\nu - \theta_{\nu'}\|_2 > 2r \qquad \forall \nu \ne \nu'.$$

Let $\{P_\nu : \nu \in V\}$ be distributions on $\mathbb{R}^d$, and let $\widehat{\theta} : (\mathbb{R}^d)^n \to \mathbb{R}^d$ be any estimator. Define the nearest-neighbor decoder

$$\psi_{\widehat{\theta}}(Z) \in \arg\min_{\nu \in V} \|\widehat{\theta}(Z) - \theta_\nu\|_2.$$

If for every $\nu \in V$,

$$P_\nu^n\big(\|\widehat{\theta}(Z) - \theta_\nu\|_2 \le r\big) \ge 1 - \zeta,$$

then

$$\frac{1}{|V|}\sum_{\nu \in V} P_\nu^n\big(\psi_{\widehat{\theta}}(Z) \ne \nu\big) \le \zeta.$$

Proof. Fix $\nu \in V$. On the event $\|\widehat{\theta}(Z) - \theta_\nu\|_2 \le r$, we have

$$\|\widehat{\theta}(Z) - \theta_{\nu'}\|_2 \ge \|\theta_\nu - \theta_{\nu'}\|_2 - \|\widehat{\theta}(Z) - \theta_\nu\|_2 > r$$

for every $\nu' \ne \nu$. Hence nearest-neighbor decoding returns ν. Therefore

$$P_\nu^n(\psi_{\widehat{\theta}}(Z) \ne \nu) \le P_\nu^n(\|\widehat{\theta}(Z) - \theta_\nu\|_2 > r) \le \zeta.$$

Averaging over $\nu \in V$ gives the claim. □

Proposition F.4 (Pure-DP high-probability private term)**.** *Let $k \ge 2$, $0 < G_2 \le G_k$, $\varepsilon \in (0, 1]$, $\zeta \in (0, 1/4]$, $n \in \mathbb{N}$, and $d \ge 1$. Every ε-DP estimator $\widehat{\mu}$ satisfies*

$$R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}; \mathcal{P}_k^{\mathrm{pack}}(G_2, G_k)) \ge c \min\Bigg\{G_2^2,\; G_2^2\frac{d + \log(1/\zeta)}{n\varepsilon},\; G_k^2\left(\frac{d + \log(1/\zeta)}{n\varepsilon}\right)^{2-2/k}\Bigg\},$$

for an absolute constant $c > 0$. Here $\mathcal{P}_k^{\mathrm{pack}}(G_2, G_k) = \bigcup_{V,p} \mathcal{P}_k^{\mathrm{pack}}(G_2, G_k; V, p)$, where $p \in (0, 1]$ and $V \subseteq \mathbb{S}^{d-1}$ ranges over finite $1/2$-packings with $|V| \ge 2^{cd}$.

Proof. Fix a $1/2$-packing $V \subseteq \mathbb{S}^{d-1}$ with $M := |V| \geq 2$ and $\log M \geq cd$. Set

$$p = \min\left\{\frac{1}{8n\varepsilon}\log\frac{M-1}{2\zeta},\, 1\right\}, \qquad a = \min\{G_2 p^{-1/2}, G_k p^{-1/k}\},$$
$$P_\nu = (1-p)\delta_0 + p\delta_{a\nu}.$$

These distributions satisfy both moment bounds. Their means $\mu_\nu = pa\nu$ are separated by at least $pa/2$. If the quantile squared-error risk were less than $(pa)^2/64$, Lemma F.3 would give an ε-DP decoder $\widehat{\nu}$ with average error at most ζ.

Generate $Z \sim P_\nu^n$ using independent Bernoulli(p) indicators, and let H be their sum. Fix an indicator pattern with $H = h$. The datasets for two packing elements differ in h entries, so the group-privacy argument of [BD14] gives

$$1 \geq Q_{\nu'}(\nu') + e^{-\varepsilon h}\sum_{\nu \neq \nu'} Q_\nu(\nu),$$

where Q_ν is the decoder's output distribution conditional on this pattern. Averaging over ν' yields

$$\frac{1}{M}\sum_{\nu\in V} Q_\nu(\nu) \leq \frac{1}{1+(M-1)e^{-\varepsilon h}}.$$

Since $\mathbb{E}H = np$, Markov's inequality gives $\Pr(H \leq 4np) \geq 3/4$. The choice of p also gives

$$(M-1)e^{-4n\varepsilon p} \geq \sqrt{2\zeta(M-1)} \geq \sqrt{2\zeta}.$$

Thus the average decoding error satisfies

$$\begin{aligned}\frac{1}{M}\sum_{\nu\in V}\Pr_{Z\sim P_\nu^n}(\widehat{\nu}(Z) \neq \nu) &\geq \frac{3}{4}\frac{(M-1)e^{-4n\varepsilon p}}{1+(M-1)e^{-4n\varepsilon p}} \\ &\geq \frac{3}{4}\frac{\sqrt{2\zeta}}{1+\sqrt{2\zeta}} > \zeta.\end{aligned}$$

The last inequality follows from $\sqrt{2\zeta} + 2\zeta \leq 1/\sqrt{2} + 1/2 < 3/2$. This contradicts the decoder guarantee, proving

$$R^{\mathrm{mean}}_{n,\varepsilon,\zeta}(\widehat{\mu}; \mathcal{P}_k^{\mathrm{pack}}(G_2, G_k)) \geq \frac{(pa)^2}{64} = \frac{1}{64}\min\{G_2^2 p, G_k^2 p^{2-2/k}\}.$$

Finally,

$$p \gtrsim \min\left\{\frac{d+\log(1/\zeta)}{n\varepsilon},\, 1\right\}.$$

Together with $G_2 \leq G_k$, this gives the stated bound. □

SCO lower bound. Finally, we translate the mean estimation lower bounds into an SCO lower bound.

Theorem F.5 (SCO lower bound – Precise version of Theorem 4.1)**.** *Let $k \geq 2$, $0 < G_2 \leq G_k$, $\varepsilon \in (0,1]$, $\zeta \in (0, 1/4]$, $n \in \mathbb{N}$, and $d \geq 1$. For every ε-DP algorithm $\mathcal{A} : (\mathbb{R}^d)^n \to \mathcal{W}$, there is a $P \in \mathcal{P}(G_2, G_k)$ such that*

$$\Pr_{Z\sim P^n}\left(F_P(\mathcal{A}(Z)) - F_P^* \geq cD\min\left\{G_2,\; G_2\sqrt{\frac{\log(1/\zeta)}{n}}\right.\right.$$
$$\left.\left. + \min\left\{G_2\sqrt{\frac{d+\log(1/\zeta)}{n\varepsilon}},\; G_k\left(\frac{d+\log(1/\zeta)}{n\varepsilon}\right)^{1-1/k}\right\}\right\}\right) \geq \zeta,$$

where $c > 0$ is an absolute constant and $F_P^* = \inf_{w\in\mathcal{W}} F_P(w)$.

Proof. For the non-private term, take the family $\{P_+, P_-\}$ from the proof of Proposition F.2. Both distributions satisfy $\|z\|_2 = G_2$ almost surely and hence belong to $\mathcal{P}(G_2, G_k)$. Their means have norm

$$\|\mu_P\|_2 \asymp G_2\min\left\{\sqrt{\frac{\log(1/\zeta)}{n}},\, 1\right\}.$$

The mean-estimation lower bound in Proposition F.2 is of order $\|\mu_P\|_2^2$. Applying Lemma F.1 gives an SCO lower bound

$$cDG_2 \min\left\{\sqrt{\frac{\log(1/\zeta)}{n}},\, 1\right\}$$

with probability at least ζ for one of these distributions.

For the private term, use the family from Proposition F.4. Its mean norm is pa, and its quantile squared-error lower bound is $(pa)^2/64$. By Lemma F.1, the SCO lower bound is of order

$$\begin{aligned} Dpa &= D\min\{G_2\sqrt{p}, G_k p^{1-1/k}\} \\ &\gtrsim D\min\left\{G_2,\ G_2\sqrt{\frac{d+\log(1/\zeta)}{n\varepsilon}},\ G_k\left(\frac{d+\log(1/\zeta)}{n\varepsilon}\right)^{1-1/k}\right\}. \end{aligned}$$

For every algorithm, one of the two families gives the larger of the sampling and privacy lower bounds. Since

$$\max\{\min(G_2, a), \min(G_2, b)\} \geq \tfrac{1}{2}\min\{G_2, a+b\} \qquad (a, b \geq 0),$$

this proves the theorem, after reducing the absolute constant if needed to pass from quantile risk to the probability statement. □

Remark F.6 (Tightness of Theorem F.5). The upper and lower bounds hold over the same class $\mathcal{P}(G_2, G_k)$. Since $\log(1/\zeta) > 1$ and

$$d\log(1/\zeta) \leq \log(1/\zeta)\big(d + \log(1/\zeta)\big),$$

each upper-bound privacy term is at most $\log(1/\zeta)^{1-1/k}$ times the corresponding lower-bound term. The sampling term and the cap DG_2 are the same. Thus the bounds match up to this logarithmic factor for every $0 < G_2 \leq G_k$. At constant confidence, the minimax rate is

$$\Theta\Bigg(D\min\Bigg\{G_2,\ \frac{G_2}{\sqrt{n}} + \min\left\{G_2\sqrt{\frac{d}{n\varepsilon}},\ G_k\left(\frac{d}{n\varepsilon}\right)^{1-1/k}\right\}\Bigg\}\Bigg).$$

G Limitations

We highlight several limitations of the present work.

First, our results are purely theoretical. We do not provide an empirical implementation or experimental evaluation of the proposed algorithms. In particular, although our main algorithm is polynomial-time, the polynomial dependence and hidden constants may be too large for practical machine-learning workloads. Thus, our results should be interpreted primarily as a characterization of the statistical and computational landscape of pure-DP heavy-tailed SCO, rather than as an immediately deployable training method.

Second, our guarantees rely on structural assumptions that may not hold in all applications. We assume convex losses, a compact convex projection-friendly domain, polynomial-time first-order access to the loss, and publicly known moment bounds G_2, G_k. We also use additional assumptions in some appendices, such as access to deterministic approximate subgradients of the Lipschitz extension for certain structured subclasses. These assumptions are natural for our theoretical setting but may be restrictive in certain practical applications.

Third, while our main algorithm runs in polynomial time with high probability, polynomial-time runtime with probability 1 is obtained only under extra conditions: either a finite polynomially bounded worst-case Lipschitz parameter, or special structured subclasses for which the Lipschitz extension admits efficient deterministic approximate subgradient computation. Obtaining optimal pure-DP heavy-tailed SCO in deterministic polynomial time for general convex losses remains open.

Fourth, our upper and lower bounds match only up to logarithmic factors. In particular, there remains a gap between the $d\log(1/\delta)$ dependence in the private optimization term of our upper bound and the $d + \log(1/\delta)$ dependence in our lower bound.

Finally, our work studies excess population risk under differential privacy, but does not address several other important properties of learning systems, including security or robustness to adversarial examples or poisoning (c.f. [GLZW25]). Pure differential privacy protects individual data contributions, but it is not by itself a complete notion of responsible or safe deployment.

H Broader Impacts

This work is theoretical and studies differentially private stochastic convex optimization under a heavy-tailed gradient assumption. Its primary positive impact is conceptual: it advances the foundations of privacy-preserving learning in settings where standard worst-case Lipschitz assumptions lead to overly pessimistic or vacuous excess risk guarantees. Heavy-tailed data models are relevant in applications involving outliers, unbounded features, or heterogeneous populations, so sharper private learning guarantees in such regimes may eventually help broaden the range of settings in which privacy-preserving methods can be used responsibly.

A second positive impact is methodological. Our results clarify the distinctions between techniques for approximate and pure differential privacy in heavy-tailed optimization, and provide new tools for understanding what can and cannot be achieved efficiently under the stronger pure-DP requirement.

At the same time, these results should not be overinterpreted. Our algorithms are primarily of theoretical interest, and we do not claim practical readiness for large-scale machine-learning deployment.

Because the paper develops stronger privacy-preserving optimization methods, one possible downstream benefit is enabling the analysis of sensitive datasets with reduced privacy risk. A possible downstream risk is that theoretical privacy guarantees may be cited in ways that obscure the gap between minimax guarantees and practical deployment realities. For this reason, we emphasize that our contribution is a foundational step in the theory of private heavy-tailed optimization, not a complete solution to the broader challenges of safe and socially responsible machine learning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions precisely: upper bounds, lower bounds, and polynomial-time algorithms for pure-DP heavy-tailed SCO, all up to logarithmic factors; see the Abstract, Section 1, and in particular the contribution paragraphs in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations, including the theoretical nature of the results, structural assumptions, runtime caveats, and remaining logarithmic gaps; see Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main assumptions are stated in Assumption 1.1, additional appendix-specific assumptions are introduced locally where needed, and complete proofs for all stated theoretical results are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include experiments; the main claims are purely theoretical.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments or empirical artifacts requiring data/code release to support the main claims.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work is a theoretical study of differentially private optimization and, to the best of our knowledge, conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative potential impacts, including improved privacy-preserving learning in heavy-tailed settings and the limits of what privacy guarantees do and do not provide; see Appendix H.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release datasets, trained models, or other high-risk assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

Justification: The paper does not rely on external code, datasets, or pretrained models as research assets; it uses only cited prior literature.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release new assets such as datasets, code packages, or models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used as a research-assistance tool to help formalize and typeset some proof steps and routine lemmas from author-provided proof roadmaps/outlines/drafts. They were not part of the paper's core methodology or theoretical claims, and all statements, assumptions, and proofs were independently checked and verified by the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.