# GROWTH GEOMETRY REDUCES GRADIENT COMPLEXITY IN DIFFERENTIALLY PRIVATE EMPIRICAL RISK MINIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As a core problem in differentially private (DP) machine learning, DP empirical risk minimization (DP-ERM) has been intensively studied in the last decade. However, from the efficiency perspective, most existing work focuses on the running time or gradient complexity required to achieve optimal statistical utility, which does not fully characterize the gradient complexity for a general target optimization accuracy. Recently, Menart & Nikolov (2026) introduced the private-proxy oracle model to study this question for general nonsmooth Lipschitz convex objectives. It remains unclear how additional geometry of the objective changes the oracle complexity. In this work, we study convex finite-sum objectives with globally $L$-Lipschitz convex components whose average satisfies the $(\theta, \lambda)$ growth condition, also called the Tsybakov noise condition, with $\theta \geq 2$. Let $d$ denote the dimension, $R$ bound the distance from feasible points to the minimizer set, $\rho > 0$ the conditional per-reply zero-concentrated differential privacy (zCDP) parameter of the proxy, $\bar{m}$ the maximum batch size, and $\alpha > 0$ the target expected excess empirical risk. In the growth-active regime $\alpha \leq \lambda R^\theta$, and under additional accuracy and dimension conditions, we prove a lower bound with a non-private optimization term $\Omega(L^2\lambda^{-2/\theta}\alpha^{-2(\theta-1)/\theta})$ and privacy-dependent terms $\Omega(\sqrt{d\log(2 + L\lambda^{-1/\theta}\alpha^{-(\theta-1)/\theta})/\rho})$ and $\Omega(d\log(2 + L\lambda^{-1/\theta}\alpha^{-(\theta-1)/\theta})/(\bar{m}\rho))$. We then give a restarted private-proxy method using $O(L^2\lambda^{-2/\theta}\alpha^{-2(\theta-1)/\theta}\max\{1, \sqrt{d/\rho}, d/(\bar{m}\rho)\})$ component first-order evaluations. Under the same conditions, these bounds match up to constant factors when $\rho \gtrsim d$. Thus, growth improves the optimization dependence from $\alpha^{-2}$ to $\alpha^{-2(\theta-1)/\theta}$. In particular, quadratic growth $(\theta = 2)$ gives an $\alpha^{-1}$ dependence.

## 1 INTRODUCTION

Differential privacy (DP) (Dwork et al., 2006b) has become a standard framework for learning from sensitive data while limiting the influence of any individual sample on the output of an algorithm. Many learning problems can be formulated as *empirical risk minimization* (ERM). Given $n$ data points $z_1, \ldots, z_n$, a parameter domain $\mathcal{W} \subseteq \mathbb{R}^d$ in dimension $d$, and a per-example loss function $f$, ERM minimizes the empirical objective

$$F(w) = \frac{1}{n}\sum_{i=1}^{n} f(w; z_i), \qquad w \in \mathcal{W} \subseteq \mathbb{R}^d.$$

ERM under differential privacy constraints, commonly referred to as *differentially private ERM* (DP-ERM), has been studied extensively. For convex problems, matching or near-matching privacy–utility guarantees are known in many settings (Bassily et al., 2014; Talwar et al., 2015; Song et al., 2021).

Privacy–utility guarantees, however, do not characterize the computation required to obtain them. Let $\alpha > 0$ denote the target excess optimization error. A natural measure of first-order computation is the *gradient complexity*: the number of individual gradient or subgradient evaluations required to produce a point whose expected excess error is at most $\alpha$. A substantial line of work has developed increasingly efficient private optimization algorithms (Wang et al., 2017; Feldman et al.,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2020; Kulkarni et al., 2021; Asi et al., 2021a). These results mainly study the gradient complexity required to attain optimal statistical utility, and therefore do not determine the gradient complexity for a general target $\alpha$.

Menart & Nikolov (2026) address this question through the *private-proxy oracle* model, in which an optimizer accesses component first-order information through a privacy-preserving proxy. Each proxy reply satisfies $\rho$-zero-concentrated differential privacy ($\rho$-zCDP) with respect to the first-order answers in the current query batch, where $\rho > 0$ is the privacy parameter. A maximum batch size $\bar{m}$ limits the number of component evaluations used in one query batch. For general nonsmooth Lipschitz convex objectives, their upper bound has the generic $\alpha^{-2}$ optimization dependence together with dimension-, privacy-, and batch-dependent costs. They also state a product-form lower bound under additional conditions. Our lower-bound analysis is independent of this result; Appendix B.5 examines the information-demand arguments used in its proof.

The result of Menart & Nikolov (2026), however, is established for the general class of nonsmooth Lipschitz convex objectives. It therefore does not determine how the gradient complexity changes when the objective has additional geometric structure. This raises a natural question: *how does favorable growth geometry change the gradient complexity of private-proxy optimization?* We study this question under the $(\theta, \lambda)$ growth condition, also known as the Tsybakov noise condition (TNC) (Ramdas & Singh, 2013), where $\theta > 1$ is the growth exponent and $\lambda > 0$ controls the strength of the condition. This condition holds for strongly convex objectives and for certain formulations of support vector machines, $\ell_1$-regularized stochastic optimization, and linear regression (Xu et al., 2017; Liu et al., 2018). It has also been studied in private optimization (Asi et al., 2021b; Su et al., 2024; Xu et al., 2025). Intuitively, the growth condition converts a smaller optimization error into a smaller distance to the solution set, allowing first-order methods to optimize over progressively smaller effective radii. In the non-private setting, Ramdas & Singh (2013) show that this structure improves the accuracy dependence of stochastic first-order complexity from $\alpha^{-2}$ for general nonsmooth convex objectives to $\alpha^{-2(\theta-1)/\theta}$, up to logarithmic factors, with the Lipschitz and growth parameters held fixed.

**Contributions.** We focus on nonsmooth convex objectives satisfying $(\theta, \lambda)$ growth with $\theta \geq 2$. Let $L$ bound component subgradients and let $R$ bound the distance from any feasible point to the minimizer set, and define

$$K_{\theta,\lambda} := \min\left\{\frac{L^2R^2}{\alpha^2}, L^2\lambda^{-2/\theta}\alpha^{-\frac{2(\theta-1)}{\theta}}\right\}.$$

Our results separate the effect of geometry from the cost of privacy. In the growth-active regime $\alpha \leq \lambda R^\theta$, $K_{\theta,\lambda} = L^2\lambda^{-2/\theta}\alpha^{-2(\theta-1)/\theta}$, so growth changes the optimization scale from the general convex $\alpha^{-2}$ dependence to $\alpha^{-2(\theta-1)/\theta}$.

**1. Lower bound under growth.** Under conditions on the accuracy and dimension, we prove that every admissible optimizer–proxy pair requires

$$\Omega\left(\max\left\{K_{\theta,\lambda}, \sqrt{\frac{d\log(2+\sqrt{K_{\theta,\lambda}})}{\rho}}, \frac{d\log(2+\sqrt{K_{\theta,\lambda}})}{\bar{m}\rho}\right\}\right)$$

component first-order evaluations. The first term is a non-private optimization lower bound, while the remaining two terms are privacy-dependent information lower bounds. Under additional conditions, Theorem B7 in Appendix B.4 gives a refined privacy-dependent lower bound including $\Omega(\sqrt{K_{\theta,\lambda}d/\rho})$.

**2. Upper bound by localization and restarting.** We develop a restarted private-proxy projected-subgradient method that, for every dimension and every maximum batch size, achieves

$$O\left(K_{\theta,\lambda}\max\left\{1, \sqrt{\frac{d}{\rho}}, \frac{d}{\bar{m}\rho}\right\}\right)$$

component first-order evaluations. Each private update costs $O(\max\{1, \sqrt{d/\rho}, d/(\bar{m}\rho)\})$ component evaluations, while growth and restarting reduce the number of required updates to $K_{\theta,\lambda}$. Under the hypotheses of Theorem 4.1, the upper and lower bounds match in the regime (18).

## 2 Related Work and Positioning

**Gradient complexity in private optimization.** A substantial literature develops computationally efficient algorithms for private ERM and stochastic convex optimization (Bassily et al., 2014; Wang et al., 2017; Feldman et al., 2020; Kulkarni et al., 2021; Asi et al., 2021a). These works primarily give upper bounds on the first-order computation sufficient to attain private statistical guarantees. Closer to our setting, Menart & Nikolov (2026) introduce the private-proxy oracle model and establish upper bounds for general nonsmooth convex objectives, for which the optimization scale is $\alpha^{-2}$ and privacy introduces additional dimension- and batch-dependent costs. They also state a product-form lower bound under additional conditions; Appendix B.5 examines the information-demand arguments used in its proof. We study how this picture changes when the objective additionally satisfies $(\theta, \lambda)$ growth. In the growth-active regime, we show that the optimization scale becomes $L^2\lambda^{-2/\theta}\alpha^{-2(\theta-1)/\theta}$, while the restarted upper bound retains the same per-update privacy-, dimension-, and batch-dependent cost.

**Growth and first-order complexity.** Growth conditions are known to improve optimization rates. For stochastic nonsmooth convex optimization under TNC, Ramdas & Singh (2013) establish, up to logarithmic factors, the optimal accuracy dependence $\alpha^{-2(\theta-1)/\theta}$, improving the general convex $\alpha^{-2}$ rate. Related acceleration under uniformly convex and error-bound conditions is studied by Juditsky & Nesterov (2014); Bolte et al. (2017); Roulet & d'Aspremont (2017). These results characterize the benefit of growth without privacy, but not the additional first-order cost imposed by a private proxy. Our upper bound uses the same localization principle through restarting, while our lower bound adapts the information-theoretic construction of Braun et al. (2017) to the prescribed growth class and combines it with a normalized convex private-proxy information lower bound.

**Private optimization under growth.** Growth can also improve privacy–utility guarantees (Asi et al., 2021b; Su et al., 2024; Xu et al., 2025). Most directly related, Su et al. (2024) study DP-SCO under TNC. For smooth losses, their localization method has essentially linear gradient complexity in the sample size; solving their statistical rate for a target error $\alpha$ yields, up to logarithmic and problem-dependent factors, $\widetilde{O}(\alpha^{-2(\theta-1)/\theta} + \sqrt{d\log(1/\delta)}\,\epsilon^{-1}\alpha^{-(\theta-1)/\theta})$, where $\epsilon$ and $\delta$ are the dataset-level privacy parameters. For nonsmooth losses, they replace their Phased-SGD subroutine by Phased-ERM (Feldman et al., 2020); this preserves the improved utility rate but does not yield the same linear gradient-complexity guarantee. Asi et al. (2021b) derive their $(\epsilon, \delta)$-DP lower bound under growth by a reduction from general convex ERM; our private lower bound also uses a reduction, but for component-evaluation complexity in the private-proxy model (Appendix D). Our work instead studies the component first-order complexity of nonsmooth finite-sum optimization in the private-proxy model.

## 3 Preliminaries and Problem Formulation

We study convex empirical risk minimization with first-order information accessed through a private proxy. Throughout, $\|\cdot\|$ denotes the Euclidean norm, $n$ is the number of data points, $[n] = \{1, \dots, n\}$ is the index set, and $d$ is the dimension. We use the standard asymptotic notation $O$, $\Omega$, $\Theta$, $\lesssim$ and $\asymp$. A tilde additionally suppresses logarithmic factors.

**Definition 3.1** (Differential privacy (Dwork et al., 2006b;a))**.** Two datasets of size $n$ are *neighboring* if they differ in one entry. For $\epsilon \geq 0$ and $\delta \in [0, 1)$, a randomized mechanism $\mathcal{M}$ is $(\epsilon, \delta)$-differentially private if, for every pair of neighboring datasets $D, D'$ and every measurable output event $S$,

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta. \tag{1}$$

The private-proxy model studied in this paper measures privacy using zero-concentrated differential privacy (zCDP).

**Definition 3.2** (Zero-concentrated differential privacy (Bun & Steinke, 2016))**.** For $\rho \geq 0$, a randomized mechanism $\mathcal{M}$ satisfies $\rho$-zCDP if, for every pair of neighboring inputs $D, D'$ and every Rényi order $\beta > 1$,

$$D_\beta(\mathcal{M}(D) \,\|\, \mathcal{M}(D')) \leq \rho\beta, \tag{2}$$

where $D_\beta$ denotes Rényi divergence of order $\beta$.

Smaller values of $\rho$ correspond to stronger privacy. A $\rho$-zCDP mechanism is also $(\epsilon, \delta)$-DP for every $\delta \in (0, 1)$ with $\epsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$ (Bun & Steinke, 2016). The explicit formula for Rényi divergence is recalled in Appendix A.2.

**Definition 3.3** (Differentially private empirical risk minimization (DP-ERM))**.** Given a dataset $D = (z_1, \ldots, z_n)$ and a public closed convex set $\mathcal{W} \subseteq \mathbb{R}^d$, let $f_i(w) = f(w; z_i)$ and define the empirical objective

$$F(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w), \qquad w \in \mathcal{W}.$$

Assume that $F$ attains its minimum on $\mathcal{W}$, and define

$$F^\star := \min_{w \in \mathcal{W}} F(w), \qquad W^\star := \arg\min_{w \in \mathcal{W}} F(w).$$

A DP-ERM algorithm returns $\widehat{w} \in \mathcal{W}$ such that the map $D \mapsto \widehat{w}$ satisfies Definition 3.1 under replacement of one training example. Its utility is measured by the expected excess error $\mathbb{E}[F(\widehat{w}) - F^\star]$.

**Assumption 3.4** (Convex Lipschitz losses)**.** *Each component $f_i : \mathbb{R}^d \to \mathbb{R}$ is convex and globally $L$-Lipschitz for a known $L > 0$; that is,*

$$|f_i(u) - f_i(v)| \le L\|u - v\|, \qquad \textit{for all } u, v \in \mathbb{R}^d.$$

*No differentiability or smoothness is assumed.*

**Definition 3.5** (Selected first-order oracle)**.** For each $i \in [n]$, fix before optimization begins a Borel-measurable selection $g_i : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $g_i(w) \in \partial f_i(w)$. By Assumption 3.4, $\|g_i(w)\| \le L$ for every $w \in \mathbb{R}^d$. The selected first-order oracle returns

$$\big(f_i(w), g_i(w)\big).$$

Each query counts as one component first-order evaluation. Queries may lie anywhere in $\mathbb{R}^d$, which is why the Lipschitz assumption is global; only the final output is required to belong to $\mathcal{W}$.

**Definition 3.6** ($(\theta, \lambda)$ growth condition / TNC)**.** For $w \in \mathbb{R}^d$, let

$$\operatorname{dist}(w, W^\star) := \inf_{v \in W^\star} \|w - v\|.$$

For $\theta \ge 2$ and $\lambda > 0$, the objective $F$ satisfies the $(\theta, \lambda)$ growth condition, also called the Tsybakov noise condition (TNC), if

$$F(w) - F^\star \ge \lambda \operatorname{dist}(w, W^\star)^\theta, \qquad w \in \mathcal{W}.$$

**Remark 3.7** (Quadratic growth)**.** The case $\theta = 2$ is quadratic growth. Every $\mu$-strongly convex objective satisfies quadratic growth with $\lambda = \mu/2$, but the converse need not hold.

**Assumption 3.8** (Distance to the minimizers)**.** *There is a known $R > 0$ such that*

$$\sup_{w \in \mathcal{W}} \operatorname{dist}(w, W^\star) \le R. \tag{3}$$

We next specify how the optimizer accesses first-order information.

**Definition 3.9** (Private proxy (Menart & Nikolov, 2026))**.** Fix $\rho > 0$ and $\bar{m} \in \mathbb{N}_{\ge 1} \cup \{\infty\}$. At each round $t$, based on the previously released proxy messages, the optimizer $\mathcal{A}$ chooses an integer $M_t \le \bar{m}$ and a batch

$$(w_{t,1}, i_{t,1}), \ldots, (w_{t,M_t}, i_{t,M_t})$$

before any current first-order response is evaluated. The randomized proxy $\mathcal{P}$ evaluates these queries and releases a message $Y_t$. Conditional on the preceding interaction and the current query batch, the map from the $M_t$ current first-order responses to $Y_t$ must satisfy $\rho$-zCDP under replacement of one current response by another admissible response, with all other responses fixed. An admissible response is any $(a, g) \in \mathbb{R} \times \mathbb{R}^d$ satisfying $\|g\| \le L$, with $a \in \mathbb{R}$ unrestricted. The proxy may depend on the current responses and the public interaction history, but on no additional private input. Thus the optimizer may adapt across rounds, while the batch is fixed within each round. Throughout, optimizer and proxy random-seed spaces and proxy reply spaces are standard Borel, and all decision, output, and reply kernels are measurable; see Appendix A.

**Remark 3.10** (Privacy model)**.** The privacy requirement in Definition 3.9 is imposed separately on each proxy reply under replacement of one current first-order response. In contrast, Definition 3.1 imposes privacy on the map from the entire dataset $D$ to the final output under replacement of one training example. We do not require the complete optimizer–proxy procedure to satisfy this dataset-level privacy condition.

We now collect the preceding assumptions into the problem class over which our upper and lower bounds are stated.

**Definition 3.11** (Problem class)**.** Fix the public parameters $n, d, L, R, \theta$, and $\lambda$. Let $\mathfrak{F}_{\theta,\lambda}(L, R)$ denote the class of instances consisting of an empirical objective $F$ and its fixed selected first-order oracle, together with a public nonempty closed convex feasible set $\mathcal{W}$, a feasible starting point, and Euclidean projection onto $\mathcal{W}$, satisfying Assumptions 3.4 and 3.8 and Definition 3.6.

**Definition 3.12** (Minimax gradient oracle complexity)**.** Let $\mathfrak{F} = \mathfrak{F}_{\theta,\lambda}(L, R)$ and fix $\rho > 0$, $\bar{m} \in \mathbb{N}_{\geq 1} \cup \{\infty\}$, and $\alpha > 0$. For every instance $H \in \mathfrak{F}$ and optimizer–proxy pair $(\mathcal{A}, \mathcal{P})$, define

$$\mathrm{Cost}_H(\mathcal{A}, \mathcal{P}) := \mathbb{E}_H\left[\sum_t M_t\right],$$

where $\mathbb{E}_H$ denotes expectation over the randomness of $\mathcal{A}$ and $\mathcal{P}$ on the fixed instance $H$. Repeated component queries are counted separately; projection, communication, and other computation are not charged. The pair $(\mathcal{A}, \mathcal{P})$ is *$\alpha$-accurate* on $\mathfrak{F}$ if, for every $H \in \mathfrak{F}$, it terminates almost surely, returns $\widehat{w} \in \mathcal{W}$ almost surely, and

$$\mathbb{E}_H[F(\widehat{w}) - F^\star] \leq \alpha,$$

with integrable excess error. The minimax gradient oracle complexity is

$$\mathfrak{T}_{\rho,\bar{m}}(\alpha; \mathfrak{F}) := \inf_{\substack{(\mathcal{A},\mathcal{P}) \\ \text{satisfying Definition 3.9}}} \sup_{\alpha\text{-accurate on } \mathfrak{F}} \sup_{H \in \mathfrak{F}} \mathrm{Cost}_H(\mathcal{A}, \mathcal{P}). \tag{4}$$

If no such optimizer–proxy pair exists, the infimum is defined to be $+\infty$.

**Remark 3.13** (Interpretation)**.** $\mathfrak{T}_{\rho,\bar{m}}(\alpha; \mathfrak{F})$ is the smallest worst-case expected number of component first-order evaluations required to attain expected excess error at most $\alpha$ uniformly over the problem class.

## 4 Lower Bounds for Gradient Complexity under Growth

We first establish a lower bound on the number of component first-order evaluations required by private-proxy methods under $(\theta, \lambda)$ growth. The result separates the intrinsic optimization cost from the privacy-dependent information cost. It applies to every admissible optimizer–proxy pair, including adaptive optimizers and proxies that process each current batch jointly.

Fix $\theta \geq 2$, $L, R, \lambda > 0$, and a target excess error $\alpha > 0$. Define

$$K_{\theta,\lambda} := \min\left\{\frac{L^2R^2}{\alpha^2}, L^2\lambda^{-2/\theta}\alpha^{-\frac{2(\theta-1)}{\theta}}\right\}, \qquad \ell_\alpha := \log\left(2 + \sqrt{K_{\theta,\lambda}}\right). \tag{5}$$

The quantity $K_{\theta,\lambda}$ is the optimization scale induced by the growth condition.

**Theorem 4.1** (Lower bound under the growth condition)**.** *There exist constants $c_0, c_1, C > 0$ such that the following holds. Fix $\theta \geq 2$, $n, d \in \mathbb{N}_{\geq 1}$, $L, R, \lambda, \rho > 0$, and $\bar{m} \in \mathbb{N}_{\geq 1} \cup \{\infty\}$. Suppose*

$$0 < \alpha \leq c_0 \min\left\{LR, \left(\frac{L^\theta}{\lambda}\right)^{1/(\theta-1)}\right\}, \qquad d \geq CK_{\theta,\lambda}.$$

*Then*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L, R)\big) \geq c_1 \max\left\{K_{\theta,\lambda}, \sqrt{\frac{d\ell_\alpha}{\rho}}, \frac{d\ell_\alpha}{\bar{m}\rho}\right\}, \tag{6}$$

*where the term $d\ell_\alpha/(\bar{m}\rho)$ is omitted when $\bar{m} = \infty$.*

The first term in (6) is the non-private optimization cost, while the remaining two terms are privacy-dependent.

When $\alpha \le \lambda R^\theta$, $K_{\theta,\lambda} = L^2\lambda^{-2/\theta}\alpha^{-\frac{2(\theta-1)}{\theta}}$, so the optimization scale changes from the generic $\alpha^{-2}$ dependence to $\alpha^{-2(\theta-1)/\theta}$. Appendix B.4 gives a refined lower bound under additional dimension and batch-size conditions, including $\Omega(\sqrt{K_{\theta,\lambda}d/\rho})$. For quadratic growth, $K_{2,\lambda} = L^2/(\lambda\alpha)$; moreover, if $\alpha \le \lambda^3 R^4/L^2$, the refined bound also gives a term of order $\sqrt{K_{2,\lambda}}\, d/(\bar{m}\rho)$.

**Proof overview.** The main difficulty is to construct hard instances satisfying the prescribed growth condition while retaining the relevant optimization and privacy obstructions. We prove the optimization and privacy-dependent terms separately.

*Non-private optimization hardness.* We first prove that attaining expected excess error at most $\alpha$ requires $\Omega(K_{\theta,\lambda})$ component first-order evaluations, even without privacy. Although Ramdas & Singh (2013) establish the corresponding $\Omega(\alpha^{-2(\theta-1)/\theta})$ lower bound under TNC for stochastic first-order oracles, their oracle model does not directly yield a component-evaluation lower bound for the finite-sum setting considered here.

Define

$$D_\alpha := \min\left\{R, \left(\frac{\alpha}{\lambda}\right)^{1/\theta}\right\}, \qquad K_{\theta,\lambda} = \frac{L^2 D_\alpha^2}{\alpha^2}.$$

We adapt the String Guessing construction of Braun et al. (2017) to the scale $D_\alpha$ while preserving convexity, global $L$-Lipschitzness, and the growth condition. The construction contains $\Theta(K_{\theta,\lambda})$ hidden signs. Sufficiently small excess error determines these signs, and the String Guessing lower bound therefore yields $\Omega(K_{\theta,\lambda})$ component evaluations. The complete construction is given in Appendix B.2.

*Privacy-dependent hardness.* For the privacy-dependent terms, we first prove a normalized convex private-proxy information lower bound. We then map the normalized class to $\mathfrak{F}_{\theta,\lambda}(L, R)$ by scaling the objective and adding a public Huberized quadratic. The reduction preserves the private-proxy interaction and component-evaluation cost and gives

$$\Omega\left(\max\left\{\sqrt{\frac{d\ell_\alpha}{\rho}}, \frac{d\ell_\alpha}{\bar{m}\rho}\right\}\right).$$

Combining this with the independent $\Omega(K_{\theta,\lambda})$ optimization lower bound proves Theorem 4.1. The complete argument is given in Appendix B.1–B.3.

## 5 Upper Bounds for Gradient Complexity under Growth

We next establish upper bounds on the number of component first-order evaluations under growth. Define

$$\Psi_{\rho,\bar{m}} := \max\left\{1, \sqrt{\frac{d}{\rho}}, \frac{d}{\bar{m}\rho}\right\}, \tag{7}$$

where the term involving $\bar{m}$ is omitted when $\bar{m} = \infty$.

Our main algorithm is a restarted projected-subgradient method. It has two ingredients. First, each private first-order update uses $O(\Psi_{\rho,\bar{m}})$ component evaluations. Second, after each constant-factor reduction in the optimization error, growth converts the new error guarantee into a smaller distance bound to the solution set. Restarting from this smaller radius reduces the number of updates from the generic $\alpha^{-2}$ scale to $K_{\theta,\lambda}$. The method is given in Algorithm 1.

**Public error certificate.** At the beginning of each phase, the restart schedule uses a deterministic upper bound on the expected optimization error. Every feasible point satisfies

$$\Delta_{\max} := \min\left\{LR, \left(\frac{L^\theta}{\lambda}\right)^{1/(\theta-1)}\right\}. \tag{8}$$

Indeed, Lipschitzness and the radius assumption give $F(w) - F^\star \le LR$. On the other hand, if $r = \mathrm{dist}(w, W^\star)$, then Lipschitzness gives $F(w) - F^\star \le Lr$, while growth gives $F(w) - F^\star \ge \lambda r^\theta$. Combining the two yields $r \le (L/\lambda)^{1/(\theta-1)}$, and hence the second term in (8). We initialize the algorithm with $\Delta_0 = \Delta_{\max}$.

**Private first-order estimation.** We next explain how one projected-subgradient update is implemented through the private proxy. At a fixed query point $u$, take $q$ separate proxy replies, each based on a minibatch of size $m$:

$$Y_b(u) = \frac{1}{m}\sum_{\ell=1}^{m} g_{I_{b,\ell}}(u) + Z_b, \qquad Z_b \sim \mathcal{N}\left(0, \frac{2L^2}{\rho m^2} I_d\right), \quad b = 1, \dots, q, \tag{9}$$

where the component indices are independent uniform draws from $[n]$ and the Gaussian noises are independent. The query point $u$ is fixed across the $q$ replies, and we set

$$\widehat{g}(u) := \frac{1}{q}\sum_{b=1}^{q} Y_b(u).$$

Replacing one current first-order answer changes a minibatch average by at most $2L/m$, so each reply in (9) satisfies the conditional $\rho$-zCDP requirement of Definition 3.9. The replies remain separate proxy messages; their joint transcript is not treated as a single $\rho$-zCDP release. The following lemma gives the two properties used by the optimization algorithm.

**Lemma 5.1** (Private first-order estimator)**.** *Fix $u \in \mathcal{W}$, and let $\mathcal{G}$ denote the interaction history before the fresh proxy replies used at $u$. There is a choice of $m$ and $q$ such that*

$$\mathbb{E}[\widehat{g}(u) \mid \mathcal{G}] = \frac{1}{n}\sum_{i=1}^{n} g_i(u) \in \partial F(u), \qquad \mathbb{E}[\|\widehat{g}(u)\|^2 \mid \mathcal{G}] \le 2L^2, \tag{10}$$

*using $O(\Psi_{\rho,\bar{m}})$ component first-order evaluations. Every proxy reply used by the estimator satisfies Definition 3.9.*

Concretely, we use

$$m := \min\left\{\bar{m}, \left\lceil\sqrt{\frac{2d}{\rho}}\right\rceil\right\}, \qquad q := \max\left\{1, \left\lceil\frac{2d}{\rho m^2}\right\rceil\right\}, \tag{11}$$

with the first minimum omitted when $\bar{m} = \infty$. If $\bar{m}$ is at least order $\sqrt{d/\rho}$, one reply is enough. Otherwise, we set $m = \bar{m}$ and average $q = \Theta(d/(\rho\bar{m}^2))$ replies. Hence

$$qm = O\left(\max\left\{1, \sqrt{\frac{d}{\rho}}, \frac{d}{\bar{m}\rho}\right\}\right) = O(\Psi_{\rho,\bar{m}}).$$

The complete second-moment calculation is given in Appendix C.1.

**Why restarting helps.** The algorithm proceeds in phases, with the output of one phase initializing the next. Suppose

$$\mathbb{E}[F(\widehat{w}) - F^\star] \le \Delta.$$

The growth condition and Jensen's inequality give

$$\mathbb{E}\big[\mathrm{dist}(\widehat{w}, W^\star)^2\big] \le \lambda^{-2/\theta}\Delta^{2/\theta}. \tag{12}$$

Thus a smaller error guarantee yields a smaller effective radius around the solution set. A single projected-subgradient phase would use only the initial radius and retain an $\alpha^{-2}$ dependence on the target accuracy. Restarting updates (12) after each constant-factor reduction in error and recalibrates the next phase to the smaller radius.

**Algorithm 1** Restarted Private-Proxy Subgradient Method

1. **Input:** a feasible point $w^{(0)}$ with a public certificate $\Delta_0 > 0$ satisfying $\mathbb{E}[F(w^{(0)}) - F^\star] \le \Delta_0$, and a target excess error $\alpha > 0$.
2. If $\alpha \ge \Delta_0$, **return** $w^{(0)}$. Otherwise, set $J := \lceil \log_2 \frac{\Delta_0}{\alpha} \rceil$ and $\widehat{w}^{(0)} := w^{(0)}$.
3. **For** $j = 0, \dots, J-1$:
   (a) Set $\Delta_j := 2^{-j}\Delta_0$, $\eta_j := \frac{\Delta_j}{8L^2}$, and $T_j := \left\lceil 16L^2\lambda^{-2/\theta}\Delta_j^{-\frac{2(\theta-1)}{\theta}} \right\rceil$.
   (b) Starting from $u_0 = \widehat{w}^{(j)}$, for $s = 0, \dots, T_j - 1$, construct $\widehat{g}_s$ as in Lemma 5.1 at $u_s$ and update $u_{s+1} = \Pi_{\mathcal{W}}\big(u_s - \eta_j \widehat{g}_s\big)$.
   (c) Set the stage output to the average $\widehat{w}^{(j+1)} := \frac{1}{T_j}\sum_{s=0}^{T_j-1} u_s$.
4. **Return** $\widehat{w}^{(J)}$.

**Lemma 5.2** (One-stage contraction)**.** *If phase $j$ starts from*

$$\mathbb{E}[F(\widehat{w}^{(j)}) - F^\star] \le \Delta_j,$$

*then Algorithm 1 satisfies*

$$\mathbb{E}[F(\widehat{w}^{(j+1)}) - F^\star] \le \frac{\Delta_j}{2}. \tag{13}$$

By (12), the phase-start squared distance is at most $\lambda^{-2/\theta}\Delta_j^{2/\theta}$. The standard projected-subgradient bound therefore gives

$$\mathbb{E}[F(\widehat{w}^{(j+1)}) - F^\star] \lesssim \frac{\lambda^{-2/\theta}\Delta_j^{2/\theta}}{\eta_j T_j} + \eta_j L^2.$$

Choosing $\eta_j \asymp \Delta_j/L^2$ so that both terms are a constant fraction of $\Delta_j$ gives

$$T_j = O\left(L^2\lambda^{-2/\theta}\Delta_j^{-\frac{2(\theta-1)}{\theta}}\right).$$

Appendix C.3 gives the complete calculation.

Because $\Delta_j$ decreases geometrically, the phase lengths increase geometrically, and

$$\sum_{j=0}^{J-1} T_j \le 32L^2\lambda^{-2/\theta}\alpha^{-\frac{2(\theta-1)}{\theta}} + J. \tag{14}$$

With $\Delta_0 = \Delta_{\max}$, $J \le 1 + \log_2 K_{\theta,\lambda} \le 2K_{\theta,\lambda}$. Hence

$$\sum_{j=0}^{J-1} T_j = O\Big(L^2\lambda^{-2/\theta}\alpha^{-\frac{2(\theta-1)}{\theta}}\Big).$$

Multiplying by the $O(\Psi_{\rho,\bar{m}})$ component evaluations required for one update gives

$$O\Big(L^2\lambda^{-2/\theta}\alpha^{-\frac{2(\theta-1)}{\theta}}\Psi_{\rho,\bar{m}}\Big). \tag{15}$$

**Recovering the full optimization scale.** When $\alpha < \lambda R^\theta$, the restarted schedule gives the growth-dependent branch of $K_{\theta,\lambda}$. When $\lambda R^\theta \le \alpha < \Delta_{\max}$, one projected-subgradient phase using the public radius $R$ costs

$$O\left(\frac{L^2R^2}{\alpha^2}\Psi_{\rho,\bar{m}}\right).$$

If $\alpha \ge \Delta_{\max}$, the initial certificate already meets the target and no first-order evaluation is required. Unlike Theorem 4.1, the upper bound holds for every dimension.

**Theorem 5.3** (Upper bound under convex growth). *Fix $\theta \geq 2$, $n, d \in \mathbb{N}_{\geq 1}$, $L, R, \lambda, \rho > 0$, and $\bar{m} \in \mathbb{N}_{\geq 1} \cup \{\infty\}$. For every $0 < \alpha \leq \Delta_{\max}$,*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) \leq O(K_{\theta,\lambda}\Psi_{\rho,\bar{m}})\,. \tag{16}$$

**Comparison with the general private-proxy rate.** For general nonsmooth convex objectives, the corresponding optimization scale is $L^2R^2/\alpha^2$ (Menart & Nikolov, 2026). The private-update cost of our method has the same dependence on $d$, $\rho$, and $\bar{m}$ through $\Psi_{\rho,\bar{m}}$. Growth changes the number of required updates. On the growth-active branch, restarting replaces the fixed radius by progressively smaller phase-dependent radii and changes the accuracy dependence from $\alpha^{-2}$ to $\alpha^{-2(\theta-1)/\theta}$. In particular, quadratic growth ($\theta = 2$) gives an $\alpha^{-1}$ dependence.

The distinction between one phase and restarting is important. If $\mathbb{E}[F(w^{(0)}) - F^\star] \leq \Delta_0$, then (12) gives $D_{\Delta_0} := \min\left\{R, (\Delta_0/\lambda)^{1/\theta}\right\}$. A single projected-subgradient phase based on $D_{\Delta_0}$ still has $\alpha^{-2}$ dependence on the final target. The improved exponent arises from updating the distance certificate after every phase.

**Accelerated refinement.** Appendix C.5 gives a second growth-localized method based on smoothing and stochastic acceleration. Combining it with Theorem 5.3 gives (Theorem C13)

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) \leq O\left(K_{\theta,\lambda} + \min\left\{K_{\theta,\lambda}\sqrt{\frac{d}{\rho}}, \frac{K_{\theta,\lambda}^{3/4}d^{5/8}}{\sqrt{\rho}}\right\} + \frac{K_{\theta,\lambda}d}{\bar{m}\rho}\right), \tag{17}$$

where the term involving $\bar{m}$ is omitted when $\bar{m} = \infty$. The accelerated method is used when $d \leq K_{\theta,\lambda}^2$ and the restarted method otherwise.

Under the hypotheses of Theorem 4.1, if

$$\rho \gtrsim \max\left\{\min\left\{d, \frac{d^{5/4}}{\sqrt{K_{\theta,\lambda}}}\right\}, \frac{d}{\bar{m}}\right\}, \tag{18}$$

where the term $d/\bar{m}$ is omitted when $\bar{m} = \infty$, then

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) = \Theta(K_{\theta,\lambda}). \tag{19}$$

See Corollary C14. A privacy-dependent gap remains outside this certified regime.

**Experimental protocols and results.** We compare three schedules: Generic uses the public radius throughout, One-shot uses the smaller radius implied by growth only once, and Restart updates this radius after each stage. We evaluate them on controlled growth objectives with $\theta \in \{2, 3, 4\}$ and on regularized hinge ERM using `a9a` and `rcv1.binary`. We also vary the maximum batch size. The results show the predicted dependence on the maximum batch size and that Restart reaches tighter target errors with fewer component evaluations. These target crossings are empirical diagnostics rather than stopping guarantees. Since most crossings occur before the first restart, the experiments do not separately measure the benefit of repeated restarting. The accelerated method above is not evaluated in these experiments. Full protocols and results are given in Appendix E.

## 6 Conclusion

We studied the gradient complexity of nonsmooth convex optimization under the private-proxy oracle model when the objective satisfies a $(\theta, \lambda)$ growth condition with $\theta \geq 2$. In the growth-active regime, growth changes the accuracy dependence from the general nonsmooth convex rate $\alpha^{-2}$ to $\lambda^{-2/\theta}\alpha^{-2(\theta-1)/\theta}$; in particular, quadratic growth yields an $\alpha^{-1}$ dependence. Our restarted private-proxy projected-subgradient method achieves this improved optimization scale by repeatedly converting excess-error guarantees into smaller distance bounds. On the lower-bound side, we establish a non-private optimization lower bound together with privacy-dependent information lower bounds. Under the hypotheses of Theorem 4.1, the upper and lower bounds match in the regime (18). Overall, our results show that favorable growth geometry can substantially reduce the optimization cost of private-proxy methods while leaving the privacy-dependent information constraints explicit.

### AI USE STATEMENT

Generative AI tools were used for writing and editing and for software assistance. The authors determined and verified the research questions, assumptions, theorem statements, proofs, experimental design, and conclusions; reviewed all AI-assisted material; and take responsibility for the paper.

### ETHICS STATEMENT

This work is a theoretical and empirical study of first-order complexity under a private-proxy interface. It uses only public benchmark datasets (`a9a`, `rcv1.binary`), involves no human subjects, and releases no new data.

The privacy requirement analyzed here is *conditional per-reply* $\rho$-zCDP under replacement of one admissible component response in a proxy interaction. No dataset-level privacy guarantee for a deployed training algorithm is claimed.

### REPRODUCIBILITY STATEMENT

The theoretical claims are stated with their assumptions in Section 3, Section 4, and Section 5, and complete proofs are given in Appendices B and C.

The empirical material in Appendix E is exploratory. Appendices E.1.2 and E.1.1 describe the benchmark and controlled protocols, including the geometry certificates, privacy calibration, evaluation accounting and censoring rules. The archived measurements, configurations and analysis scripts document those exploratory runs. The reported Restart experiments use Algorithm 1. All three methods use the uniform second-moment bound $2L^2$ in their schedules. The measurements illustrate these schedules and are not used to establish the theoretical guarantees.

## REFERENCES

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in $\ell_1$ geometry. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 393–403, 2021a.

Hilal Asi, Daniel Levy, and John C. Duchi. Adapting to function difficulty and growth conditions in private optimization. In *Advances in Neural Information Processing Systems 34*, pp. 19069–19081, 2021b.

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

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7):4709–4724, 2017.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC 2016-B)*, volume 9985 of *Lecture Notes in Computer Science*, pp. 635–658, 2016.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology – EUROCRYPT 2006*, volume 4004 of *Lecture Notes in Computer Science*, pp. 486–503, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284, 2006b.

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

Anatoli Juditsky and Yurii Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth ERM and SCO in subquadratic steps. In *Advances in Neural Information Processing Systems 34*, pp. 4053–4064, 2021.

Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, and Tianbao Yang. Fast rates of ERM and stochastic approximation: Adaptive to error bound conditions. In *Advances in Neural Information Processing Systems*, volume 31, pp. 4678–4689, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/716e1b8c6cd17b771da77391355749f3-Abstract.html.

Michael Menart and Aleksandar Nikolov. On the gradient complexity of private optimization with private oracles. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5115–5158, 2026.

Aaditya Ramdas and Aarti Singh. Optimal rates for stochastic convex optimization under Tsybakov noise condition. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 365–373, 2013.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems 30*, pp. 1119–1129, 2017.

Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private GLMs. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2638–2646, 2021.

Jinyan Su, Lijie Hu, and Di Wang. Faster rates of differentially private stochastic convex optimization. *Journal of Machine Learning Research*, 25(114):1–41, 2024.

Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private LASSO. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems 30*, 2017.

Difei Xu, Meng Ding, Zihang Xiang, Jinhui Xu, and Di Wang. Beyond ordinary Lipschitz constraints: Differentially private optimization with TNC. *Transactions on Machine Learning Research*, 2025.

Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 2017. URL https://proceedings.mlr.press/v70/xu17a.html.

# A Additional Details on the Private-Proxy Model

This section supplements Section 3 with the formal interaction notation used in the proofs and the explicit Rényi-divergence formula deferred from Definition 3.2.

## A.1 Formal private-proxy interaction

We use the standard measurable-protocol convention. Message and random-seed spaces are standard Borel, query and stopping rules are measurable, and proxy response laws are measurable probability kernels. The optimizer's and proxy's random sources are independent of the instance and of one another, and any information shared between them is part of the released transcript. Each conditional reply kernel may be realized using a fresh independent seed. This is only a realization convention and does not restrict repeated queries or reuse of previous query points. Outputs and oracle selections are also measurable. These conventions ensure the existence of regular conditional laws and justify the conditional information chain rules used in Appendices B.3.1 and B.4.2.

Fix an instance $H \in \mathfrak{F}_{\theta,\lambda}(L,R)$ and an optimizer–proxy pair $(\mathcal{A},\mathcal{P})$. At round $t$, let $\mathcal{G}_{t-1}$ denote the sigma-field generated by the released transcript, the optimizer randomness used before round $t$, and the public information available before that round. The optimizer chooses an integer $M_t \le \bar{m}$ and a batch

$$\widehat{Q}_t = \big((w_{t,1}, i_{t,1}), \dots, (w_{t,M_t}, i_{t,M_t})\big) \in (\mathbb{R}^d \times [n])^{M_t}$$

before receiving any current first-order response.

Conditional on $\mathcal{G}_{t-1}$ and the current query batch, the proxy evaluates the corresponding first-order responses and releases a message $Y_t$. The conditional message kernel must satisfy $\rho$-zCDP under replacement of one current response by another admissible response, as in Definition 3.9. The proxy may depend on the current responses and the public interaction history, but on no additional private input. Thus information from previous batches is available only through the released transcript. The optimizer may adapt across rounds, while each current batch is fixed before its responses are evaluated.

Queries may lie outside $\mathcal{W}$, but the final output $\widehat{w}$ must belong to $\mathcal{W}$ almost surely. Each component query counts as one first-order evaluation, including repeated queries to the same component or query point.

For a fixed instance $H$, the expected evaluation cost is

$$\mathrm{Cost}_H(\mathcal{A},\mathcal{P}) := \mathbb{E}_H\left[\sum_t M_t\right].$$

We write $\mathrm{Cost}$ when the instance and optimizer–proxy pair are clear from context. We also let $\overline{T}$ denote the number of distinct query vectors among the $w_{t,j}$. Since each distinct query point requires at least one component evaluation,

$$\overline{T} \le \sum_t M_t$$

pointwise. The quantity $\overline{T}$ is not part of the complexity measure. It is used only in Appendix B.4.2 and in the discussion of Menart & Nikolov (2026) in Appendix B.5. Our lower bounds impose no distinct-query restriction.

All minimax statements use the class and accuracy criterion of Definition 3.12. In particular, admissible procedures terminate almost surely, return a feasible output almost surely, and have integrable excess error.

## A.2 Rényi divergence

For probability measures $P, Q$ on a common measurable space and $\beta > 1$, the Rényi divergence in Definition 3.2 is

$$D_\beta(P \,\|\, Q) := \frac{1}{\beta - 1} \log \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\beta \mathrm{d}Q, \tag{20}$$

with value $+\infty$ if $P$ is not absolutely continuous with respect to $Q$. Throughout, adjacency for a proxy reply means replacement of one current first-order response by another admissible response, as in Definition 3.9.

# B Proofs for the Lower Bound

This section proves Theorem 4.1. The proof has two independent components. Appendix B.2 establishes the non-private optimization cost under growth. Appendix B.3 establishes the privacy-dependent cost by combining a normalized convex information lower bound with a class-wide reduction to the growth class. The two bounds are then combined. Appendix B.4 proves a refined lower bound under additional conditions, and Appendix B.5 discusses the Menart–Nikolov product argument.

## B.1 Main lower bound

Recall

$$D_\alpha := \min\left\{R, \left(\frac{\alpha}{\lambda}\right)^{1/\theta}\right\}, \qquad K_{\theta,\lambda} := \frac{L^2 D_\alpha^2}{\alpha^2}, \tag{21}$$

and define

$$\ell_\alpha := \log\left(2 + \sqrt{K_{\theta,\lambda}}\right). \tag{22}$$

As usual, terms involving $\bar{m}$ are omitted when $\bar{m} = \infty$.

**Theorem B1** (Combined lower bound)**.** *There exist universal constants $c_0, c_1, C_0 > 0$ such that the following holds. Fix $\theta \geq 2$, $n, d \in \mathbb{N}_{\geq 1}$, $L, R, \lambda, \rho > 0$, and $\bar{m} \in \mathbb{N}_{\geq 1} \cup \{\infty\}$. If*

$$0 < \alpha \leq c_0 \min\left\{LR, \left(\frac{L^\theta}{\lambda}\right)^{1/(\theta-1)}\right\}, \qquad d \geq C_0 K_{\theta,\lambda},$$

*then*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) \geq c_1 \max\left\{K_{\theta,\lambda}, \sqrt{\frac{d\ell_\alpha}{\rho}}, \frac{d\ell_\alpha}{\bar{m}\rho}\right\}. \tag{23}$$

Theorem B1 restates Theorem 4.1 with named constants. It combines the non-private lower bound of Appendix B.2 with the privacy-dependent lower bound of Appendix B.3 by a maximum. We prove these two components first and complete the proof of Theorem B1 at the end of Appendix B.3.

## B.2 Non-private lower bound

This subsection establishes the optimization term in Theorem 4.1. We construct a finite family of objectives satisfying the prescribed growth condition and reduce accurate optimization to recovering a hidden sign string. Recall

$$K_{\theta,\lambda} = \frac{L^2 D_\alpha^2}{\alpha^2}, \qquad D_\alpha = \min\left\{R, \left(\frac{\alpha}{\lambda}\right)^{1/\theta}\right\}.$$

**Theorem B2** (Non-private lower bound under growth)**.** *Fix $\theta \geq 2$, $n, d \in \mathbb{N}_{\geq 1}$, and $L, R, \lambda > 0$. Suppose*

$$0 < \alpha \leq \frac{\Delta_{\max}}{11520}, \qquad d \geq K_{\theta,\lambda}.$$

*Then there exists a finite distribution $\Pi$ over instances in $\mathfrak{F}_{\theta,\lambda}(L,R)$ such that every optimizer–proxy pair whose expected excess risk, averaged over $H \sim \Pi$ and its own randomness, is at most $\alpha$ satisfies*

$$\mathbb{E}_{H\sim\Pi}[\mathrm{Cost}_H(\mathcal{A},\mathcal{P})] \geq \frac{K_{\theta,\lambda}}{18432}.$$

*The constants are uniform over $\theta \geq 2$, and the conclusion is independent of $\rho$ and $\bar{m}$.*

### B.2.1 THE FIXED ORACLE AND SIGN RECOVERY

We first construct a fixed selected first-order oracle for which accurate optimization reveals a hidden sign string and each answer can be simulated by one query to a String Guessing Oracle. For $r > 0$, let

$$\mathbb{B}_2(r) := \{w \in \mathbb{R}^d : \|w\| \le r\}.$$

For a hidden string $S \in \{-1, +1\}^M$, the *String Guessing Oracle* of Braun et al. (2017, Section III) takes a guess $a \in \{-1, +1\}^M$ and an ordering $\sigma$ of a subset of the coordinates, and returns the first position in this ordering at which $a$ differs from $S$, or EQUAL if no such position exists.

**Lemma B3** (String guessing with a fixed first-order oracle)**.** *Let $\varepsilon_{\mathrm{sg}} := 1/\sqrt{2}$, and suppose*

$$0 < \varepsilon \le \varepsilon_{\mathrm{sg}}, \qquad M = \lfloor \varepsilon^{-2} \rfloor - 1, \qquad d \ge \varepsilon^{-2}.$$

*For each $s \in \{-1, +1\}^M$, define*

$$b_s(x) := \max_{i \le M} s_i x_i, \qquad x \in \mathbb{R}^d.$$

*Over $\mathbb{B}_2(1)$, its optimum is $b_s^\star = -M^{-1/2}$. There is a deterministic selected first-order oracle $\mathcal{O}_s^{\mathrm{sg}}$ for $b_s$, defined on all of $\mathbb{R}^d$, whose selected subgradient has norm one and whose response to one query can be simulated by one call to the String Guessing Oracle. If $S$ is uniform on $\{-1, +1\}^M$, every randomized algorithm that accesses $S$ only through this oracle, returns $X \in \mathbb{B}_2(1)$, and satisfies*

$$\mathbb{P}\big(b_S(X) - b_S^\star < \varepsilon\big) \ge \frac{3}{4}$$

*uses at least $((3/4)M - 1)/2$ oracle calls in expectation.*

*Proof of Lemma B3.* The hypotheses give $\varepsilon^{-2} \ge 2$, hence $M \ge 1$, and $M + 1 \le \varepsilon^{-2} \le d$, hence $M < d$; so the first $M$ coordinates are available in $\mathbb{R}^d$. For $p = 2$, Theorem V.1 and Lemma V.3 of Braun et al. (2017) use exactly the family $b_s(x) = \max_{i \le M} s_i x_i$ with $M = \lfloor \varepsilon^{-2} \rfloor - 1$. The point with first $M$ coordinates $-s_i/\sqrt{M}$ is feasible and has value $-1/\sqrt{M}$. Conversely, if every $s_i x_i < -1/\sqrt{M}$, then the squared norm of the first $M$ coordinates exceeds one; hence the displayed value is optimal. Since $M = \lfloor \varepsilon^{-2} \rfloor - 1$ forces $M + 1 \le \varepsilon^{-2}$, we have $M\varepsilon^2 < 1$, hence $\varepsilon < 1/\sqrt{M}$ and $b_s^\star + \varepsilon < 0$; the $-1$ in the choice of $M$ ensures this strict inequality. Consequently $s_i x_i \le b_s(x) < b_s^\star + \varepsilon < 0$ for every $i \le M$ at any $x$ with $b_s(x) - b_s^\star < \varepsilon$, so every strict $\varepsilon$-minimum satisfies $s_i = -\operatorname{sgn}(x_i)$ for all $i \le M$ and therefore identifies $s$. Proposition III.3 of that paper states that identifying a uniform $M$-bit string with error probability at most $P_e$ requires at least $((1 - P_e)M - 1)/2$ expected String Guessing calls.

It remains to define the selector for every query outside the feasible set. Given $x \in \mathbb{R}^d$, order the first $M$ coordinates by decreasing $|x_i|$, breaking ties by index, and write the resulting order as $\sigma(1), \dots, \sigma(M)$. Retain all $M$ coordinates, including every zero coordinate. Define the coordinate-indexed guess vector by

$$a_i = \begin{cases} +1, & x_i \le 0, \\ -1, & x_i > 0, \end{cases} \qquad i = 1, \dots, M.$$

Thus $a_i = -\operatorname{sgn}(x_i)$ at every nonzero coordinate, and $a_i = +1$ at zero. Submit $(a, \sigma)$ to the String Guessing Oracle. If its first mismatch occurs at position $j$, put $J = \sigma(j)$; the compared signs are $a_J$ and $S_J$. If $x_J \ne 0$, then $S_J = -a_J = \operatorname{sgn}(x_J)$ and $S_J x_J = |x_J|$. If $x_J = 0$, then $a_J = +1$ and $S_J = -1$, again giving $S_J x_J = 0 = |x_J|$. Every earlier coordinate matched its guess, so its product is $-|x_i|$, including zero coordinates. Every later coordinate satisfies $S_i x_i \le |x_i| \le |x_J|$ by the decreasing-magnitude order. Hence $J$ is active.

If the answer is EQUAL, take $J = \sigma(M)$. Every coordinate matched its guess, so $S_i = a_i$ and $S_i x_i = -|x_i|$ for all $i$. The last coordinate has the smallest magnitude, so its product is largest and $J$ is active. This includes the case of zero coordinates: then $x_J = 0$ and $S_J = a_J = +1$. In both reply cases the selected sign $S_J$ is determined by the reply and the public query point. The response

$$\big(b_S(x), S_J e_J\big) = \big(S_J x_J, S_J e_J\big)$$

is therefore a valid first-order response, has unit subgradient, and uses one String Guessing call. The ordering and selected sign depend on finitely many measurable sign and magnitude comparisons,

so this fixes a Borel-measurable oracle on all of $\mathbb{R}^d$. Apply Proposition III.3 with $P_e = 1/4$. The algorithm's random coins are independent of $S$ and may be included in its side information. □

Lemma B3 gives a hard problem on the Euclidean unit ball without imposing growth. We next rescale this construction and add a public regularizer so that the resulting objective belongs to $\mathfrak{F}_{\theta,\lambda}(L, R)$ while preserving the sign-recovery lower bound.

### B.2.2 Growth construction and proof of Theorem B2

We now scale the sign-encoding construction to the target geometry. Set

$$D := \frac{D_\alpha}{4}, \qquad L_h := \frac{L}{2}, \qquad \mu := 2\lambda(2D)^{\theta-2}. \tag{24}$$

Here $D$ is the construction radius, and the Lipschitz budget $L$ is split between the sign-encoding term and a public regularizer. The class-wide reduction of Appendix B.3 uses the same $D$ and $\mu$. Define

$$q_D(w) := \begin{cases} \dfrac{\mu}{2}\|w\|^2, & \|w\| \le D, \\ \mu D\|w\| - \dfrac{\mu D^2}{2}, & \|w\| > D. \end{cases} \tag{25}$$

Its gradient is

$$\nabla q_D(w) = \begin{cases} \mu w, & \|w\| \le D, \\ \mu D\, w/\|w\|, & \|w\| > D, \end{cases}$$

so $q_D$ is convex, continuously differentiable, and globally $\mu D$-Lipschitz. On the public domain $\mathcal{W}_D := \mathbb{B}_2(D)$, define

$$F_s(w) := L_h D\, b_s(w/D) + q_D(w) = L_h \max_{i\le M} s_i w_i + q_D(w), \tag{26}$$

where

$$\varepsilon_{\mathrm{np}} := \frac{6\alpha}{L_h D}, \qquad M := \lfloor \varepsilon_{\mathrm{np}}^{-2} \rfloor - 1. \tag{27}$$

On $\mathcal{W}_D$, $F_s$ is $\mu$-strongly convex. We next verify that this family belongs to the prescribed growth class and that $\alpha$-accuracy recovers the hidden signs.

*Proof of Theorem B2.* We verify class membership and then reduce accurate optimization to sign recovery.

*Parameter relations.* Write $K = K_{\theta,\lambda}$ and

$$A := \left(\frac{L^\theta}{\lambda}\right)^{1/(\theta-1)}, \qquad p_\theta := \frac{2(\theta-1)}{\theta} \in [1, 2),$$

so that $\Delta_{\max} = \min\{LR, A\}$, $A^{p_\theta} = L^2\lambda^{-2/\theta}$, and

$$K = \min\left\{\left(\frac{LR}{\alpha}\right)^2, \left(\frac{A}{\alpha}\right)^{p_\theta}\right\}.$$

Since $\alpha \le \Delta_{\max}/11520$ and $p_\theta \ge 1$, both terms are at least 11520, so $K \ge 11520$. By (21) and (24), $L_h D = LD_\alpha/8 = \alpha\sqrt{K}/8$. Hence

$$\varepsilon_{\mathrm{np}} = \frac{48}{\sqrt{K}} \le \frac{1}{\sqrt{5}} < \varepsilon_{\mathrm{sg}}, \qquad d \ge K \ge \varepsilon_{\mathrm{np}}^{-2}, \qquad M = \left\lfloor \frac{K}{2304} \right\rfloor - 1 \ge \frac{K}{4608}, \tag{28}$$

where the last inequality uses $x := K/2304 \ge 5$ and $\lfloor x \rfloor - 1 \ge x - 2 \ge x/2$; the same bound gives $M \ge 4$. Thus the hypotheses of Lemma B3 hold with $\varepsilon = \varepsilon_{\mathrm{np}}$.

*Class membership.* We have $2D = D_\alpha/2 \le R$. Also $D_\alpha \le (\alpha/\lambda)^{1/\theta}$, so $\lambda D_\alpha^\theta \le \alpha$ and

$$\max_{w\in\mathcal{W}_D} q_D(w) = \frac{\mu D^2}{2} = 2^{\theta-2}\lambda D^\theta = 2^{-\theta-2}\lambda D_\alpha^\theta \le \frac{\alpha}{16}.$$

Since $\alpha \le A$,

$$\lambda D_\alpha^{\theta-1} \le \lambda^{1/\theta}\alpha^{(\theta-1)/\theta} \le L,$$

and therefore

$$\mu D = \lambda(2D)^{\theta-1} = 2^{1-\theta}\lambda D_\alpha^{\theta-1} \le \frac{L}{2}. \tag{29}$$

On $\mathcal{W}_D$, the regularizer is $\mu$-strongly convex. Hence $F_s$ is $\mu$-strongly convex there. Let $w_s^\star$ be its constrained minimizer. The following argument also applies when $F_s$ is nonsmooth. For $w \in \mathcal{W}_D$ and $a \in (0,1]$ the segment point $aw + (1-a)w_s^\star$ lies in $\mathcal{W}_D$, so strong convexity and constrained minimality give

$$F_s^\star \le F_s\big(aw + (1-a)w_s^\star\big) \le aF_s(w) + (1-a)F_s^\star - \frac{a(1-a)\mu}{2}\left\|w - w_s^\star\right\|^2.$$

Rearranging, dividing by $a > 0$ and letting $a \downarrow 0$ yields

$$F_s(w) - F_s^\star \ge \frac{\mu}{2}\left\|w - w_s^\star\right\|^2.$$

Every two points in $\mathcal{W}_D$ are at distance at most $2D$, and therefore

$$\frac{\mu}{2}\left\|w - w_s^\star\right\|^2 = \lambda(2D)^{\theta-2}\left\|w - w_s^\star\right\|^2 \ge \lambda\left\|w - w_s^\star\right\|^\theta. \tag{30}$$

Thus the objective satisfies the required growth condition.

Use the global selector in Lemma B3. The selected subgradient of (26) is

$$L_h g_s(w/D) + \nabla q_D(w), \qquad \|g_s(w/D)\| = 1.$$

The regularizer has global gradient norm at most $\mu D$, so (29) bounds the selected subgradient by $L$ at every query point in $\mathbb{R}^d$. For any public $n$, let $H_s$ be the instance with $n$ identical components $f_i = F_s$, this fixed oracle, public domain $\mathcal{W}_D$, public starting point $0$, and Euclidean projection onto $\mathcal{W}_D$. Each $H_s$ belongs to $\mathfrak{F}_{\theta,\lambda}(L,R)$; the radius condition follows from $\sup_{w\in\mathcal{W}_D}\|w - w_s^\star\| \le 2D \le R$. Let $\Pi$ be the law of $H_S$ for $S$ uniform on $\{-1,+1\}^M$.

It remains to convert $\alpha$-accuracy into recovery of the hidden signs and to count evaluations.

*Sign recovery and evaluation cost.* Suppose now that the expected excess risk, averaged over $H \sim \Pi$ and the randomness of the pair, is at most $\alpha$. Markov's inequality gives

$$\mathbb{P}\big(F_S(\widehat{w}) - F_S^\star < 4\alpha\big) \ge \frac{3}{4}.$$

The point $v = -DM^{-1/2}\sum_{i\le M} S_i e_i$ lies in $\mathcal{W}_D$ and satisfies $b_S(v/D) = b_S^\star$. Hence

$$F_S^\star \le L_h D\, b_S^\star + \max_{\mathcal{W}_D} q_D.$$

Together with $q_D \ge 0$, this bounds the unregularized error on the event above:

$$b_S(\widehat{w}/D) - b_S^\star < \frac{4\alpha + \max_{\mathcal{W}_D} q_D}{L_h D} \le \frac{5\alpha}{L_h D} < \varepsilon_{\mathrm{np}}. \tag{31}$$

The last inequality is strict because $\alpha > 0$ and $\varepsilon_{\mathrm{np}} = 6\alpha/(L_h D)$.

If $\mathbb{E}_{H\sim\Pi}[\mathrm{Cost}_H(\mathcal{A},\mathcal{P})]$ is infinite, the theorem is immediate. Otherwise the first-order evaluation count is finite almost surely. Simulate the optimizer–proxy interaction using the String Guessing Oracle. For each component query $w$, one String Guessing call returns

$$(b, g) = \mathcal{O}_S^{\mathrm{sg}}(w/D).$$

Feed the original proxy kernel the exact first-order answer

$$\big(L_h D\, b + q_D(w),\ L_h g + \nabla q_D(w)\big).$$

All queries within one proxy batch are chosen before any current first-order answer is observed, so these calls can be serialized without changing the proxy transcript law. The proxy has no hidden instance input beyond the batch of first-order answers. Consequently, one String Guessing call is used per component first-order evaluation, including for queries outside $\mathcal{W}_D$.

At termination $\widehat{w}/D \in \mathbb{B}_2(1)$, and on (31) the packing decoder

$$\widehat{S}_i = -\operatorname{sgn}(\widehat{w}_i), \qquad i \le M,$$

recovers $S$. This decoder is made total by setting $\widehat{S}_i = +1$ whenever $\widehat{w}_i = 0$; the successful packing event itself has no zero coordinate, so the convention cannot change success. On the good event of (31), which has probability at least $3/4$, all $M$ hidden signs are therefore recovered. Since $M \ge 4$, Lemma B3 and (28) yield

$$\mathbb{E}_{H\sim\Pi}[\mathrm{Cost}_H(\mathcal{A},\mathcal{P})] \ge \frac{(3/4)M - 1}{2} \ge \frac{M}{4} \ge \frac{K_{\theta,\lambda}}{18432}.$$

□

### B.3 Private lower bound

This subsection establishes the privacy-dependent term in Theorem 4.1. We first prove a normalized convex private-proxy information lower bound and then transfer it to $\mathfrak{F}_{\theta,\lambda}(L,R)$ through a class-wide growth-preserving reduction. At the end of the subsection, we combine this result with Theorem B2 to prove Theorem B1.

The argument is independent of the product-form lower bound of Menart & Nikolov (2026); Appendix B.5 discusses that result.

#### B.3.1 Normalized convex private-proxy lower bound

For fixed public integers $n, d \ge 1$, let $\mathfrak{C}$ denote the class of finite-sum objectives

$$G(z) := \frac{1}{n}\sum_{i=1}^{n} h_i(z), \qquad z \in \mathcal{U},$$

where $\mathcal{U} \subseteq \mathbb{B}_2(1)$ is a public nonempty compact convex set, each $h_i : \mathbb{R}^d \to \mathbb{R}$ is globally convex and 1-Lipschitz, and each component is equipped with a fixed selected first-order oracle. A feasible starting point and Euclidean projection onto $\mathcal{U}$ are public. Optimizer–proxy pairs follow the protocol of Appendix A, with unit Lipschitz parameter. Accuracy and cost use the conventions of Definition 3.12. The result below continues to hold under the stronger requirement that privacy hold under arbitrary replacement in $\mathbb{R} \times \mathbb{R}^d$.

**Theorem B4** (Normalized convex private-proxy information lower bound)**.** *There exist universal constants $a_0, b_0 > 0$ and $\varepsilon_0 \in (0,1)$ such that the following holds. Fix $\rho > 0$, $\bar{m} \in \mathbb{N}_{\ge 1} \cup \{\infty\}$, and $0 < \varepsilon \le \varepsilon_0$. If $d \ge b_0$, then every optimizer–proxy pair that is uniformly $\varepsilon$-accurate over $\mathfrak{C}$ satisfies*

$$\sup_{H\in\mathfrak{C}} \mathrm{Cost}_H(\mathcal{A},\mathcal{P}) \ge a_0 \max\left\{ \sqrt{\frac{d\log(2+1/\varepsilon)}{\rho}}, \frac{d\log(2+1/\varepsilon)}{\bar{m}\rho} \right\}. \tag{32}$$

*The term involving $\bar{m}$ is omitted when $\bar{m} = \infty$.*

The proof combines a packing of possible minimizers with transcript information bounds for the private-proxy interaction. It does not require a per-direction $\Omega(d)$ information lower bound.

We measure mutual information and relative entropy in nats. The argument uses only the conditional current-batch privacy requirement; it places no restriction on the number of distinct query vectors.

*Proof of Theorem B4.* It suffices to take $\varepsilon_0 = 1/256$ and $b_0 = 1$. Fix a uniformly $\varepsilon$-accurate pair and write $T = \sup_{H\in\mathfrak{C}} \mathrm{Cost}_H(\mathcal{A},\mathcal{P})$. The case $T = +\infty$ is immediate.

*A packing of minimizers.* Let $v_1, \dots, v_N$ be a maximal finite packing of $\mathbb{B}_2(1)$ with pairwise distances strictly greater than $8\varepsilon$. Maximality implies that the balls of radius $8\varepsilon$ centered at these points cover $\mathbb{B}_2(1)$, so comparison of volumes gives

$$N \ge (8\varepsilon)^{-d}.$$

For every $j \in [N]$ take the same public domain $\mathcal{U} = \mathbb{B}_2(1)$, starting point 0, and components

$$h_{i,j}(z) = \|z - v_j\|, \qquad i \in [n].$$

Use the selected subgradient $(z - v_j)/\|z - v_j\|$ away from $v_j$ and $0$ at $v_j$. These are valid instances of $\mathfrak{C}$ for every fixed $n$. If $J$ is uniform on $[N]$, accuracy implies $\mathbb{E}\|\widehat{z} - v_J\| \le \varepsilon$. Nearest-center decoding therefore recovers $J$ with probability at least $3/4$, by Markov's inequality at radius $4\varepsilon$.

*A deterministic evaluation budget.* If $T = 0$, no evaluations occur almost surely on any of these finitely many instances. The output is then independent of $J$, contradicting the preceding decoding guarantee since $N > 2$. Thus assume $T > 0$. Set $B = \lfloor 4T \rfloor$. Modify the interaction by stopping before any batch that would bring the total evaluation count above $B$; on this event output the public point $0$. The batch size is selected before its answers are evaluated, so this stopping rule uses only the optimizer's information and preserves the conditional privacy guarantee. Couple the original and modified procedures using the same randomness. If $C$ is the original total evaluation count, the probability of a forced stop is

$$\mathbb{P}(C > B) = \mathbb{P}(C > 4T) \le \frac{\mathbb{E}C}{4T} \le \frac{1}{4},$$

where the equality uses that $C$ is integer valued. Consequently the modified output still decodes $J$ with probability at least $1/2$.

Let $\mathsf{Tr}_B$ include the modified transcript and the optimizer's random seed, which is independent of $J$ before the interaction. The output is a function of $\mathsf{Tr}_B$. Fano's inequality gives

$$I(J; \mathsf{Tr}_B) \ge \frac{1}{2} \log N - \log 2 \ge c_{\mathrm{pack}}\, d \log(2 + 1/\varepsilon) \tag{33}$$

with a universal $c_{\mathrm{pack}} > 0$. For example, $c_{\mathrm{pack}} = 1/10$ works for $d \ge 1$ and $\varepsilon \le 1/256$: writing $x = \log(1/\varepsilon) \ge 8 \log 2$, the middle expression is at least $d(x/2 - (5/2)\log 2) \ge 3dx/16$, whereas $\log(2 + 1/\varepsilon) \le x + \log 2 \le 9x/8$. Indeed, the displayed Fano bound follows from $H(J \mid \mathsf{Tr}_B) \le \log 2 + \mathbb{P}(\widehat{J} \ne J) \log N$.

*Information in the private interaction.* Fix a history just before a nonempty batch of size $M_t$ is answered. The query batch and the conditional proxy kernel are now fixed. The current answer arrays for any two values of $J$ differ in at most $M_t$ positions. All intermediate arrays obtained by replacing these positions are admissible. Group privacy for zCDP (Bun & Steinke, 2016, Proposition 1.9), followed by the limit of Rényi divergence as its order decreases to 1, gives

$$D_{\mathrm{KL}}(P_j \,\|\, P_k) \le \rho M_t^2,$$

where $P_j$ is the conditional law of the reply on instance $j$. For any conditional distribution $(p_j)$ of $J$, convexity of relative entropy in its second argument yields

$$\sum_j p_j D_{\mathrm{KL}}\left(P_j \,\middle\|\, \sum_k p_k P_k\right) \le \sum_{j,k} p_j p_k D_{\mathrm{KL}}(P_j \,\|\, P_k) \le \rho M_t^2.$$

This is a bound on conditional mutual information under the actual posterior, so it applies to adaptive batches as well. The proxy has no additional private state, as required in Appendix A; its conditional law is determined by the current answers and the released history.

Query choices and stopping decisions, conditional on the previous transcript and optimizer seed, reveal no additional information. After padding terminated runs by constant messages, the chain rule therefore gives

$$I(J; \mathsf{Tr}_B) \le \rho\, \mathbb{E}\left[\sum_t M_t^2\right] \le \rho B^2 \le 16 \rho T^2. \tag{34}$$

Here $\sum_t M_t \le B$ holds on every run. There are at most $B$ nonempty batches. If empty rounds are allowed, apply the chain rule to finite transcript prefixes and pass to the limit; these rounds have no private input and contribute zero conditional information. If $\bar{m} < \infty$, the same calculation also gives

$$I(J; \mathsf{Tr}_B) \le \rho \bar{m}\, \mathbb{E}\left[\sum_t M_t\right] \le \rho \bar{m} B \le 4 \rho \bar{m} T. \tag{35}$$

Combining (33) with (34) and (35) proves the theorem, for example with $a_0 = 1/40$. □

### B.3.2 CLASS-WIDE REDUCTION TO THE GROWTH CLASS

We next show that every normalized source instance can be transformed into an instance of $\mathfrak{F}_{\theta,\lambda}(L,R)$.

**Lemma B5** (Growth-preserving class-wide reduction)**.** *Fix $\theta \geq 2$, $L, R, \lambda > 0$, and $0 < \alpha \leq \Delta_{\max}$. Let*

$$D_\alpha := \min\left\{R, \left(\frac{\alpha}{\lambda}\right)^{1/\theta}\right\}, \qquad D := \frac{D_\alpha}{4},$$

*and set*

$$s := \frac{LD}{2}, \qquad \mu := 2\lambda(2D)^{\theta-2}.$$

*Define the public regularizer*

$$q_D(w) := \begin{cases} \dfrac{\mu}{2}\|w\|^2, & \|w\| \leq D, \\ \mu D\|w\| - \dfrac{\mu D^2}{2}, & \|w\| > D. \end{cases} \tag{36}$$

*For any source instance $H \in \mathfrak{C}$, define*

$$f_i(w) := s h_i(w/D) + q_D(w), \qquad \mathcal{W} := D\mathcal{U}, \tag{37}$$

*and let*

$$F(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w).$$

*Then:*

1. *the transformed instance belongs to $\mathfrak{F}_{\theta,\lambda}(L,R)$;*
2. *if $\widehat{w} \in \mathcal{W}$ satisfies*
$$\mathbb{E}[F(\widehat{w}) - F^\star] \leq \alpha,$$
*then $\widehat{z} := \widehat{w}/D \in \mathcal{U}$ satisfies*
$$\mathbb{E}[G(\widehat{z}) - G^\star] \leq \frac{17}{2\sqrt{K_{\theta,\lambda}}}; \tag{38}$$
3. *the transformation preserves the conditional $\rho$-zCDP requirement, batch sizes, the maximum batch size $\bar{m}$, and total component-evaluation cost.*

*Proof. Class membership.*

Since $D_\alpha \leq (\alpha/\lambda)^{1/\theta}$ and $\alpha \leq \Delta_{\max}$,

$$\lambda D_\alpha^{\theta-1} \leq \lambda^{1/\theta}\alpha^{(\theta-1)/\theta} \leq L.$$

Therefore

$$\sup_{w\in\mathbb{R}^d}\|\nabla q_D(w)\| = \mu D = \lambda(2D)^{\theta-1} = 2^{1-\theta}\lambda D_\alpha^{\theta-1} \leq \frac{L}{2}.$$

The scaled source term in (37) is $L/2$-Lipschitz, so every $f_i$ is globally convex and $L$-Lipschitz.

Because $\mathcal{W} \subseteq \mathbb{B}_2(D)$, the regularizer is quadratic on the entire feasible set and $F$ is $\mu$-strongly convex there. Let $w^\star$ be its constrained minimizer. Strong convexity and constrained minimality give

$$F(w) - F^\star \geq \frac{\mu}{2}\|w - w^\star\|^2.$$

Since any two points of $\mathcal{W}$ are at distance at most $2D$,

$$F(w) - F^\star \geq \lambda(2D)^{\theta-2}\|w - w^\star\|^2 \geq \lambda\|w - w^\star\|^\theta.$$

Moreover,

$$\sup_{w\in\mathcal{W}} \operatorname{dist}(w, W^\star) \leq 2D \leq R.$$

Thus the transformed instance belongs to $\mathfrak{F}_{\theta,\lambda}(L,R)$.

*Accuracy transfer.* For every $w \in \mathcal{W}$,

$$0 \le q_D(w) \le \frac{\mu D^2}{2} = 2^{\theta-2}\lambda D^\theta = 2^{-\theta-2}\lambda D_\alpha^\theta \le \frac{\alpha}{16}.$$

If $z^\star$ minimizes $G$ over $\mathcal{U}$, then

$$F^\star \le sG(z^\star) + q_D(Dz^\star),$$

and hence, for every $w \in \mathcal{W}$,

$$s\big(G(w/D) - G^\star\big) \le F(w) - F^\star + \frac{\alpha}{16}.$$

Since

$$K_{\theta,\lambda} = \frac{L^2 D_\alpha^2}{\alpha^2}, \qquad s = \frac{\alpha\sqrt{K_{\theta,\lambda}}}{8},$$

taking expectations proves (38).

*Preservation of the private-proxy interaction.*

At a transformed query $w \in \mathbb{R}^d$, query the source oracle at $z = w/D$. Map each source response $(a, g) \in \mathbb{R} \times \mathbb{R}^d$ entrywise to

$$(a, g) \longmapsto \Big(sa + q_D(w),\, \frac{s}{D}\Pi_{\mathbb{B}_2(1)}(g) + \nabla q_D(w)\Big). \tag{39}$$

Genuine source answers satisfy $\|g\| \le 1$, so the projection leaves them unchanged, and the transformed subgradient is $(s/D)g_i(w/D) + \nabla q_D(w) \in \partial f_i(w)$. It is a fixed Borel-measurable selection on all of $\mathbb{R}^d$. For every $(a, g) \in \mathbb{R} \times \mathbb{R}^d$, the transformed gradient has norm at most

$$\frac{s}{D} + \|\nabla q_D(w)\| \le \frac{L}{2} + \frac{L}{2} = L,$$

so (39) maps every element of $\mathbb{R} \times \mathbb{R}^d$ to an admissible response. The map is deterministic and acts separately on each response. Replacing one source response by an arbitrary element of $\mathbb{R} \times \mathbb{R}^d$ therefore replaces at most one transformed response by another admissible response. Precomposing the original proxy kernel with this map therefore gives a source proxy whose replies are $\rho$-zCDP in the sense required of source pairs. The proxy may still process the entire transformed batch jointly; the simulator uses only current source answers and the released history.

The coordinate map $w \leftrightarrow w/D$ is one-to-one, so batch sizes and component-evaluation counts are unchanged. The public starting point and Euclidean projection also transport under scaling:

$$\Pi_{D\mathcal{U}}(w) = D\Pi_{\mathcal{U}}(w/D).$$

Using the same optimizer and proxy random seeds couples the two interactions round by round. Stopping decisions and the final output transport under scaling, preserving almost-sure termination and feasibility. The accuracy inequality above also preserves integrability of the excess error. This proves the final claim. □

### B.3.3 Private lower bound under growth

**Theorem B6** (Private lower bound under growth)**.** *There exist universal constants $c_0, c_{\mathrm{priv}}, C_0 > 0$ such that the following holds. Fix $\theta \ge 2$, $n, d \in \mathbb{N}_{\ge 1}$, $L, R, \lambda, \rho > 0$, and $\bar{m} \in \mathbb{N}_{\ge 1} \cup \{\infty\}$. If*

$$0 < \alpha \le c_0 \min\left\{LR, \left(\frac{L^\theta}{\lambda}\right)^{1/(\theta-1)}\right\}, \qquad d \ge C_0 K_{\theta,\lambda},$$

*then every optimizer–proxy pair that is uniformly $\alpha$-accurate on $\mathfrak{F}_{\theta,\lambda}(L,R)$ satisfies*

$$\sup_{H \in \mathfrak{F}_{\theta,\lambda}(L,R)} \mathrm{Cost}_H(\mathcal{A}, \mathcal{P}) \ge c_{\mathrm{priv}} \max\left\{\sqrt{\frac{d\ell_\alpha}{\rho}}, \frac{d\ell_\alpha}{\bar{m}\rho}\right\}. \tag{40}$$

*The term involving $\bar{m}$ is omitted when $\bar{m} = \infty$.*

*Proof.* Fix an $\alpha$-accurate optimizer–proxy pair $(\mathcal{A}, \mathcal{P})$ and write

$$T := \sup_{H \in \mathfrak{F}_{\theta,\lambda}(L,R)} \mathrm{Cost}_H(\mathcal{A}, \mathcal{P}).$$

If $T = +\infty$, there is nothing to prove.

Apply Lemma B5 to each source instance in $\mathfrak{C}$. Since $(\mathcal{A}, \mathcal{P})$ is uniformly $\alpha$-accurate on $\mathfrak{F}_{\theta,\lambda}(L, R)$, the induced source procedure is uniformly $\varepsilon$-accurate on $\mathfrak{C}$, where

$$\varepsilon := \frac{17}{2\sqrt{K_{\theta,\lambda}}}.$$

The transformation preserves component-evaluation cost.

For $c_0 \le 1$, the accuracy assumption gives $K_{\theta,\lambda} \ge c_0^{-1}$ uniformly over $\theta \ge 2$: the two terms defining $K_{\theta,\lambda}$ are at least $c_0^{-2}$ and $c_0^{-2(\theta-1)/\theta}$, respectively. Thus taking $c_0 \le (2\varepsilon_0/17)^2$ ensures $\varepsilon \le \varepsilon_0$. The assumption $d \ge C_0 K_{\theta,\lambda}$, with $C_0$ sufficiently large, implies the dimension requirement of Theorem B4.

Therefore

$$T \ge a_0 \max\left\{ \sqrt{\frac{d \log(2 + 1/\varepsilon)}{\rho}}, \frac{d \log(2 + 1/\varepsilon)}{\bar{m}\rho} \right\}.$$

Since

$$\frac{1}{\varepsilon} = \frac{2\sqrt{K_{\theta,\lambda}}}{17},$$

and $K_{\theta,\lambda}$ is bounded below by a universal constant under the accuracy assumption, there is a universal constant $c_{\log} > 0$ such that

$$\log\left(2 + \frac{2\sqrt{K_{\theta,\lambda}}}{17}\right) \ge c_{\log} \log\left(2 + \sqrt{K_{\theta,\lambda}}\right) = c_{\log} \ell_\alpha.$$

Absorbing $c_{\log}$ into the universal constant proves (40). □

We can now complete the proof of Theorem B1.

*Proof of Theorem B1.* Fix an optimizer–proxy pair $(\mathcal{A}, \mathcal{P})$ that is uniformly $\alpha$-accurate on $\mathfrak{F}_{\theta,\lambda}(L, R)$, and let

$$T := \sup_{H \in \mathfrak{F}_{\theta,\lambda}(L,R)} \mathrm{Cost}_H(\mathcal{A}, \mathcal{P}).$$

By Theorem B2, after decreasing $c_0$ and increasing $C_0$ by universal constants if necessary, there exists a finite distribution $\Pi$ supported on $\mathfrak{F}_{\theta,\lambda}(L, R)$ such that

$$\mathbb{E}_{H\sim\Pi}\mathbb{E}_H[F_H(\widehat{w}) - F_H^\star] \le \alpha.$$

Hence

$$T \ge \mathbb{E}_{H\sim\Pi}[\mathrm{Cost}_H(\mathcal{A}, \mathcal{P})] \ge c_{\mathrm{np}} K_{\theta,\lambda} \tag{41}$$

for a universal constant $c_{\mathrm{np}} > 0$.

Since $(\mathcal{A}, \mathcal{P})$ is uniformly $\alpha$-accurate, it satisfies the averaged accuracy condition of Theorem B2 for the hard distribution $\Pi$ provided by that theorem. Therefore, after decreasing $c_0$ and increasing $C_0$ by universal constants if necessary,

$$T \ge \mathbb{E}_{H\sim\Pi}[\mathrm{Cost}_H(\mathcal{A}, \mathcal{P})] \ge c_{\mathrm{np}} K_{\theta,\lambda} \tag{42}$$

for a universal constant $c_{\mathrm{np}} > 0$.

The two lower bounds may be witnessed by different hard instances, but both bound the same worst-case cost $T$. Therefore (23) holds with

$$c_1 := \min\{c_{\mathrm{np}}, c_{\mathrm{priv}}\}.$$

Taking the infimum over all uniformly accurate optimizer–proxy pairs completes the proof. □

### B.4 A REFINED LOWER BOUND

Theorem B1 gives the class-wide lower bound used in the main result. We next derive a refined bound under additional dimension and batch-size conditions by analyzing a growth-scaled version of the Menart–Nikolov hard construction. The proof establishes its own geometric and information estimates and does not invoke the product theorem or the information-demand lemmas discussed in Appendix B.5.

Write

$$K := K_{\theta,\lambda}, \qquad \ell_\alpha := \log(2+\sqrt{K}),$$

and define

$$\begin{aligned} p := \frac{2(\theta-1)}{\theta}, \qquad Q := \left(\frac{LR}{\alpha}\right)^2, \qquad K_g := L^2\lambda^{-2/\theta}\alpha^{-p}, \\ K = \min\{Q, K_g\}, \qquad \kappa_\theta := \min\left\{Q^{1/3}, K_g^{\theta/(3\theta-2)}\right\}. \end{aligned} \tag{43}$$

The auxiliary scale $\kappa_\theta$ determines the batch-size term in the refined lower bound. Since

$$1 \le p < 2, \qquad \frac{\theta}{3\theta-2} = \frac{1}{1+p},$$

for $K \ge 1$,

$$K^{1/3} \le \kappa_\theta \le \sqrt{K}. \tag{44}$$

For the upper inequality, use $Q^{1/3} \le \sqrt{Q}$ when $K = Q$ and $K_g^{1/(1+p)} \le \sqrt{K_g}$ when $K = K_g$.

**Theorem B7** (Refined lower bound under growth)**.** *Fix $\theta \ge 2$, $n, d \in \mathbb{N}_{\ge 1}$, $L, R, \lambda, \rho > 0$, and $\bar{m} \in \mathbb{N}_{\ge 1} \cup \{\infty\}$. Under the measurable-protocol convention of Appendix A, suppose*

$$0 < \alpha \le \frac{\Delta_{\max}}{16384}, \qquad d \ge 2\max\left\{K\ell_\alpha, \frac{K\ell_\alpha^2}{\rho}\right\}, \qquad \rho\bar{m} \ge \ell_\alpha \quad \text{if } \bar{m} < \infty. \tag{45}$$

*Then*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) \ge \frac{1}{8046}\max\left\{K, \sqrt{\frac{Kd}{\rho}}, \frac{\kappa_\theta d}{\bar{m}\rho}\right\}. \tag{46}$$

*The term involving $\bar{m}$ is omitted when $\bar{m} = \infty$. The constants are universal, and the result imposes neither a distinct-query restriction nor an upper bound on $\rho$.*

#### B.4.1 HARD CONSTRUCTION UNDER GROWTH

Let $d_V = \lfloor d/2 \rfloor$ and take an integer $k \ge 1$ with $d_V + k < d$. Draw a Haar $d_V$-dimensional subspace $V$ and, conditional on $V$, a Haar orthonormal frame $X_1, \dots, X_k$ in $V^\perp$. For $D > 0$, set $a = D/\sqrt{k}$ and $b = L/2$, and define identical components on the public domain $\mathcal{W} = \mathbb{B}_2(D)$ by

$$f_i(w) = b\,h(w), \qquad h(w) = \max\left\{\max_{j\le k}|\langle w, X_j\rangle - a|, 2\|\Pi_V w\|\right\}, \qquad i \in [n]. \tag{47}$$

The starting point is 0 and projection onto the ball is public. The oracle selects the smallest active coordinate index, when one exists, and otherwise the norm branch. Its coordinate subgradient is $b\,\mathrm{sgn}(\langle w, X_j\rangle - a)X_j$, with $\mathrm{sgn}(0) = 1$; its norm-branch subgradient is $2b\Pi_V w/\|\Pi_V w\|$. A strictly active norm branch has nonzero projection, so this formula is well defined. These rules give a fixed Borel selected first-order oracle on all of $\mathbb{R}^d$, including ties and points outside $\mathcal{W}$.

The next lemma verifies membership in the growth class and shows that sufficiently accurate outputs must correlate with every hidden direction.

**Lemma B8** (Growth and directional accuracy)**.** *The instance* (47) *belongs to $\mathfrak{F}_{\theta,\lambda}(L,R)$ whenever*

$$2D \le R, \qquad \lambda(2D)^{\theta-1}\sqrt{k} \le b. \tag{48}$$

*Equivalently parameterize $D = 2\alpha P\sqrt{k}/L$, so $P = ab/\alpha$. The following conditions suffice uniformly over $\theta \ge 2$:*

$$kP^2 \le Q/16, \qquad kP^p \le K_g/16. \tag{49}$$

*For $P > 4$, put $\delta = 4/P$ and*

$$E_j = (V + \operatorname{span}\{X_i : i \neq j\})^{\perp}, \qquad m = d - d_V - k + 1, \qquad t_k = \frac{1-\delta}{\sqrt{1 + (k-1)(2\delta - \delta^2)}}.$$

*Every feasible $w$ with $F(w) - F^\star \leq 4\alpha$ satisfies*

$$\Pi_{E_j} w \neq 0, \qquad \left\langle X_j, \frac{\Pi_{E_j} w}{\|\Pi_{E_j} w\|} \right\rangle \geq t_k, \qquad j \in [k]. \tag{50}$$

*Proof.* The objective $h$ is globally convex and 2-Lipschitz, so each component is globally convex and $L$-Lipschitz. Its unique constrained minimizer is $w_* = a\sum_j X_j$, with $\|w_*\| = D$ and $F(w_*) = 0$. Indeed, zero loss fixes all $k$ coordinates to $a$; their squared norms already sum to $D^2$, leaving no other feasible component. For any $w \in \mathcal{W}$,

$$\|w - w_*\|^2 \leq 2a \sum_j (a - \langle w, X_j \rangle) \leq 2D\sqrt{k}\, h(w).$$

Together with $\|w - w_*\| \leq 2D$, this gives

$$F(w) - F^\star \geq \frac{b}{(2D)^{\theta-1}\sqrt{k}} \|w - w_*\|^\theta.$$

The radius bound and growth now follow from (48). Substituting $D$ gives $kP^2 \leq Q/16$ for the radius and $kP^p \leq 2^{-(4-2/\theta)} K_g$ for growth. Since $2^{-(4-2/\theta)} \geq 1/16$, (49) suffices.

Accuracy gives $a(1-\delta) \leq \langle w, X_i \rangle \leq a(1+\delta)$ for every $i$. For $v = \Pi_{E_j} w$ this implies

$$\langle X_j, v \rangle \geq a(1-\delta) > 0, \qquad \|v\|^2 \leq D^2 - (k-1)a^2(1-\delta)^2 = a^2[1 + (k-1)(2\delta - \delta^2)].$$

Dividing proves (50). □

The directional condition in (50) identifies a spherical cap for each hidden direction. We next bound its probability under the conditional prior.

*Conditional sphere probabilities.* Given $C_j = (\Pi_V, (X_i)_{i \neq j})$, the conditional prior of $X_j$ is uniform on the unit sphere of $E_j$, which has dimension $m$. One way to see this precisely is to generate a Haar orthogonal matrix by Gram–Schmidt applied to independent Gaussian columns. Column exchangeability allows the columns specifying $V$ and $X_{-j}$ to be generated first. The next direction has law $\Pi_{E_j} Z / \|\Pi_{E_j} Z\|$ for a fresh standard Gaussian $Z$. This kernel depends only on $C_j$, so the tower property gives the asserted conditional law. This is a statement about the conditional prior, before observing the optimizer's transcript.

If $U$ is uniform on the sphere in $\mathbb{R}^m$, then, for a fixed unit vector $v$ and $0 < t < 1$,

$$\mathbb{P}(\langle U, v \rangle \geq t) \leq \min\left\{ e^{-mt^2/2}, \frac{(1-t^2)^{(m-1)/2}}{t\sqrt{2\pi(m-1)}} \right\}. \tag{51}$$

For completeness, the even moments of $U_1$ are $\mathbb{E}U_1^{2r} = (1 \cdot 3 \cdots (2r-1))/(m(m+2)\cdots(m+2r-2))$. Thus $\mathbb{E}e^{sU_1} \leq e^{s^2/(2m)}$, and the Chernoff bound gives the first estimate. For the second, the density is $c_m(1-x^2)^{(m-3)/2}$ on $(-1,1)$, where $c_m = 1/\mathrm{B}\left(\frac{1}{2}, \frac{m-1}{2}\right) \leq \sqrt{(m-1)/(2\pi)}$ by log-convexity of the Gamma function. Integrating and using $x \geq t$ gives $\int_t^1 (1-x^2)^{(m-3)/2}\, dx \leq (1-t^2)^{(m-1)/2}/(t(m-1))$, as required.

For a fixed decoder output, the event in (50) consequently has conditional-prior probability at most $e^{-\Gamma}$, where the two choices used below are

| $P$ | accuracy implication | $\Gamma$ |
|---|---|---|
| 12 | $t_k^2 \geq 4/(9k)$ | $2m/(9k)$ |
| $32k$ | $1 - t_k^2 \leq 1/4$ | $(m-1)\log 2$. |

(52)

The second row uses $m \geq 2$, for which $t_k\sqrt{2\pi(m-1)} > 1$.

### B.4.2 PROOF OF THEOREM B7

We first convert the directional requirement above into an expected-evaluation lower bound. The next lemma is used for both choices of $(P, \Gamma)$ in (52).

**Lemma B9** (Cost of learning a frame)**.** *Consider a valid family* (47)*, with $P > 4$ and a common conditional-prior cap exponent $\Gamma > 0$ as above. For any uniformly $\alpha$-accurate optimizer–proxy pair, let $T$ be its expected evaluation count averaged over this frame prior, and set $h_k = 1 + \log k$. If $J = \Gamma/2 - \log 2 > 0$, then*

$$T \geq \frac{kJ}{h_k + 4\min\{\sqrt{\rho J}, \rho\bar{m}\}}, \tag{53}$$

*with only $\sqrt{\rho J}$ in the minimum when $\bar{m} = \infty$. Independently of privacy,*

$$T \geq \frac{k(3\Gamma/4 - \log 2)}{\Gamma + h_k}. \tag{54}$$

*Proof.* If $T = \infty$ there is nothing to prove. Label each current oracle answer by its selected coordinate $j \in [k]$, or by $\perp$ for the norm branch. Let $Z_j$ be the total number of slots labelled $j$, $a_j = \mathbb{E}Z_j$, and $\overline{T}$ the number of distinct query vectors. Every repeated slot is counted in $Z_j$ and in the evaluation cost, so

$$\sum_j a_j \leq T, \qquad \mathbb{E}\overline{T} \leq T. \tag{55}$$

Given $C_j$, a query, and the bit specifying whether its label is $j$, the full label is known: if the bit is zero, maximize the other coordinate branches and the norm branch using the same tie rule. In that case the entire first-order answer, including its value, is also known. This locality holds for all query points.

*Records stopped at a direction budget.* Fix a deterministic threshold $N_j \geq 0$ and a finite round horizon $H$. Record the optimizer seed, the batches, and all their labels until the first batch that would make the cumulative $j$-labelled count exceed $N_j$. Keep the labels of this triggering batch, omit its reply, and pad all later entries. Include all earlier replies. Call the resulting record $R_j$. Its decoder simulates the optimizer from the recorded seed and replies, returning the optimizer's output if it stops within the horizon without requesting an omitted reply, and returning $0$ otherwise. On inconsistent records also return $0$. This is a total measurable map and agrees with the original output on $\{Z_j \leq N_j,\ N_{\mathrm{stop}} \leq H\}$.

Let $\Lambda_j(N_j, H)$ be the sum of conditional label entropies in this record, including the triggering batch, given $C_j$ and the preceding recorded data. For an included reply, condition also on all its current labels. Replacing its $s_j$ entries labelled $j$ by $(0, 0)$ leaves a reference array determined by these conditioning variables. Group zCDP bounds the KL divergence to this reference reply law by $\rho s_j^2$. The identity $\mathbb{E}D_{\mathrm{KL}}(P_{Y|X}\|Q) = I(X;Y) + D_{\mathrm{KL}}(P_Y\|Q)$ bounds conditional information by the same quantity for every posterior of $X_j$. No independence of reused answers is needed. The information chain rule and the pathwise budget therefore give

$$I(X_j; R_j \mid C_j) \leq \Lambda_j(N_j, H) + \rho\min\{N_j^2, \bar{m}N_j\}. \tag{56}$$

At infinity omit $\bar{m}N_j$. When $N_j = 0$, every included answer array is determined by the conditioning variables, so the reply contribution is zero even for an arbitrary non-private proxy. Queries, stopping tags, and padding add no information beyond the recorded seed, labels, and replies.

*A shared label-entropy budget.* Consider the full augmented execution, exposing all batch labels in slot order before each reply. Just before one label, let $\mathcal{G}$ be this common preceding history, including the optimizer seed and the whole current batch, and put

$$p_j = \mathbb{P}(\text{label} = j \mid \mathcal{G}, C_j), \qquad S = \sum_j p_j.$$

The side information differs with $j$, so $S \leq 1$ need not hold pointwise. The tower property does give $\mathbb{E}[S \mid \mathcal{G}] \leq 1$. Locality identifies the conditional label entropy with $h_2(p_j)$, where $h_2$ is binary entropy. For any $p_j \in [0, 1]$,

$$\sum_j h_2(p_j) \leq S(1 + \log k) - S\log S =: \phi_k(S), \qquad \phi_k(0) = 0.$$

Here $\phi_k$ is concave on $[0,k]$ and nondecreasing on $[0,1]$. Conditional Jensen consequently bounds the expected sum of these entropies by $\phi_k(1) = h_k$. At a repeated query vector, its fixed label is already in the augmented history, so all these entropies vanish.

While a direction record remains active, its preceding history equals the full augmented history. The two conditioning sigma-fields have the same traces on this measurable activity event, so their conditional label laws agree there. Its entropy contribution is therefore the full-history contribution multiplied by its activity indicator. Discarding these indicators only increases the sum. Keeping every label in the triggering batch makes this argument valid throughout that batch. Summing nonnegative contributions over directions and slots, using Tonelli for random batch lengths, proves

$$\sum_{j=1}^{k} \Lambda_j(N_j, H) \le h_k \, \mathbb{E}\overline{T} \le h_k T. \tag{57}$$

*Information required by the decoder.* For any decoded output, apply (50) to its projection on $E_j$, declaring the cap event false if the projection is zero. For each fixed record and $C_j$, this event has probability at most $e^{-\Gamma}$ under the conditional prior of $X_j$. Under the conditional product of the prior and record marginal, it has the same upper bound. Binary data processing for KL therefore gives

$$I(X_j; R_j \mid C_j) \ge \mathbb{P}(\text{cap event})\Gamma - \log 2.$$

Indeed, binary divergence with success probabilities $q_1, q_0$ is at least $q_1 \log(1/q_0) - \log 2$. This uses a prior probability, not a posterior-uniformity assertion.

Take $N_j = 4a_j$. Accuracy, Markov's inequality for $Z_j$, and decoder agreement give cap-event probability at least $1/2 - \delta_H$, where $\delta_H = \mathbb{P}(N_{\text{stop}} > H) \to 0$. If $a_j = 0$, then $Z_j = 0$ almost surely and no division is needed. Combining demand, (56), and (57), then letting $H \to \infty$, gives

$$\sum_j (J - 16\rho a_j^2)_+ \le h_k T, \qquad \sum_j (J - 4\rho\bar{m} a_j)_+ \le h_k T$$

for the second inequality when $\bar{m} < \infty$. For $\beta, x \ge 0$ and $J > 0$, $(J - \beta x^2)_+ \ge J - \sqrt{\beta J}\, x$. Using (55) now gives both $kJ \le (h_k + 4\sqrt{\rho J})T$ and $kJ \le (h_k + 4\rho\bar{m})T$, proving (53). Only finite records and deterministic thresholds were used; no second moment of the original cost is assumed.

Finally, take $N_j = 0$. The probability of omitting a needed reply is at most $\mathbb{P}(Z_j \ge 1) \le a_j$, so the same argument gives $\Lambda_j(0, H) \ge (3/4 - a_j - \delta_H)\Gamma - \log 2$. Sum and let $H \to \infty$ to obtain $(\Gamma + h_k)T \ge k(3\Gamma/4 - \log 2)$, which is (54). □

*Proof of Theorem B7.* Fix a uniformly $\alpha$-accurate pair and write $T_* = \sup_{H \in \mathfrak{F}_{\theta,\lambda}(L,R)} \mathrm{Cost}_H(\mathcal{A}, \mathcal{P})$. Every hard-prior mean cost below is at most $T_*$. If $T_* = \infty$, the conclusion is immediate. Put $s = \Delta_{\max}/\alpha$. The definitions give $s \le K \le s^2$, so (45) implies

$$K \ge 16384, \qquad \ell_\alpha > 1, \qquad \log K \le 2\ell_\alpha, \qquad d \ge K.$$

*Non-private and square-root terms.* Take $P = 12$ and $k = \lfloor K/2304 \rfloor$ in (47). Since $12^p \le 144$, (49) holds. Also $d_V + k < d$ and

$$\frac{K}{2682} \le k \le \frac{K}{2304}.$$

For the lower inequality, $\lfloor x \rfloor \ge 7x/8$ for $x \ge 64/9$, and $7/18432 > 1/2682$. Use $\Gamma = 2m/(9k)$ from (52). Writing $h_k = 1 + \log k \le \log K \le 2\ell_\alpha$, the first dimension condition yields

$$\Gamma \ge \frac{d}{9k} - \frac{2}{9} \ge 512\ell_\alpha - \frac{2}{9} \ge 3h_k + 3.$$

It follows that $(3\Gamma/4 - \log 2)/(\Gamma + h_k) \ge 1/3$, and (54) gives

$$T_* \ge k/3 \ge K/8046. \tag{58}$$

For the private estimate, $d/k \ge 2304$ implies

$$J = \frac{\Gamma}{2} - \log 2 \ge \frac{d}{18k} - \frac{1}{9} - \log 2 \ge \frac{d}{18.2k} > 0.$$

Using the denominator with $\sqrt{\rho J}$ in (53), and $x/(u+v) \geq \frac{1}{2}\min\{x/u, x/v\}$ for positive $u, v$, gives

$$T_* \geq \min\left\{\frac{d}{37\log K}, \frac{1}{1768}\sqrt{\frac{Kd}{\rho}}\right\}.$$

The constants follow from $2(18.2) < 37$ and $64(18.2)(2682) < 1768^2$. The second dimension condition removes this minimum:

$$\rho d \geq 2K\ell_\alpha^2 \geq K(\log K)^2(37/1768)^2.$$

Consequently

$$T_* \geq \frac{1}{1768}\sqrt{\frac{Kd}{\rho}}. \tag{59}$$

*The batch-size term.* Suppose $\bar{m} < \infty$, and now choose the different hard prior with

$$k = \left\lfloor 2^{-14/3}\kappa_\theta \right\rfloor, \qquad P = 32k.$$

By (44), $\kappa_\theta \geq 16384^{1/3} = 2^{14/3}$. Thus the unrounded value is at least one, and $\lfloor x \rfloor \geq x/2$ for $x \geq 1$ gives

$$\frac{\kappa_\theta}{50.8} \leq k \leq 2^{-14/3}\kappa_\theta < \frac{\sqrt{K}}{25} \leq \frac{\sqrt{d}}{25}. \tag{60}$$

In particular this construction has at least one direction even at the smallest allowed $\kappa_\theta$. Membership follows from

$$kP^2 = 1024k^3 \leq Q/16$$

and, since $\kappa_\theta^{1+p} \leq K_g$,

$$kP^p = 32^p k^{1+p} \leq 2^{5p-14(1+p)/3}K_g = 2^{(p-14)/3}K_g \leq K_g/16.$$

Also $d_V + k < d$ and $m \geq 3$. Use $\Gamma = (m-1)\log 2$ from (52). Then

$$J \geq (d/4 - k/2 - 1)\log 2 \geq 0.17d.$$

For the last inequality, $d \geq 16384$ and (60) give $k \leq d/3200$; $(1/4 - 1/6400 - 1/16384)\log 2 > 0.17$. The denominator with $\rho\bar{m}$ in (53) gives

$$T_* \geq \min\left\{\frac{\kappa_\theta d}{600\log(e\kappa_\theta)}, \frac{\kappa_\theta d}{2400\rho\bar{m}}\right\},$$

because $h_k \leq \log(e\kappa_\theta)$, $0.17/(2\cdot 50.8) > 1/600$, and $0.17/(8\cdot 50.8) > 1/2400$. Finally,

$$\log(e\kappa_\theta) \leq 1 + \ell_\alpha \leq 4\ell_\alpha \leq 4\rho\bar{m}.$$

The second entry of the minimum is therefore no larger than the first, proving

$$T_* \geq \frac{\kappa_\theta d}{2400\rho\bar{m}}. \tag{61}$$

*Combining the bounds.* The hard priors used for (58)–(61) may differ, but all three inequalities lower-bound the same worst-case cost $T_*$. Since $8046 > \max\{1768, 2400\}$, taking their maximum gives (46). Taking the infimum over all uniformly accurate admissible pairs completes the proof. □

### B.4.3 Quadratic-growth specialization

Within the accuracy range of Theorem B7, the auxiliary scale $\kappa_\theta$ simplifies under the additional condition

$$\alpha \leq \frac{\lambda^3 R^{3\theta-2}}{L^2}. \tag{62}$$

Indeed, substituting the definitions in (43) gives

$$K_g^{\theta/(3\theta-2)} \leq Q^{1/3} \quad \Longleftrightarrow \quad L^2\alpha \leq \lambda^3 R^{3\theta-2}.$$

Under this condition,

$$K = K_g, \qquad \kappa_\theta = K^{\theta/(3\theta-2)}.$$

The condition $K = K_g$ alone does not imply the second identity.

**Corollary B10** (Refined lower bound under quadratic growth)**.** *Assume all hypotheses of Theorem B7, with $\theta = 2$, and additionally $\alpha \le \lambda^3 R^4/L^2$. Then $K = L^2/(\lambda\alpha)$, $\kappa_2 = \sqrt{K}$, and*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{2,\lambda}(L,R)\big) \ge \frac{1}{8046} \max\left\{K, \sqrt{\frac{Kd}{\rho}}, \frac{\sqrt{K}\,d}{\bar{m}\rho}\right\}. \tag{63}$$

*The last term is omitted when $\bar{m} = \infty$.*

*Proof.* At $\theta = 2$, $K_g = L^2/(\lambda\alpha)$ and $\kappa_2 = \min\{(LR/\alpha)^{2/3}, \sqrt{K_g}\}$. The positive-power comparison $\sqrt{K_g} \le Q^{1/3}$ is equivalent to $L^2\alpha \le \lambda^3 R^4$. Equation (62) therefore gives $K = K_g$ and $\kappa_2 = \sqrt{K}$. Substitute these identities into Theorem B7. □

### B.5 On the Menart–Nikolov product lower bound

The lower bounds proved above do not use the product-form lower bound of Menart & Nikolov (2026). We explain here why. In the published version, the proof of this result uses an information-demand lemma and a geometric lemma. Appendices B.5.1 and B.5.2 give counterexamples to these two auxiliary statements, respectively. Appendix B.5.3 discusses the implications for the product argument and gives a separate counterexample to the published private-oracle theorem in its distinct-query form.

The private-oracle product statement is Theorem 5 in the published version of Menart & Nikolov (2026), where the two auxiliary results used in its proof are Lemmas 16 and 17. We refer to these below as the information-demand lemma and the geometric lemma, respectively.

#### B.5.1 A counterexample to the information-demand lemma

The information-demand lemma asserts, under its stated query-budget and dimension conditions, a per-direction lower bound on the information contained in the algorithm's final output.

We construct an optimizer that is uniformly accurate over the entire class of globally convex, 2-Lipschitz objectives on public compact convex subsets of $\mathbb{B}_2(1)$. We use $n = 1$ and the identity proxy, which is allowed by the auxiliary information-demand lemma. This proxy is not a private proxy.

*The source family and an accuracy certificate.* In this subsection, let $K \ge 2$ count the source's hidden directions, set $C = 480$, $a = K^{-1/2}$, and $\eta = a/C$, and take $d$ even. Draw a Haar $d/2$-dimensional subspace $V$ and, conditional on $V$, a Haar orthonormal frame $X_1, \dots, X_K$ in $V^\perp$. The source objective and its minimizer are

$$F(w) = \max\left\{\max_{j\le K} |\langle X_j, w\rangle - a|, 2\|\Pi_V w\|\right\}, \qquad w_* = a\sum_{j=1}^{K} X_j \in \mathbb{B}_2(1). \tag{64}$$

Here $F(w_*) = 0$ and $\|w_*\| = 1$. Use the source oracle's smallest-index tie breaking among active coordinate branches, and the norm branch only when no coordinate branch is active. For the auxiliary gradient of $2\|\Pi_V w\|$ at $\Pi_V w = 0$, use 0.

The adaptive singleton queries $r_j = a\sum_{i<j} X_i$, $j = 1, \dots, K$, return $(a, -X_j)$ on this family. The same protocol can be run on an arbitrary objective $f$: set $X_j$ equal to the negative observed gradient and continue only if the observed values are $a$ and the vectors are orthonormal. Any failure invokes the fallback below. On a recognized transcript, the supporting planes imply, for every $w \in \mathbb{B}_2(1)$,

$$f(w) \ge \max_j\{a - \langle X_j, w\rangle\} \ge a - \frac{1}{K}\left\langle \sum_j X_j, w\right\rangle \ge 0. \tag{65}$$

Thus a feasible candidate whose queried value is at most $\eta$ is certified $\eta$-accurate for every objective consistent with the transcript.

On any failed recognition or verification, run projected subgradient descent for $N_{\mathrm{fb}} = \lceil 16/\eta^2 \rceil$ steps and output the average iterate. With step size $1/\sqrt{N_{\mathrm{fb}}}$, the projection inequality, diameter at most

2, and subgradient norm at most 2 give error at most $4/\sqrt{N_{\mathrm{fb}}} \le \eta$. On public domains other than $\mathbb{B}_2(1)$, use this fallback directly. These rules ensure uniform accuracy over the whole class and all fixed selected oracles.

*A noisy final output with little information per direction.* After recognizing the prefix, draw $z \sim N(0, (c/d)I_d)$, where $c > 0$, and query $r = w_* + z$. If its gradient $g$ has norm 2, form $p = f(r)g/4$ and the feasible candidate $\Pi_{\mathbb{B}_2(1)}(r-p)$. Query this candidate and return it if its value is at most $\eta$. If either check fails, query $w_*$ and return it if its value is at most $\eta$; otherwise invoke the fallback. The preceding certificate proves uniform accuracy. The total number of queries is at most $K + 3 + N_{\mathrm{fb}}$ on every instance.

On the source family, define

$$\xi = \Pi_{V^\perp} z, \qquad Y = \Pi_{\mathbb{B}_2(1)}(w_* + \xi).$$

Whenever the norm branch is active at $r$, its response gives $p = \Pi_V r$. The actual output $W$ is therefore $Y$ when that branch is active and $F(Y) \le \eta$, and is $w_*$ otherwise. Let $A$ denote this latter event and $\delta = \mathbb{P}(A)$. Writing $s_0 = \max\{1, \|w_* + \xi\|\}$, we have $s_0 - 1 \le |\langle w_*, \xi\rangle| + \|\xi\|^2/2$, and hence

$$F(Y) \le a|\langle w_*, \xi\rangle| + \frac{a}{2}\|\xi\|^2 + \max_j |\langle X_j, \xi\rangle|.$$

The Gaussian moment bounds $\mathbb{E}\|\xi\|^2 = c/2$, $\mathbb{E}|\langle w_*, \xi\rangle| \le \sqrt{c/d}$, and $\mathbb{E}\max_j |\langle X_j, \xi\rangle| \le \sqrt{(2c/d)\log(2K)}$, followed by Markov's inequality, bound the verification failure. Moreover, $\|\Pi_V z\|^2$ has mean $c/2$ and variance $c^2/d$. Chebyshev's inequality and a union bound for the coordinate projections show that the norm branch fails to be active with probability at most $(K+16)/d$. Consequently

$$\delta \le \frac{K+16}{d} + C\sqrt{\frac{cK}{d}}\left(a + \sqrt{2\log(2K)}\right) + \frac{Cc}{4}. \tag{66}$$

Fix $j$ and condition on $D_j = (X_{\neq j}, V)$. Before projection, $Y$ is a known vector plus $aX_j$ and Gaussian noise with covariance $(c/d)I$ on $V^\perp$. Comparing this Gaussian channel with the same Gaussian whose mean omits $aX_j$, and then applying data processing, gives

$$I(X_j; Y \mid D_j) \le \frac{d}{2cK}.$$

Indeed, for a reference distribution $Q$, $\mathbb{E}_X D_{\mathrm{KL}}(P_{Z|X}\|Q) = I(X;Z) + D_{\mathrm{KL}}(P_Z\|Q)$, and the Gaussian relative entropy here is $a^2/(2c/d)$. Independence of the frame columns is not needed.

Let $\widehat{X}_j$ be the quantization used in the source, whose support has cardinality at most $(3C)^d = 1440^d$. Augment $Y$ by the bit $\mathbf{1}_A$ and, only on $A$, the fallback point $w_*$. The actual output is a function of this augmentation. The bit contributes at most $\log 2$ nats, and on $A$ the information about $\widehat{X}_j$ is at most its finite-support entropy. Thus

$$\frac{I(\widehat{X}_j; W \mid D_j)}{d} \le \frac{1}{2cK} + \frac{\log 2}{d} + \delta \log 1440. \tag{67}$$

Take $c = K^{-1/2}$ and $d = 2K^3$, with $K \to \infty$. Equations (66) and (67) give $I(\widehat{X}_j; W \mid D_j)/d = O(K^{-1/2}) \to 0$, for every $j$. Meanwhile $K + 3 + N_{\mathrm{fb}} = O(C^2 K)$, so eventually

$$K + 3 + N_{\mathrm{fb}} \le \frac{d}{640\log(2/\eta)}, \qquad d\eta^2 \longrightarrow \infty.$$

Thus the source's dimension and distinct-query hypotheses hold, even with the stricter divisor $640$, while its claimed lower bound $I(\widehat{X}_j; W \mid D_j) \ge d/160$ fails for every direction. The example also rules out replacing $1/160$ by any fixed positive constant. The transcript may reveal every $X_j$; the conclusion is specifically about the final output.

### B.5.2 A COUNTEREXAMPLE TO THE GEOMETRIC LEMMA

The geometric lemma asserts that, under its stated accuracy and distinct-query conditions, the residual component of the output is small relative to its projection onto the hidden subspace.

We show that the geometric conclusion can fail even when the algorithm satisfies the required accuracy and distinct-query conditions. Use the same recognized prefix and uniform-accuracy certificate, but consider the following continuation.

Set

$$\beta = 1 - \frac{1}{2C}, \qquad \gamma = \sqrt{1-\beta^2}, \qquad b = \gamma/4,$$

and fix a public unit vector $v$. Query $r = w_* + bv$, which may lie outside the feasible ball. If the response has gradient norm 2, form $p = f(r)g/4$, $o = r - p$, and $\lambda_0 = \beta a/\langle o, X_K\rangle$ when this is defined. The candidate is

$$W = \lambda_0 o + \sum_{j<K} \big(\beta a - \lambda_0\langle o, X_j\rangle\big)X_j. \tag{68}$$

Return it only after checking feasibility and querying a value at most $\eta$; otherwise use the fallback. Uniform accuracy follows from (65), with at most $K + 2 + N_{\mathrm{fb}}$ queries on every objective.

For the source distribution, put $U_0 = (\mathrm{span}(X_1, \dots, X_K) + V)^\perp$ and $u = \Pi_{U_0} v$. Consider

$$E = \left\{ \|\Pi_V v\|^2 \ge \frac{1}{4},\ \|u\|^2 \ge \frac{1}{4},\ \max_j |\langle v, X_j\rangle| \le \frac{1}{4},\ |\langle v, X_K\rangle| \le \frac{a}{2b} \right\}.$$

On $E$, the norm branch strictly dominates at $r$, so $p = \Pi_V r$, $2\beta/3 \le \lambda_0 \le 2\beta$, and

$$W = \beta w_* + \lambda_0 b u.$$

It is feasible because $\|W\|^2 \le \beta^2 + \beta^2\gamma^2/4 \le 1$, and $F(W) = (1-\beta)a = \eta/2$. The verification succeeds.

We now account for every query, including the verification, in the source's auxiliary spaces. Its $Z$-vectors are the gradients of $2\|\Pi_V w\|$, including when that branch is not returned, and lie in $V$. Its $O$-vector at a query is the normalized component orthogonal to the corresponding $Z$-vector in their spanned plane, or zero in the degenerate case. The prefix $O$-vectors lie in $\mathrm{span}(X_1, \dots, X_{K-1})$; the extra query contributes a vector proportional to $o$; and the verification point lies in the same combined span. Therefore

$$\mathrm{span}(O) \subseteq \mathrm{span}(X_1, \dots, X_{K-1}, o).$$

Since $\langle o, X_K\rangle \neq 0$, the source's space $S = \mathrm{span}(O) \cap X_K^\perp$ is contained in $\mathrm{span}(X_1, \dots, X_{K-1})$. Writing $\Pi^\perp_{X,S}$ for projection orthogonal to $\mathrm{span}(X_1, \dots, X_K) + S$, its residual is exactly

$$\mathcal{R} := \Pi^\perp_{X,S}\Pi^\perp_Z W = \lambda_0 b u, \qquad \Pi_V \mathcal{R} = 0, \qquad \|\mathcal{R}\| \ge \frac{\beta\gamma}{12} > 2\eta. \tag{69}$$

The strict inequality holds for $K \ge 2$ and $C = 480$.

For a Haar $r$-plane $J$ and fixed unit vector $v$, $\mathbb{E}\|\Pi_J v\|^2 = r/d$ and $\mathrm{Var}(\|\Pi_J v\|^2) = 2r(d-r)/(d^2(d+2)) \le 1/(2(d+2))$. For $d \ge 8K$, Chebyshev bounds the first two failures defining $E$ by $8/(d+2)$ and $32/(d+2)$. Each $X_j$ is marginally uniform on the sphere, so the remaining failures have total probability at most $16K/d + 4b^2K/d \le 17K/d$. For $d \ge 512K$ this gives

$$\mathbb{P}(E) \ge 1 - \frac{40}{d+2} - \frac{17K}{d} \ge 1 - \frac{37K}{d} > \frac{9}{10}.$$

Choose $d$ even and still larger if necessary so that $d \ge C_2/\eta^2$ for the source's proposed constant $C_2$ and $K + 2 + N_{\mathrm{fb}} \le d/[640\log(2/\eta)]$. Equation (69) therefore makes the event

$$\|\mathcal{R}\| \ge 2\|\Pi_V \mathcal{R}\| + 2\eta$$

occur with probability greater than $9/10$, whereas the geometric lemma bounds this probability by $1/10$. Thus the construction gives a counterexample to that lemma while satisfying its distinct-query premise.

### B.5.3 CONSEQUENCES FOR THE PRODUCT LOWER BOUND

The preceding two constructions use the identity proxy. They give counterexamples to the auxiliary information-demand and geometric lemmas, which themselves impose no privacy or finite-capacity restriction. Consequently, the published proof based on these lemmas does not establish the product lower bound. These constructions do not by themselves disprove a private-oracle or information-limited minimax theorem. Replacing final-output information by transcript information avoids the first counterexample, but a separate geometric argument would still be needed.

For the private-oracle statement, we next give a direct counterexample to the published theorem in its distinct-query form (Theorem 5 of the published version). The construction is uniformly accurate over the normalized convex class $\mathfrak{C}$, for every fixed $n$.

**Proposition B11** (Uniform private-proxy counteralgorithm)**.** *For $d, n \geq 1$, $0 < \varepsilon \leq 1$, $\rho > 0$, and $\bar{m} = \infty$, there is a uniformly $\varepsilon$-accurate optimizer–proxy pair on $\mathfrak{C}$ with deterministic cost*

$$O\left(\varepsilon^{-2} + \frac{d^{1/4}}{\varepsilon} + \frac{d^{5/8}}{\sqrt{\rho}\,\varepsilon^{3/2}}\right)$$

*and at most $O(\varepsilon^{-2} + d^{1/4}/\varepsilon)$ distinct query vectors. Its replies satisfy conditional $\rho$-zCDP even under arbitrary replacement in $\mathbb{R} \times \mathbb{R}^d$.*

*Proof.* We give the construction and its convergence inequality. *Smoothing.* For any $G$ in $\mathfrak{C}$, set

$$\tau = \frac{\varepsilon}{4\sqrt{d}}, \qquad f(x) = \mathbb{E}_{Z\sim N(0,I_d)} G(x + \tau Z), \qquad H = \frac{1}{\tau}.$$

Convexity and global Lipschitzness imply $G \leq f \leq G + \varepsilon/4$. Lipschitz functions are differentiable almost everywhere, so the fixed selections satisfy $\nabla f(x) = \mathbb{E}_{I,Z} g_I(x + \tau Z)$, for uniform $I \in [n]$. Differentiating the Gaussian density in a unit direction $u$ gives $D_u \nabla f(x) = \tau^{-1}\mathbb{E}[(u \cdot Z) g_I(x + \tau Z)]$. Its norm is at most $H$, hence $f$ is convex and $H$-smooth. *Private first-order estimator.* Choose public integers and a target batch length by

$$T = \left\lceil 8\sqrt{H/\varepsilon} \right\rceil, \quad q = \left\lceil \frac{2T}{H\varepsilon} \right\rceil, \quad m_* = \sqrt{\frac{4dT}{\rho H \varepsilon}}, \quad b = \max\{1, \lceil m_*/q \rceil\}, \quad M = qb.$$

At center $x_t$, independently sample $q$ pairs $(I, Z)$ and put $b$ identical copies of each query $(x_t + \tau Z, I)$ in one batch. All $M$ evaluations are charged. The proxy discards values, clips each gradient to $\mathbb{B}_2(1)$, and releases their average plus independent $N(0, 2I_d/(\rho M^2))$ noise. The sensitivity is $2/M$, so the Gaussian divergence formula gives conditional $\rho$-zCDP on all replacement arrays. On genuine answers, repetitions agree and clipping has no effect. Conditional on the history before the fresh pairs are drawn,

$$\mathbb{E}[\widehat{g}_t] = \nabla f(x_t), \qquad \mathbb{E}\|\widehat{g}_t - \nabla f(x_t)\|^2 \leq \frac{1}{q} + \frac{2d}{\rho M^2} \leq \frac{H\varepsilon}{T} =: V_0. \tag{70}$$

The sampling variance is $1/q$; the repetitions only reduce the privacy noise.

*Accelerated optimization.*

Start from a public $y_0 = z_0 \in \mathcal{U}$ and set $a_t = t$, $A_t = t(t+1)/2$, $A_0 = 0$, and $\beta_0 = 4H$. For $t = 1, \ldots, T$, use

$$x_t = \frac{A_{t-1}y_{t-1} + a_t z_{t-1}}{A_t}, \quad z_t = \Pi_{\mathcal{U}}\left(z_{t-1} - \frac{a_t}{\beta_0}\widehat{g}_t\right), \quad y_t = \frac{A_{t-1}y_{t-1} + a_t z_t}{A_t}.$$

These iterates are feasible; the Gaussian-perturbed queries may be outside $\mathcal{U}$, as the protocol permits. For a minimizer $w$ of $f$ on $\mathcal{U}$, write $e_t = \widehat{g}_t - \nabla f(x_t)$ and $v_t = z_t - z_{t-1}$. Convexity, smoothness, and the projection inequality imply

$$A_t(f(y_t) - f(w)) \leq A_{t-1}(f(y_{t-1}) - f(w)) + \frac{\beta_0}{2}\left(\|z_{t-1} - w\|^2 - \|z_t - w\|^2\right) - H\|v_t\|^2 - a_t\langle e_t, z_t - w\rangle,$$

where $a_t^2/A_t < 2$ was used. Split the last inner product at $z_{t-1}$ and use $-a_t\langle e_t, v_t\rangle \le H\|v_t\|^2 + a_t^2\|e_t\|^2/(4H)$. The other noise term has zero expectation by (70). Telescoping, with $\|z_0 - w\| \le 2$, yields

$$\mathbb{E}[f(y_T) - f(w)] \le \frac{16H}{T^2} + \frac{V_0(2T+1)}{12H} \le \frac{\varepsilon}{2}.$$

Thus $\mathbb{E}[G(y_T) - \min_{\mathcal{U}} G] \le 3\varepsilon/4$.

*Evaluation and distinct-query complexity.*

Finally,

$$M \le q + m_*, \qquad Tq \le T + \frac{2T^2}{H\varepsilon}, \qquad T \le \frac{17d^{1/4}}{\varepsilon}.$$

There are at most $Tq$ distinct vectors and $TM \le Tq + Tm_*$ evaluations. Substitution gives both asserted bounds, uniformly over the class and the selected oracles. □

For example, let $s \to \infty$ range over the positive integers and set

$$\varepsilon = s^{-8}, \qquad d = s^{24}, \qquad \rho = 1.$$

Proposition B11 then has cost $O(s^{27})$ and uses $O(s^{16})$ distinct query vectors. For all sufficiently large $s$, it satisfies both the dimension condition and the distinct-query premise $d/[640\log(1/\varepsilon)]$ of the published theorem, whereas the claimed product lower bound is

$$\Omega(\varepsilon^{-2}\sqrt{d}) = \Omega(s^{28}).$$

Thus this sequence gives a counterexample to the published private-oracle statement in its distinct-query form.

This does not contradict the unconditional truncated lower bound, and neither this counteralgorithm nor the preceding identity-proxy constructions resolve the separate theorem for information-limited oracles. Theorem B1 uses only the independently proved normalized bound of Theorem B4. The refined bound of Theorem B7 instead uses the growth geometry and the information estimates proved in Appendices B.4.1 and B.4.2; it does not use either auxiliary lemma.

# C Proofs for the Upper Bound

This section proves Lemmas 5.1 and 5.2 and Theorem 5.3. We first analyze one private first-order update, then prove the projected-subgradient guarantee used at each stage, and finally combine it with growth localization and restarting. Appendix C.5 develops an accelerated alternative, and Appendix C.6 combines the two methods and identifies a regime in which the resulting upper bound matches the lower bound.

For brevity, write

$$p_\theta := \frac{2(\theta-1)}{\theta}.$$

## C.1 Private first-order estimator

This subsection proves Lemma 5.1 of the main text: a Gaussian private-proxy estimator that is conditionally unbiased, has conditional second moment at most $2L^2$, and consumes $O(\Psi_{\rho,\bar{m}})$ component evaluations per update.

An update averages $q$ separate proxy replies at a fixed query point before applying one projected-subgradient step.

We first recall the Gaussian zCDP calibration used to construct the estimator. If a map has Euclidean sensitivity at most $\Delta_2$ and $\sigma > 0$, then adding $\mathcal{N}(0, \sigma^2 I_d)$ gives $\Delta_2^2/(2\sigma^2)$-zCDP (Bun & Steinke, 2016). The adjacency relation is the one fixed in Definition 3.9: replacement of one current first-order answer by another admissible answer.

**Lemma C1** (One Gaussian minibatch message)**.** *Fix an integer $m \geq 1$. Conditional on the complete preceding sigma-field $\mathcal{G}_{t-1}$, let the current point $w$ be measurable. The optimizer draws fresh*

$$I_1, \ldots, I_m \overset{\text{iid}}{\sim} \text{Unif}([n]).$$

*The proxy discards the function-value coordinates, forms*

$$\overline{g}(w) = \frac{1}{m} \sum_{\ell=1}^{m} g_{I_\ell}(w),$$

*and releases*

$$Y(w) = \overline{g}(w) + Z, \qquad Z \sim \mathcal{N}\left(0, \frac{2L^2}{\rho m^2} I_d\right). \tag{71}$$

*This message is conditionally $\rho$-*zCDP *under replacement of one current first-order answer,*

$$\mathbb{E}[Y(w) \mid \mathcal{G}_{t-1}] = g_F(w) := \frac{1}{n} \sum_{i=1}^{n} g_i(w) \in \partial F(w),$$

*and*

$$\mathbb{E}[\|Y(w)\|^2 \mid \mathcal{G}_{t-1}] \leq L^2 \left(1 + \frac{2d}{\rho m^2}\right).$$

*Proof.* After conditioning on the current query batch, replacing one first-order answer changes the averaged selected subgradient by at most $2L/m$. The Gaussian variance $2L^2/(\rho m^2)$ therefore gives $\rho$-zCDP. Before drawing the component indices, fresh uniform sampling gives $\mathbb{E}[Y(w) \mid \mathcal{G}_{t-1}] = n^{-1} \sum_i g_i(w) \in \partial F(w)$. The sample mean has norm at most $L$, and the independent centered noise has expected squared norm $2dL^2/(\rho m^2)$. Adding these contributions proves the second-moment bound. □

The estimator averages $q$ replies at a fixed query point to reduce noise. The next lemma bounds the second moment and shows that the conditional mean and per-reply privacy are preserved.

**Lemma C2** (Aggregation of separately private messages)**.** *Fix integers $m, q \geq 1$, with $m \leq \bar{m}$, before the first reply of an update. Hold $w$ fixed for $q$ consecutive message rounds, each using Lemma C1 with fresh component samples and independent Gaussian noise. The optimizer observes all $q$ messages before forming*

$$\widehat{g}(w) := \frac{1}{q} \sum_{j=1}^{q} Y_j(w).$$

*Every constituent message is conditionally $\rho$-*zCDP *and consumes $m \leq \bar{m}$ component evaluations. The averaged estimator is conditionally unbiased and satisfies*

$$\mathbb{E}[\|\widehat{g}(w)\|^2 \mid \mathcal{G}] \leq G_{m,q}^2 := L^2 \left(1 + \frac{2d}{\rho m^2 q}\right), \tag{72}$$

*where $\mathcal{G}$ is the sigma-field at the start of the update. One update costs exactly $qm$ component evaluations.*

*Proof.* Each release is a genuine message round, so Lemma C1 applies after conditioning on the transcript that precedes that release. We do not claim that the joint $q$-message transcript is one $\rho$-zCDP message. The iterate, $m$, and $q$ remain fixed, and all samples and noises are fresh, so iterated conditional expectation gives

$$\mathbb{E}[\widehat{g}(w) \mid \mathcal{G}] = g_F(w).$$

For an exact moment calculation, put

$$S^2(w) := \frac{1}{n} \sum_{i=1}^{n} \|g_i(w)\|^2.$$

Jensen's inequality and the component bound give $\|g_F(w)\|^2 \leq S^2(w) \leq L^2$. Also $qm \geq 1$, so $1 - 1/(qm)$ and $1/(qm)$ are nonnegative weights. Hence the sampling part $\|g_F(w)\|^2 + [S^2(w) -$

$\|g_F(w)\|^2]/(qm)$ is at most $L^2$. Independence across all $qm$ component samples and all Gaussian noises gives

$$\mathbb{E}[\|\widehat{g}(w)\|^2 \mid \mathcal{G}] = \|g_F(w)\|^2 + \frac{S^2(w) - \|g_F(w)\|^2}{qm} + \frac{2dL^2}{\rho m^2 q}$$
$$\leq L^2 \left(1 + \frac{2d}{\rho m^2 q}\right).$$

□

It remains to choose $m$ and $q$. We use

$$m := \min\left\{\bar{m}, \left\lceil\sqrt{\frac{2d}{\rho}}\right\rceil\right\}, \qquad q := \max\left\{1, \left\lceil\frac{2d}{\rho m^2}\right\rceil\right\}, \tag{73}$$

omitting the first minimum when $\bar{m} = \infty$, and set $G^2 := G^2_{m,q}$ and $a := 2d/\rho$. This choice gives $G^2 \leq 2L^2$ and $mq = O(\Psi_{\rho,\bar{m}})$. Each update holds its query point fixed for $q$ separate replies, averages them, and performs one projection onto $\mathcal{W}$.

*Calibration and cost per update.* Lemmas C1 and C2 show that every reply satisfies Definition 3.9, and that the averaged estimator is conditionally unbiased with second moment at most $G^2$. The choice of $q$ gives $G^2 \leq 2L^2$. If $m \geq \sqrt{a}$, then $q = 1$ and $m = \Theta(\max\{1, \sqrt{a}\})$. If $m < \sqrt{a}$, then $m = \bar{m}$ and $q = \Theta(a/m^2)$. Thus in both cases

$$mq = O(\Psi_{\rho,\bar{m}}), \qquad mq + \frac{a}{m} = O(\Psi_{\rho,\bar{m}}).$$

These relations include all integer-rounding cases.

Collecting these facts gives exactly Lemma 5.1 of the main text: with $m, q$ as in (73), the averaged estimator $\widehat{g}(u)$ satisfies

$$\mathbb{E}[\widehat{g}(u) \mid \mathcal{G}] = \frac{1}{n}\sum_{i=1}^{n} g_i(u) \in \partial F(u), \qquad \mathbb{E}[\|\widehat{g}(u)\|^2 \mid \mathcal{G}] \leq 2L^2,$$

every constituent reply satisfies Definition 3.9, and one update consumes $mq = O(\Psi_{\rho,\bar{m}})$ component first-order evaluations.

### C.2 Projected-subgradient bound

The next lemma gives the projected-subgradient guarantee used both for the global-radius branch and within each restarted stage. In the former case the initial squared-distance bound is $R^2$; under growth it is replaced by the shrinking certificate derived in Appendix C.3. Throughout the remainder of this section, let $a := 2d/\rho$, as in Appendix C.1.

**Lemma C3** (Generic convex update bound)**.** *Assume the convex Lipschitz losses and radius bound of Assumptions 3.4 and 3.8. For fixed integers $1 \leq m \leq \bar{m}$ and $q \geq 1$, start from any $u_0 \in \mathcal{W}$, use the averaged estimator of Lemma C2, and set*

$$\eta = \frac{\alpha}{G^2_{m,q}}, \qquad N_{\mathrm{g}} = \max\left\{1, \left\lceil\frac{G^2_{m,q}R^2}{\alpha^2}\right\rceil\right\}.$$

*Projected stochastic subgradient descent, averaged into $\overline{u} = N_{\mathrm{g}}^{-1}\sum_{s=0}^{N_{\mathrm{g}}-1} u_s$, has expected gap $\mathbb{E}[F(\overline{u}) - F^\star] \leq \alpha$ and*

$$\mathrm{Cost}_{\mathrm{g}} \leq \frac{L^2R^2}{\alpha^2}\left(qm + \frac{a}{m}\right) + qm. \tag{74}$$

*Proof.* Fix a nearest minimizer $w^\star = \Pi_{W^\star}(u_0)$. If $u_0$ is random, this projection is measurable, and we condition on $u_0$ as part of the history throughout the stage. Nonexpansiveness of $\Pi_{\mathcal{W}}$ gives

$$\|u_{s+1} - w^\star\|^2 \leq \|u_s - \eta\widehat{g}_s - w^\star\|^2$$
$$= \|u_s - w^\star\|^2 - 2\eta\langle\widehat{g}_s, u_s - w^\star\rangle + \eta^2\|\widehat{g}_s\|^2.$$

Conditional on the full history before update $s$, the estimator has mean $g_F(u_s) \in \partial F(u_s)$ and second moment at most $G_{m,q}^2$. The subgradient inequality therefore bounds the expected inner product below by $F(u_s) - F^\star$. Summing the resulting inequalities and using convexity for the average $\overline{u}$ yields

$$\mathbb{E}[F(\overline{u}) - F^\star] \le \frac{\mathbb{E}\|u_0 - w^\star\|^2}{2\eta N_{\mathrm{g}}} + \frac{\eta G_{m,q}^2}{2}. \tag{75}$$

All terms are integrable: the initial squared distance is bounded by $R^2$, and the estimator has bounded second moment. Nonexpansiveness then bounds every subsequent squared distance by an integrable finite sum of these terms.

Substitute $\mathbb{E}\|u_0 - w^\star\|^2 \le R^2$, $\eta = \alpha/G_{m,q}^2$, and $N_{\mathrm{g}} \ge G_{m,q}^2 R^2/\alpha^2$ into (75). Both terms are at most $\alpha/2$. Each update costs $qm$ component evaluations. Since

$$qmG_{m,q}^2 = L^2(qm + a/m),$$

the ceiling in $N_{\mathrm{g}}$ gives (74). □

### C.3 Growth localization and one-stage guarantee

We now prove Lemma 5.2. The key consequence of growth is that an expected excess-error guarantee yields an expected squared distance bound to the minimizer set. The next lemma records this localization property together with the public initial certificate $\Delta_{\max}$.

**Lemma C4** (Initial error and distance under growth)**.** *Every instance in $\mathfrak{F}_{\theta,\lambda}(L,R)$ satisfies*

$$0 \le F(w) - F^\star \le \Delta_{\max} \qquad \textit{for every } w \in \mathcal{W}.$$

*For any random $U \in \mathcal{W}$ with $\mathbb{E}[F(U) - F^\star] \le \alpha$,*

$$\mathbb{E}[\mathrm{dist}(U, W^\star)^2] \le \lambda^{-2/\theta}\alpha^{2/\theta}.$$

*Proof.* The minimizer set $W^\star$ is nonempty, closed, and convex. For $w^\star = \Pi_{W^\star}(w)$, convexity and $\|g_i(w)\| \le L$ for the selected subgradients give

$$F(w) - F^\star \le \left\langle \frac{1}{n}\sum_i g_i(w), w - w^\star \right\rangle \le L\,\mathrm{dist}(w, W^\star) \le LR.$$

Combining the middle bound with growth gives $\mathrm{dist}(w, W^\star) \le (L/\lambda)^{1/(\theta-1)}$ whenever the distance is positive. Substitution yields the other term in $\Delta_{\max}$; points in $W^\star$ have zero gap.

Growth also gives $\mathrm{dist}(U, W^\star)^2 \le \lambda^{-2/\theta}(F(U) - F^\star)^{2/\theta}$. Since $\theta \ge 2$, the power $x \mapsto x^{2/\theta}$ is concave on $\mathbb{R}_+$. The gap is bounded by the first claim, so Jensen's inequality applies:

$$\mathbb{E}[(F(U) - F^\star)^{2/\theta}] \le \big(\mathbb{E}[F(U) - F^\star]\big)^{2/\theta} \le \alpha^{2/\theta}. \tag{76}$$

This proves the distance bound. □

The first claim gives the initial certificate $\Delta_0 = \Delta_{\max}$. The second is applied at each stage with $\alpha = \Delta_j$ to obtain the shrinking distance bound used by the restart schedule. We next prove the one-stage contraction for an arbitrary deterministic second-moment bound $\bar{G}^2$.

**Lemma C5** (One restarted stage)**.** *Let $\Delta > 0$ and let $u_0 \in \mathcal{W}$ be random with $\mathbb{E}[F(u_0) - F^\star] \le \Delta$. Use conditionally unbiased estimators satisfying $\mathbb{E}[\widehat{g}_s \mid \mathcal{G}_s] = g_F(u_s) \in \partial F(u_s)$, where $\mathcal{G}_s$ contains the stage start and the history before update $s$. Let $\bar{G}^2 > 0$ be a deterministic bound satisfying $\mathbb{E}[\|\widehat{g}_s\|^2 \mid \mathcal{G}_s] \le \bar{G}^2$ at every update of the stage. With*

$$\eta_\Delta = \frac{\Delta}{4\bar{G}^2}, \qquad N_\Delta = \left\lceil 8\bar{G}^2\lambda^{-2/\theta}\Delta^{-p_\theta} \right\rceil,$$

*run*

$$u_{s+1} = \Pi_{\mathcal{W}}\big(u_s - \eta_\Delta \widehat{g}_s\big), \qquad s = 0, \dots, N_\Delta - 1,$$

*and return $\overline{u} = N_\Delta^{-1}\sum_{s=0}^{N_\Delta-1} u_s$. Then*

$$\mathbb{E}[F(\overline{u}) - F^\star] \le \frac{\Delta}{2}.$$

*Proof of Lemma C5.* For every realization of $u_0$, let $w^\star(u_0) = \Pi_{W^\star}(u_0)$. Projection onto the closed convex set $W^\star$ is continuous and hence measurable. Lemma C4 gives

$$\mathbb{E}\,\|u_0 - w^\star(u_0)\|^2 \le \lambda^{-2/\theta}\Delta^{2/\theta}. \tag{77}$$

Projection keeps all subsequent iterates in $\mathcal{W}$, where the growth condition holds, while the selected-subgradient bound is global. Localization is therefore needed only for the stage-start distance in (77). Apply (75) with this fixed, realization-dependent minimizer throughout the stage. After conditioning first on $u_0$ and then taking total expectation,

$$\mathbb{E}[F(\overline{u}) - F^\star] \le \frac{\lambda^{-2/\theta}\Delta^{2/\theta}}{2\eta_\Delta N_\Delta} + \frac{\eta_\Delta \bar{G}^2}{2}.$$

By the definitions of $\eta_\Delta$ and $N_\Delta$, the first term is at most

$$\frac{1}{4}\Delta^{\,2/\theta-1+p_\theta} = \frac{\Delta}{4},$$

because $2/\theta - 1 + p_\theta = 1$. The second term is exactly $\Delta/8$. Their sum is therefore $3\Delta/8 \le \Delta/2$, as required by the schedule $\Delta_{j+1} = \Delta_j/2$. □

Instantiating Lemma C5 at $\bar{G}^2 = 2L^2$, which is admissible by (73) and Lemma C2, gives

$$\eta_\Delta = \frac{\Delta}{8L^2}, \qquad N_\Delta = \left\lceil 16L^2\lambda^{-2/\theta}\Delta^{-p_\theta} \right\rceil,$$

which are the $\eta_j$ and $T_j$ of Algorithm 1 at $\Delta = \Delta_j$. This proves Lemma 5.2.

*One-stage use of growth.* For comparison, we record the complexity obtained by using the growth certificate only once. This result is not needed for Theorem 5.3.

**Corollary C6** (One-stage use of growth)**.** *Let $0 < \alpha < \Delta_0 \le \Delta_{\max}$ and let $w^{(0)} \in \mathcal{W}$ satisfy $\mathbb{E}[F(w^{(0)}) - F^\star] \le \Delta_0$. Set $D_{\Delta_0} := \min\{R, (\Delta_0/\lambda)^{1/\theta}\}$. One projected-subgradient stage with the private estimator parameters in* (73) *achieves expected error at most $\alpha$ using $O(L^2 D_{\Delta_0}^2 \alpha^{-2}\Psi_{\rho,\bar{m}})$ evaluations.*

*Proof of Corollary C6.* The class-radius bound and TNC, followed by Jensen's inequality as in (76), give

$$\mathbb{E}\Big[\mathrm{dist}(w^{(0)}, W^\star)^2\Big] \le \min\{R^2, \lambda^{-2/\theta}\Delta_0^{2/\theta}\} = D_{\Delta_0}^2.$$

Repeat the proof of Lemma C3 with the measurable projection of $w^{(0)}$ onto $W^\star$ and replace its initial $R^2$ term by this expected squared distance. The choices $\eta = \alpha/G^2$ and $N_{\mathrm{g}} = \lceil G^2 D_{\Delta_0}^2/\alpha^2 \rceil$ then give expected error at most $\alpha$ for the stage average. By (73), $G^2 \le 2L^2$ and $mq = O(\Psi_{\rho,\bar{m}})$. Finally, $\Delta_0 \le \Delta_{\max}$ implies both $\Delta_0 \le LR$ and $\Delta_0 \le L(\Delta_0/\lambda)^{1/\theta}$, so $\alpha < \Delta_0 \le LD_{\Delta_0}$. The ceiling in $N_{\mathrm{g}}$ is therefore absorbed in $mqN_{\mathrm{g}} = O(L^2 D_{\Delta_0}^2 \alpha^{-2}\Psi_{\rho,\bar{m}})$. □

For fixed $\Delta_0$, this bound retains an $\alpha^{-2}$ dependence, because $D_{\Delta_0}$ does not depend on the final target $\alpha$. The improved exponent in Algorithm 1 comes from refreshing the distance certificate between stages.

### C.4 Proof of the main upper bound

We now prove Theorem 5.3. The argument follows the order of Section 5: sum the geometrically increasing stage lengths of the restarted branch, multiply by the $O(\Psi_{\rho,\bar{m}})$ component evaluations that one update costs, compare against the global-radius branch, and observe that the public choice between the two schedules realizes the minimum defining $K_{\theta,\lambda}$.

*Proof of Theorem 5.3.* If $\alpha = \Delta_{\max}$, Lemma C4 shows that any feasible starting point already has expected excess error at most $\alpha$, so no first-order evaluation is needed. Hence assume $0 < \alpha < \Delta_{\max}$.

*Private estimator.* Choose $m$ and $q$ as in (73). By Section C.1, every proxy reply satisfies Definition 3.9, while the averaged estimator is conditionally unbiased and has conditional second moment at most $2L^2$. Moreover,

$$mq = O(\Psi_{\rho,\bar{m}}), \qquad mq + \frac{a}{m} = O(\Psi_{\rho,\bar{m}}),$$

where $a = 2d/\rho$.

*Global-radius branch.*

Suppose $\alpha \geq \lambda R^\theta$. Applying Lemma C3 with the public radius $R$ gives expected excess error at most $\alpha$ and cost

$$O\left(\frac{L^2R^2}{\alpha^2}\Psi_{\rho,\bar{m}}\right).$$

In this regime,

$$K_{\theta,\lambda} = \frac{L^2R^2}{\alpha^2},$$

so the required bound follows.

*Growth-active branch.*

Now suppose $\alpha < \lambda R^\theta$. By Lemma C4, every feasible starting point satisfies the initial certificate

$$\mathbb{E}[F(w^{(0)}) - F^\star] \leq \Delta_{\max}.$$

Run Algorithm 1 with $\Delta_0 = \Delta_{\max}$. Lemma C5 halves the expected error certificate at every stage, so

$$J := \left\lceil \log_2 \frac{\Delta_{\max}}{\alpha} \right\rceil$$

stages suffice. Since $\Delta_j = 2^{-j}\Delta_{\max}$,

$$\begin{aligned}\sum_{j=0}^{J-1} T_j &\leq 16L^2\lambda^{-2/\theta}\sum_{j=0}^{J-1}\Delta_j^{-p_\theta} + J \\ &\leq 16A_\theta L^2\lambda^{-2/\theta}\alpha^{-p_\theta} + J,\end{aligned}$$

where

$$A_\theta := \frac{2^{p_\theta}}{2^{p_\theta}-1} \leq 2.$$

Indeed, $\Delta_{J-1} > \alpha$, and summing backward from the last stage gives a geometric-series factor with ratio $2^{-p_\theta}$. This factor is at most $A_\theta \leq 2$ because $p_\theta \geq 1$. Since every update costs $O(\Psi_{\rho,\bar{m}})$ component evaluations,

$$\text{Cost} \leq O\left(\left(L^2\lambda^{-2/\theta}\alpha^{-p_\theta} + J\right)\Psi_{\rho,\bar{m}}\right).$$

On this branch,

$$K_{\theta,\lambda} = L^2\lambda^{-2/\theta}\alpha^{-p_\theta}.$$

It remains only to control the additive number of stages. From (8),

$$\Delta_{\max} \leq \left(\frac{L^\theta}{\lambda}\right)^{1/(\theta-1)} = \left(L^2\lambda^{-2/\theta}\right)^{1/p_\theta},$$

and hence

$$\frac{\Delta_{\max}}{\alpha} \leq K_{\theta,\lambda}^{1/p_\theta}.$$

Therefore

$$J \leq \left\lceil \frac{\log_2 K_{\theta,\lambda}}{p_\theta} \right\rceil \leq 1 + \log_2 K_{\theta,\lambda} \leq 2K_{\theta,\lambda},$$

where the last two inequalities use $p_\theta \geq 1$ and $K_{\theta,\lambda} \geq 1$; the latter holds because $\alpha < \Delta_{\max}$. Hence $\sum_{j=0}^{J-1} T_j \leq 34K_{\theta,\lambda}$, and the restarted branch has cost

$$O(K_{\theta,\lambda}\Psi_{\rho,\bar{m}}).$$

The branch condition depends only on the public parameters $\alpha, \lambda$, and $R$. Hence the same procedure is valid uniformly over $\mathfrak{F}_{\theta,\lambda}(L,R)$, and its output is feasible because all iterates are projected onto $\mathcal{W}$ and each stage returns their average. Consequently,

$$\sup_{H\in\mathfrak{F}_{\theta,\lambda}(L,R)} \mathrm{Cost}_H(\mathcal{A},\mathcal{P}) \le O(K_{\theta,\lambda}\Psi_{\rho,\bar{m}})\,.$$

Taking the infimum over admissible optimizer–proxy pairs proves

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha;\mathfrak{F}_{\theta,\lambda}(L,R)\big) \le O(K_{\theta,\lambda}\Psi_{\rho,\bar{m}})\,.$$

□

### C.5 Accelerated upper bound

The restarted method of Theorem 5.3 performs $\Theta(K_{\theta,\lambda})$ private first-order updates. We next reduce the number of adaptive updates by smoothing the objective and applying an accelerated stochastic method. Fewer updates allow more component evaluations to be averaged within each private reply, which can reduce the privacy-dependent cost.

Throughout this subsection and the next, write

$$K := K_{\theta,\lambda}, \qquad \Psi := \Psi_{\rho,\bar{m}},$$

and omit terms involving $\bar{m}$ when $\bar{m}=\infty$.

**Theorem C7** (Accelerated upper bound)**.** *There are universal constants $C_{\mathrm{acc}}, C'_{\mathrm{acc}} > 0$ such that the following holds. Fix $\theta \ge 2$, $n, d \in \mathbb{N}_{\ge 1}$, $L, R, \lambda, \rho, \alpha > 0$, and $\bar{m} \in \mathbb{N}_{\ge 1} \cup \{\infty\}$. There is an optimizer–proxy pair $(\mathcal{A},\mathcal{P})$ satisfying Definition 3.9 that is $\alpha$-accurate on $\mathfrak{F}_{\theta,\lambda}(L,R)$ and whose worst-case cost satisfies*

$$\sup_{H\in\mathfrak{F}_{\theta,\lambda}(L,R)} \mathrm{Cost}_H(\mathcal{A},\mathcal{P}) \le C_{\mathrm{acc}}\left(K + \frac{K^{3/4}d^{5/8}}{\sqrt{\rho}} + \frac{Kd}{\bar{m}\rho} + \sqrt{K}\,d^{1/4}\,\Psi\right). \tag{78}$$

*If in addition $d \le K^2$, then*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha;\mathfrak{F}_{\theta,\lambda}(L,R)\big) \le C'_{\mathrm{acc}}\left(K + \frac{K^{3/4}d^{5/8}}{\sqrt{\rho}} + \frac{Kd}{\bar{m}\rho}\right). \tag{79}$$

The last term in (78) has a simple source. The method performs $O(\sqrt{K}\,d^{1/4})$ adaptive updates, and each update uses at least one private reply, whose cost is $O(\Psi)$. When $d \le K^2$, this term is dominated by the other three. This is also the regime in which acceleration improves on Theorem 5.3; see Appendix C.6.

The proof has three steps. Appendix C.5.1 smooths the objective and constructs a private unbiased estimator of the smoothed gradient. Appendix C.5.2 proves an accelerated one-stage guarantee. Appendix C.5.3 restarts this stage under growth, exactly as in Algorithm 1.

#### C.5.1 Smoothing and private stochastic gradient estimation

Let $\mathsf{B} := \{z \in \mathbb{R}^d : \|z\| \le 1\}$. For $u > 0$ and $Z \sim \mathrm{Unif}(\mathsf{B})$, define the ball smoothing

$$F_u(x) := \mathbb{E}[F(x+uZ)], \qquad x \in \mathbb{R}^d.$$

For $k \ge 0$, let $\omega_k$ denote the volume of the unit ball in $\mathbb{R}^k$, with $\omega_0 = 1$.

**Lemma C8** (Ball smoothing)**.** *Under Assumption 3.4, the following hold for every $u > 0$.*

*(i) $F_u$ is convex and $L$-Lipschitz on $\mathbb{R}^d$, and $F(x) \le F_u(x) \le F(x) + Lu$ for every $x \in \mathbb{R}^d$.*

*(ii) $F_u$ is differentiable. For independent $I \sim \mathrm{Unif}([n])$ and $Z \sim \mathrm{Unif}(\mathsf{B})$, and the selected subgradients $g_i$ of Definition 3.5,*

$$\nabla F_u(x) = \mathbb{E}[g_I(x+uZ)], \qquad x \in \mathbb{R}^d.$$

*(iii) $\nabla F_u$ is $\beta_u$-Lipschitz on $\mathbb{R}^d$, where $\beta_u := 2L\sqrt{d}/u$.*

*Proof.* (i) $F_u$ is an average of the convex functions $x \mapsto F(x+uz)$, hence convex. Also $|F_u(x) - F_u(y)| \le \mathbb{E}|F(x+uZ) - F(y+uZ)| \le L\,\|x-y\|$. Since $\mathbb{E}Z = 0$, Jensen's inequality gives $F_u(x) \ge F(x)$. Since $\|uZ\| \le u$, Lipschitzness gives $F(x+uZ) \le F(x) + Lu$, and taking expectations gives the upper bound.

(ii) Fix $i \in [n]$ and $x, v \in \mathbb{R}^d$. By Rademacher's theorem, $f_i$ is differentiable outside a Lebesgue-null set $N_i$. At every point of differentiability, the subdifferential of the convex function $f_i$ is $\{\nabla f_i\}$, so $g_i = \nabla f_i$ outside $N_i$. Since $Z$ has a density, $x + uZ \notin N_i$ almost surely. For $t \neq 0$, the difference quotients

$$\frac{f_i(x+tv+uZ) - f_i(x+uZ)}{t}$$

are bounded by $L\,\|v\|$ and converge almost surely to $\langle g_i(x+uZ), v\rangle$ as $t \to 0$. Put $\gamma_i(x) := \mathbb{E}[g_i(x+uZ)]$. By dominated convergence, $(f_i)_u(x) := \mathbb{E}[f_i(x+uZ)]$ has two-sided directional derivative $\langle \gamma_i(x), v\rangle$ at every $x$ and in every direction $v$. The proof of (iii) below shows that $\gamma_i$ is Lipschitz, hence continuous. The Lipschitz function $t \mapsto (f_i)_u(x+tv)$ has derivative $\langle \gamma_i(x+tv), v\rangle$ at every $t$, so

$$|(f_i)_u(x+v) - (f_i)_u(x) - \langle \gamma_i(x), v\rangle| = \left| \int_0^1 \langle \gamma_i(x+tv) - \gamma_i(x), v\rangle\, dt \right| = o(\|v\|).$$

Thus $(f_i)_u$ is differentiable with $\nabla (f_i)_u = \gamma_i$. Averaging over $i$ and using the independence of $I$ and $Z$ proves the claim.

(iii) Fix $i$ and write $\gamma_i(x) = (\omega_d u^d)^{-1} \int_{\|y-x\| \le u} g_i(y)\, dy$. For $x \neq x'$, let $s := \|x - x'\|$ and $e := (x - x')/s$. Every line parallel to $e$ meets the ball of radius $u$ around $x$ in an interval, and meets the ball around $x'$ in the same interval translated by $-s$ along $e$. Hence the part of the first interval outside the second has length at most $s$. The lines that meet the first ball are indexed by its orthogonal projection onto $e^\perp$, a $(d-1)$-dimensional ball of radius $u$. By Fubini's theorem, the volume of the first ball outside the second is at most $s\,\omega_{d-1} u^{d-1}$, and the same bound holds with $x$ and $x'$ exchanged. Since $\|g_i\| \le L$,

$$\|\gamma_i(x) - \gamma_i(x')\| \le \frac{L \cdot 2s\,\omega_{d-1} u^{d-1}}{\omega_d u^d} = \frac{2Ls}{u} \cdot \frac{\Gamma(d/2+1)}{\sqrt{\pi}\,\Gamma(d/2+1/2)}.$$

Log-convexity of the Gamma function gives $\Gamma(y+1/2) \le \Gamma(y)^{1/2}\Gamma(y+1)^{1/2} = \sqrt{y}\,\Gamma(y)$ for $y > 0$. With $y = (d+1)/2$, the last ratio is at most $\sqrt{(d+1)/(2\pi)} \le \sqrt{d}$. Averaging over $i$ proves (iii). □

The private estimator samples perturbed query points. It is otherwise the estimator of Lemma C2.

**Lemma C9** (Smoothed Gaussian messages)**.** *Fix $u > 0$ and integers $1 \le m \le \bar{m}$ and $q \ge 1$. Let $x \in \mathbb{R}^d$ be measurable with respect to the sigma-field $\mathcal{G}$ at the start of an update, and hold $x$ fixed for $q$ consecutive message rounds. In round $j \in [q]$, the optimizer draws fresh independent*

$$I_{j,1}, \dots, I_{j,m} \overset{\text{iid}}{\sim} \mathrm{Unif}([n]), \qquad Z_{j,1}, \dots, Z_{j,m} \overset{\text{iid}}{\sim} \mathrm{Unif}(\mathsf{B}),$$

*and queries the batch $(x + uZ_{j,\ell}, I_{j,\ell})_{\ell \le m}$. The proxy discards the function values and releases*

$$Y_j = \frac{1}{m} \sum_{\ell=1}^{m} g_{I_{j,\ell}}(x + uZ_{j,\ell}) + \xi_j, \qquad \xi_j \sim \mathcal{N}\left(0, \frac{2L^2}{\rho m^2} I_d\right),$$

*with noises independent of each other and of everything else. The optimizer forms $\widehat{g}_u(x) := q^{-1} \sum_{j=1}^{q} Y_j$. Then every reply satisfies Definition 3.9, the update costs exactly $qm$ component evaluations, and*

$$\mathbb{E}[\widehat{g}_u(x) \mid \mathcal{G}] = \nabla F_u(x), \qquad \mathbb{E}\Big[\|\widehat{g}_u(x) - \nabla F_u(x)\|^2 \,\Big|\, \mathcal{G}\Big] \le \frac{L^2}{qm} + \frac{2dL^2}{\rho m^2 q}. \tag{80}$$

*Proof.* Each batch has $m \le \bar{m}$ queries and is fixed before its responses are evaluated. Its query points may lie outside $\mathcal{W}$, which Definition 3.5 permits. Conditional on the preceding interaction and the current batch, replacing one current response by another admissible response changes the

average of the $m$ selected subgradients by at most $2L/m$. As in Lemma C1, the Gaussian variance $2L^2/(\rho m^2)$ therefore gives $\rho$-zCDP.

Conditional on $\mathcal{G}$, the $qm$ sampled subgradients are independent and identically distributed. By Lemma C8(ii), each has mean $\nabla F_u(x)$, and each has squared norm at most $L^2$. Their average therefore has conditional mean $\nabla F_u(x)$ and conditional variance at most $L^2/(qm)$. The averaged noise $q^{-1}\sum_j \xi_j$ is independent and centered, with expected squared norm $2dL^2/(\rho m^2 q)$. Adding the two contributions proves (80). □

We calibrate the estimator to a target variance $L^2/b$, where $b \geq 1$ is an integer. Set

$$m_b := \min\left\{\bar{m},\, \max\left\{2b,\, \left\lceil 2\sqrt{db/\rho}\,\right\rceil\right\}\right\}, \qquad q_b := \max\left\{1,\, \left\lceil \frac{2b}{m_b}\right\rceil,\, \left\lceil \frac{4db}{\rho m_b^2}\right\rceil\right\}. \tag{81}$$

**Lemma C10** (Cost of one smoothed update)**.** *With $(m, q) = (m_b, q_b)$, the estimator of Lemma C9 has conditional variance at most $L^2/b$, and one update costs*

$$q_b m_b \leq 5b + 2\sqrt{\frac{db}{\rho}} + \frac{4db}{\rho\bar{m}}.$$

*Proof.* Since $q_b m_b \geq 2b$ and $\rho m_b^2 q_b \geq 4db$, both terms in (80) are at most $L^2/(2b)$. Let $m_\star := \max\{2b, \lceil 2\sqrt{db/\rho}\,\rceil\}$, so that $m_\star \leq 2b + 2\sqrt{db/\rho} + 1$. If $m_\star \leq \bar{m}$, then $m_b = m_\star$, both ratios in the definition of $q_b$ are at most one, and $q_b m_b = m_\star$. Otherwise $m_b = \bar{m} < m_\star$ and

$$q_b m_b \leq \bar{m} + 2b + \frac{4db}{\rho\bar{m}} < m_\star + 2b + \frac{4db}{\rho\bar{m}}.$$

In both cases $b \geq 1$ gives the claim; the second case cannot occur when $\bar{m} = \infty$. □

### C.5.2 Accelerated one-stage guarantee

We first state an accelerated stochastic gradient bound for a smooth convex function, with a complete proof. The method is the accelerated stochastic approximation scheme with the standard weights $a_t = 2/(t+1)$. The bound depends on the initial point only through an expected squared distance, which is the quantity that growth controls.

**Lemma C11** (Accelerated stochastic gradient bound)**.** *Let $h : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, with $\beta$-Lipschitz gradient for $\beta > 0$. Let $\mathcal{G}_0$ be a sigma-field, and let $x_0 \in \mathcal{W}$ and $x^\diamond \in \mathcal{W}$ be $\mathcal{G}_0$-measurable with $\mathbb{E}\,\|x_0 - x^\diamond\|^2 \leq \mathcal{D}^2$ for a constant $\mathcal{D} > 0$. Assume also $\mathbb{E}\,\|\nabla h(x^\diamond)\|^2 < \infty$. Fix an integer $N \geq 1$ and $\sigma \geq 0$, set*

$$\gamma := \min\left\{\frac{1}{2\beta}, \sqrt{\frac{12\mathcal{D}^2}{\sigma^2 N(N+1)(2N+1)}}\right\}, \qquad a_t := \frac{2}{t+1}, \qquad \gamma_t := \frac{\gamma t}{2},$$

*where the second entry of the minimum is $+\infty$ if $\sigma = 0$, and set $x_0^{\mathrm{ag}} := x_0$. For $t = 1, \ldots, N$, let*

$$\begin{aligned} x_t^{\mathrm{md}} &:= (1 - a_t)x_{t-1}^{\mathrm{ag}} + a_t x_{t-1}, \\ x_t &:= \Pi_{\mathcal{W}}\big(x_{t-1} - \gamma_t G_t\big), \\ x_t^{\mathrm{ag}} &:= (1 - a_t)x_{t-1}^{\mathrm{ag}} + a_t x_t. \end{aligned}$$

*Let $\mathcal{G}_{t-1} \supseteq \mathcal{G}_0$ be the sigma-field generated by the history before $G_t$ is formed, so that $x_{t-1}$, $x_{t-1}^{\mathrm{ag}}$, and $x_t^{\mathrm{md}}$ are $\mathcal{G}_{t-1}$-measurable. Assume $\mathbb{E}[G_t \mid \mathcal{G}_{t-1}] = \nabla h(x_t^{\mathrm{md}})$ and $\mathbb{E}[\big\|G_t - \nabla h(x_t^{\mathrm{md}})\big\|^2 \mid \mathcal{G}_{t-1}] \leq \sigma^2$. Then $x_N^{\mathrm{ag}} \in \mathcal{W}$ and*

$$\mathbb{E}\big[h(x_N^{\mathrm{ag}}) - h(x^\diamond)\big] \leq \frac{4\beta\mathcal{D}^2}{N(N+1)} + \frac{2\mathcal{D}\sigma}{\sqrt{N}}. \tag{82}$$

*Proof.* All iterates lie in $\mathcal{W}$, because $\mathcal{W}$ is convex and each $x_t^{\mathrm{md}}$ and $x_t^{\mathrm{ag}}$ is a convex combination of points of $\mathcal{W}$. Write $\delta_t := G_t - \nabla h(x_t^{\mathrm{md}})$, $D_t := \|x_t - x^\diamond\|^2$, and $\nabla_t := \nabla h(x_t^{\mathrm{md}})$. Since $x_t^{\mathrm{ag}} - x_t^{\mathrm{md}} = a_t(x_t - x_{t-1})$, smoothness gives

$$h(x_t^{\mathrm{ag}}) \le h(x_t^{\mathrm{md}}) + \langle \nabla_t, x_t^{\mathrm{ag}} - x_t^{\mathrm{md}} \rangle + \frac{\beta a_t^2}{2} \|x_t - x_{t-1}\|^2 .$$

The first two terms on the right are affine in $x_t^{\mathrm{ag}}$. Split $x_t^{\mathrm{ag}}$ into its two convex-combination parts and use convexity at $x_{t-1}^{\mathrm{ag}}$ and at $x^\diamond$:

$$h(x_t^{\mathrm{ag}}) \le (1-a_t)h(x_{t-1}^{\mathrm{ag}}) + a_t\big[h(x^\diamond) + \langle \nabla_t, x_t - x^\diamond \rangle\big] + \frac{\beta a_t^2}{2} \|x_t - x_{t-1}\|^2 .$$

The projection step and the identity $2\langle a-b, b-c\rangle = \|a-c\|^2 - \|a-b\|^2 - \|b-c\|^2$ give

$$\gamma_t \langle G_t, x_t - x^\diamond \rangle \le \frac{1}{2}\big(D_{t-1} - D_t - \|x_t - x_{t-1}\|^2\big).$$

Substitute $\nabla_t = G_t - \delta_t$ and write $\langle \delta_t, x_t - x^\diamond\rangle = \langle \delta_t, x_{t-1} - x^\diamond\rangle + \langle \delta_t, x_t - x_{t-1}\rangle$. With $e_t := h(x_t^{\mathrm{ag}}) - h(x^\diamond)$, this yields

$$\begin{aligned} e_t \le (1-a_t)e_{t-1} + \frac{a_t}{2\gamma_t}(D_{t-1} - D_t) - a_t\langle \delta_t, x_{t-1} - x^\diamond\rangle \\ - \left(\frac{a_t}{2\gamma_t} - \frac{\beta a_t^2}{2}\right)\|x_t - x_{t-1}\|^2 + a_t \|\delta_t\| \, \|x_t - x_{t-1}\| . \end{aligned}$$

Since $\beta a_t \gamma_t = \beta\gamma t/(t+1) \le 1/2$, the coefficient of $\|x_t - x_{t-1}\|^2$ is at most $-a_t/(4\gamma_t)$. Maximizing $-a_t s^2/(4\gamma_t) + a_t\|\delta_t\| s$ over $s \ge 0$ bounds the second line by $a_t\gamma_t\|\delta_t\|^2$.

Let $\Gamma_t := 2/(t(t+1))$. Then $\Gamma_1 = 1$, $(1-a_t)/\Gamma_t = 1/\Gamma_{t-1}$ for $t \ge 2$, and $a_1 = 1$. Also $a_t/\Gamma_t = t$, $a_t/(2\gamma_t\Gamma_t) = 1/\gamma$, and $a_t\gamma_t/\Gamma_t = \gamma t^2/2$. Dividing by $\Gamma_t$ and summing over $t = 1, \dots, N$ gives

$$\frac{e_N}{\Gamma_N} \le \frac{D_0}{\gamma} + \sum_{t=1}^{N} \frac{\gamma t^2}{2}\|\delta_t\|^2 - \sum_{t=1}^{N} t\,\langle \delta_t, x_{t-1} - x^\diamond\rangle.$$

The initial distance bound, the square-integrability of $\nabla h(x^\diamond)$, the Lipschitz gradient, and the conditional variance bound imply inductively that $x_t - x^\diamond$, $x_t^{\mathrm{ag}} - x^\diamond$, and $G_t$ are square-integrable. Smoothness then makes $e_t$ integrable, so every term above is integrable. The vector $x_{t-1} - x^\diamond$ is $\mathcal{G}_{t-1}$-measurable and $\mathbb{E}[\delta_t \mid \mathcal{G}_{t-1}] = 0$, so the last sum has mean zero. Using $\sum_{t\le N} t^2 = N(N+1)(2N+1)/6$,

$$\mathbb{E}[e_N] \le \frac{2\mathcal{D}^2}{\gamma N(N+1)} + \frac{\gamma\sigma^2(2N+1)}{6}.$$

Let $\gamma_\star$ denote the second entry of the minimum defining $\gamma$. Then $1/\gamma \le 2\beta + 1/\gamma_\star$ and $\gamma \le \gamma_\star$, so

$$\mathbb{E}[e_N] \le \frac{4\beta\mathcal{D}^2}{N(N+1)} + 2\mathcal{D}\sigma\sqrt{\frac{2N+1}{3N(N+1)}}.$$

Finally, $(2N+1)/(N+1) \le 2$ proves (82). □

We now apply Lemma C11 to the smoothed objective, with the private estimator of Appendix C.5.1.

**Lemma C12** (Accelerated private stage)**.** *Let $\varepsilon, \mathcal{D} > 0$ with $\nu := L^2\mathcal{D}^2/\varepsilon^2 \ge 1$. Let $x_0 \in \mathcal{W}$ be measurable with respect to the stage-start sigma-field $\mathcal{G}_0$, and assume $\mathbb{E}[\mathrm{dist}(x_0, W^\star)^2] \le \mathcal{D}^2$. Set*

$$u := \frac{\varepsilon}{4L}, \qquad N := \left\lceil 12\sqrt{\nu}\, d^{1/4} \right\rceil, \qquad b := \left\lceil \frac{64\nu}{N} \right\rceil, \qquad \sigma^2 := \frac{L^2}{b}.$$

*Run the method of Lemma C11 with $h = F_u$, $\beta = \beta_u$, $x^\diamond := \Pi_{W^\star}(x_0)$, and $G_t := \widehat{g}_u(x_t^{\mathrm{md}})$, the estimator of Lemma C9 with $(m, q) = (m_b, q_b)$ from* (81)*. Every reply satisfies Definition 3.9, the output $x_N^{\mathrm{ag}} \in \mathcal{W}$ satisfies $\mathbb{E}[F(x_N^{\mathrm{ag}}) - F^\star] \le 3\varepsilon/4$, and the stage costs at most*

$$320\nu + \frac{58\,\nu^{3/4}d^{5/8}}{\sqrt{\rho}} + \frac{256\,\nu d}{\bar{m}\rho} + 143\sqrt{\nu}\, d^{1/4}\,\Psi \tag{83}$$

*component evaluations.*

*Proof. Privacy and estimator.* Each update of Lemma C11 evaluates one estimator at $x_t^{\mathrm{md}}$, which is measurable with respect to the sigma-field at the start of the update. Lemmas C9 and C10 show that every reply satisfies Definition 3.9, that $\mathbb{E}[G_t \mid \mathcal{G}_{t-1}] = \nabla F_u(x_t^{\mathrm{md}})$, and that the conditional variance is at most $\sigma^2$.

*Accuracy.* The projection onto the nonempty closed convex set $W^\star$ is continuous, so $x^\diamond \in W^\star \subseteq \mathcal{W}$ is $\mathcal{G}_0$-measurable, and $\mathbb{E}\,\|x_0 - x^\diamond\|^2 \le \mathcal{D}^2$. By Lemma C8, $F_u$ is convex with $\beta_u = 2L\sqrt{d}/u = 8L^2\sqrt{d}/\varepsilon$. Since $F_u$ is $L$-Lipschitz, $\|\nabla F_u\| \le L$, so the moment assumption of Lemma C11 holds. Lemma C11 and $N(N+1) \ge N^2$ give

$$\mathbb{E}\big[F_u(x_N^{\mathrm{ag}}) - F_u(x^\diamond)\big] \le \frac{32\,\varepsilon\nu\sqrt{d}}{N^2} + \frac{2\mathcal{D}L}{\sqrt{bN}} \le \frac{2\varepsilon}{9} + \frac{\varepsilon}{4},$$

since $N^2 \ge 144\nu\sqrt{d}$, $bN \ge 64\nu$, and $\sqrt{\nu} = L\mathcal{D}/\varepsilon$. By Lemma C8(i), $F \le F_u$ and $F_u(x^\diamond) \le F^\star + Lu = F^\star + \varepsilon/4$. Hence $\mathbb{E}[F(x_N^{\mathrm{ag}}) - F^\star] \le 2\varepsilon/9 + \varepsilon/2 \le 3\varepsilon/4$.

*Cost.* By Lemma C10, the stage costs at most

$$N\left(5b + 2\sqrt{\frac{db}{\rho}} + \frac{4db}{\rho\bar{m}}\right).$$

Now $Nb \le 64\nu + N$ and $N\sqrt{b} \le \sqrt{N}\sqrt{64\nu + N} \le 8\sqrt{N\nu} + N$. Since $\nu \ge 1$ and $d \ge 1$, also $N \le 13\sqrt{\nu}\,d^{1/4}$ and $\sqrt{N\nu} \le \sqrt{13}\,\nu^{3/4}d^{1/8}$. Therefore the cost is at most

$$320\nu + 16\sqrt{13}\,\frac{\nu^{3/4}d^{5/8}}{\sqrt{\rho}} + \frac{256\,\nu d}{\rho\bar{m}} + N\left(5 + 2\sqrt{\frac{d}{\rho}} + \frac{4d}{\rho\bar{m}}\right).$$

Finally, $16\sqrt{13} \le 58$ and $5 + 2\sqrt{d/\rho} + 4d/(\rho\bar{m}) \le 11\Psi$. This proves (83). □

### C.5.3 Growth localization across stages

We restart the accelerated stage with the certificate schedule of Algorithm 1. Only the stage method changes; the localization argument is that of Lemma C5.

*Proof of Theorem C7.* If $\alpha \ge \Delta_{\max}$, Lemma C4 shows that the public starting point is already $\alpha$-accurate, and no evaluation is needed. Hence assume $0 < \alpha < \Delta_{\max}$. All parameters below depend only on the public quantities $\theta, n, d, L, R, \lambda, \rho, \bar{m}$, and $\alpha$, so the construction is uniform over $\mathfrak{F}_{\theta,\lambda}(L, R)$. Every stage performs a deterministic number of updates, so the procedure terminates, and its output lies in $\mathcal{W}$ by Lemma C11.

*Global-radius branch.* Suppose $\alpha \ge \lambda R^\theta$, so that $K = L^2R^2/\alpha^2$. By (8), $\alpha < \Delta_{\max} \le LR$, hence $K \ge 1$. By Assumption 3.8, the public starting point satisfies $\mathrm{dist}(w^{(0)}, W^\star) \le R$. Lemma C12 with $\varepsilon = \alpha$ and $\mathcal{D} = R$ has $\nu = K$. It returns a point with expected excess error at most $3\alpha/4$, at the cost (83) with $\nu = K$.

*Growth-active branch.* Now suppose $\alpha < \lambda R^\theta$, so that $K = L^2\lambda^{-2/\theta}\alpha^{-p_\theta}$. Let $\Delta_j := 2^{-j}\Delta_{\max}$ and $J := \lceil \log_2(\Delta_{\max}/\alpha) \rceil$. Stage $j = 0, \dots, J-1$ starts from $w^{(j)}$ with $\mathbb{E}[F(w^{(j)}) - F^\star] \le \Delta_j$. For $j = 0$ this is Lemma C4. By the same lemma, $\mathbb{E}[\mathrm{dist}(w^{(j)}, W^\star)^2] \le \mathcal{D}_j^2 := \lambda^{-2/\theta}\Delta_j^{2/\theta}$. Apply Lemma C12 with $\varepsilon_j := \Delta_j/2$ and $\mathcal{D} = \mathcal{D}_j$. It returns $w^{(j+1)}$ with

$$\mathbb{E}[F(w^{(j+1)}) - F^\star] \le \frac{3\Delta_j}{8} \le \Delta_{j+1},$$

which is the invariant for the next stage. After $J$ stages, the expected excess error is at most $\Delta_J \le \alpha$.

Stage $j$ has

$$\nu_j := \frac{L^2\mathcal{D}_j^2}{\varepsilon_j^2} = 4L^2\lambda^{-2/\theta}\Delta_j^{-p_\theta}.$$

The bound $\Delta_j \le \Delta_{\max} \le (L^2\lambda^{-2/\theta})^{1/p_\theta}$, established in the proof of Theorem 5.3, gives $\nu_j \ge 4$. Since $\Delta_{J-1} > \alpha$, we have $\Delta_j > 2^{J-1-j}\alpha$. Together with $p_\theta \ge 1$, this gives $\nu_j < 4K\,2^{-(J-1-j)}$. Summing the geometric series,

$$\sum_{j<J} \nu_j \le 8K, \qquad \sum_{j<J} \nu_j^{3/4} \le \frac{2\sqrt{2}\,K^{3/4}}{1-2^{-3/4}} \le 7K^{3/4}, \qquad \sum_{j<J} \sqrt{\nu_j} \le \frac{2\sqrt{K}}{1-2^{-1/2}} \le 7\sqrt{K}.$$

Summing (83) over the stages bounds the total cost by

$$2560K + \frac{406\,K^{3/4}d^{5/8}}{\sqrt{\rho}} + \frac{2048\,Kd}{\bar{m}\rho} + 1001\sqrt{K}\,d^{1/4}\,\Psi.$$

Both branches therefore satisfy (78) with $C_{\mathrm{acc}} = 2560$.

*Simplification when $d \le K^2$.* Then $d^{1/4} \le \sqrt{K}$ and $d^{1/8} \le K^{1/4}$, so

$$\sqrt{K}\,d^{1/4} \le K, \qquad \sqrt{K}\,d^{1/4}\sqrt{\frac{d}{\rho}} \le \frac{K^{3/4}d^{5/8}}{\sqrt{\rho}}, \qquad \sqrt{K}\,d^{1/4}\,\frac{d}{\bar{m}\rho} \le \frac{Kd}{\bar{m}\rho}.$$

Since $\Psi$ is the maximum of 1, $\sqrt{d/\rho}$, and $d/(\bar{m}\rho)$, the last term of (78) is at most the sum of the other three. This proves (79) with $C'_{\mathrm{acc}} = 2C_{\mathrm{acc}}$, after taking the infimum over admissible pairs. □

### C.6 Combined upper bound and matching regime

The restarted and accelerated methods depend only on public parameters. We therefore choose between them before optimization according to the relation between $d$ and $K$.

**Theorem C13** (Combined upper bound)**.** *There is a universal constant $C_{\mathrm{up}} > 0$ such that, for every $\theta \ge 2$, $n, d \in \mathbb{N}_{\ge 1}$, $L, R, \lambda, \rho, \alpha > 0$, and $\bar{m} \in \mathbb{N}_{\ge 1} \cup \{\infty\}$,*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) \le C_{\mathrm{up}} \left( K + \min\left\{ K\sqrt{\frac{d}{\rho}}, \frac{K^{3/4}d^{5/8}}{\sqrt{\rho}} \right\} + \frac{Kd}{\bar{m}\rho} \right). \tag{84}$$

*The right-hand side is at most $3C_{\mathrm{up}}K\Psi$.*

*Proof.* If $\alpha \ge \Delta_{\max}$, return the public starting point without any evaluations, by Lemma C4. Hence assume $0 < \alpha < \Delta_{\max}$. The quantities $K$ and $d$ are public. If $d \le K^2$, run the pair of Theorem C7. Otherwise run the pair of Theorem 5.3. Both pairs are $\alpha$-accurate and satisfy Definition 3.9.

If $d \le K^2$, then (79) applies. Moreover $d^{1/8} \le K^{1/4}$ gives $K^{3/4}d^{5/8}/\sqrt{\rho} \le K\sqrt{d/\rho}$, so the accelerated term is the minimum in (84).

If $d > K^2$, Theorem 5.3 gives cost $O(K\Psi) \le O(K + K\sqrt{d/\rho} + Kd/(\bar{m}\rho))$. Now $d^{1/8} > K^{1/4}$ gives $K\sqrt{d/\rho} < K^{3/4}d^{5/8}/\sqrt{\rho}$, so the restarted term is the minimum.

The final claim follows because the minimum is at most $K\sqrt{d/\rho}$, and $K + K\sqrt{d/\rho} + Kd/(\bar{m}\rho) \le 3K\Psi$. □

The switch at $d = K^2$ is exactly where the two privacy terms in (84) cross. The public choice therefore has an upper bound no worse in order than the restarted bound; its privacy term without $\bar{m}$ is smaller when $d < K^2$.

**Corollary C14** (Matching regime)**.** *Assume the hypotheses of Theorem B1. If*

$$\rho \gtrsim \max\left\{ \min\left\{ d, \frac{d^{5/4}}{\sqrt{K}} \right\}, \frac{d}{\bar{m}} \right\}, \tag{85}$$

*where the term $d/\bar{m}$ is omitted when $\bar{m} = \infty$, then*

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha; \mathfrak{F}_{\theta,\lambda}(L,R)\big) = \Theta(K).$$

*Proof.* The lower bound

$$\mathfrak{T}_{\rho,\bar{m}}\big(\alpha;\mathfrak{F}_{\theta,\lambda}(L,R)\big) \geq c_1 K$$

is the first term of Theorem B1.

For the upper bound, the first term inside the maximum in (85) implies either

$$\rho \gtrsim d$$

or

$$\rho \gtrsim \frac{d^{5/4}}{\sqrt{K}},$$

according to which term realizes the minimum. Hence the privacy term

$$\min\left\{K\sqrt{\frac{d}{\rho}}, \frac{K^{3/4}d^{5/8}}{\sqrt{\rho}}\right\}$$

is $O(K)$. The second term in (85) gives $\rho\bar{m} \gtrsim d$, and therefore

$$\frac{Kd}{\bar{m}\rho} = O(K).$$

Theorem C13 now gives the matching $O(K)$ upper bound. □

Since $d^{5/4}/\sqrt{K} \leq d$ exactly when $d \leq K^2$, the first term in (85) equals $d^{5/4}/\sqrt{K}$ for $d \leq K^2$ and $d$ for $d > K^2$. Thus acceleration enlarges the certified matching regime when $d < K^2$. In this regime, the privacy-dependent terms are absorbed by the optimization scale $K$. A privacy-dependent gap remains outside the certified regime (85). Under the hypotheses of Theorem B7, the privacy term of (84) exceeds $\sqrt{Kd/\rho}$ by the factor

$$\min\{\sqrt{K}, K^{1/4}d^{1/8}\},$$

while the batch-size term $Kd/(\bar{m}\rho)$ exceeds $\kappa_\theta d/(\bar{m}\rho)$ in (46) by

$$\frac{K}{\kappa_\theta} \in [\sqrt{K}, K^{2/3}].$$

# D Extended Comparison to Related Work

This section expands the comparisons in Section 2.

*Resource and privacy model.* DP-ERM and DP-SCO optimize different statistical objectives—empirical and population risk, respectively—while the private-proxy guarantee studied here is different from end-to-end dataset-level differential privacy. We therefore use results from DP-ERM and DP-SCO only as computational and statistical points of comparison.

*Smooth DP-SCO under growth.* For convex, $\beta$-smooth, and Lipschitz losses whose population risk satisfies $(\theta,\lambda)$-TNC, with a known lower bound $\bar{\theta} > 1$ on $\theta$, Su et al. (2024, Theorem 14) obtain an $(\epsilon,\delta)$-DP excess population-risk bound of

$$\widetilde{O}\left(\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)^{\frac{\theta}{\theta-1}}\right),$$

with the Lipschitz and growth factors suppressed. The result requires a sufficiently large sample size, $\epsilon \leq 2\log(1/\delta)$, and step sizes at most $1/\beta$. Their method applies Phased-SGD (Feldman et al., 2020) to disjoint subsets of the sample and therefore uses at most $n$ stochastic-gradient evaluations. Solving the preceding bound for the sample size needed to attain excess risk $\alpha$ gives

$$\widetilde{O}\left(\alpha^{-\frac{2(\theta-1)}{\theta}} + \frac{\sqrt{d\log(1/\delta)}}{\epsilon}\alpha^{-\frac{\theta-1}{\theta}}\right)$$

stochastic-gradient evaluations.

Thus TNC improves the accuracy dependence in smooth DP-SCO. For nonsmooth losses, Su et al. (2024) obtain the corresponding privacy–utility improvement by replacing Phased-SGD with Phased-ERM. This extension does not directly yield the same linear gradient-complexity guarantee. Our setting instead concerns component first-order complexity for nonsmooth finite-sum optimization in the private-proxy model.

*Lower bounds under growth.* Asi et al. (2021b) study private stochastic convex optimization under $\kappa$-growth. They give localization-based algorithms that adapt to the growth exponent together with matching lower bounds on the excess risk of dataset-level private algorithms. For $(\epsilon, \delta)$-DP and $\kappa \geq 2$, their lower bound is obtained through a reduction from general convex ERM: an algorithm for growth ERM can be used to solve a sequence of problems regularized toward the previous iterate (Asi et al., 2021b, Proposition 3).

Our private lower bound also transfers hardness from a general convex class to the growth class, but for component-evaluation complexity in the private-proxy model. Lemma B5 uses a single public regularizer while preserving conditional per-reply privacy, joint processing within each batch, the maximum batch size, and the total evaluation count. The source hardness is the normalized convex private-proxy information lower bound of Appendix B.3.1. Combining this with the independent non-private construction gives the maximum of the optimization and information costs in Theorem B1. Under additional dimension and batch-size conditions, Theorem B7 gives the refined lower bound by directly analyzing a growth-constrained hard construction. None of these results relies on the Menart–Nikolov product lower bound; Appendix B.5 discusses that result separately.

# E Experimental Protocols and Results

The experiments illustrate the private-proxy calibration and the optimization schedules analyzed in Section 5. They are not intended to test the lower-bound regime.

## E.1 Experimental setup

### E.1.1 Controlled objectives

We use a family whose growth constants are known exactly. Let $\mathcal{W} = \mathbb{B}_2(1/2)$ and $w^\star = u/2$, where $u = 0.8e_1 + 0.6v$ and $v$ is a fixed public unit vector orthogonal to $e_1$. The component losses are

$$f_i(w) = \phi_\theta(\|w - w^\star\|) + \langle a_i, w\rangle, \qquad \phi_\theta(r) = \begin{cases} r^\theta/\theta, & r \leq 1, \\ 1/\theta + r - 1, & r > 1. \end{cases} \tag{86}$$

where the vectors $a_i$ occur in opposite pairs with $\|a_i\| = 1/8$. Their average vanishes, so $F(w) - F^\star = \|w - w^\star\|^\theta/\theta$ on $\mathcal{W}$. Every component is globally $L = 9/8$-Lipschitz, growth holds with $\lambda = 1/\theta$, and the distance bound is $R = 1$. All methods start at $w_0 = -e_1/2$ with the initial error bound $\Delta_0 = 0.9^{\theta/2}/\theta$. This is the exact initial gap and provides the initial certificate for Algorithm 1; it is smaller than the generic public bound $\Delta_{\max} = 9/8$.

### E.1.2 Regularized hinge-loss ERM

Let $\ell(z) = \max\{0, 1 - z\}$ be the hinge loss, with labels $y_i \in \{-1, +1\}$ and examples normalized so that $x_{\max} = \max_i \|x_i\| \leq 1$. With the Huberized ridge

$$H_{\mu,S}(w) = \begin{cases} \frac{\mu}{2}\|w\|^2, & \|w\| \leq S, \\ \mu S\|w\| - \frac{\mu S^2}{2}, & \|w\| > S, \end{cases} \tag{87}$$

we minimize

$$F(w) = \frac{1}{n}\sum_{i=1}^n \ell\big(y_i\langle x_i, w\rangle\big) + H_{\mu,S}(w), \qquad \mathcal{W} = \{w : \|w\| \leq R_{\mathrm{dom}}\}, \tag{88}$$

where $\mu > 0$ and the Huber knee satisfies $S \geq R_{\mathrm{dom}}$. On $\mathcal{W}$ the Huberized ridge equals $\frac{\mu}{2}\|w\|^2$, so $F$ is $\mu$-strongly convex there and satisfies Definition 3.6 with

$$\theta = 2, \qquad \lambda = \frac{\mu}{2}. \tag{89}$$

Outside $\mathcal{W}$, the Huberized extension keeps the subgradients bounded, which is needed because the proxy model allows queries outside the feasible set. Since the hinge loss has subgradients of magnitude at most one,

$$\|g_i(w)\| \le x_{\max} + \mu S \le 1 + \mu S =: L \qquad \text{for all } w \in \mathbb{R}^d. \tag{90}$$

All methods start at $w_0 = 0$ with $\Delta_0 = \ell(0) = 1$, which bounds $F(0) - F^\star$ because the loss and the regularizer are nonnegative. We use the class radius $R_{\text{class}} = 2R_{\text{dom}}$, since any two points in the feasible ball are at distance at most $2R_{\text{dom}}$. We set $\mu = 10^{-2}$ and $R_{\text{dom}} = S = 15$, so that $\lambda = 1/200$, $L = 23/20$, and $R_{\text{class}} = 30$ on both datasets. Growth at $w_0$ gives $\|w^\star\| \le \sqrt{\Delta_0/\lambda} = \sqrt{200} < 15$ for the constrained minimizer $w^\star$, so $w^\star$ is interior and no reference solution is needed to choose the domain.

#### E.1.3 Methods

All methods use the private estimator of Lemma C2, with $m$ and $q$ chosen by (73), and the common second-moment bound $G_{m,q}^2 \le 2L^2$. We compare three schedules.

- *Generic* runs one projected-subgradient phase with step size $\alpha/(2L^2)$ and horizon $\lceil 2L^2R^2/\alpha^2 \rceil$, using the global radius $R$ and the projected-stage bound in the proof of Lemma C3.
- *One-shot* uses the growth condition once, replacing $R$ by $D_{\Delta_0}$ of Corollary C6.
- *Restart* runs Algorithm 1: $\eta_j = \Delta_j/(8L^2)$ and $T_j = \lceil 16L^2\lambda^{-2/\theta}\Delta_j^{-p_\theta} \rceil$.

Each stage output averages the pre-update iterates of the stage, as in the analysis.

#### E.1.4 Private proxy and evaluation cost

Each update consists of $q$ separate proxy replies at a fixed query point, with $m$ component evaluations per reply, and is therefore charged $qm$ component evaluations. Objective evaluations used only for logging or reference solutions are not charged. For computational efficiency, the GPU implementation draws the same $qm$ component indices and samples directly from the distribution of the average Gaussian noise. This produces the same averaged estimator and iterate distribution as $q$ separate private replies; the underlying protocol and charged cost remain $q$ replies of size $m$.

#### E.1.5 Evaluation protocol

We record the running average at every stage end and at up to 128 log-spaced points within each stage, fixed before the runs. The stage lemmas bound the expected excess error of the output of a complete stage; the figures show these outputs as filled markers and the running average within a stage as a faint curve. We also report the *first logged target crossing*: the first recorded evaluation count at which the running average reaches the target $\alpha$. It is an empirical diagnostic rather than a stopping guarantee, and every run executes its complete schedule. Each configuration is repeated with 20 random seeds (10 on the ERM tasks and for $\rho \in \{1/4, 4\}$), paired across methods, and we report medians and interquartile ranges. A run that never reaches its target would be recorded as unreached rather than discarded; every run reached its target.

#### E.1.6 Tested regimes

All configurations lie on the growth-active branch: $\Delta_0 < \lambda R^\theta$. Indeed, $\lambda R^\theta = 1/\theta$ for the controlled objectives, while $\lambda R_{\text{class}}^2 = 9/2$ for the ERM tasks. The ratio $\Delta_0/\alpha$ determines the number of Restart stages. In our sweep these ratios are powers of two, so

$$J = \log_2(\Delta_0/\alpha).$$

When $\Delta_0/\alpha = 2$, Restart has a single stage and therefore performs no restart.

In every configuration, $\sqrt{d/\rho} \ge 4$, so a privacy-dependent term in $\Psi_{\rho,\bar{m}}$ dominates its non-private term 1. When $\bar{m} = \infty$, as in Appendices E.2 and E.3, each update uses one reply of size $\lceil \sqrt{2d/\rho} \rceil$.

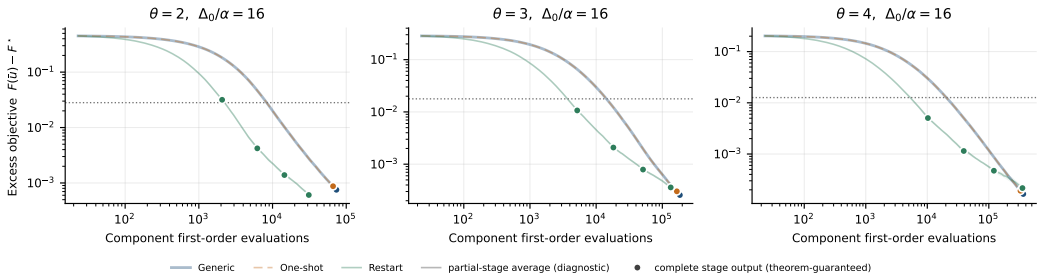


Figure 1: Controlled objectives at the tightest target $\Delta_0/\alpha = 16$. Curves are medians and bands are interquartile ranges over 20 runs; markers and curves are as described in Appendix E.1.5. The dotted line is the target.

The sweep of Appendix E.4 also includes $\bar{m} < \sqrt{d/\rho}$, where $q > 1$ replies are averaged and the per-update cost scales as $d/(\bar{m}\rho)$.

The sufficient conditions of Theorem 4.1 are not satisfied by the tested configurations. With the explicit constants in its proof, the accuracy condition implies $K_{\theta,\lambda} \geq 11520$, whereas the largest tested value is $K_{\theta,\lambda} = 2116$. The experiments therefore illustrate the upper-bound behavior of Theorem 5.3, which applies to all tested configurations, rather than the regime covered by the lower-bound theorem.

### E.2 Controlled growth results

We use $\theta \in \{2, 3, 4\}$, $d = 256$, $\rho = 1$, $\bar{m} = \infty$, and targets $\Delta_0/\alpha \in \{2, 4, 8, 16\}$. Table 2 reports the median first logged target crossings, and Figure 1 shows the trajectories at the tightest target. Generic and One-shot use the same step size and paired seeds, so their trajectories coincide until their horizons differ.

The advantage of Restart increases as the target tightens. Its median crossing cost is 1.94–2.04× smaller than that of Generic at $\Delta_0/\alpha = 8$ and 3.75–3.97× smaller at $\Delta_0/\alpha = 16$, across the three values of $\theta$. At $\Delta_0/\alpha = 2$, Restart is slower. These trends are consistent with the initial step-size ratio: the first Restart step is $\Delta_0/(8L^2)$, whereas Generic and One-shot use $\alpha/(2L^2)$, giving the ratio $(\Delta_0/\alpha)/4$. The measured Generic-to-Restart cost ratios track this quantity closely (Table 1).

Table 1: Step-size ratios and measured cost ratios for the controlled objectives.

| $\Delta_0/\alpha$ | step ratio | measured cost ratio |
|---|---|---|
| 2 | 0.50 | 0.52–0.56 |
| 4 | 1.00 | 1.00–1.01 |
| 8 | 2.00 | 1.94–2.04 |
| 16 | 4.00 | 3.75–3.97 |

In 11 of the 12 Restart configurations, all runs reach the target no later than the end of the first stage. These experiments therefore reflect the calibrated schedules as a whole; they do not isolate the effect of repeated restarting, from which Theorem 5.3 derives the improved accuracy dependence.

### E.3 Results on regularized hinge-loss ERM

We use the training splits of `a9a` ($n = 32,561$, $d = 123$, density 0.11276) and `rcv1.binary` ($n = 20,242$, $d = 47,236$, density 0.00157), with labels in $\{-1, +1\}$, no intercept, preserved sparsity, and per-example normalization. We set $\rho = 1$, $\bar{m} = \infty$, and $\alpha = 1/8$; the target was fixed before any reference solution was computed. Reference optima are used only for evaluation. Their weak-duality gaps are $1.2e-13$ and $0.0e+00$, and their norms, 2.680 and 5.678, are below $R_{\mathrm{dom}} = 15$. Table 3 and Figure 2 report the results.

Table 2: Controlled objectives: median first logged target crossing over 20 runs.

| $\theta$ | $\Delta_0/\alpha$ | Generic | One-shot | Restart | Restart crossings in stage 1 |
|---|---|---|---|---|---|
| 2 | 2 | 230 | 230 | 414 | 20/20 |
| 2 | 4 | 874 | 874 | 874 | 20/20 |
| 2 | 8 | 2,921 | 2,852 | 1,472 | 20/20 |
| 2 | 16 | 8,487 | 8,349 | 2,254 | 0/20 |
| 3 | 2 | 276 | 276 | 529 | 20/20 |
| 3 | 4 | 1,265 | 1,242 | 1,265 | 20/20 |
| 3 | 8 | 4,462 | 4,393 | 2,300 | 20/20 |
| 3 | 16 | 14,398 | 14,329 | 3,841 | 20/20 |
| 4 | 2 | 322 | 322 | 598 | 20/20 |
| 4 | 4 | 1,587 | 1,564 | 1,564 | 20/20 |
| 4 | 8 | 6,279 | 6,233 | 3,082 | 20/20 |
| 4 | 16 | 21,735 | 21,758 | 5,474 | 20/20 |

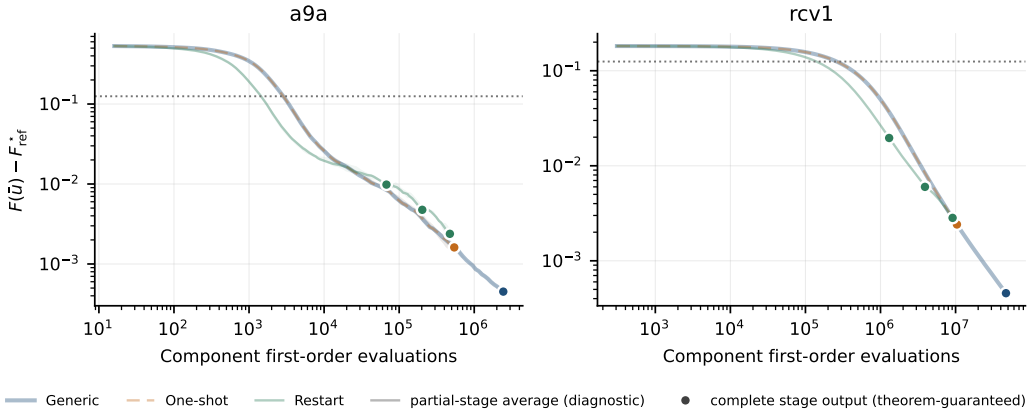


Figure 2: Regularized hinge-loss ERM. Markers and curves as in Figure 1; the vertical axis is the excess objective with respect to the evaluation-only reference optimum.

Restart has the smaller median first logged target crossing on both datasets: $1,488$ against $3,088$ for Generic on `a9a` ($2.08\times$), and $148,764$ against $267,344$ on `rcv1` ($1.80\times$). As on the controlled objectives, every Restart crossing occurs no later than the end of the first stage ($10/10$ and $10/10$ runs), and at this target the first Restart step is twice the Generic step. The complete schedules have different lengths: Generic runs about five times longer than Restart, so the terminal errors in Table 3 are attained at unequal budgets and do not rank the methods. On `rcv1`, the target is only $1.46\times$ below the initial gap $0.1820$, compared with $4.25\times$ on `a9a`, so the `rcv1` comparison is weaker evidence.

### E.4 Effect of the maximum batch size

We fix $\theta = 2$ and $\alpha = \Delta_0/8$ on the controlled objectives and vary $d \in \{64, 256, 1024, 4096\}$, $\rho \in \{1, 1/4, 4\}$, and $\bar{m} \in \{1, \lceil s/8 \rceil, \lceil s/4 \rceil, \lceil s/2 \rceil, \lceil s \rceil, \lceil s_{\mathrm{cal}} \rceil, \lceil 2s \rceil, \lceil 4s \rceil, \infty\}$, where $s = \sqrt{d/\rho}$ is the threshold in $\Psi_{\rho,\bar{m}}$ and $s_{\mathrm{cal}} = \sqrt{2d/\rho}$ determines the largest batch size chosen by (73). All 104 configurations lie on the growth-active branch and use Restart. Figure 3 shows the results at $\rho = 1$.

The measured cost follows the shape predicted by $\Psi_{\rho,\bar{m}}$. Below the threshold $\sqrt{d/\rho}$, one update uses $q = \Theta(d/(\rho\bar{m}^2))$ replies of size $\bar{m}$, and hence $\Theta(d/(\rho\bar{m}))$ component evaluations (Lemma 5.1). Above this threshold, the dependence on $\bar{m}$ flattens. At $d = 4096$ and $\rho = 1$, the median crossing cost decreases by approximately $90\times$ over the tested batch sizes.

Table 3: Regularized hinge-loss ERM: medians over 10 runs.

| Dataset | Method | $m$ | $qm$ | Crossing (diag.) | Scheduled cost | Terminal excess |
|---|---|---|---|---|---|---|
| a9a | Generic | 16 | 16 | 3,088 | 2,437,632 | 4.513e-04 |
| a9a | One-shot | 16 | 16 | 3,072 | 541,696 | 1.611e-03 |
| a9a | Restart | 16 | 16 | 1,488 | 473,984 | 2.383e-03 |
| rcv1 | Generic | 308 | 308 | 267,344 | 46,924,416 | 4.563e-04 |
| rcv1 | One-shot | 308 | 308 | 281,204 | 10,427,648 | 2.412e-03 |
| rcv1 | Restart | 308 | 308 | 148,764 | 9,124,192 | 2.830e-03 |

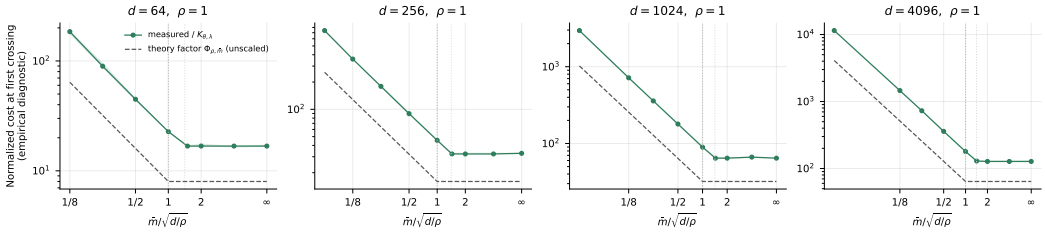


Figure 3: Effect of the maximum batch size at $\rho = 1$: median first logged target-crossing cost, normalized by $K_{\theta,\lambda}$, against $\bar{m}/\sqrt{d/\rho}$. The dashed curve is $\Psi_{\rho,\bar{m}}$ with no fitted constant. The rightmost marker is $\bar{m} = \infty$, and the dotted vertical lines mark $\bar{m}/s = 1$ and $\bar{m}/s = \sqrt{2}$.

All configurations use the same optimization schedule, including the same stage lengths, step sizes, and checkpoint update indices; changing $\bar{m}$ changes the evaluation cost $qm$ and the estimator distribution. Across all configurations, the median number of updates at the first logged crossing lies between 61 and 67.

The measured cost flattens near $s_{\mathrm{cal}} = \sqrt{2d/\rho}$, the threshold used by the estimator calibration in (73). This differs by a factor $\sqrt{2}$ from the threshold $\sqrt{d/\rho}$ appearing in $\Psi_{\rho,\bar{m}}$.

### E.5 Implementation details

We use two implementations with a common schedule module: a CPU reference that explicitly generates the $q$ replies of $m$ component evaluations, and a batched GPU implementation using the equivalent sampling procedure of Appendix E.1.4. With identical randomness, they agree exactly on all discrete quantities (updates, stage lengths, step sizes, reply and evaluation counts, and branch selection) and to within $1.2e-15$ relative error on the trajectories. The experiments comprise 146 configurations and 2170 seeded runs with 0 failures, use 1,293,253,180 component evaluations in total, and took 5.4 minutes on four GPUs.