High-Dimensional Private Empirical Risk Minimization by Greedy Coordinate Descent

Paul Mangold[1], Aurélien Bellet[1], Joseph Salmon[2,3], and Marc Tommasi[4]

[1]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[2]IMAG, Univ Montpellier, CNRS, Montpellier, France
[3]Institut Universitaire de France (IUF)
[4]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

ABSTRACT

In this paper, we study differentially private empirical risk minimization (DP-ERM). It has been shown that the (worst-case) utility of DP-ERM reduces as the dimension increases. This is a major obstacle to privately learning large machine learning models. In high dimension, it is common for some model's parameters to carry more information than others. To exploit this, we propose a differentially private greedy coordinate descent (DP-GCD) algorithm. At each iteration, DP-GCD privately performs a coordinate-wise gradient step along the gradients' (approximately) greatest entry. We show theoretically that DP-GCD can improve utility by exploiting structural properties of the problem's solution (such as sparsity or quasi-sparsity), with very fast progress in early iterations. We then illustrate this numerically, both on synthetic and real datasets. Finally, we describe promising directions for future work.

1 Introduction

Machine learning crucially relies on data, which can be sensitive or confidential. Unfortunately, trained models are susceptible of leaking information about specific training points (Shokri et al., 2017). A standard approach for training models while provably controlling the amount of leakage is to solve an empirical risk minimization (ERM) problem under a differential privacy (DP) constraint (Chaudhuri et al., 2011). In this work, we consider the following generic problem formulation:

$$w^* \in \operatorname*{arg\,min}_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) \ , \tag{1}$$

where $D = (d_1, \ldots, d_n)$ is a dataset of n samples drawn from a universe $\mathcal{X}$, and $\ell(\cdot, d) : \mathbb{R}^p \to \mathbb{R}$ is a loss function which is convex and smooth for all $d \in D$.

The DP constraint in DP-ERM induces a trade-off between the precision of the solution (utility) and privacy. Bassily et al. (2014) proved lower bounds on utility under a fixed DP budget. These lower bounds scale polynomially with the dimension p. Since machine learning models are often high-dimensional (*e.g.,* $n \approx p$ or even $n \ll p$), this is a massive drawback for the use of DP-ERM.

In high-dimensional models, some parameters are typically more significant than others. It is notably (but not only) the case when models are sparse, which is often a desired property in high dimension (Tibshirani, 1996). Private learning algorithms could thus be designed to take advantage of this by focusing on the most significant parameters of the problem.

Several works have tried to exploit such high-dimensional models' structure to reduce the dependence on the dimension, *e.g.,* from polynomial to logarithmic. Talwar et al. (2015), Bassily et al. (2021), and Asi

et al. (2021) proposed a DP Frank-Wolfe algorithm (DP-FW) that exploits the solution's sparsity. However, their algorithm only works on ℓ_1-constrained DP-ERM, restricting its range of application. For sparse linear regression, Kifer et al. (2012) proposed to first identify some support and then solve the DP-ERM problem on the restricted support. Unfortunately, their approach requires implicit knowledge of the solution's sparsity. Finally, Kairouz et al. (2021) and Zhou et al. (2021) used public data to estimate lower-dimensional subspaces, where the gradient can be computed at a reduced privacy cost. A key limitation is that such public data set, from the same domain as the private data, is typically not available in many learning scenarios involving sensitive data.

In this work, we propose a private algorithm that does not have these pitfalls: the differentially private greedy coordinate descent algorithm (DP-GCD). At each iteration, DP-GCD privately determines the gradient's greatest coordinate, and performs a gradient step in this direction. Hence, it avoids wasting privacy budget on updating non-significant parameters, focusing on the truly useful ones. Our algorithm works on any smooth, unconstrained DP-ERM problem, and is adaptive to the sparsity of the solution. It can also ignore small (but non-zero) parameters, improving utility even on non-sparse problems.

We show formally that DP-GCD can achieve an excess risk of $\widetilde{O}(\log(p)/\mu_1^2 n^2 \epsilon^2)$ on problems with μ_1-strongly-convex objective w.r.t. the ℓ_1-norm. When $\mu_1 = O(1)$, DP-GCD is thus the first algorithm (to our knowledge) whose utility scales with $\log(p)/n^2$. We derive rates for the convex case as well. Furthermore, we prove that on problems with quasi-sparse solutions (including but not limited to sparse problems, see Definition 3.5), DP-GCD progresses particularly fast in the first iterations. This property is very appealing in private optimization, where performing more iterations requires increasing the noise to guarantee privacy. These theoretical observations are then confirmed by our experiments on real and synthetic datasets. Our contributions can be summarized as follows:

1. We propose differentially private greedy coordinate descent (DP-GCD), a coordinate method that, at each iteration, performs updates along the (approximately) greatest entry of the gradient. We formally prove its private nature, and derive corresponding high probability utility upper bounds.
2. We prove that DP-GCD exploits underlying structural properties of the problem (*e.g.,* quasi-sparse solutions) to improve utility beyond the worst-case. Importantly, DP-GCD does not require prior knowledge of this structure to exploit it.
3. We confirm our theoretical results numerically on a variety of synthetic and real datasets, showing that DP-GCD indeed outperforms existing private optimization algorithms when the problem's solution is quasi-sparse.

The rest of the paper is organized as follows. First, we present the relevant mathematical background in Section 2. Section 3 then introduces DP-GCD, and formally analyzes its privacy and utility. The relation to existing work is discussed in Section 4. We confirm our theoretical results numerically in Section 5. Finally, we conclude and discuss the limitations of our results in Section 6.

2 Preliminaries

In this section, we introduce important technical notions that will be used throughout the paper.

Norms. We start by defining two conjugate norms that will allow to keep track of coordinate-wise quantities. Let $M = \operatorname{diag}(M_1, \dots, M_p)$ with $M_1, \dots, M_p > 0$, and

$$\|w\|_{M,1} = \sum_{j=1}^{p} M_j^{\frac{1}{2}} |w_j| \ , \qquad \|w\|_{M^{-1},\infty} = \max_{j \in [p]} M_j^{-\frac{1}{2}} |w_j| \ .$$

When M is the identity matrix I, $\|\cdot\|_{M,1}$ is the standard ℓ_1-norm and $\|\cdot\|_{M^{-1},\infty}$ is the ℓ_∞-norm. We also define the Euclidean dot product $\langle u, v \rangle = \sum_{j=1}^{p} u_i v_i$ and corresponding norms $\|\cdot\|_{M,2} = \langle \cdot, M \cdot \rangle^{\frac{1}{2}}$ and $\|\cdot\|_{M^{-1},2} = \langle \cdot, M^{-1} \cdot \rangle^{\frac{1}{2}}$. Similarly, we recover the standard ℓ_2-norm when $M = I$.

Regularity assumptions. We recall classical regularity assumptions along with ones specific to the coordinate-wise setting. We denote by ∇f the gradient of a differentiable function f, and by $\nabla_j f$ its j-th coordinate. We denote by e_j the j-th vector of $\mathbb{R}^p$'s standard basis.

(Strong)-convexity. For $q \in \{1, 2\}$, a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is $\mu_{M,q}$-strongly-convex w.r.t. the norm $\|\cdot\|_{M,q}$ if for all $v, w \in \mathbb{R}^p$, $f(w) \geq f(v) + \langle \nabla f(v), w - v\rangle + \frac{\mu_M}{2}\|w - v\|^2_{M,q}$. The case $M_{1,q} = \cdots = M_{p,q} = 1$ recovers standard $\mu_{I,q}$-strong convexity w.r.t. the ℓ_q-norm. When $\mu_{M,q} = 0$, the function is just said to be *convex*.

Component Lipschitzness. A function $f : \mathbb{R}^p \to \mathbb{R}$ is L-component-Lipschitz for $L = (L_1, \ldots, L_p)$ with $L_1, \ldots, L_p > 0$ if for $w \in \mathbb{R}^p$, $t \in \mathbb{R}$ and $j \in [p]$, $|f(w + te_j) - f(w)| \leq L_j|t|$. For $q \in \{1, 2\}$, f is Λ_q-Lipschitz w.r.t. $\|\cdot\|_q$ if for $v, w \in \mathbb{R}^p$, $|f(v) - f(w)| \leq \Lambda_q\|v - w\|_q$.

Component smoothness. A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is M-component-smooth for $M_1, \ldots, M_p > 0$ if for $v, w \in \mathbb{R}^p$, $f(w) \leq f(v) + \langle \nabla f(v), w - v\rangle + \frac{1}{2}\|w - v\|^2_{M,2}$. When $M_1 = \cdots = M_p = \beta$, f is said to be β-smooth.

Component-wise regularity assumptions are not restrictive: for $q \in \{1, 2\}$, Λ_q-Lipschitzness w.r.t. $\|\cdot\|_q$ implies $(\Lambda_q, \ldots, \Lambda_q)$-component-Lipschitzness and β-smoothness implies $(\beta, \ldots, \beta)$-component-smoothness. Yet, the actual component-wise constants of a function can be much lower than what can be deduced from their global counterparts. In the following of this paper, we will use $M_{\min} = \min_{j\in[p]} M_j$, $M_{\max} = \max_{j\in[p]} M_j$, and their Lipschitz counterparts $L_{\min}$ and $L_{\max}$.

Differential privacy (DP). Let $\mathcal{D}$ be a set of datasets and $\mathcal{F}$ a set of possible outcomes. Two datasets $D, D' \in \mathcal{D}$ are said *neighboring* (denoted by $D \sim D'$) if they differ on at most one element.

Definition 2.1 (Differential Privacy, Dwork 2006)**.** A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is (ϵ, δ)-differentially private if, for all neighboring datasets $D, D' \in \mathcal{D}$ and all $S \subseteq \mathcal{F}$ in the range of $\mathcal{A}$:

$$\Pr\left[\mathcal{A}(D) \in S\right] \leq e^\epsilon \cdot \Pr\left[\mathcal{A}(D') \in S\right] + \delta \ .$$

In this paper, we consider the classic central model of DP, where a trusted curator has access to the raw dataset and releases a model trained on this dataset[1].

A common principle for releasing a private estimate of a function $h : \mathcal{D} \to \mathbb{R}^p$ is to perturb it with Laplace or Gaussian noise. To ensure privacy, the noise is scaled with the sensitivity $\Delta_q(h) = \sup_{D\sim D'}\|h(D) - h(D')\|_q$ of h, with $q = 1$ for Laplace, and $q = 2$ for Gaussian mechanism. In coordinate descent methods, we release coordinate-wise gradients. The j-th coordinate of a loss function's gradient $\nabla_j \ell : \mathbb{R}^p \to \mathbb{R}$ has sensitivity $\Delta_1(\nabla_j f) = \Delta_2(\nabla_j f)$ (since $\nabla_j f$ is scalar). For a L-component-Lipschitz loss, these sensitivities are upper bounded by $2L_j$ (Mangold et al., 2022).

In our algorithm, we will also need to compute the index of the gradient's maximal entry privately. To this end, we use the report-noisy-argmax mechanism (Dwork and Roth, 2013). This mechanism perturbs each entry of a vector with Laplace noise, calibrated to its *coordinate-wise* sensitivities, and releases the index of a maximal entry of this noisy vector. Revealing only this index allows to greatly reduce the noise, in comparison to releasing the full gradient. This will be the cornerstone of our greedy algorithm.

3 Private Greedy CD

In this section we present our main contribution: the differentially private greedy coordinate descent algorithm (DP-GCD). DP-GCD updates only one parameter per iteration. This coordinate is selected greedily as the (approximately) largest entry of the gradient, hoping to maximize the improvement in utility at each iteration. We establish privacy and utility guarantees for DP-GCD and show that it is well-suited for high-dimensional problems with a *quasi-sparse* solution (*i.e.,* with a fraction of the parameters dominating the others).

[1] In fact, our privacy guarantees hold even if all intermediate iterates are released (not just the final model).

3.1 The Algorithm

At each iteration, DP-GCD (Algorithm 1) updates the parameter with the greatest gradient value (rescaled by the inverse square root of the coordinate-wise smoothness constant). This corresponds to the Gauss-Southwell-Lipschitz rule (Nutini et al., 2015). To guarantee privacy, this selection is done using the report-noisy-max mechanism (Dwork and Roth, 2013) with noise scales λ'_j along j-th entry ($j \in [p]$). DP-GCD then performs a gradient step with step size $\gamma_j > 0$ along this direction. The gradient is privatized using the Laplace mechanism (Dwork and Roth, 2013) with scale λ_j.

Algorithm 1 DP-GCD: Private Greedy CD

Input: initial $w^0 \in \mathbb{R}^p$, iteration count $T > 0, \forall j \in [p]$, noise scales λ_j, λ'_j, step sizes $\gamma_j > 0$.
for $t = 0$ to $T - 1$ **do**
  $j_t = \arg\max_{j' \in [p]} \frac{|\nabla_{j'} f(w^t) + \chi^t_{j'}|}{\sqrt{M_{j'}}}$, with $\chi^t_{j'} \sim \mathrm{Lap}(\lambda'_{j'})$. ▷ Choose j_t using report-noisy-max.
  $w^{t+1} = w^t - \gamma_j(\nabla_j f(w^t) + \eta^t_{j_t})e_{j_t}$, with $\eta^t_{j_t} \sim \mathrm{Lap}(\lambda_{j_t})$. ▷ Update the chosen coordinate.
return w^T.

In Section 3.2 we describe how to set λ_j, λ'_j (for $j \in [p]$) to ensure (ϵ, δ)-differential privacy. We then give high-probability utility results in Section 3.3. We further show in Section 3.4 that DP-GCD can progress very fast in its first iterations when the solutions are quasi-sparse. Finally, we discuss DP-GCD's computational complexity in Section 3.5.

3.2 Privacy Guarantees

The privacy guarantees of DP-GCD depends on the noise scales λ_j and λ'_j. In Theorem 3.1, we describe how to set these values so as to ensure that DP-GCD is (ϵ, δ)-differentially private.

Theorem 3.1. *Let $\epsilon, \delta \in (0, 1]$. Algorithm 1 with $\lambda_j = \lambda'_j = \frac{8L_j}{n\epsilon}\sqrt{T\log(1/\delta)}$ is (ϵ, δ)-DP.*

Sketch of Proof. (Detailed proof in Appendix A) Let $\epsilon' = \epsilon/\sqrt{16T\log(1/\delta)}$. At an iteration t, data is accessed twice. First, to compute the index j_t of the coordinate to update. It is obtained as the index of the largest noisy entry of f's gradient, with noise $\mathrm{Lap}(\lambda'_j)$. By the report-noisy-argmax mechanism, j_t is ϵ'-DP. Second, to compute the gradient's j_t's entry, which is released with noise $\mathrm{Lap}(\lambda_j)$.The Laplace mechanism ensures that this computation is also ϵ'-DP. Algorithm 1 is thus the $2T$-fold composition of ϵ'-DP mechanisms, and the result follows from DP's advanced composition theorem (Dwork and Roth, 2013). □

Remark 3.2. The assumption $\epsilon \in (0, 1]$ is only used to give a closed-form expression for the noise scales λ, λ's. In practice, we tune them numerically so that we obtain the desired value of ϵ by the advanced composition theorem (see eq. (2) in Appendix A).

Computing the greedy update requires injecting Laplace noise that scales with the coordinate-wise Lipschitz constants $L_1, \ldots, L_p$ of the loss. These constants are typically smaller than their global counterpart. This allows DP-GCD to inject less noise on smaller-scaled coordinates.

3.3 Utility Guarantees

We now state utility results for DP-GCD on smooth DP-ERM. For some problems, DP-GCD can provide better utility than the worst-case lower bounds. For clarity, we state asymptotic results in Theorem 3.3, (where $\widetilde{O}$ ignores non-significant logarithmic terms). Complete non-asymptotic results are given in Appendix B. In the rest of this paper, we denote $f(w) = \frac{1}{n}\sum_{i=1}^n \ell(w; d_i)$.

Theorem 3.3. *Let $\epsilon, \delta \in (0, 1]$. Assume $\ell(\cdot; d)$ is a convex and L-component-Lipschitz loss function for all $d \in \mathcal{X}$, and f is M-component-smooth. Define $\mathcal{W}^*$ the set of minimizers of f, and f^* the minimum of f. Let $w_{priv} \in \mathbb{R}^p$ be the output of Algorithm 1 with step sizes $\gamma_j = 1/M_j$, and noise scales $\lambda_1, \ldots, \lambda_p$, and $\lambda'_1, \ldots, \lambda'_p$ set as in Theorem 3.1 (with T chosen below) to ensure (ϵ, δ)-DP. Then, the following holds for $\zeta \in (0, 1]$:*

1. *When f is convex, assume $f(w^0) - f^* \geq 16 L_{\max}\sqrt{T\log(1/\delta)\log(2Tp/\zeta)}/M_{\min} n\epsilon$, and set $T = O(n^{2/3}\epsilon^{2/3} R_{M,1}^{2/3} M_{\min}^{1/3}/L_{\max}^{2/3}\log(1/\delta)^{1/3})$. Then we have, with probability at least $1-\zeta$,*

$$f(w_{priv}) - f^* = \widetilde{O}\left(\frac{R_{M,1}^{4/3} L_{\max}^{2/3}\log(1/\delta)\log(p/\zeta)}{n^{2/3}\epsilon^{2/3} M_{\min}^{1/3}}\right) ,$$

where $R_{M,1} = \max_{w\in\mathbb{R}^p} \min_{w^*\in\mathcal{W}^*} \left\{\|w - w^*\|_{M,1} \mid f(w) \leq f(w^0)\right\}$.

2. *When f is $\mu_{M,1}$-strongly convex w.r.t. $\|\cdot\|_{M,1}$, set $T = O\left(\frac{1}{\mu_{M,1}}\log(\frac{M_{\min}\mu_{M,1} n\epsilon(f(w^0)-f(w^*))}{L_{\max}\log(1/\delta)\log(2p/\zeta)})\right)$. Then we have, with probability at least $1-\zeta$,*

$$f(w_{priv}) - f^* = \widetilde{O}\left(\frac{L_{\max}^2\log(1/\delta)\log(2p/\mu_M\zeta)}{M_{\min}\mu_{M,1}^2 n^2\epsilon^2}\right) .$$

Sketch of Proof. (Detailed proof in Appendix B). To prove our theorem, we start by proving that the objective function decreases (up to noise) at each iteration. The smoothness of f gives, for a coordinate update with step size $1/M_j$,

$$f(w^{t+1}) \leq f(w^t) - \frac{1}{2M_j}\nabla_j f(w^t)^2 + \frac{1}{2M_j}(\eta_j^t)^2 .$$

We then use the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$ (see Lemma B.1) and the greedy rule to obtain $-\frac{1}{2M_j}\nabla_j f(w^t)^2 \leq \frac{1}{4M_j}(\nabla_j f(w^t) + \chi_j^t)^2 + \frac{1}{2M_j}(\chi_j^t)^2 \leq \frac{1}{4}\|\nabla f(w^t) + \chi^t\|_{M^{-1},\infty}^2 + \frac{1}{2M_j}(\chi_j^t)^2 = \frac{1}{4M_{j^*}}(\nabla_{j^*} f(w^t) + \chi_{j^*}^t)^2 + \frac{1}{2M_j}(\chi_j^t)^2$ for some $j^* \in [p]$. We then use Lemma B.1 again to upper bound $-\frac{1}{4M_{j^*}}(\nabla_{j^*} f(w^t) + \chi_{j^*}^t)^2 \leq -\frac{1}{8}\|\nabla f(w^t)\|_{M^{-1},1}^2 + \frac{1}{4M_{j^*}}(\chi_{j^*}^t)^2$.

To prove the result for convex functions, we further upper bound $-\frac{1}{8}\|\nabla f(w^t)\|_{M^{-1},1}^2$ using f's convexity: $f(w^t) - f(w^*) \leq \langle \nabla f(w^t), w^t - w^*\rangle \leq \|\nabla f(w^t)\|_{M^{-1},\infty}\|w^t - w^*\|_{M,1}$, where we also used Hölder's inequality. Squaring this inequality and replacing in our descent inequality gives

$$f(w^{t+1}) - f(w^*) \leq f(w^t) - f(w^*) - \frac{(f(w^t) - f(w^*))^2}{8\|w^t - w^*\|_{M,1}^2} + \frac{|\eta_j^t|^2}{2M_j} + \frac{|\chi_j^t|^2}{2M_j} + \frac{|\chi_{j^*}^t|^2}{4M_{j^*}} .$$

When f is $\mu_{M,1}$-strongly-convex w.r.t. $\|\cdot\|_{M,1}$, we use the strong convexity inequality and minimize it on both sides of to get $f(w^*) \geq f(w^t) - \sup_{w\in\mathbb{R}^p}\left\{-\langle\nabla f(w^t), w^t - w\rangle - \frac{\mu_{M,1}}{2}\|w^t - w\|_{M,1}^2\right\} = f(w^t) - \frac{1}{2\mu_{M,1}}\|\nabla f(w^t)\|_{M^{-1},\infty}^2$. Combining this inequality with the descent inequality above yields

$$f(w^{t+1}) - f(w^*) \leq \left(1 - \frac{\mu_{M,1}}{4}\right)(f(w^t) - f(w^*)) + \frac{|\eta_j^t|^2}{2M_j} + \frac{|\chi_j^t|^2}{2M_j} + \frac{|\chi_{j^*}^t|^2}{4M_{j^*}} .$$

We then convert these inequalities to high-probability bounds using concentration inequalities. To conclude the proof, we need to solve the two aforementioned recursions. This requires particular attention (especially for general convex functions), as the noise can prevent the function f from decreasing between two iterations. We describe the proof in details in Appendix B.1. □

Remark 3.4. The lower bound on $f(w^0) - f^*$ in Theorem 3.3 is a standard assumption in the analysis of inexact coordinate descent methods: it ensures that sufficient decrease is possible despite the noise. A similar assumption is made by Tappenden et al. (2016), see Theorem 5.1 therein.

In Theorem 3.3, the explicit dependence on the dimension p in the utility of DP-GCD is *logarithmic*. This proves that DP-GCD achieves nearly dimension-independent utility, as soon as $R_{M,1}$ or $\mu_{M,1}$ are independent of the dimension. This can notably be the case when the M_j's are imbalanced, and most of the errors comes from one coordinate.

In the convex setting, and when $R_M = O(1)$, DP-GCD matches the utility of DP-FW (Talwar et al., 2015) with ℓ_1-constraints. Importantly, DP-GCD can achieve this rate for unconstrained, smooth ERM problems: we are not aware of previously proposed algorithms achieving such utility. We also propose and empirically evaluate a natural proximal extension of DP-GCD to handle nonsmooth regularization (such as ℓ_1) in Section 5.

In the strongly-convex setting, DP-GCD is the first algorithm (to our knowledge) that can exploit strong-convexity w.r.t. the ℓ_1-norm. Its utility can be as low as $O(\log(p)/n^2\epsilon^2)$ when $\mu_{M,1} = O(1)$. The closest result (although in a different setting) is the one of Asi et al. (2021), who achieve $O(\log(p)^2/n^{4/3}\epsilon^{4/3})$ utility using a Frank-Wolfe algorithm. We also note that their algorithm is built on a reduction to the convex setting which, unlike DP-GCD, can be impractical.

3.4 Fast Initial Convergence on Quasi-Sparse Problems

In this section, we show that when problem (1) is strongly-convex, DP-GCD can progress very fast during its first iterations. For non-private greedy coordinate descent, Fang et al. (2020) established that, for strongly-convex objectives, sparse solutions induce fast initial convergence. We generalize their results to the private (noisy) setting and to solutions that are not necessarily sparse, but have few large entries. We call these vectors *quasi-sparse* (Definition 3.5).

Definition 3.5 ((α,τ)-quasi-sparsity)**.** A vector $w \in \mathbb{R}^p$ is (α,τ)-quasi-sparse if it has at most τ entries superior to α (in modulus). When $\alpha = 0$, we call the vector τ-sparse.

Note that any vector in $\mathbb{R}^p$ is (α, p)-quasi-sparse (for some $\alpha > 0$). In fact, α and τ are linked: $\tau(\alpha)$ can be seen as a function of α. Of course, quasi-sparsity will only yield meaningful improvements when α and τ are small simultaneously. We now state the main result of this section. Theorem 3.6 shows that DP-GCD progresses fast on strongly-convex problems with a quasi-sparse solution.

Theorem 3.6 (Proof in Appendix C)**.** *Let f satisfy the hypotheses of Theorem 3.3, where Algorithm 1 is initialized with $w^0 = 0$ and outputs w^T. Assume that, for $\tau' \in [p]$, f is $\mu_{M,1}^{(\tau')}$-strongly-convex w.r.t. $\|\cdot\|_{M,1}$ when restricted to τ'-sparse vectors and $\mu_{M,2}$-strongly-convex w.r.t. $\|\cdot\|_{M,2}$. Assume that the (unique) solution of problem (1) is (α,τ)-quasi-sparse for some $\alpha, \tau \geq 0$. Let $0 \leq T \leq p - \tau$ and $\zeta \in [0,1]$. Then with probability at least $1-\zeta$:*

$$
\begin{aligned}
f(w^T) - f^* &\leq \prod_{t=1}^{T} \Big(1 - \frac{\mu_{M,1}^{(t+\tau)}}{4}\Big)(f(w^0) - f^*) + \widetilde{O}\Big((T+\tau)(p-\tau)\alpha^2 + \frac{L_{\max}^2 T(T+\tau)}{M_{\min}\mu_{M,2} n^2\epsilon^2}\Big) \\
&\leq \prod_{t=1}^{T} \Big(1 - \frac{\mu_{M,2}}{4(t+\tau)}\Big)(f(w^0) - f^*) + \widetilde{O}\Big((T+\tau)(p-\tau)\alpha^2 + \frac{L_{\max}^2 T(T+\tau)}{M_{\min}\mu_{M,2} n^2\epsilon^2}\Big) \ .
\end{aligned}
$$

For problems with τ-sparse solutions (*i.e.,* $\alpha = 0$), Theorem 3.6 recovers the non-private results of Fang et al. (2020), up to the additive term due to privacy. However, unlike Theorem 3.6, our result shows that improvements are still possible in some regimes where the solution is not sparse (*i.e.,* , $\alpha > 0$). If τ is small, the first term $1 - \mu_{M,1}^{(t+\tau)}/4$ (for small t) can be very low. If α is also small (*i.e.,* , w^* is not perfectly sparse), important progress happens in the first iterations. Typically, when $\alpha = O(\sqrt{L_{\max}^2 T/M_{\min}\mu_{M,2} p n^2\epsilon^2})$, the residual term from the quasi-sparsity is smaller than the term due to the noise. For these values of α, the result is thus the same for τ-sparse and (α,τ)-quasi-sparse vectors.

Performing many updates is very costly in private optimization, since the scale of the noise grows with the number of released gradients. Thus, in high dimensional problems where other private algorithms are unable to make any progress, DP-GCD's fast initial convergence may yield better solutions. We will illustrate this in our experiments.

3.5 Computational Cost

Each iteration of DP-GCD requires computing a full gradient, but only uses one of its coordinates. In non-private optimization, one would generally be better off performing the full update to avoid wasting computation. This is not the case when gradients are private. Using the full gradient would require privatizing p coordinates

of the gradient, even though only few of them may be significant or useful at all. Conversely, the report noisy max mechanism (Dwork and Roth, 2013) allows selecting these entries *without paying the full price of dimension*. Hence, the greedy updates of DP-GCD allow to limit the noise at the cost of more computation.

In practice, the higher computational cost of each iteration may not translate in a significantly larger cost overall: as shown by our theoretical results, DP-GCD is able to exploit the *quasi-sparsity* of the solution to progress fast and only a handful of iterations may be needed to reach a good private solution. In contrast, most updates of classic private optimization algorithms (like DP-SGD) may not be worth doing, and simply lead to unnecessary injection of noise. We illustrate this phenomenon numerically in Section 5.

4 Related Work

DP-ERM. Differentially private empirical risk minimization (DP-ERM) was first formulated by Chaudhuri et al. (2011), who proposed solvers based on output and objective perturbation. Song et al. (2013) proposed a differentially private stochastic gradient descent (DP-SGD) algorithm and studied it empirically. Bassily et al. (2014) analyzed the theoretical properties of DP-SGD on non-smooth objectives, and derived matching utility lower bounds for DP-ERM. Faster algorithms (based on proximal SVRG, see Johnson and Zhang, 2013; Xiao and Zhang, 2014) were designed by Wang et al. (2017) for composite problems. Abadi et al. (2016) introduced gradient clipping for private optimization, which is ubiquitous in practical implementations of DP-SGD. Wu et al. (2017) studied a variant of DP-SGD with output perturbation, that is efficient when only few passes on the data are possible. Iyengar et al. (2019) conducted a large experimental study of private algorithms.

DP-SCO. Closely related to DP-ERM is the differentially private stochastic convex optimization (DP-SCO) problem. Bassily et al. (2019) used algorithmic stability arguments (following work from Hardt et al., 2016; Bassily et al., 2020) to show that the population risk of DP-SCO is the same as in non-private SCO. Feldman et al. (2020) and Wang et al. (2022) then developed efficient (linear-time) algorithm to solve this problem. The work of Dwork et al. (2015), Bassily et al. (2016), and Jung et al. (2021) provides a way to convert results from DP-ERM to DP-SCO.

High-dimensional DP-ML. Restricted to ℓ_1-constrained ERM problems, Talwar et al. (2015) used a differentially private Frank-Wolfe algorithm (DP-FW) (Frank and Wolfe, 1956; Jaggi, 2013) to achieve utility that scales logarithmically with the dimension. Asi et al. (2021) and Bassily et al. (2021) proposed stochastic DP-FW algorithms, extending the above results to DP-SCO. Still for constrained optimization, Kasiviswanathan and Jin (2016) randomly project the data on a smaller-dimensional space, and lift the result back onto the higher-dimension space. The dependence in the dimension is encoded in the Gaussian width of the parameter space, leading to $O(\log p)$ error for the ℓ_1 ball or the simplex. Kifer et al. (2012) propose an approach specific to sparse linear regression which first identifies some support, and then solves DP-ERM on this restricted support. Their approach achieves an error of $O(\log p)$ but relies either on prior knowledge on the solution's sparsity, or on the tuning of an additional hyperparameter. Kairouz et al. (2021) and Zhou et al. (2021) estimate low-dimensional subspaces, in which the gradients live, using public data. This reduces noise addition, but in practice, public data is only rarely available. Finally, Wang and Xu (2021) show that for sparse linear regression in the local model of DP, the estimation error must be polynomial in p.

Coordinate descent. Randomized and Cyclic Coordinate descent (CD) algorithms have a long history in optimization. Luo and Tseng (1992), Tseng (2001), and Tseng and Yun (2009) have shown convergence results for (block) CD algorithms for nonsmooth optimization. Nesterov (2010) later proved a global non-asymptotic $1/T$ convergence rate for CD with random choice of coordinates for a convex, smooth objective. Parallel, proximal variants were developed by Richtárik and Takáč (2014) and Fercoq and Richtárik (2014), while Hanzely et al. (2018) considered non-separable non-smooth parts, and Tappenden et al. (2016) studied an inexact version of CD. Shalev-Shwartz and Zhang (2013) introduced Dual CD algorithms for smooth ERM, showing performance similar to SVRG.

Greedy coordinate descent methods, which update the coordinate with greatest gradient entry, have been studied by Luo and Tseng (1992) and Tseng and Yun (2009). Dhillon et al. (2011) proved convergence results for convex objective functions. Nutini et al. (2015) showed improved convergence rate for smooth, strongly-convex functions, by measuring strong convexity in the ℓ_1-norm. Karimireddy et al. (2019) extended

Table 1: Number of records and features in each dataset.

	log1, log2	sparse	mtp	dexter	california	madelon
Records	$1,000$	$1,000$	$4,450$	600	$20,640$	$2,600$
Features	100	$1,000$	202	$11,035$	8	501

these results for ℓ_1- and box-regularized problems, by using a modified greedy CD algorithm. Dhillon et al. (2011), Nutini et al. (2015), and Karimireddy et al. (2019) proposed nearest-neighbors-based schemes to efficiently compute the (approximate) greedy update. Stich et al. (2017) keep an estimate of the full gradient to compute the greedy update efficiently. We refer to Wright (2015) and Shi et al. (2017) for detailed reviews on CD.

Private coordinate descent. Differentially Private Coordinate Descent (DP-CD) was studied by Mangold et al. (2022), who analyzed its utility and derived corresponding lower bounds. Damaskinos et al. (2021) proposed a dual coordinate descent algorithm for generalized linear models. Private CD has also been used by Bellet et al. (2018) in a decentralized setting. All these works use random selection rule, which fail to exploit key problem's properties such as sparsity. Our private greedy coordinate selection rule reduces the dependence on dimension in such settings.

5 Experiments

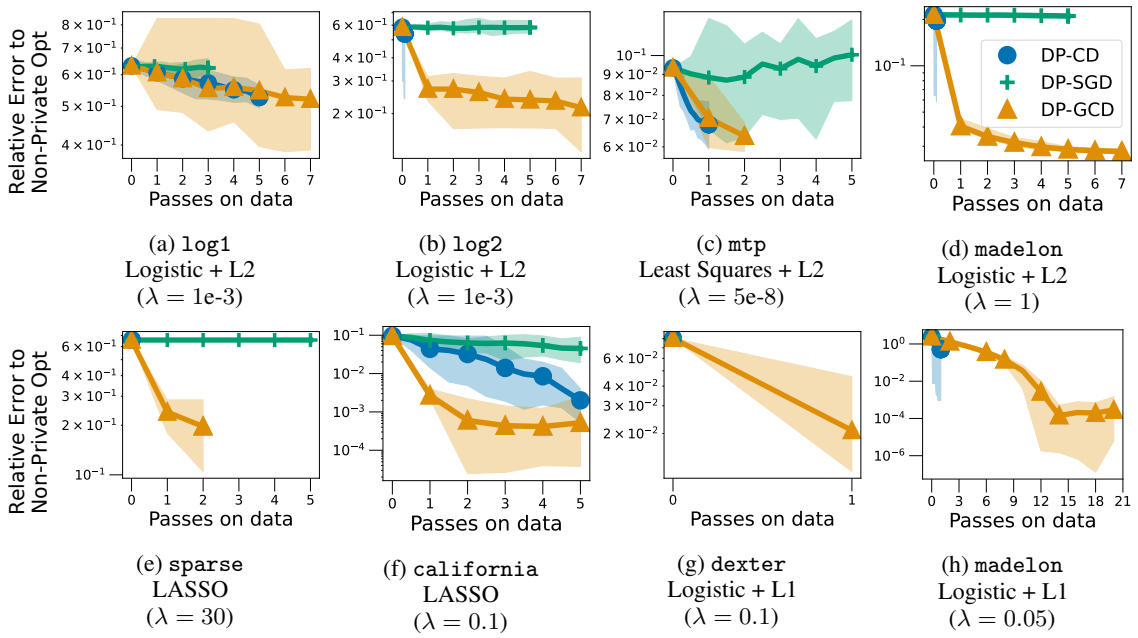
In this section, we evaluate the practical performance of DP-GCD on linear models with the logistic and squared loss. We compare DP-GCD with stochastic gradient descent with default batch size 1 (DP-SGD), and with stochastic coordinate descent with uniform choice of coordinates (DP-CD). The datasets we use are of various sizes. The first two datasets, coined `log1` and `log2`, are synthetic. We generated a design matrix $X \in \mathbb{R}^{1,000\times 100}$ with unit-variance, normally-distributed, columns. Labels are computed as $y = Xw^{(true)} + \varepsilon$, where ε is normally-distributed noise and $w^{(true)}$ is drawn from a log-normal distribution of parameters $\mu = 0$ and σ equal to 1 and 2 respectively (the larger σ, the more quasi-sparse the solution). The `sparse` dataset is generated with the same process, but with $X \in \mathbb{R}^{1,000\times 1,000}$ and $w^{(true)}$ having only 10 non-zero values. The `california` dataset can be downloaded from `scikit-learn` (Pedregosa et al., 2011) while `mtp`, `madelon` and `dexter` are available in `OpenML` repositories (Vanschoren et al., 2014). Information on datasets is summarized in Table 1, and we discuss levels of (quasi)-sparsity of each problem's solution in Appendix E.

Algorithmic setup. *(Privacy.)* For each algorithm, the tightest noise scales are computed numerically to guarantee $(1, 1/n^2)$-DP, where n is the number of records in the dataset. This is generally regarded as a good privacy guarantee. For DP-CD and DP-SGD, we privatize the gradients with the Gaussian mechanism (Dwork and Roth, 2013), and account for privacy tightly using Rényi differential privacy (RDP) (Mironov, 2017). For DP-SGD, we use RDP amplification for the subsampled Gaussian mechanism (Mironov et al., 2019).

(Hyperparameters.) For DP-SGD, we use constant step sizes and standard gradient clipping (Abadi et al., 2016). For DP-GCD and DP-CD, we set the step sizes to $\eta_j = \gamma / M_j$, and adapt the coordinate-wise clipping thresholds from one hyperparameter, as proposed by Mangold et al. (2022). Thus, for each algorithm, we tune two hyperparameters: one for step sizes and one for clipping thresholds. The complete hyperparameter grids are detailed in Appendix E.

(Plots.) In all experiments, we plot the relative error to the *non-private* optimal objective value for the best set of hyperparameters, averaged over 10 runs. We plot this value as a function of the number of passes on the data. Note that each pass on the data corresponds to p iterations of DP-CD, n iterations of DP-SGD and 1 iteration of DP-GCD. This guarantees that the same amount of computation is done by each algorithm for each tick in the x-axis.

Quasi-sparse solutions. On the `log1` dataset (Figure 1a), $w^{(true)}$ is not imbalanced enough: DP-GCD cannot exploit structure too much and only slightly improves over existing algorithms. The same phenomenon happens on the `mtp` dataset (Figure 1c). Conversely, on `log2` (where most entries of the solution are very



(a) `log1` Logistic + L2 (λ = 1e-3)
(b) `log2` Logistic + L2 (λ = 1e-3)
(c) `mtp` Least Squares + L2 (λ = 5e-8)
(d) `madelon` Logistic + L2 ($\lambda = 1$)
(e) `sparse` LASSO ($\lambda = 30$)
(f) `california` LASSO ($\lambda = 0.1$)
(g) `dexter` Logistic + L1 ($\lambda = 0.1$)
(h) `madelon` Logistic + L1 ($\lambda = 0.05$)

Figure 1: Relative error to non-private optimal for DP-CD, DP-GCD and DP-SGD on different problems. On the x-axis, 1 tick represents a full access to the data: p iterations of DP-CD, n iterations of DP-SGD and 1 iteration of DP-GCD. Number of iterations, clipping thresholds and step sizes are tuned simultaneously for each algorithm. We report min/mean/max values over 10 runs.

small), DP-GCD vastly improves over all algorithms (Figure 1b). This does not mean that DP-GCD finds the true solution to the non-private problem, but it finds as good a private solution as possible by looking for sparse approximations. In other words, the private greedy rule (even if noisy) is able select the most efficient updates and thus spends its privacy budget efficiently. This is also what we observe on `madelon` (Figure 1d), whose solution has many small values.

Extension to ℓ_1 regularization. Our results from Section 3.4 show that DP-GCD's first iterations are especially good when the problem's solution is (quasi) sparse. This leads us to consider problems with an additional sparsity-inducing regularization term, such as the ℓ_1 norm of w (Tibshirani, 1996). Unfortunately, this makes the objective function non-smooth, which is not covered by Algorithm 1, nor by our utility analysis. We however propose a proximal version of DP-GCD (for which the same privacy guarantees hold), building upon the multiple greedy rules have been proposed for the nonsmooth setting (see *e.g.,* Tseng and Yun, 2009; Nutini et al., 2015). We describe this extension in Appendix D and report the results with ℓ_1 regularization on `sparse`, `california`, `dexter` and `madelon` in Figures 1e to 1h. On these problems, DP-GCD consistently outperforms other methods. In fact, it is the only algorithm able to make any progress on the `sparse`, `dexter` and `madelon` datasets (Figures 1e, 1g and 1h). On the small-dimensional dataset `california`, we again observe the fast initial convergence of DP-GCD, although it finishes on par with DP-CD (Figure 1f).

Computational complexity. As raised in Section 3.5, one iteration of DP-GCD requires a full pass on the data. This is as costly as p iterations of DP-CD or n iterations of DP-SGD. Nonetheless, on many problems, DP-GCD requires just as many passes on the data as DP-CD and DP-SGD (Figures 1a and 1c to 1f). When more computation is required, it also provides significantly better solutions than DP-CD and DP-SGD (Figures 1b, 1g and 1h). This is in line with our theoretical results from Section 3.4.

6 Conclusion and Discussion

We proposed DP-GCD, a greedy coordinate descent algorithm for DP-ERM. We analyzed DP-GCD and showed that it is particularly fit for minimizing high-dimensional objective functions whose solutions have a few parameters that dominate the others. On such problems, DP-GCD can bypass the ambient dimension to achieve better privacy-utility trade-offs than the worst case.

We showed that DP-GCD's first iterations are particularly effective. This is confirmed experimentally, and allows DP-GCD to vastly outperform existing private optimization methods on high-dimensional problems. While DP-GCD's updates are costly to compute, this is generally compensated by the small number of iterations needed to reach a satisfying private solution. This suggests that, combined with sublinear-time approximations of the greedy updates (Dhillon et al., 2011; Stich et al., 2017; Karimireddy et al., 2019), DP-GCD could yield a very efficient algorithm for DP-ERM.

Finally, we explored numerically the behavior of DP-GCD on ℓ_1-regularized problems. These problems enforce the sparsity of the solution, but introduce a non-smooth term to the objective. Our analysis does not cover this case, which is known to be hard to analyze even for non-private GCD. The best existing analysis is the one of Karimireddy et al. (2019), which is restricted to ℓ_1-regularized and box-constrained problems. In the private setting, their splitting of iterations into `good` and `bad` steps ones does not work anymore, since it is not guaranteed that `bad` steps are followed by `good` ones, nor that `bad` steps do not increase the objective too much. Extending such results to the differentially private setting is an exciting direction for future work.

Acknowledgments

This work was supported in part by the Inria Exploratory Action FLAMED and by the French National Research Agency (ANR) through grant ANR-20-CE23-0015 (Project PRIDE) and ANR-20-CHIA-0001-01 (Chaire IA CaMeLOt).

References

Abadi, M., A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang (Oct. 2016). "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. New York, NY, USA: Association for Computing Machinery, pp. 308–318.

Asi, H., V. Feldman, T. Koren, and K. Talwar (Mar. 2021). "Private Stochastic Convex Optimization: Optimal Rates in ℓ_1 Geometry". In: *arXiv:2103.01516 [cs, math, stat]*.

Bassily, R., V. Feldman, C. Guzmán, and K. Talwar (2020). "Stability of Stochastic Gradient Descent on Nonsmooth Convex Losses". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 4381–4391.

Bassily, R., V. Feldman, K. Talwar, and A. Guha Thakurta (2019). "Private Stochastic Convex Optimization with Optimal Rates". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc.

Bassily, R., C. Guzman, and A. Nandi (July 2021). "Non-Euclidean Differentially Private Stochastic Convex Optimization". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. PMLR, pp. 474–499.

Bassily, R., K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman (June 2016). "Algorithmic Stability for Adaptive Data Analysis". In: *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*. STOC '16. New York, NY, USA: Association for Computing Machinery, pp. 1046–1059.

Bassily, R., A. Smith, and A. Thakurta (Oct. 2014). "Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds". In: *arXiv:1405.7085 [cs, stat]*.

Bellet, A., R. Guerraoui, M. Taziki, and M. Tommasi (Mar. 2018). "Personalized and Private Peer-to-Peer Machine Learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 473–481.

Boyd, S. P. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge, UK ; New York: Cambridge University Press.

Chaudhuri, K., C. Monteleoni, and A. D. Sarwate (2011). "Differentially Private Empirical Risk Minimization". In: *Journal of Machine Learning Research* 12.29, pp. 1069–1109.

Damaskinos, G., C. Mendler-Dünner, R. Guerraoui, N. Papandreou, and T. Parnell (May 2021). "Differentially Private Stochastic Coordinate Descent". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35, pp. 7176–7184.

Dhillon, I., P. Ravikumar, and A. Tewari (2011). "Nearest Neighbor Based Greedy Coordinate Descent". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc.

Dwork, C. (2006). "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 1–12.

Dwork, C., V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth (June 2015). "Preserving Statistical Validity in Adaptive Data Analysis". In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. STOC '15. New York, NY, USA: Association for Computing Machinery, pp. 117–126.

Dwork, C. and A. Roth (2013). "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends® in Theoretical Computer Science* 9.3-4, pp. 211–407.

Fang, H., Z. Fan, Y. Sun, and M. Friedlander (June 2020). "Greed Meets Sparsity: Understanding and Improving Greedy Coordinate Descent for Sparse Optimization". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 434–444.

Feldman, V., T. Koren, and K. Talwar (June 2020). "Private Stochastic Convex Optimization: Optimal Rates in Linear Time". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. New York, NY, USA: Association for Computing Machinery, pp. 439–449.

Fercoq, O. and P. Richtárik (Mar. 2014). "Accelerated, Parallel and Proximal Coordinate Descent". In: *arXiv:1312.5799 [cs, math, stat]*.

Frank, M. and P. Wolfe (Mar. 1956). "An Algorithm for Quadratic Programming". In: *Naval Research Logistics Quarterly* 3.1-2, pp. 95–110.

Hanzely, F., K. Mishchenko, and P. Richtarik (Dec. 2018). "SEGA: Variance Reduction via Gradient Sketching". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., pp. 2086–2097.

Hardt, M., B. Recht, and Y. Singer (June 2016). "Train Faster, Generalize Better: Stability of Stochastic Gradient Descent". In: *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, pp. 1225–1234.

Iyengar, R., J. P. Near, D. Song, O. Thakkar, A. Thakurta, and L. Wang (May 2019). "Towards Practical Differentially Private Convex Optimization". In: *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 299–316.

Jaggi, M. (Feb. 2013). "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization". In: *International Conference on Machine Learning*. PMLR, pp. 427–435.

Johnson, R. and T. Zhang (2013). "Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc.

Jung, C., K. Ligett, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and M. Shenfeld (June 2021). "A New Analysis of Differential Privacy’s Generalization Guarantees (Invited Paper)". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. New York, NY, USA: Association for Computing Machinery, p. 9.

Kairouz, P., M. R. Diaz, K. Rush, and A. Thakurta (July 2021). "(Nearly) Dimension Independent Private ERM with AdaGrad Rates via Publicly Estimated Subspaces". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. PMLR, pp. 2717–2746.

Karimireddy, S. P., A. Koloskova, S. U. Stich, and M. Jaggi (Apr. 2019). "Efficient Greedy Coordinate Descent for Composite Problems". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2887–2896.

Kasiviswanathan, S. P. and H. Jin (2016). "Efficient Private Empirical Risk Minimization for High-dimensional Learning". In: p. 10.

Kifer, D., A. Smith, and A. Thakurta (2012). "Private Convex Empirical Risk Minimization and High-dimensional Regression". In: p. 40.

Luo, Z.-Q. and P. Tseng (Jan. 1992). "On the Convergence of the Coordinate Descent Method for Convex Differentiable Minimization". In: *Journal of Optimization Theory and Applications* 72.1, pp. 7–35.

Mangold, P., A. Bellet, J. Salmon, and M. Tommasi (Oct. 2022). "Differentially Private Coordinate Descent for Composite Empirical Risk Minimization". In: *International Conference on Machine Learning*. PMLR.

Mironov, I. (Aug. 2017). "Renyi Differential Privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275.

Mironov, I., K. Talwar, and L. Zhang (Aug. 2019). "Rényi Differential Privacy of the Sampled Gaussian Mechanism". In: *arXiv:1908.10530 [cs, stat]*.

Nesterov, Y. (Jan. 2010). "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems". In: *SIAM Journal on Optimization* 22.2, pp. 341–362.

Nutini, J., M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke (June 2015). "Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection". In: *International Conference on Machine Learning*. PMLR, pp. 1632–1641.

Parikh, N. and S. Boyd (Jan. 2014). "Proximal Algorithms". In: *Foundations and Trends in Optimization* 1.3, pp. 127–239.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau (2011). "Scikit-Learn: Machine Learning in Python". In: *MACHINE LEARNING IN PYTHON*, p. 6.

Richtárik, P. and M. Takáč (Apr. 2014). "Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function". In: *Mathematical Programming* 144.1-2, pp. 1–38.

Shalev-Shwartz, S. and T. Zhang (Feb. 2013). "Stochastic Dual Coordinate Ascent Methods for Regularized Loss". In: *The Journal of Machine Learning Research* 14.1, pp. 567–599.

Shi, H.-J. M., S. Tu, Y. Xu, and W. Yin (Jan. 2017). "A Primer on Coordinate Descent Algorithms". In: *arXiv:1610.00040 [math, stat]*.

Shokri, R., M. Stronati, C. Song, and V. Shmatikov (May 2017). "Membership Inference Attacks Against Machine Learning Models". In: *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18.

Song, S., K. Chaudhuri, and A. D. Sarwate (Dec. 2013). "Stochastic Gradient Descent with Differentially Private Updates". In: *2013 IEEE Global Conference on Signal and Information Processing*. Austin, TX, USA: IEEE, pp. 245–248.

Stich, S. U., A. Raj, and M. Jaggi (July 2017). "Approximate Steepest Coordinate Descent". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, pp. 3251–3259.

Talwar, K., A. Guha Thakurta, and L. Zhang (2015). "Nearly Optimal Private LASSO". In: *Advances in Neural Information Processing Systems* 28.

Tappenden, R., P. Richtárik, and J. Gondzio (July 2016). "Inexact Coordinate Descent: Complexity and Preconditioning". In: *Journal of Optimization Theory and Applications* 170.1, pp. 144–176.

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tseng, P. (June 2001). "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization". In: *Journal of Optimization Theory and Applications* 109.3, pp. 475–494.

Tseng, P. and S. Yun (Mar. 2009). "A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization". In: *Mathematical Programming* 117.1, pp. 387–423.

Vanschoren, J., J. N. rijnvan Rijn, B. Bischl, and L. Torgo (June 2014). "OpenML: Networked Science in Machine Learning". In: *ACM SIGKDD Explorations Newsletter* 15.2, pp. 49–60.

Wang, D. and J. Xu (Feb. 2021). "On Sparse Linear Regression in the Local Differential Privacy Model". In: *IEEE Transactions on Information Theory* 67.2, pp. 1182–1200.

Wang, D., M. Ye, and J. Xu (2017). "Differentially Private Empirical Risk Minimization Revisited: Faster and More General". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Wang, P., Y. Lei, Y. Ying, and H. Zhang (Jan. 2022). "Differentially Private SGD with Non-Smooth Losses". In: *Applied and Computational Harmonic Analysis* 56, pp. 306–336.

Wright, S. J. (June 2015). "Coordinate Descent Algorithms". In: *Mathematical Programming* 151.1, pp. 3–34.

Wu, X., F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton (May 2017). "Bolt-on Differential Privacy for Scalable Stochastic Gradient Descent-based Analytics". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: Association for Computing Machinery, pp. 1307–1322.

Xiao, L. and T. Zhang (Jan. 2014). "A Proximal Stochastic Gradient Method with Progressive Variance Reduction". In: *SIAM Journal on Optimization* 24.4, pp. 2057–2075.

Zhou, Y., Z. S. Wu, and A. Banerjee (2021). "Bypassing the Ambiant Dimension: Private SGD with Gradient Subspace Identification". In: p. 28.

A Proof of Privacy

Theorem 3.1. *Let $\epsilon, \delta \in (0, 1]$. Algorithm 1 with $\lambda_j = \lambda'_j = \frac{8L_j}{n\epsilon}\sqrt{T\log(1/\delta)}$ is (ϵ, δ)-DP.*

Proof. In each iteration of Algorithm 1, the data is accessed twice: once to choose the coordinate and once to compute the private gradient. In total, data is thus queried $2T$ times.

Let $\lambda_j = \lambda'_j = \frac{2L_j}{n\epsilon'}$. For $j \in [p]$, the gradient's j-th entry has sensitivity $2L_j$. Thus, by the report noisy max mechanism (Dwork and Roth, 2013), the greedy choice of j is ϵ'-DP. By the Laplace mechanism (Dwork and Roth, 2013), computing the corresponding gradient coordinate is also ϵ'-DP.

The advanced composition theorem for differential privacy thus ensures that the $2T$-fold composition of these mechanisms is (ϵ, δ)-DP for $\delta > 0$ and

$$\epsilon = \sqrt{4T\log(1/\delta)}\epsilon' + 2T\epsilon'(\exp(\epsilon') - 1) \ , \tag{2}$$

where we recall that $\epsilon' = \frac{2L_j}{n\lambda_j} = \frac{2L_j}{n\lambda'_j}$ for all $j \in [p]$. When $\epsilon \leq 1$, we can give a simpler expression (see Corollary 3.21 of Dwork and Roth, 2013): with $\epsilon' = \epsilon/4\sqrt{T\log(1/\delta)}$, Algorithm 1 is (ϵ, δ)-DP for $\lambda_j = \lambda'_j = 8L_j\sqrt{T\log(1/\delta)}/n\epsilon$. □

B Proof of Utility

To prove Theorem 3.3, we first prove a technical lemma (Section B.1) and a descent lemma (Section B.2). We then use these results to prove the two statements of Theorem 3.3: the general convex case (Section B.3) and the strongly convex case (Section B.4).

B.1 Technical Lemmas

To prove our results, we will use three technical lemmas. In Appendix B.1.1, we state a classical inequality. In Appendix B.1.2, we prove a lemma that will be used in the utility proof for convex functions, once we will have a recursion between $f(w^{t+1}) - f(w^*)$ and $f(w^t) - f(w^*)$. In Appendix B.1.3 we prove a link between f's largest gradient entry and its suboptimality gap $f(w) - f(w^*)$, under a strong convexity assumption. This inequality will be used to prove the utility of DP-GCD for strongly-convex functions, and to assess its fast initial convergence for problems whose solution is quasi-sparse.

B.1.1 A Classical Inequality

We start by stating and proving the following inequality in Lemma B.1. While very simple, it will play a crucial role in our proofs.

Lemma B.1. *Let $a, b \in \mathbb{R}$, it holds that*

$$-(a+b)^2 \leq -\frac{1}{2}a^2 + b^2 \ , \tag{3}$$

$$-a^2 \leq -\frac{1}{2}(a+b)^2 + b^2 \ . \tag{4}$$

Proof. First, remark that for $\alpha, \beta \in \mathbb{R}$, it holds that

$$(\alpha+\beta)^2 - 2\alpha^2 - 2\beta^2 = \alpha^2 + 2\alpha\beta + \beta^2 - 2\alpha^2 - 2\beta^2 \tag{5}$$

$$= -\alpha^2 + 2\alpha\beta - \beta^2 = -(\alpha-\beta)^2 \leq 0 \ , \tag{6}$$

hence $(\alpha+\beta)^2 \leq 2\alpha^2 + \beta^2$. Now, take $\alpha = a+b$ and $\beta = -b$, we obtain $a^2 \leq 2(a+b)^2 + 2b^2$, which, after reorganizing, gives $-(a+b)^2 \leq -\frac{1}{2}a^2 + b^2$. Similarly, taking $\alpha = a$ and $\beta = b$, gives $(a+b)^2 \leq 2a^2 + 2b^2$ and consequently $-a^2 \leq -\frac{1}{2}(a+b)^2 + b^2$. We proved our lemma. □

We now state a useful concentration inequality on Laplace random variables.

Lemma B.2. *Let $K > 0$ and $\lambda_1, \dots, \lambda_K > 0$. Define $X_k \sim \mathrm{Lap}(\lambda_k)$ and $\lambda_{\max} = \max_{k\in[K]} \lambda_k$. For any $\beta > 0$, it holds that*

$$\Pr\left[\sum_{k=1}^{K} X_k^2 \geq \beta\right] \leq 2^K \exp\left(-\frac{\sqrt{\beta}}{2\lambda_{\max}}\right) \ . \tag{7}$$

Proof. We first remark that $(\sum_{k=1}^K |X_k|)^2 = \sum_{k=1}^K \sum_{k'=1}^K |X_k||X_{k'}| \geq \sum_{k=1}^K X_k^2$. Therefore

$$\Pr\left[\sum_{k=1}^{K} X_k^2 \geq a^2\right] \leq \Pr\left[\Big(\sum_{k=1}^{K} |X_k|\Big)^2 \geq a^2\right] = \Pr\left[\Big(\sum_{k=1}^{K} |X_k|\Big) \geq a\right] \ . \tag{8}$$

Chernoff's inequality now gives, for any $\gamma > 0$,

$$\Pr\left[\sum_{k=1}^{K} |X_k| \geq a\right] \leq \exp(-\gamma a)\mathbb{E}\left[\exp(\gamma \sum_{k=1}^{K} |X_k|)\right] \ . \tag{9}$$

By the properties of the exponential and the X_k's independence, we can rewrite the inequality as

$$\Pr\left[\sum_{k=1}^{K} |X_k| \geq a\right] \leq \exp(-\gamma a)\mathbb{E}\left[\prod_{k=1}^{K} \exp\Big(\gamma|X_k|\Big)\right] = \exp(-\gamma a)\prod_{k=1}^{K} \mathbb{E}\Big[\exp\Big(\gamma|X_k|\Big)\Big] \ . \tag{10}$$

We can now compute the expectation of $\exp(\gamma|X_k|)$ for $k \in [K]$,

$$\mathbb{E}\Big[\exp\Big(\gamma|X_k|\Big)\Big] = \frac{1}{2\lambda_k}\int_{-\infty}^{+\infty} \exp(\gamma|x|)\exp(-\frac{|x|}{\lambda_k})dx = \frac{1}{\lambda_k}\int_{0}^{+\infty} \exp\Big((\gamma - \frac{1}{\lambda_k})x\Big)dx \ . \tag{11}$$

We choose $\gamma = 1/2\lambda_{\max}$, such that $\gamma \leq 1/2\lambda_k$ for all $k \in [K]$ and obtain

$$\mathbb{E}\Big[\exp\Big(\gamma|X_k|\Big)\Big] = \frac{1}{\lambda_k}\frac{1}{\frac{1}{\lambda_k} - \gamma} = \frac{1}{1 - \gamma\lambda_k} \leq 2 \ . \tag{12}$$

Plugging everything together, we have proved that

$$\Pr\left[\sum_{k=1}^{K} X_k^2 \geq a^2\right] \leq \Pr\left[\sum_{k=1}^{K} |X_k| \geq a\right] \leq 2^K \exp(-\frac{a}{2\lambda_{\max}}) \ , \tag{13}$$

and taking $a = \sqrt{\beta}$ gives the result. □

B.1.2 For Convex Functions

In this section, we prove Lemma B.3, that will be used to solve the recursion we obtain for convex functions. In this lemma, one should think of ξ_t as $f(w^t) - f(w^*)$ and of β as a kind of variance term.

Lemma B.3. *Let $\{c_t\}_{t\geq 0}$ and $\{\xi_t\}_{t\geq 0}$ be two sequences of positive values that satisfy, for all $t \geq 0$,*

$$\xi_{t+1} \leq \xi_t - \frac{\xi_t^2}{c_t} + \beta, \tag{14}$$

such that if $\xi_t \leq \xi_0$ then $c_t \leq c_0$. Assume that $\beta \leq c_0$ and $\xi_0 \geq 2\sqrt{\beta c_0}$. Then:

1. *for all $t > 0$, $c_t \leq c_0$, and there exists $t^* > 0$ such that $\xi_{t+1} \leq \xi_t$ if $t < t^*$ and $\xi_t \leq 2\sqrt{\beta c_0}$ if $t \geq t^*$.*
2. *for all $t \geq 1$, $\xi_t \leq \frac{c_0}{t} + 2\sqrt{\beta c_0}$.*

Proof. 1. Assume that for $t \geq 0$, $\sqrt{\beta c_0} \leq \xi_t \leq \xi_0$. Then,

$$\xi_{t+1} \leq \xi_t - \frac{\xi_t^2}{c_t} + \beta \leq \xi_t - \frac{\sqrt{\beta c_0}^2}{c_0} + \beta = \xi_t \ , \tag{15}$$

where the second inequality comes from $\xi_t \geq \sqrt{\beta c_0}$ and $\xi_t \leq \xi_0$ (which implies $c_t \leq c_0$). We now define the following value t^*, which defines the point of rupture between two regimes for ξ_t:

$$t^* = \min \left\{ t \geq 0 \,\middle|\, \xi_t \leq \sqrt{\beta c_0} \right\} \ . \tag{16}$$

Let $t < t^*$, assume that $\xi_t \leq \xi_0$, then (15) holds, that is $\xi_{t+1} \leq \xi_t \leq \xi_0$. By induction, it follows that for all $t < t^*$, $\xi_{t+1} \leq \xi_t \leq \xi_0$ and $c_t \leq c_0$.

Remark now that $\xi_{t^*} \leq \sqrt{\beta c_0}$, we prove by induction that ξ_t stays under $2\sqrt{\beta c_0}$ for $t \geq t^*$. Assume that for $t \geq t^*$, $\xi_t \leq 2\sqrt{\beta c_0}$. Then, there are two possibilities. If $\xi_t \leq \sqrt{\beta c_0}$, then

$$\xi_{t+1} \leq \xi_t - \frac{\xi_t^2}{c_t} + \beta \leq \sqrt{\beta c_0} + \beta \leq 2\sqrt{\beta c_0} \ , \tag{17}$$

and $\xi_{t+1} \leq 2\sqrt{\beta c_0}$. Otherwise, $\sqrt{\beta c_0} \leq \xi_t \leq 2\sqrt{\beta c_0} \leq \xi_0$ and (15) holds, which gives $\xi_{t+1} \leq \xi_t \leq 2\sqrt{\beta c_0}$. We proved that for $t \geq t^*$, $\xi_t \leq 2\sqrt{\beta c_0}$, which concludes the proof of the first part of the lemma.

2. We start by proving this statement for $0 < t < t^* - 1$. Define $\omega = \frac{2u}{c_0}$ and $u = \sqrt{\beta c_0}$. The assumption on ξ_t implies, by the first part of the lemma, $\xi_{t+1} \leq \xi_t - \frac{\xi_t^2}{c_t} + \beta \leq \xi_t - \frac{\xi_t^2}{c_0} + \beta$, which can be rewritten

$$\xi_{t+1} - u \leq (1-\omega)(\xi_t - u) - \frac{(\xi_t - u)^2}{c_0} \ , \tag{18}$$

since $(1-\omega)(\xi_t - u) - \frac{(\xi_t - u)^2}{c_0} = \xi_t - \omega\xi_t - u + \omega u - \frac{\xi_t^2}{c_0} - \frac{2\xi_t u}{c_0} - \frac{u^2}{c_0} = \xi_t - \frac{\xi_t^2}{c_0} - u + \omega u - \frac{u^2}{c_0}$, and $\omega u - \frac{u^2}{c_0} = \frac{u^2}{c_0} = \beta$. Since $t < t^* - 1$, $\xi_{t+1} - u > 0$ and $\xi_t - u > 0$, we can thus divide (18) by $(\xi_{t+1} - u)(\xi_t - u)$ to obtain

$$\frac{1}{\xi_t - u} \leq \frac{1-\omega}{\xi_{t+1} - u} - \frac{\xi_t - u}{(\xi_{t+1} - u)c_0} \leq \frac{1-\omega}{\xi_{t+1} - u} - \frac{1}{c_0} \leq \frac{1}{\xi_{t+1} - u} - \frac{1}{c_0} \ , \tag{19}$$

where the second inequality comes from $\xi_{t+1} - u \leq \xi_t - u$ from the first part of the lemma. By applying this inequality recursively and taking the inverse of the result, we obtain the desired resuld $\xi_t \leq \frac{c_0}{t} + \sqrt{\beta c_0} \leq \frac{c_0}{t} + 2\sqrt{\beta c_0}$ for all $0 < t < t^*$.

For $t \geq t^*$, we have already proved that $\xi_t \leq 2\sqrt{\beta c_0} \leq \frac{c_0}{t} + 2\sqrt{\beta c_0}$, which concludes our proof.

□

B.1.3 For Strongly-Convex Functions

We now prove a link between f's largest gradient entry and the suboptimality gap, under the assumption that there exists a unique minimizer w^* of f that is (α, τ)-quasi-sparse (which is not restrictive). We will denote by $\mathcal{W}_{\tau,\alpha} \subseteq \mathbb{R}^p$ the set of (α, τ)-quasi-sparse vectors of $\mathbb{R}^p$:

$$\mathcal{W}_{\tau,\alpha} = \{ w \in \mathbb{R}^p \mid |\{ j \in [p] \mid |w_j| \geq \alpha \}| \leq \tau \} \ . \tag{20}$$

When $\alpha = 0$, we simply write $\mathcal{W}_\tau = \mathcal{W}_{\tau,0}$, that is the set of τ-sparse vectors. We also define the associated thresholding operator π_α, that puts to 0 the coordinates that are smaller than α, "projecting" vectors from $\mathcal{W}_{\tau,\alpha}$ to $\mathcal{W}_\tau$, *i.e.,* for $w \in \mathbb{R}^p$,

$$\pi_\alpha(w) = \begin{cases} 0 & \text{if } |w_j| \leq \alpha \ , \\ w_j & \text{otherwise} \ . \end{cases} \tag{21}$$

We are ready to prove Lemma B.4.

Lemma B.4. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function that is M-component-smooth, and $\mu_{M,1}^{(\tau)}$-strongly-convex w.r.t. $\|\cdot\|_{M,1}$ when restricted to τ-sparse vectors, for $\tau \in [p]$. Assume that the unique minimizer w^* of f is (τ, α)-quasi-sparse, for $\alpha, \tau \geq 0$. Let $w^t \in \mathbb{R}^p$ be a t-sparse vector for $0 \leq t \leq p - \tau$. Then we have*

$$-\frac{1}{2}\|\nabla f(w^t)\|_{M^{-1},\infty} \leq -\mu_{M,1}^{(t+\tau)}(f(w^t) - f(w^*)) + \frac{1}{2}M_{\max}\mu_{M,1}^{(t+\tau)}(p-\tau)\alpha^2 \ . \tag{22}$$

Proof. Let $w^t \in \mathbb{R}^p$ be a t-sparse vector. Remark that w^* is (α, τ)-quasi-sparse, meaning that $\pi_\alpha(w^*)$ is τ-sparse. The union of w^t and $\pi_\alpha(w^*)$'s supports ($\mathrm{supp}(w^t)$ and $\mathrm{supp}(\pi_\alpha(w^*))$) thus satisfies $|\mathrm{supp}(w) \cup \mathrm{supp}(\pi_\alpha(w^*))| \leq t + \tau$. As the function f is $\mu_{M,1}^{(t+\tau)}$-strongly-convex with respect to $\|\cdot\|_{M,1}$ and $t+\tau$ sparse vector,

$$f(\pi_\alpha(w)) \geq f(w^t) + \langle \nabla f(w^t), \pi_\alpha(w) - w^t \rangle + \frac{\mu_{M,1}^{(t+\tau)}}{2}\|\pi_\alpha(w) - w^t\|_{M,1}^2 \ . \tag{23}$$

Since $\pi_\alpha : \mathcal{W}_{\tau,\alpha} \to \mathcal{W}_{\tau,0}$ is surjective, minimizing this equation for $w \in \mathcal{W}_{\tau,\alpha}$ on both sides gives

$$\inf_{w \in \mathcal{W}_\tau} f(w) \geq f(w^t) - \sup_{w \in \mathcal{W}_{\tau,\alpha}} \left\{ \langle -\nabla f(w^t), w^t - \pi_\alpha(w) \rangle - \frac{\mu_{M,1}^{(t+\tau)}}{2}\|\pi_\alpha(w) - w^t\|_{M,1}^2 \right\} \tag{24}$$

$$\geq f(w^t) - \sup_{w \in \mathbb{R}^p} \left\{ \langle -\nabla f(w^t), w^t - w \rangle - \frac{\mu_{M,1}^{(t+\tau)}}{2}\|w - w^t\|_{M,1}^2 \right\} \ . \tag{25}$$

We thus made appear the conjugate of the function $\frac{1}{2}\|\cdot\|_{M,1}^2$, that is $\frac{1}{2}\|\cdot\|_{M^{-1},\infty}^2$ (Boyd and Vandenberghe, 2004). This gives

$$\inf_{w \in \mathcal{W}_\tau} f(w) \geq f(w^t) - \left(\frac{\mu_{M,1}^{(t+\tau)}}{2}\|\cdot\|_1^2 \right)^* (-\nabla f(w')) \tag{26}$$

$$= f(w^t) - \frac{1}{2\mu_{M,1}^{(t+\tau)}}\|\nabla f(w')\|_{M^{-1},\infty}^2 \ . \tag{27}$$

Finally, w^* is the minimizer of f (which is convex), thus $\nabla f(w^*) = 0$. The smoothness of f gives, for any $w \in \mathbb{R}^p$, $f(w) \leq f(w^*) + \frac{1}{2}\|w - w^*\|_{M,2}^2$. Hence

$$\inf_{w \in \mathcal{W}_\tau} f(w) \leq f(w^*) + \inf_{w \in \mathcal{W}_\tau} \frac{1}{2}\|w - w^*\|_{M,2}^2. \leq f(w^*) + \frac{1}{2}\|\pi_\alpha(w^*) - w^*\|_{M,2}^2 \ , \tag{28}$$

where the second inequality comes from $\pi_\alpha(w^*) \in \mathcal{W}_\tau$, since $w^* \in \mathcal{W}_{\tau,\alpha}$. It remains to remark that $\|\pi_\alpha(w^*) - w^*\|_{M,2}^2 \leq M_{\max}(p-\tau)\alpha^2$ to get the result. □

Corollary B.5. *For τ-sparse vectors, we have $\alpha = 0$ and consequently $(p-\tau)\alpha = 0$. Lemma B.4 can thus be simplified as*

$$-\frac{1}{2}\|\nabla f(w^t)\|_{M^{-1},\infty}^2 \leq -\mu_{M,1}^{(t+\tau)}(f(w^t) - f(w^*)) \ . \tag{29}$$

B.2 Descent Lemma

In this section, we prove our key (noisy) descent lemma. This lemma links the value of $f(w^{t+1}) - f(w^*)$ at a given iteration to the value of largest gradient entry. We will then link this gradient value to the one of $f(w^t) - f(w^*)$, using either convexity or strong convexity.

Lemma B.6. *Let $t \geq 0$ and $w^t, w^{t+1} \in \mathbb{R}^p$ two consecutive iterates of Algorithm 1, with $\gamma_j = 1/M_j$ and λ_j, λ_j' chosen as in Theorem 3.1 to ensure ϵ, δ-DP. We denote by $j \in [p]$ the coordinate chosen at this step t and by $j^* = \arg\max_{j \in [p]} |\nabla_j f(w^t)|/M_j$ the coordinate that would have been chosen without noise. The following inequality holds*

$$f(w^{t+1}) - f(w^*) \leq f(w^t) - f(w^*) - \frac{1}{8}\|\nabla f(w^t)\|_{M^{-1},\infty}^2 + \frac{1}{2M_j}|\eta_j^t|^2 + \frac{1}{2M_j}|\chi_j^t|^2 + \frac{1}{4M_{j^*}}|\chi_{j^*}^t|^2 \ . \tag{30}$$

Proof. The smoothness of f gives a first inequality

$$f(w^{t+1}) \leq f(w^t) + \langle \nabla f(w^t), w^{t+1} - w^t \rangle + \frac{1}{2}\|w^{t+1} - w^t\|_M^2 \tag{31}$$

$$= f(w^t) - \frac{1}{M_j}\nabla_j f(w^t)(\nabla_j f(w^t) + \eta_j^t) + \frac{1}{2M_j}(\nabla_j f(w^t) + \eta_j^t)^2 \tag{32}$$

$$= f(w^t) - \frac{1}{M_j}\nabla_j f(w^t)^2 - \frac{1}{M_j}\nabla_j f(w^t)\eta_j^t + \frac{1}{2M_j}(\nabla_j f(w^t))^2 + \frac{1}{M_j}\nabla_j f(w^t)\eta_j^t + \frac{1}{2M_j}(\eta_j^t)^2 \tag{33}$$

$$= f(w^t) - \frac{1}{2M_j}\nabla_j f(w^t)^2 + \frac{1}{2M_j}(\eta_j^t)^2 \ . \tag{34}$$

We will make the noisy gradient appear, so as to use the noisy greedy rule. To do so, we apply Lemma B.1 and get

$$-\frac{1}{2M_j}\nabla_j f(w^t)^2 \leq -\frac{1}{4M_j}(\nabla_j f(w^t) + \chi_j^t)^2 + \frac{1}{2M_j}(\chi_j^t)^2 \ . \tag{35}$$

And the noisy greedy rule gives $\frac{1}{\sqrt{M_{j^*}}}|\nabla_{j^*} f(w^t) + \chi_{j^*}^t| \leq \frac{1}{\sqrt{M_j}}|\nabla_j f(w^t) + \chi_j^t|$. We replace in (35) to obtain

$$-\frac{1}{2M_j}\nabla_j f(w^t)^2 \leq -\frac{1}{4M_{j^*}}(\nabla_{j^*} f(w^t) + \chi_{j^*}^t)^2 + \frac{1}{2M_j}(\chi_j^t)^2 \tag{36}$$

$$\leq -\frac{1}{8M_{j^*}}(\nabla_{j^*} f(w^t))^2 + \frac{1}{4M_{j^*}}(\chi_{j^*}^t)^2 + \frac{1}{2M_j}(\chi_j^t)^2 \ . \tag{37}$$

The result follows from (34) and $\frac{1}{M_{j^*}}(\nabla_{j^*} f(w^t))^2 = \|\nabla f(w^t)\|^2_{M^{-1},\infty}$. □

B.3 Utility for General Convex Functions

For convex functions, we will use convexity go upper bound the gradient's largest entry, and convert our inequality into a high-probability one, allowing the use of our technical Lemma B.3.

Lemma B.7. *Under the hypotheses of Lemma B.6, for a convex objective function f, we have*

$$f(w^{t+1}) - f(w^*) \leq f(w^t) - f(w^*) - \frac{(f(w^t) - f(w^*))^2}{8\|w^t - w^*\|_{M,1}^2} + \frac{1}{2M_j}|\eta_j^t|^2 + \frac{1}{2M_j}|\chi_j^t|^2 + \frac{1}{4M_{j^*}}|\chi_{j^*}^t|^2 \ . \tag{38}$$

Proof. Since f is convex, it holds that

$$f(w^*) \geq f(w^t) + \langle \nabla f(w^t), w^* - w^t \rangle \ . \tag{39}$$

After reorganizing the terms, we can upper bound them using Hölder's inequality

$$f(w^t) - f(w^*) \leq \langle \nabla f(w^t), w^t - w^* \rangle \tag{40}$$

$$\leq \|\nabla f(w^t)\|_{M^{-1},\infty}\|w^t - w^*\|_{M,1} \ , \tag{41}$$

where the second inequality holds since $\|\cdot\|_{M,1}$ and $\|\cdot\|_{M^{-1},\infty}$ are conjugate norms. We now divide (41) by $\|w^t - w^*\|_{M,1}$, square it and reorganize to get $-\|\nabla f(w^t)\|^2_{M^{-1},\infty} \leq -\frac{(f(w^t)-f(w^*))^2}{\|w^t - w^*\|^2_{M,1}}$. Replacing in Lemma B.6 gives the result. □

Theorem 3.3. *(Convex Case) Let $\epsilon, \delta \in (0,1]$. Assume $\ell(\cdot; d)$ is a convex and L-component-Lipschitz loss function for all $d \in \mathcal{X}$, and f is M-component-smooth. Define $\mathcal{W}^*$ the set of minimizers of f, and f^* the minimum of f. Let $w_{priv} \in \mathbb{R}^p$ be the output of Algorithm 1 with step sizes $\gamma_j = 1/M_j$, and noise scales $\lambda_1, \dots, \lambda_p, \lambda'_1, \dots, \lambda'_p$ set as in Theorem 3.1 (with T chosen below) to ensure (ϵ, δ)-DP. Then, the following holds for $\zeta \in (0,1]$:*

$$f(w_{priv}) - f(w^*) \leq \frac{8R_M^2}{T} + \sqrt{32R_M^2\beta} \ , \tag{42}$$

where $\beta = \frac{2\lambda_{\max}^2}{M_{\min}} \log(\frac{8Tp}{\zeta})^2$, *and* $R_M = \max_{w \in \mathbb{R}^p} \min_{w^* \in \mathcal{W}^*} \{\|w - w^*\|_{M,1} \mid f(w) \leq f(w^0)\}$. *If we set* $T = \left(\frac{n^2\epsilon^2 R_M^2 M_{\min}}{2^7 L_{\max}^2 \log(1/\delta)}\right)^{1/3}$, *then with probability at least* $1 - \zeta$,

$$f(w^T) - f(w^0) = \widetilde{O}\Big(\frac{R_M^{4/3} L_{\max}^{2/3} \log(p/\zeta)}{M_{\min}^{1/3} n^{2/3} \epsilon^{2/3}}\Big) \ . \tag{43}$$

Proof. Let $\xi_t = f(w^t) - f(w^*)$. We upper bound the following probability by the union bound, and the fact that for $t \geq 0$, the events E_j^t : "coordinate j is updated at step t" for $j \in [p]$ partition the probability space:

$$\Pr\left[\exists t, \xi_{t+1} \geq \xi_t - \frac{\xi_t^2}{8\|w^t - w^*\|_{M,1}^2} + \beta\right] \leq \sum_{t=0}^{T-1} \Pr\left[\xi_{t+1} \geq \xi_t - \frac{\xi_t^2}{8\|w^t - w^*\|_{M,1}^2} + \beta\right] \tag{44}$$

$$= \sum_{t=0}^{T-1} \sum_{j=1}^{p} \Pr\left[\xi_{t+1} \geq \xi_t - \frac{\xi_t^2}{8\|w^t - w^*\|_{M,1}^2} + \beta \ \wedge \ E_j^t\right] \ . \tag{45}$$

Lemma B.7 gives $\xi_{t+1} \leq \xi_t - \frac{\xi_t^2}{8\|w^t - w^*\|_{M,1}^2} + \frac{1}{2M_j}|\eta_j^t|^2 + \frac{1}{2M_j}|\chi_j^t|^2 + \frac{1}{4M_{j^*}}|\chi_{j^*}^t|^2$. We thus have the following upper bound:

$$\Pr\left[\exists t, \xi_{t+1} \geq \xi_t - \frac{1}{8\|w^t - w^*\|_{M,1}^2}\xi_t^2 + \beta\right] \leq \sum_{t=0}^{T-1} \sum_{j=1}^{p} \Pr\left[\frac{|\eta_j^t|^2}{2M_j} + \frac{|\chi_j^t|^2}{2M_j} + \frac{|\chi_{j^*}^t|^2}{4M_{j^*}} \geq \beta\right] \tag{46}$$

$$\leq \sum_{t=0}^{T-1} \sum_{j=1}^{p} \Pr\left[|\eta_j^t|^2 + |\chi_j^t|^2 + |\chi_{j^*}^t|^2 \geq 2M_{\min}\beta\right] \ . \tag{47}$$

By Lemma B.2 with $X_1 = \eta_j^t \sim \mathrm{Lap}(\lambda_j)$, $X_2 = \chi_j^t \sim \mathrm{Lap}(\lambda'_j)$ and $X_3 = \chi_{j^*}^t \sim \mathrm{Lap}(\lambda'_{j^*})$, it holds that

$$\Pr\left[|\eta_j^t|^2 + |\chi_j^t|^2 + |\chi_{j^*}^t|^2 \geq 2M_{\min}\beta\right] \leq 8\exp\left(-\frac{\sqrt{2M_{\min}\beta}}{2\lambda_{\max}}\right) = \frac{\zeta}{Tp} \ , \tag{48}$$

where the last equality comes from $\beta = \frac{2\lambda_{\max}^2}{M_{\min}} \log(\frac{8Tp}{\zeta})^2$. We have proved that

$$\Pr\left[\exists t, \xi_{t+1} \geq \xi_t - \frac{\xi_t^2}{8\|w^t - w^*\|_{M,1}^2} + \beta\right] \leq \sum_{t=0}^{T-1} \sum_{j=1}^{p} \frac{\zeta}{Tp} = \zeta \ . \tag{49}$$

We now use our Lemma B.3, with $\xi_t = f(w^t) - f(w^*)$; $c_0 = 8R_M^2$ and $c_t = 8\|w^t - w^*\|_{M,1}^2$ for $t > 0$; and $\beta = \frac{2\lambda_{\max}^2}{M_{\min}} \log(\frac{8Tp}{\zeta})^2$. These values satisfies the assumptions of Lemma B.3 since, by the definition of R_M, it holds that $c_t \leq c_0$ whenever $\xi_t \leq \xi_0$ (*i.e.,* $f(w^t) - f(w^*) \leq f(w^0) - f(w^*)$). Additionally, $f(w^0) - f(w^*) \geq \sqrt{32R_M^2\beta}$, therefore $f(w^0) - f(w^*) \geq 2\sqrt{\beta c_0}$, and $\beta \leq c_0$.

We obtain the result, with probability at least $1 - \zeta$:

$$f(w^t) - f(w^0) \leq \frac{c_0}{t} + 2\sqrt{\beta c_0} = \frac{8R_M^2}{t} + \frac{64R_M L_{\max} \log(8Tp/\zeta)\sqrt{T\log(1/\delta)}}{\sqrt{M_{\min}} n\epsilon} \ . \tag{50}$$

For $T = \frac{R_M^{2/3} M_{\min}^{1/3} n^{2/3} \epsilon^{2/3}}{4 L_{\max}^{2/3} \log(1/\delta)^{1/3}}$, we obtain that, with probability at least $1-\zeta$,

$$f(w^t) - f(w^0) \leq \frac{64 R_M^{4/3} L_{\max}^{2/3} \log(1/\delta)^{1/3}}{M_{\min}^{1/3} n^{2/3} \epsilon^{2/3}} \log\Big(\frac{p R_M^{2/3} M_{\min}^{1/3} n^{2/3} \epsilon^{2/3}}{4\zeta L_{\max}^{2/3} \log(1/\delta)^{1/3}}\Big) \ , \tag{51}$$

which is the result of the lemma. □

B.4 Utility for Strongly-Convex Functions

Theorem 3.3. *(Strongly-Convex Case) Let $\epsilon, \delta \in (0,1]$. Assume $\ell(\cdot; d)$ is a $\mu_{M,1}$-strongly-convex w.r.t. $\|\cdot\|_{M,1}$ and L-component-Lipschitz loss function for all $d \in \mathcal{X}$, and f is M-component-smooth. Define $\mathcal{W}^*$ the set of minimizers of f, and f^* the minimum of f. Let $w_{priv} \in \mathbb{R}^p$ be the output of Algorithm 1 with step sizes $\gamma_j = 1/M_j$, and noise scales $\lambda_1, \ldots, \lambda_p, \lambda'_1, \ldots, \lambda'_p$ set as in Theorem 3.1 (with T chosen below) to ensure (ϵ, δ)-DP. Then, the following holds for $\zeta \in (0,1]$:*

$$f(w^T) - f(w^*) \leq (1 - \frac{\mu_M}{2})^T (f(w^0) - f(w^*)) + \frac{64 T L_{\max}^2 \log(1/\delta)}{M_{\min} \mu_M n^2 \epsilon^2} \log(\frac{2Tp}{\zeta}) \ . \tag{52}$$

If we set $T = \frac{2}{\mu_M} \log(\frac{M_{\min} \mu_M n^2 \epsilon^2 (f(w^0) - f(w^*))}{32 L_{\max}^2 \log(1/\delta)})$, then with probability at least $1-\zeta$,

$$f(w^T) - f(w^*) = \widetilde{O}\Big(\frac{L_{\max}^2 \log(p/\zeta)}{M_{\min} \mu_M^2 n^2 \epsilon^2}\Big) \ . \tag{53}$$

Proof. When f is $\mu_{M,1}$-strongly-convex w.r.t. the norm $\|\cdot\|_{M,1}$, Corollary B.5 yields

$$-\frac{1}{2} \|\nabla f(w^t)\|^2_{M^{-1},\infty} \leq -\mu_{M,1} (f(w^t) - f(w^*)) \ . \tag{54}$$

We replace this in Lemma B.6 to obtain

$$\begin{aligned} f(w^{t+1}) - f(w^*) \leq (1 - \frac{\mu_M}{4})(f(w^t) - f(w^*)) \\ + \frac{1}{2M_j} |\eta_j^t|^2 + \frac{1}{2M_j} |\chi_j^t|^2 + \frac{1}{4M_{j^*}} |\chi_{j^*}^t|^2 \ . \end{aligned} \tag{55}$$

Analogously to the proof of Theorem 3.3, we define $\xi_t = f(w^t) - f(w^*)$ and show that $\Pr\left[\exists t, \xi_{t+1} \geq (1 - \frac{\mu_M}{4})\xi_t + \beta\right] \leq \zeta/Tp$, with $\beta = \frac{2\lambda_{\max}^2}{M_{\min}} \log(\frac{8Tp}{\zeta})^2$. This yields that, with probability at least $1-\zeta$,

$$f(w^T) - f(w^*) \leq (1 - \frac{\mu_M}{4})^T (f(w^0) - f(w^*)) + \sum_{t=0}^{T-1} (1 - \frac{\mu_M}{4})^{T-t} \beta \tag{56}$$

$$\leq (1 - \frac{\mu_M}{4})^T (f(w^0) - f(w^*)) + \frac{4}{\mu_M} \frac{32 T L_{\max}^2 \log(1/\delta)}{M_{\min} n^2 \epsilon^2} \log\Big(\frac{8Tp}{\zeta}\Big)^2 \ , \tag{57}$$

With $T = \frac{4}{\mu_M} \log\left(\frac{\mu_M M_{\min} n^2 \epsilon^2 (f(w^0) - f(w^*))}{128 L_{\max}^2 \log(1/\delta) \log(8p/\zeta)}\right)$ we have, with probability at least $1-\zeta$,

$$\begin{aligned} f(w^T) - f(w^*) \leq \frac{128 L_{\max}^2 \log(1/\delta) \log(8p/\zeta)^2}{\mu_M M_{\min} n^2 \epsilon^2} \\ + \frac{512 L_{\max}^2 \log(1/\delta) \log(8Tp/\zeta)^2}{\mu_M^2 M_{\min} n^2 \epsilon^2} \log\left(\frac{\mu_M M_{\min} n^2 \epsilon^2 (f(w^0) - f(w^*))}{128 L_{\max}^2 \log(1/\delta) \log(8p/\zeta)^2}\right) \ , \end{aligned} \tag{58}$$

which is the desired result. □

C Fast Initial Convergence

Theorem 3.6. *Let f satisfy the hypotheses of Theorem 3.3, where Algorithm 1 is initialized with $w^0 = 0$ and outputs w^T. Assume that f is $\mu^{(\tau)}_{M,1}$-strongly-convex w.r.t. $\|\cdot\|_{M,1}$ for τ-sparse vectors and $\mu_{M,2}$-strongly-convex w.r.t. $\|\cdot\|_{M,2}$. Assume that the (unique) solution of (1) is (α,τ)-quasi-sparse for some $\alpha,\tau \geq 0$. Let $0 \leq T \leq p-\tau$, $\zeta \in [0,1]$. Then with probability at least $1-\zeta$, it holds for $\beta = \frac{2\lambda^2_{\max}}{M_{\min}} \log(TP/\zeta)^2$,*

$$f(w^T) - f(w^*) \leq \Big(1 - \frac{\mu^{(T+\tau)}_{M,1}}{4}\Big)^T (f(w^0) - f(w^*)) + \frac{4(T+\tau)\beta}{\mu_{M,2}} + \frac{T+\tau}{8}(p-\tau)\alpha^2 \tag{59}$$

$$\leq \Big(1 - \frac{\mu_{M,2}}{4(T+\tau)}\Big)^T (f(w^0) - f(w^*)) + \frac{4(T+\tau)\beta}{\mu_{M,2}} + \frac{T+\tau}{8}(p-\tau)\alpha^2 \ . \tag{60}$$

Proof. First, we remark that at each iteration, we change only one coordinate. Therefore, after t iterations, the iterate w^t is t-sparse. Additionally, we assumed that w^* is (τ,α)-almost-sparse. Therefore, Lemma B.4 yields

$$-\frac{1}{2}\|\nabla f(w^t)\|_{M^{-1},\infty} \leq -\mu^{(t+\tau)}_{M,1}(f(w^t) - f(w^*)) + \frac{\mu^{(t+\tau)}_{M,1}}{2}(p-\tau)\alpha^2 \ , \tag{61}$$

and Lemma B.6 becomes

$$\begin{aligned} f(w^{t+1}) - f(w^*) \leq (1 - \frac{\mu^{(t+\tau)}_{M,1}}{4})(f(w^t) - f(w^*)) + \frac{\mu^{(t+\tau)}_{M,1}}{8}(p-\tau)\alpha^2 \\ + \frac{1}{2M_j}|\eta^t_j|^2 + \frac{1}{2M_j}|\chi^t_j|^2 + \frac{1}{4M_{j^*}}|\chi^t_{j^*}|^2 \ . \end{aligned} \tag{62}$$

Then by Chernoff's equality, we obtain (simiarly to the proof of Theorem 3.3 for the convex case) that with probability at least $1-\zeta$, for $T \leq p-\tau$,

$$\begin{aligned} f(w^T) - f(w^*) \leq \prod_{t=0}^{T}\Big(1 - \frac{\mu^{(t+\tau)}_{M,1}}{4}\Big)(f(w^0) - f(w^*)) \\ + \sum_{t=0}^{T-1}\prod_{k=T-t}^{T}\Big(1 - \frac{\mu^{(k+\tau)}_{M,1}}{4}\Big)\Big(\beta + \frac{\mu^{(t+\tau)}_{M,1}}{8}(p-\tau)\alpha^2\Big) \ . \end{aligned} \tag{63}$$

Since for $k \in [T]$, $\mu^{k+\tau}_{M,1} \geq \mu^{T+\tau}_{M,1}$, we can further upper bound $\mu^{(t+\tau)}_{M,1} \leq \mu^{(\tau)}_{M,1}$, $1 - \frac{\mu^{(t+\tau)}_{M,1}}{4} \leq 1 - \frac{\mu^{(T+\tau)}_{M,1}}{4}$ and

$$\sum_{t=0}^{T-1}\prod_{k=T-t}^{T}\Big(1 - \frac{\mu^{(k+\tau)}_{M,1}}{4}\Big) \leq \sum_{t=0}^{T-1}\Big(1 - \frac{\mu^{(T+\tau)}_{M,1}}{4}\Big)^t \leq \frac{4}{\mu^{(T+\tau)}_{M,1}} \ , \tag{64}$$

which allow simplifying the above expression to

$$f(w^T) - f(w^*) \leq \Big(1 - \frac{\mu^{(T+\tau)}_{M,1}}{4}\Big)^T (f(w^0) - f(w^*)) + \frac{4}{\mu^{(T+\tau)}_{M,1}}\Big(\beta + \frac{\mu^{(\tau)}_{M,1}}{8}(p-\tau)\alpha^2\Big) \tag{65}$$

$$\leq \Big(1 - \frac{\mu_{M,2}}{4(T+\tau)}\Big)^T (f(w^0) - f(w^*)) + \frac{4(T+\tau)}{\mu_{M,2}}\Big(\beta + \frac{\mu_{M,2}}{8}(p-\tau)\alpha^2\Big) \tag{66}$$

$$\leq \Big(1 - \frac{\mu_{M,2}}{4(T+\tau)}\Big)^T (f(w^0) - f(w^*)) + \frac{4(T+\tau)\beta}{\mu_{M,2}} + \frac{T+\tau}{8}(p-\tau)\alpha^2 \ , \tag{67}$$

where the second inequality follows from $\mu^{(T+\tau)}_{M,1} \geq \frac{\mu^{(T+\tau)}_{M,2}}{T+\tau} \geq \frac{\mu_{M,2}}{T+\tau}$ and $\mu^{(\tau)}_{M,1} \leq \mu_{M,2}$. We have proven the two inequalities of our lemma. □

D Greedy Coordinate Descent for Composite Problems

Consider the problem of privately approximating

$$w^* \in \arg\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \ell(w; d_i) + \psi(w), \tag{68}$$

where $D = (d_1, \ldots, d_n)$ is a dataset of n samples drawn from a universe $\mathcal{X}$, $\ell : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ is a loss function which is convex and smooth in w, and $\psi : \mathbb{R}^p \to \mathbb{R}$ is a convex regularizer which is separable (*i.e.,* $\psi(w) = \sum_{j=1}^{p} \psi_j(w_j)$) and typically nonsmooth (*e.g.,* ℓ_1-norm).

Algorithm 2 DP-GCD (Proximal Version): Private Proximal Greedy CD

1: **Input:** initial $w^0 \in \mathbb{R}^p$, iteration count $T > 0, \forall j \in [p]$, noise scales λ_j, λ'_j, step sizes $\gamma_j > 0$.
2: **for** $t = 0$ to $T - 1$ **do**
3: Select j_t by the noisy `GS-s`, `GS-r` or `GS-q` rule.
4: $w^{t+1} = w^t + (\operatorname{prox} \gamma_j \psi_j (w^t - \gamma_j (\nabla_j f(w^t) + \eta^t_{j_t})) - w^t_j) e_j, \qquad \eta^t_{j_t} \sim \operatorname{Lap}(\lambda_{j_t})$.
5: **return** w^T.

We propose a proximal greedy algorithm to solve (68), see Algorithm 2. The proximal operator is the following (we refer to Parikh and Boyd, 2014, for a detailed discussion on proximal operator and related algorithms):

$$\operatorname{prox}_{\gamma\psi}(v) = \arg\min_{x \in \mathbb{R}^p} \{ \frac{1}{2} \|v - x\|_2^2 + \gamma \psi(x) \} \ . \tag{69}$$

The same privacy guarantees as for the smooth DP-GCD algorithm hold since, privacy-wise, the proximal step is a post-processing step. We also adapt the greedy selection rule to incorporate the non-smooth term. We can use one of the following three rules

$$j_t = \arg\max_{j \in [p]} \min_{\xi_j \in \partial \psi_j(w_j)} \frac{1}{\sqrt{M_j}} |\nabla_j f(w^t) + \eta^t_j + \xi_j| \ , \tag{GS-s}$$

$$j_t = \arg\max_{j \in [p]} \sqrt{M_j} |\operatorname{prox}_{\frac{1}{M_j}\psi_j}(w^t_j - \frac{1}{M_j}(\nabla_j f(w^t) + \eta^t_j) - w^t_j| \ , \tag{GS-r}$$

$$j_t = \arg\max_{j \in [p]} \min_{\alpha \in \mathbb{R}} \nabla_j f(w^t)\alpha + \frac{M_j}{2}\alpha^2 + \psi_j(w^t_j + \alpha) - \psi_j(w^t_j) \ . \tag{GS-q}$$

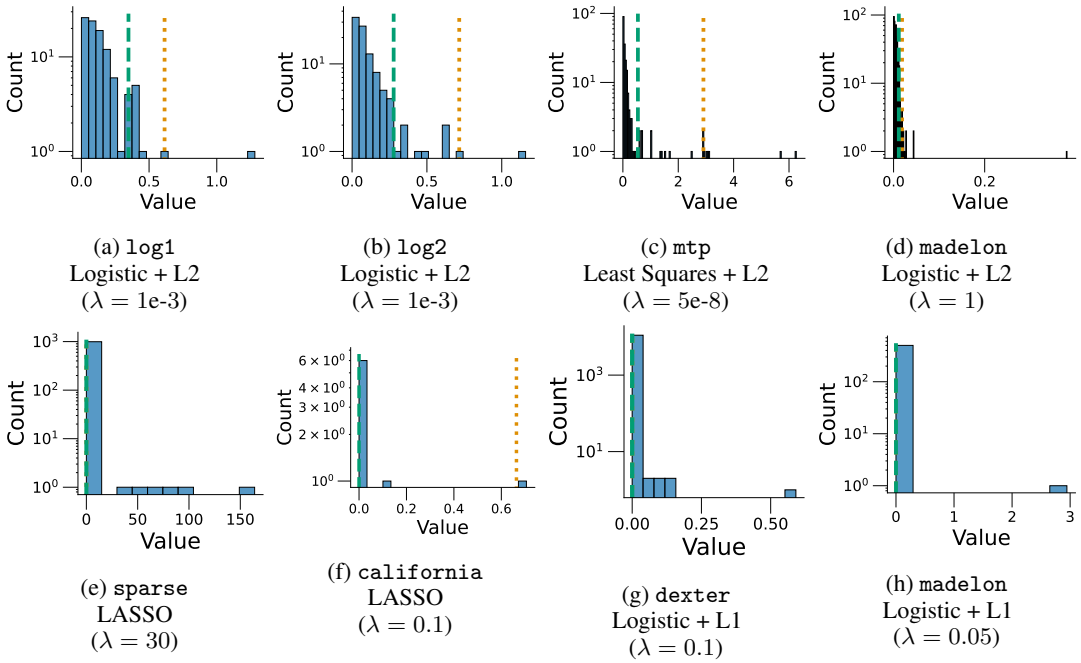
These rules are commonly considered in the non-private GCD literature (see *e.g.,* Tseng and Yun, 2009; Shi et al., 2017; Karimireddy et al., 2019), except for the noise η^t_j and the rescaling in the `GS-s` and `GS-r` rules.

E Experimental Details

In this section, we provide more details about the experiments, such as details on implementation, datasets and the hyperparameter grid we use for each algorithm. We then give the full results on our L1-regularized, non-smooth, problems, with the three greedy rules (as opposed to Section 5 where we only plotted results for the `GS-r` rule).

Code and setup. The algorithms are implemented in C++ for efficiency, together with a Python wrapper for simple use. It is provided as supplementary. Experiments are run on a computer with a Intel (R) Xeon(R) Silver 4114 CPU @ 2.20GHz and 64GB of RAM, and took about 10 hours in total to run (this includes all hyperparameter tuning).

Datasets. The datasets we use are described in Table 1. In Figure 2, we plot the histograms of the absolute value of each problem solution's parameters. On the `log1`, `log2`, `mtp` and `madelon` (l2) problems, the histograms show that many of the solution's parameters are small. On `log1` and `mtp`, this does not ensure



(a) `log1` Logistic + L2 ($\lambda =$ 1e-3)
(b) `log2` Logistic + L2 ($\lambda =$ 1e-3)
(c) `mtp` Least Squares + L2 ($\lambda =$ 5e-8)
(d) `madelon` Logistic + L2 ($\lambda = 1$)
(e) `sparse` LASSO ($\lambda = 30$)
(f) `california` LASSO ($\lambda = 0.1$)
(g) `dexter` Logistic + L1 ($\lambda = 0.1$)
(h) `madelon` Logistic + L1 ($\lambda = 0.05$)

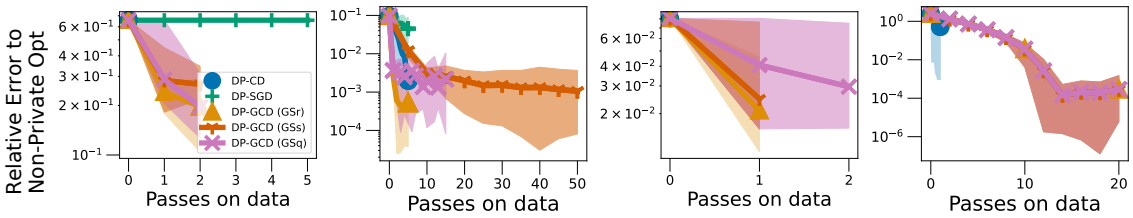
Figure 2: Histogram of the absolute value of each problem solution's parameters. Orange dotted line indicates the 99th quantile, Green dashed line indicated the α for which the plotted vector is $(\alpha, \lfloor \frac{p}{10} \rfloor)$-quasi-sparse. Y-axis is logarithmic.

Table 2: Hyperparameter Grid.

Algorithm	Parameter	Values
DP-CD	Passes on data	`[0.001, 0.01, 0.1, 1, 2, 3, 5]`
	Step sizes	`np.logspace(-2, 1, 10)`
	Clipping threshold	`np.logspace(-4, 6, 100)`
DP-SGD	Passes on data	`[0.001, 0.01, 0.1, 1, 2, 3, 5]`
	Step sizes	`np.logspace(-6, 0, 10)`
	Clipping threshold	`np.logspace(-4, 6, 100)`
DP-GCD	Passes on data	`[1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50]`
	Step sizes	`np.logspace(-2, 1, 10)`
	Clipping threshold	`np.logspace(-4, 6, 100)`

a high enough level of quasi-sparsity for DP-GCD to really outperform random selection. On `log2` and `madelon` (l2), solutions tend to be sparser (*i.e.,* (α, τ)-quasi-sparse for some τ and smaller α), which DP-GCD manages to exploit. On `sparse`, `california`, `dexter` and `madelon` (l1), the problems' solutions are actually sparse: our proximal DP-GCD algorithm uses this property to improve utility.

Hyperparameters. On all datasets, we use the same hyperparameter grid. For each algorithm, we choose between roughly the same number of hyperparameters. The number of passes on data represents p iterations of DP-CD, n iterations of DP-SGD, and 1 iteration of DP-GCD. The complete grid is described in Table 2.



(a) `sparse` LASSO ($\lambda = 30$)

(b) `california` LASSO ($\lambda = 0.1$)

(c) `dexter` Logistic + L1 ($\lambda = 0.1$)

(d) `madelon` Logistic + L1 ($\lambda = 0.05$)

Figure 3: Relative error to non-private optimal for DP-CD, proximal DP-GCD (with `GS-r`, `GS-s` and `GS-q` rules) and DP-SGD on different problems. On the x-axis, 1 tick represents a full access to the data: p iterations of DP-CD, n iterations of DP-SGD and 1 iteration of DP-GCD. Number of iterations, clipping thresholds and step sizes are tuned simultaneously for each algorithm. We report min/mean/max values over 10 runs.

Additional experiments on proximal DP-GCD. In Figure 3, we show the results of the proximal DP-GCD algorithm, after tuning the hyperparameters with the grid described above for each of the `GS-s`, `GS-r` and `GS-q` rules.

The three rules seem to behave qualitatively the same on `sparse`, `dexter` and `madelon`, our three high-dimensional non-smooth problems. There, most coordinates are chosen about one time. Thus, as described by Karimireddy et al. (2019), all the steps are "`good`" steps (along their terminology): and on such good steps, the three rules coincide. On the lower-dimensional dataset `california`, coordinates can be chosen more than one time, and "`bad`" steps are likely to happen. On these steps, the three rules differ.