Differentially Private Coordinate Descent for Composite Empirical Risk Minimization

Paul Mangold, Aurélien Bellet, Joseph Salmon, Marc Tommasi

DIFFERENTIALLY PRIVATE COORDINATE DESCENT FOR COMPOSITE EMPIRICAL RISK MINIMIZATION

Paul Mangold
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

Aurélien Bellet
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

Joseph Salmon
IMAG, Univ Montpellier, CNRS, Montpellier, France
Institut Universitaire de France (IUF)

Marc Tommasi
Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France

ABSTRACT

Machine learning models can leak information about the data used to train them. To mitigate this issue, Differentially Private (DP) variants of optimization algorithms like Stochastic Gradient Descent (DP-SGD) have been designed to trade-off utility for privacy in Empirical Risk Minimization (ERM) problems. In this paper, we propose Differentially Private proximal Coordinate Descent (DP-CD), a new method to solve composite DP-ERM problems. We derive utility guarantees through a novel theoretical analysis of inexact coordinate descent. Our results show that, thanks to larger step sizes, DP-CD can exploit imbalance in gradient coordinates to outperform DP-SGD. We also prove new lower bounds for composite DP-ERM under coordinate-wise regularity assumptions, that are nearly matched by DP-CD. For practical implementations, we propose to clip gradients using coordinate-wise thresholds that emerge from our theory, avoiding costly hyperparameter tuning. Experiments on real and synthetic data support our results, and show that DP-CD compares favorably with DP-SGD.

1 Introduction

Machine learning fundamentally relies on the availability of data, which can be sensitive or confidential. It is now well-known that preventing learned models from leaking information about individual training points requires particular attention [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 aim to design a differentially private algorithm which approximates the solution to a composite ERM problem of the form:

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

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).

Differential privacy constraints induce a trade-off between the privacy and the utility (i.e., optimization error) of the solution of (1). This trade-off was made explicit by Bassily et al. [2014], who derived lower bounds on the achievable error given a fixed privacy budget. To solve the DP-ERM problem in practice, the most popular approaches are based on Differentially Private variants of Stochastic Gradient Descent (DP-SGD) [Bassily et al., 2014, Abadi et al., 2016, Wang et al., 2017], in which random perturbations are added to the (stochastic) gradients. Bassily et al. [2014] analyzed DP-SGD in the non-smooth DP-ERM setting, and Wang et al. [2017] then proposed an efficient DP-SVRG algorithm for composite DP-ERM. Both algorithms match known lower bounds. SGD-style algorithms perform well in a wide variety of settings, but also have some flaws: they either require small (or decreasing) step sizes or variance reduction schemes to guarantee convergence, and they can be slow when gradients' coordinates are imbalanced. These flaws propagate to the private counterparts of these algorithms. Despite a few attempts at designing other differentially

private solvers for ERM under different setups [Talwar et al., 2015, Damaskinos et al., 2021], the differentially private optimization toolbox remains limited, which undoubtedly restricts the resolution of practical problems.

In this paper, we propose and analyze a Differentially Private proximal Coordinate Descent algorithm (DP-CD), which performs updates based on perturbed coordinate-wise gradients (*i.e.,* partial derivatives). Coordinate Descent (CD) methods have encountered a large success in non-private machine learning due to their simplicity and effectiveness [Liu et al., 2009, Friedman et al., 2010, Chang et al., 2008, Sardy et al., 2000], and have seen a surge of practical and theoretical interest in the last decade [Nesterov, 2012, Wright, 2015, Shi et al., 2017, Richtárik and Takáč, 2014, Fercoq and Richtárik, 2015, Tappenden et al., 2016, Hanzely et al., 2020, Nutini et al., 2015, Karimireddy et al., 2019]. In contrast to SGD, they converge with constant step sizes that adapt to the coordinate-wise smoothness of the objective. Additionally, CD updates naturally tend to have a lower sensitivity. Operating with partial gradients thus enables our private algorithm to reduce the perturbation required to guarantee privacy without resorting to amplification by subsampling [Balle et al., 2018, Mironov et al., 2019].

We propose a novel analysis of proximal CD with perturbed gradients to derive optimal upper bounds on the privacy-utility trade-off achieved by DP-CD. We prove a recursion on distances of CD iterates to an optimal point that keeps track of coordinate-wise regularity constants in a tight manner and allows to use large, constant step sizes that yield high utility. Our results highlight the fact that DP-CD can exploit imbalanced gradient coordinates to outperform DP-SGD. They also improve upon known convergence rates for inexact CD in the non-private setting [Tappenden et al., 2016]. We assess the optimality of DP-CD by deriving lower bounds that capture coordinate-wise Lipschitz regularity measures, and show that DP-CD matches those bounds up to logarithmic factors. Our lower bounds also suggest interesting perspectives for future work on DP-CD algorithms.

Our theoretical results have important consequences for practical implementations, which heavily rely on gradient clipping to achieve good utility. In contrast to DP-SGD, DP-CD requires to set *coordinate-wise* clipping thresholds, which can lead to impractical coordinate-wise hyperparameter tuning. We instead propose a simple rule for adapting these thresholds from a single hyperparameter. We also show how the coordinate-wise smoothness constants used by DP-CD can be estimated privately. We validate our theory with numerical experiments on real and synthetic datasets. These experiments further show that even in balanced problems, DP-CD can still improve over DP-SGD, confirming the relevance of DP-CD for DP-ERM.

Our main contributions can be summarized as follows:

1. We propose the first proximal CD algorithm for composite DP-ERM, formally prove its utility, and highlight regimes where it outperforms DP-SGD.
2. We show matching lower bounds under coordinate-wise regularity assumptions.
3. We give practical guidelines to use DP-CD, and show its relevance through numerical experiments.

The rest of this paper is organized as follows. We first describe some mathematical background in Section 2. In Section 3, we present our DP-CD algorithm, show that it satisfies DP, establish utility guarantees, and compare these guarantees with those of DP-SGD. In Section 4, we derive lower bounds under coordinate-wise regularity assumptions, and show that DP-CD can match them. Section 5 discusses practical questions related to gradient clipping and the private estimation of smoothness constants. Section 6 presents our numerical experiments, comparing DP-CD and DP-SGD on LASSO and ℓ_2-regularized logistic regression problems. Finally, we review existing work in Section 7, and conclude with promising lines of future work in Section 8.

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 be crucial in our analysis, for they allow to keep track of coordinate-wise quantities. Let $\langle u, v \rangle = \sum_{j=1}^{p} u_i v_i$ be the Euclidean dot product, let $M = \operatorname{diag}(M_1, \dots, M_p)$ with $M_1, \dots, M_p > 0$, and

$$\|w\|_M = \sqrt{\langle Mw, w \rangle} \ , \qquad \|w\|_{M^{-1}} = \sqrt{\langle M^{-1}w, w \rangle} \ .$$

When M is the identity matrix I, the I-norm $\|\cdot\|_I$ is the standard ℓ_2-norm $\|\cdot\|_2$.

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 canonical basis.

Convexity: a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is convex if for all $v, w \in \mathbb{R}^p$, $f(w) \geq f(v) + \langle \nabla f(v), w - v \rangle$.

Strong convexity: a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is μ_M-strongly-convex *w.r.t.* the norm $\|\cdot\|_M$ if for all $v, w \in \mathbb{R}^p$, $f(w) \geq f(v) + \langle \nabla f(v), w - v \rangle + \frac{\mu_M}{2} \|w - v\|_M^2$. The case $M_1 = \cdots = M_p = 1$ recovers standard μ_I-strong convexity *w.r.t.* the ℓ_2-norm.

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 all $w \in \mathbb{R}^p$, $t \in \mathbb{R}$ and $j \in [p]$, $|f(w + te_j) - f(w)| \leq L_j |t|$. It is Λ-Lipschitz if for all $v, w \in \mathbb{R}^p$, $|f(v) - f(w)| \leq \Lambda \|v - w\|_2$.

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 all $v, w \in \mathbb{R}^p$, $f(w) \leq f(v) + \langle \nabla f(v), w - v \rangle + \frac{1}{2} \|w - v\|_M^2$. When $M_1 = \cdots = M_p = \beta$, f is said to be β-smooth.

The above component-wise regularity hypotheses are not restrictive: Λ-Lipschitzness implies $(\Lambda, \ldots, \Lambda)$-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. This will be crucial for our analysis and in the performance of DP-CD.

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 $\mathcal{A}$:

$$\Pr[\mathcal{A}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{A}(D') \in S] + \delta \ .$$

The value of a function $h : \mathcal{D} \to \mathbb{R}^p$ can be privately released using the Gaussian mechanism, which adds centered Gaussian noise to $h(D)$ before releasing it [Dwork and Roth, 2014]. The scale of the noise is calibrated to the sensitivity $\Delta(h) = \sup_{D \sim D'} \|h(D) - h(D')\|_2$ of h. In our setting, we will perturb coordinate-wise gradients: we denote by $\Delta(\nabla_j \ell)$ the sensitivity of the j-th coordinate of gradient of the loss function ℓ with respect to the data. When $\ell(\cdot; d)$ is L-component-Lipschitz for all $d \in \mathcal{X}$, upper bounds on these sensitivities are readily available: we have $\Delta(\nabla_j \ell) \leq 2L_j$ for any $j \in [p]$ (see Appendix A). The following quantity, relating the coordinate-wise sensitivities of gradients to coordinate-wise smoothness is central in our analysis:

$$\Delta_{M^{-1}}(\nabla \ell) = \Big(\sum_{j=1}^{p} \frac{1}{M_j} \Delta(\nabla_j \ell)^2 \Big)^{\frac{1}{2}} \leq 2 \|L\|_{M^{-1}} \ . \tag{2}$$

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].

3 Differentially Private Coordinate Descent

In this section, we introduce the Differentially Private proximal Coordinate Descent (DP-CD) algorithm to solve problem (1) under (ϵ, δ)-DP constraints. We first describe our algorithm, show how to parameterize it to satisfy the desired privacy constraint, and prove corresponding utility results. Finally, we compare these utility guarantees with DP-SGD.

3.1 Private Proximal Coordinate Descent

Let $D = \{d_1, \ldots, d_n\} \in \mathcal{X}^n$ be a dataset. We denote by $f(w) = \frac{1}{n} \sum_{i=1}^n \ell(w; d_i)$ the M-component-smooth part of (1), by $\psi(w) = \sum_{j=1}^p \psi_j(w_j)$ its separable part, and let $F(w) = f(w) + \psi(w)$. Proximal coordinate descent methods Richtárik and Takáč [2014] solve problem (1) through iterative proximal gradient steps along each coordinate of F. Formally, given $w \in \mathbb{R}^p$ and $j \in [p]$, the j-th coordinate of w is updated as follows:

$$w_j^+ = \mathrm{prox}_{\gamma_j \psi_j} \left(w_j - \gamma_j \nabla_j f(w_t) \right) \ , \tag{3}$$

where $\gamma_j > 0$ is the step size and $\mathrm{prox}_{\gamma_j \psi_j}(w) = \arg\min_{v \in \mathbb{R}^p} \left\{ \frac{1}{2} \|v - w\|_2^2 + \gamma_j \psi_j(v) \right\}$ is the proximal operator associated with ψ_j [Parikh and Boyd, 2014].

Update (3) only requires the computation of the j-th entry of the gradient. To satisfy differential privacy, we perturb this gradient entry with additive Gaussian noise of variance σ_j^2. The complete DP-CD procedure is shown in Algorithm 1.

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

Algorithm 1 Differentially Private Proximal Coordinate Descent Algorithm (DP-CD).

Input: noise scales $\sigma = (\sigma_1, \dots, \sigma_p)$ for $\sigma_1, \dots, \sigma_p > 0$; step sizes $\gamma_1, \dots, \gamma_p > 0$; initial point $\bar{w}^0 \in \mathbb{R}^p$; iteration budgets $T, K > 0$.

1: **for** $t = 0, \dots, T-1$ **do**
2: Set $\theta^0 = \bar{w}^t$
3: **for** $k = 0, \dots, K-1$ **do**
4: Pick j from $\{1, \dots, p\}$ uniformly at random
5: Draw $\eta_j \sim \mathcal{N}(0, \sigma_j^2)$
6: Set $\theta^{k+1} = \theta^k$
7: Set $\theta_j^{k+1} = \mathrm{prox}_{\gamma_j \psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j))$
8: **end for**
9: Set $\bar{w}_{t+1} = \frac{1}{K} \sum_{k=1}^{K} \theta^k$
10: **end for**
11: **return** $w_{priv} = \bar{w}_T$

At each iteration, we pick a coordinate uniformly at random and update according to (3), albeit with noise addition (see line 7). For technical reasons related to our analysis, we use a periodic averaging scheme (line 9). This scheme is similar to DP-SVRG [Johnson and Zhang, 2013], although no variance reduction is required since DP-CD computes coordinate gradients over the whole dataset.

3.2 Privacy Guarantees

For Algorithm 1 to satisfy (ϵ, δ)-DP, the noise scales $\sigma = (\sigma_1, \dots, \sigma_p)$ can be calibrated as given in Theorem 3.1.

Theorem 3.1. *Assume $\ell(\cdot; d)$ is L-component-Lipschitz $\forall d \in \mathcal{X}$. Let $\epsilon \leq 1$ and $\delta < 1/3$. If $\sigma_j^2 = \frac{12 L_j^2 T K \log(1/\delta)}{n^2 \epsilon^2}$ for all $j \in [p]$, then Algorithm 1 satisfies (ϵ, δ)-DP.*

Sketch of Proof. (Complete proof in Appendix B). We track the privacy loss using Rényi differential privacy (RDP), which gives better guarantees than (ϵ, δ)-DP for the composition of Gaussian mechanisms [Mironov, 2017]. The j-th entry of ∇f has sensitivity $\Delta(\nabla_j f) = \Delta(\nabla_j \ell)/n \leq 2L_j/n$. By the Gaussian mechanism each iteration of DP-CD is $(\alpha, \frac{2L_j^2 \alpha}{n^2 \sigma_j^2})$-RDP for all $\alpha > 1$. The composition theorem for RDP gives a global RDP guarantee for DP-CD, that we convert to (ϵ, δ)-DP using Proposition 3 of Mironov [2017]. Choosing α carefully finally proves the result. □

The dependence of the noise scales on ϵ, δ, n and TK (the number of updates) in Theorem 3.1 is standard in DP-ERM. However, the noise is calibrated to the loss function's *component*-Lipschitz constants. These can be much lower their global counterpart, the latter being used to calibrate the noise in DP-SGD algorithms. This will be crucial for DP-CD to achieve better utility than DP-SGD in some regimes. We also note that, unlike DP-SGD, DP-CD does not rely on privacy amplification by subsampling [Balle et al., 2018, Mironov et al., 2019], and thereby avoids the approximations required by these schemes to bound the privacy loss.

Remark 3.2. Theorem 3.1 gives a simple closed form for the noise scales, but in practice we can numerically compute tighter values using Rényi DP formulas directly.

3.3 Utility Guarantees

We now state our central result on the utility of DP-CD for the composite DP-ERM problem. As done in previous work, we use the asymptotic notation $\widetilde{O}$ to hide non-significant logarithmic factors. Non-asymptotic utility bounds can be found in Appendix C.

Theorem 3.3. *Let $\ell(\cdot; d)$ be a convex and L-component-Lipschitz loss function for all $d \in \mathcal{X}$, and f be convex and M-component-smooth. Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be a convex and separable function. Let $\epsilon \leq 1, \delta < 1/3$ be the privacy budget. Let w^* be a minimizer of F and $F^* = F(w^*)$. Let $w_{priv} \in \mathbb{R}^p$ be the output of Algorithm 1 with step sizes $\gamma_j = 1/M_j$, and noise scales $\sigma_1, \dots, \sigma_p$ set as in Theorem 3.1 (with T and K chosen below) to ensure (ϵ, δ)-DP. Then, the following holds:*

1. *For F convex, $K = O\left(\frac{R_M \sqrt{p} n\epsilon}{\|L\|_{M^{-1}}}\right)$ and $T = 1$, then:*

$$\mathbb{E}[F(w_{priv}) - F^*] = \widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\|L\|_{M^{-1}} R_M\right) ,$$

where $R_M = \max(\sqrt{F(w^0) - F(w^*)}, \left\|w^0 - w^*\right\|_M)$ and more simply $R_M = \left\|w^0 - w^*\right\|_M$ when $\psi = 0$.

2. *For F μ_M-strongly convex w.r.t. $\|\cdot\|_M$, $K = O\left(p/\mu_M\right)$ and $T = O\left(\log(n\epsilon\mu_M/p\,\|L\|_{M^{-1}})\right)$, then:*

$$\mathbb{E}[F(w_{priv}) - F^*] = \widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\|L\|^2_{M^{-1}}}{\mu_M}\right) .$$

Expectations are over the randomness of the algorithm.

Sketch of Proof. (Complete proof in Appendix C). Existing analyses of CD fail to track the noise tightly across coordinates when adapted to the private setting. Contrary to these classical analyses, we prove a recursion on $\mathbb{E}\left\|\theta^k - w^*\right\|^2_M$, rather than on $\mathbb{E}\big[F(\theta^k) - F(w^*)\big]$. Our key technical result is a descent lemma (Lemma C.3) allowing us to obtain

$$\mathbb{E}\big[F(\theta^{k+1}) - F^*\big] - \frac{p-1}{p}\mathbb{E}\big[F(\theta^k) - F^*\big] \leq \mathbb{E}\left\|\theta^k - w^*\right\|^2_M - \mathbb{E}\left\|\theta^{k+1} - w^*\right\|^2_M + \frac{\|\sigma\|^2_M}{p} . \tag{4}$$

The above inequality shows that coordinate-wise updates leave a fraction $\frac{p-1}{p}$ of the function "unchanged", while the remaining part decreases (up to additive noise). Importantly, all quantities are measured in M-norm. When summing (4) for $k = 0, \ldots, K-1$, its left hand side simplifies and its right hand side is simplified as a telescoping sum:

$$\frac{1}{p}\sum_{k=1}^{K}\mathbb{E}\big[F(\theta^k) - F^*\big] \leq \mathbb{E}\big[F(\bar{w}^t) - F^*\big] + \mathbb{E}\left\|\bar{w}^t - w^*\right\|^2_M + \frac{K}{p}\|\sigma\|^2_{M^{-1}} , \tag{5}$$

where $\bar{w}^t$ comes from $\theta^0 = \bar{w}^t$. As $\bar{w}^{t+1} = \sum_{k=1}^K \frac{\theta^k}{K}$ and F is convex, we have $F(\bar{w}^{t+1}) - F^* \leq \frac{1}{K}\sum_{k=1}^K F(\theta^k) - F^*$. This proves the sub-linear convergence (up to an additive noise term) of the inner loop. The result in the convex case follows directly (since $T = 1$, only one inner loop is run). For strongly convex F, it further holds that $\mathbb{E}\left\|\bar{w}^t - w^*\right\|^2_M \leq \frac{2}{\mu_M}\mathbb{E}[F(\bar{w}^t) - F(w^*)]$. Replacing in (5) with large enough K gives $\mathbb{E}\big[F(\bar{w}^{t+1}) - F^*\big] \leq \frac{1}{2}\mathbb{E}[F(\bar{w}^t) - F^*] + \|\sigma\|^2_{M^{-1}}$, and linear convergence (up to an additive noise term) follows. Finally, K and T are chosen to balance the "optimization" and the "privacy" errors. □

Remark 3.4. Our novel convergence proof of CD is also useful in the non-private setting. In particular, we improve upon known convergence rates for inexact CD methods with additive error [Tappenden et al., 2016], under the hypothesis that gradients are noisy and unbiased. In their formalism, we have $\alpha = 0$ and $\beta = \|\sigma\|^2_{M^{-1}}/p$. With our analysis, the algorithm requires $2pR^2_M/(\xi - p\beta)$ (resp. $4p/\mu_M \log((F(w^0) - F^*)/(\xi - p\beta))$) iterations to achieve expected precision $\xi > p\beta$ when F is convex (resp. μ_M-strongly-convex *w.r.t.* $\|\cdot\|_M$), improving upon Tappenden et al. [2016]'s results by a factor $\sqrt{p\beta/2R^2_M}$ (resp. $\mu_M/2$). See Appendix C.3 for details. Moreover, unlike this prior work, our analysis does not require the objective to decrease at each iteration, which is essential to guarantee DP.

Our utility guarantees stated in Theorem 3.3 directly depend on precise coordinate-wise regularity measures of the objective function. In particular, the initial distance to optimal, the strong convexity parameter and the overall sensitivity of the loss function are measured in the norms $\|\cdot\|_M$ and $\|\cdot\|_{M^{-1}}$ (*i.e.,* weighted by coordinate-wise smoothness constants or their inverse). In the remainder of this section, we thoroughly compare our utility results with existing ones for DP-SGD. We will show the optimality of our utility guarantees in Section 4.

3.4 Comparison with DP-SGD and DP-SVRG

We now compare DP-CD with DP-SGD and DP-SVRG, for which Bassily et al. [2014] and Wang et al. [2017] proved utility guarantees. In this section, we assume that the loss function ℓ satisfies the hypotheses of Theorem 3.3, and is Λ-Lipschitz. We denote by μ_I the strong convexity parameter of $\ell(\cdot, d)$ *w.r.t.* $\|\cdot\|_2$ and R_I the equivalent of R_M when M is the identity matrix I. As can be seen from Table 1, comparing DP-CD and DP-SGD boils down to comparing $\|L\|_{M^{-1}} R_M$ with ΛR_I for convex functions and $\|L\|^2_{M^{-1}}/\mu_M$ with Λ^2/μ_I for strongly-convex functions. We compare these terms in two scenarios, depending on the distribution of coordinate-wise smoothness constants. To ease the comparison, we assume that $R_M = \left\|w^0 - w^*\right\|_M$ and $R_I = \left\|w^0 - w^*\right\|_I$ (which is notably the case when $\psi = 0$), and that F has a unique minimizer w^*.

Table 1: Utility guarantees for DP-CD (proposed), DP-SGD [Bassily et al., 2014] and DP-SVRG[Wang et al., 2017] for L-component-Lipschitz, Λ-Lipschitz loss.

	Convex	Strongly-convex
DP-CD	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\left\|L\right\|_{M^{-1}} R_M\right)$	$\widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\left\|L\right\|_{M^{-1}}^2}{\mu_M}\right)$
DP-SGD DP-SVRG	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\Lambda R_I\right)$	$\widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\Lambda^2}{\mu_I}\right)$

Balanced. When the smoothness constants M are all equal, $\left\|L\right\|_{M^{-1}} R_M = \left\|L\right\|_2 R_I$ and $\left\|L\right\|_{M^{-1}}^2/\mu_M = \left\|L\right\|_2^2/\mu_I$. This boils down to comparing $\left\|L\right\|_2$ to Λ. As $\Lambda \leq \left\|L\right\|_2 \leq \sqrt{p}\Lambda$, DP-CD can be up to p times worse than DP-SGD. This can only happen when features are extremely correlated, which is generally not the case in machine learning. We show empirically in Section 6.2 that, even in balanced regimes, DP-CD can still significantly outperform DP-SGD.

Unbalanced. More favorable regimes exist when smoothness constants are imbalanced. To illustrate this, consider the case where the first coordinate of the loss function ℓ dominates others. There, $M_{\max} = M_1 \gg M_{\min} = M_j$ and $L_{\max} = L_1 \gg L_{\min} = L_j$ for all $j \neq 1$, so that L_1^2/M_1 dominates the other terms of $\left\|L\right\|_{M^{-1}}^2$. This yields $\left\|L\right\|_{M^{-1}}^2 \approx L_1^2/M_1 \approx \Lambda/M_{\max}$, and $\mu_M = \mu_I M_{\min}$. Moreover, if the first coordinate of w^* is already well estimated by w^0 (which is common for sparse models), then $R_M \approx M_{\min} R_I$. We obtain that $\left\|L\right\|_{M^{-1}} R_M \approx \sqrt{M_{\min}/M_{\max}}\Lambda R_I$ for convex losses and $\frac{\left\|L\right\|_{M^{-1}}}{\mu_M} \approx \frac{M_{\min}}{M_{\max}}\frac{\Lambda^2}{\mu_I}$ for strongly-convex ones. In both cases, DP-CD can perform arbitrarily better than DP-SGD, depending on the ratio between the smallest and largest coordinate-wise smoothness constants of the loss function. This is due to the inability of DP-SGD to adapt its step size to each coordinate. DP-CD thus converges quicker than DP-SGD on coordinates with smaller-scale gradients, requiring fewer accesses to the dataset, and in turn less noise addition. We give more details on this comparison in Appendix D, and complement it with an empirical evaluation on synthetic and real-world data in Section 6.

4 Lower Bounds

We now prove a new lower bound on the error achievable for composite DP-ERM with L-component-Lipschitz loss functions. While our proof borrows some ideas from the lower bounds known for constrained ERM with Λ-Lipschitz losses [Bassily et al., 2014], deriving our lower bounds requires to address a number of specific challenges. First, we cannot use an ℓ_2 norm constraint as in Bassily et al. [2014] in the design of the worst-case problem instances: we can only rely on *separable* regularizers. Second, imbalanced coordinate-wise Lipschitz constants prevent lower-bounding the distance between an arbitrary point and the solution. This leads us to revisit the construction of a "reidentifiable dataset" from Bun et al. [2014] so that we have L-component-Lipschitzness while the sum of each column is large enough, which is crucial in our proof. The full proof is given in Appendix E.

Theorem 4.1. *Let $n, p > 0$, $\epsilon > 0$, $\delta = o(\frac{1}{n})$, $L_1, \dots, L_p > 0$, such that for all $\mathcal{J} \subseteq [p]$ of size at least $\lceil \frac{p}{75} \rceil$, $\sum_{j\in\mathcal{J}} L_j^2 = \Omega(\left\|L\right\|_2^2)$. Let $\mathcal{X} = \prod_{j=1}^p \{\pm L_j\}$ and consider any (ϵ, δ)-differentially private algorithm that outputs w^{priv}. In each of the two following cases there exists a dataset $D \in \mathcal{X}^n$, a L-component-Lipschitz loss $\ell(\cdot, d)$ for all $d \in D$ and a regularizer ψ so that, with F the objective of* (1) *minimal at $w^* \in \mathbb{R}^p$:*

1. *If F is convex:*

$$\mathbb{E}\big[F(w^{priv}; D) - F(w^*)\big] = \Omega\left(\frac{\sqrt{p}\left\|L\right\|_2 \left\|w^*\right\|_2}{n\epsilon}\right) \quad .$$

2. *If F is μ_I-strongly-convex* w.r.t. $\left\|\cdot\right\|_2$*:*

$$\mathbb{E}\big[F(w^{priv}; D) - F(w^*)\big] = \Omega\left(\frac{p\left\|L\right\|_2^2}{\mu_I n^2\epsilon^2}\right) \quad .$$

We recover the lower bounds of Bassily et al. [2014] for Λ-Lipschitz losses as a special case of ours by setting $L_1 = \dots = L_p = \Lambda/\sqrt{p}$. In this case, the loss function used in our proof is indeed $(\sum_{j=1}^p L_j^2)^{1/2} = \Lambda$-Lipschitz. To relate these lower bounds to the performance of DP-CD, consider a suboptimal version of our algorithm where the

step sizes are set to $\gamma_1 = \cdots = \gamma_p = (\max_j M_j)^{-1}$. In this setting, results from Theorem 3.3 still hold, and match the lower bounds from Theorem 4.1 up to logarithmic factors. We leave open the question of the optimality of DP-CD under the additional hypothesis of smoothness.

We note that the assumption on the sum of the L_j's over a set of indices $\mathcal{J}$ in Theorem 4.1 can be eliminated at the cost of an additional factor of $L_{\min}/L_{\max}$ for convex losses and $(L_{\min}/L_{\max})^2$ for strongly-convex losses, making the bound looser. Although the aforementioned assumption may seem solely technical, we conjecture that better utility is possible when a few coordinate-wise Lipschitz constants dominate the others. We discuss this further in Section 8.

5 DP-CD in Practice

We now discuss practical questions related to DP-CD. First, we show how to implement coordinate-wise gradient clipping using a single hyperparameter. Second, we explain how to privately estimate the smoothness constants. Finally, we discuss the possibility of standardizing the features and how this relates to estimating smoothness constants for the important problem of fitting generalized linear models.

5.1 Coordinate-wise Gradient Clipping

To bound the sensitivity of coordinate-wise gradients, our analysis of Section 3 relies on the knowledge of Lipschitz constants for the loss function $\ell(\cdot; d)$ that must hold for all possible data points $d \in \mathcal{X}$, see inequality (2) and the discussion above it. This is classic in the analysis of DP optimization algorithms [see e.g., Bassily et al., 2014, Wang et al., 2017]. In practice however, these Lipschitz constants can be difficult to bound tightly and often give largely pessimistic estimates of sensitivities, thereby making gradients overly noisy. To overcome this problem, the common practice in concrete deployments of DP-SGD algorithms is to *clip per-sample gradients* so that their norm does not exceed a fixed threshold parameter $C > 0$ [Abadi et al., 2016]:

$$\mathrm{clip}(\nabla \ell(w), C) = \min\left(1, \frac{C}{\|\nabla \ell(w)\|_2}\right) \nabla \ell(w) \ . \tag{6}$$

This effectively ensures that the sensitivity $\Delta(\mathrm{clip}(\nabla \ell, C))$ of the clipped gradient is bounded by $2C$.

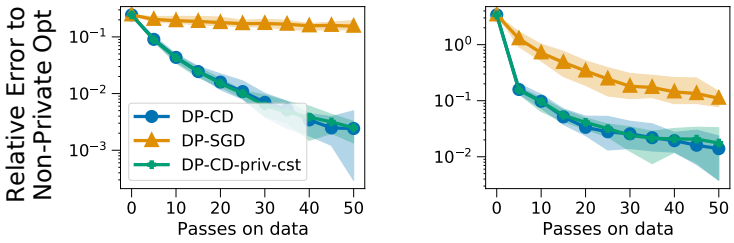
In DP-CD, gradients are released one coordinate at a time and should thus be clipped in a coordinate-wise fashion. Using the same threshold for each coordinate would ruin the ability of DP-CD to account for imbalance across gradient coordinates, whereas tuning coordinate-wise thresholds as p individual hyperparameters $\{C_j\}_{j=1}^p$ is impractical.

Instead, we leverage the results of Theorem 3.3 to adapt them from a single hyperparameter. We first remark that our utility guarantees are invariant to the scale of the matrix M. After rescaling M to $\widetilde{M} = \frac{p}{\mathrm{tr}(M)} M$ so that $\mathrm{tr}(\widetilde{M}) = \mathrm{tr}(I) = p$, as proposed by Richtárik and Takáč [2014], the key quantity $\Delta_{\widetilde{M}^{-1}}(\nabla \ell)$ as defined in (2) appears in our utility bounds instead of $\|L\|_{M^{-1}}$. This suggests to parameterize the j-th threshold as $C_j = \sqrt{M_j/\mathrm{tr}(M)}C$ for some $C > 0$, ensuring that $\Delta_{\widetilde{M}^{-1}}(\{\mathrm{clip}(\nabla_j \ell, C_j)\}_{j=1}^p) \leq 2C$. The parameter C thus controls the overall sensitivity, allowing clipped DP-CD to perform p iterations for the same privacy budget as one iteration of clipped DP-SGD.

5.2 Private Smoothness Constants

DP-CD requires the knowledge of the coordinate-wise smoothness constants $M_1, \ldots, M_p$ of f to set appropriate step sizes (see Theorem 3.3) and clipping thresholds (see above).[2] In most problems, the M_j's depend on the dataset D and must thus be estimated privately using a fraction of the overall privacy budget. Since f is an average of loss terms, its coordinate-wise smoothness constants are the average of those of $\ell(\cdot, d)$ over $d \in D$. These per-sample quantities are easy to get for typical losses (see Section 5.3 for the case of linear models). Privately estimating $M_1, \ldots, M_p$ thus reduces to a classic private mean estimation problem for which many methods exist. For instance, assuming that the practitioner knows a crude upper bound on per-sample smoothness constants, he/she can compute the smoothness constants of the $\ell(\cdot, d)$'s, clip them to the pre-defined upper bounds, and privately estimate their mean using the Laplace mechanism (see Appendix F for details). We show numerically in Section 6 that dedicating 10% of the total budget ϵ to this strategy allows DP-CD to effectively exploit the imbalance across gradients' coordinates.

[2] In fact, only $M_j / \sum_{j'} M_{j'}$ is needed, as we tune the clipping threshold and scaling factor for the step sizes; see Section 6.



(a) Electricity
Logistic, $\lambda = 1/n$.

(b) California
LASSO, $\lambda = 1.0$.

Figure 1: Relative error to non-private optimal for DP-CD (blue), DP-CD with privately estimated coordinate-wise smoothness constants (green) and DP-SGD (orange) on two *imbalanced* problems. The number of passes is tuned separately for each algorithm to achieve lowest error. We report min/mean/max values over 10 runs.

5.3 Feature Standardization

CD algorithms are very popular to solve generalized linear models [Friedman et al., 2010] and their regularized version (*e.g.,* LASSO, logistic regression). For these problems, the coordinate-wise smoothness constants are $M_j \propto \frac{1}{n} \|X_{:,j}\|_2^2$, where $X_{:,j} \in \mathbb{R}^n$ is the vector containing the value of the j-th feature. Therefore, standardizing the features to have zero mean and unit variance (a standard preprocessing step) makes coordinate-wise smoothness constants equal. However, this requires to compute the mean and variance of each feature in D, which is more costly than the smoothness constants to estimate privately.[3] Moreover, while our theory suggests that DP-CD may not be superior to DP-SGD when smoothness constants are all equal (see Section 3.4), the numerical results of Section 6 show that DP-CD often outperforms DP-SGD even when features are standardized.

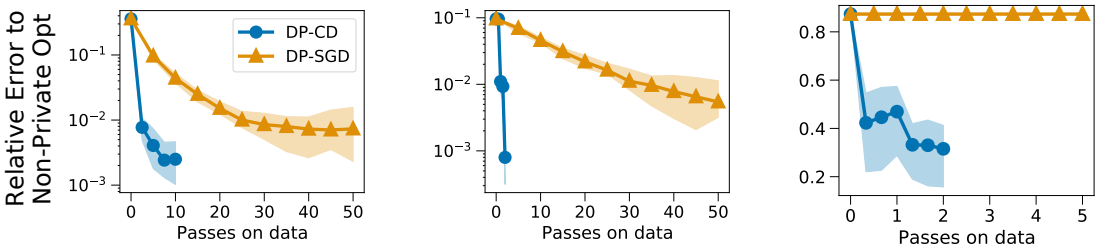
Finally, we emphasize that standardization is not always possible. This notably happens when solving the problem at hand is a subroutine of another algorithm. For instance, the Iteratively Reweighted Least Squares (IRLS) algorithm [Holland and Welsch, 1977] finds the maximum likelihood estimate of a generalized linear model by solving a sequence of linear regression problems with reweighted features, proscribing standardization. Similar situations happen when using reweighted ℓ_1 methods for non-convex sparse regression [Candès et al., 2008], relying on convex (LASSO) solvers for the inner loop. DP-CD is thus a method of choice to serve as subroutine in private versions of these algorithms.

6 Numerical Experiments

In this section, we assess the practical performance of DP-CD against (proximal) DP-SGD on LASSO[4] and ℓ_2-regularized logistic regression[5]. For LASSO, we use the California dataset [Kelley Pace and Barry, 1997], with $n = 20,640$ records and $p = 8$ features as well as a synthetic dataset (coined "Sparse LASSO") with $n = 1,000$ records and $p = 1,000$ independent features that follow a standard normal distribution. The labels are then computed as a noisy sparse linear combination of a subset of 10 active features. For logistic regression, we consider the Electricity dataset [Electricity] with $45,312$ records and 8 features. On California and Electricity, we set $\epsilon = 1$ and $\delta = 1/n^2$, which is generally seen as a rather high privacy regime. The Sparse LASSO dataset corresponds to a challenging setting for privacy ($n = p$), so we consider a low privacy regime with $\epsilon = 10$, $\delta = 1/n^2$. Privacy accounting for DP-SGD is done by numerically evaluating the Rényi DP formula given by the sampled Gaussian mechanism [Mironov et al., 2019]. Similarly for DP-CD, we do not use the closed-form formula of Theorem 3.1 but rather numerically evaluate the tighter Rényi DP formula given in Appendix B.

For DP-SGD, we use constant step sizes and standard gradient clipping. For DP-CD, we adapt the coordinate-wise clipping thresholds from one hyperparameter, as described in Section 5.1. Similarly, coordinate-wise step sizes are set to $\gamma_j = \gamma/M_j$, where γ is a hyperparameter. When the coordinate-wise smoothness constants are not all equal, we also consider DP-CD with privately computed M_j's, as described in Section 5.2. For each dataset and each algorithm, we simultaneously tune three hyperparameters: the step size, the clipping threshold and the number of passes over the dataset. After tuning these parameters, we report the relative error to the (non-private) optimal objective value.

[3] We note that the privacy cost of standardization is rarely accounted for in practical evaluations.
[4] *i.e.,* $\ell(w,(x,y)) = (w^\top x - y)^2$, $\psi(w) = \lambda \|w\|_1$.
[5] *i.e.,* $\ell(w,(x,y)) = \log(1+\exp(-yw^\top x))$, $\psi(w) = \frac{\lambda}{2}\|w\|_2^2$.



(a) Electricity
Logistic, $\lambda = 1/n$.

(b) California
LASSO, $\lambda = 0.1$.

(c) Sparse
LASSO, $\lambda = 30$.

Figure 2: Relative error to non-private optimal for DP-CD (blue) and DP-SGD (orange) on three *balanced* problems. The number of passes is tuned separately for each algorithm to achieve lowest error. We report min/mean/max values over 10 runs.

The complete tuning procedure is described in Appendix G.1, where we also give the best error for various numbers of passes for each algorithm and dataset. The code used to obtain all our results can be found in the supplementary material.

6.1 Imbalanced Datasets

In the Electricity and California datasets, features are naturally imbalanced. DP-CD can exploit this through the use of coordinate-wise smoothness constants. We also consider a variant of DP-CD (DP-CD-priv-cst) which dedicates 10% of the privacy budget ϵ to estimate these constants (see Section 5.2) from a crude upper bound on each feature (twice their maximal absolute value). It then uses the resulting private smoothness constants in step sizes and clipping thresholds. Figure 1 shows that DP-CD outperforms DP-SGD by an order of magnitude on both datasets, even when the smoothness constants are estimated privately.

6.2 Balanced Datasets

To assess the performance of DP-CD when coordinate-wise smoothness constants are balanced, we standardize the Electricity and California datasets (see Section 5.3). As standardization is done for both DP-CD and DP-SGD, we do not account for it in the privacy budget. On standardized datasets, coordinate-wise smoothness constants are all equal, removing the need of estimating them privately. We report the results in Figure 2. Although our theory suggests that DP-CD may do worse than DP-SGD in balanced regimes, we observe that it still improves over DP-SGD in practice. Similar observations hold in our challenging Sparse LASSO problem, where DP-SGD is barely able to make any progress. We believe these results are in part due to the beneficial effect of clipping in DP-CD, and the fact that DP-SGD relies on amplification by subsampling, for which privacy accounting is not perfectly tight. Additionally, CD methods are known to perform well on fitting linear models: our results show that this transfers well to private optimization.

6.3 Running Time

The results above showed that DP-CD yields better utility than DP-SGD. We also observe that DP-CD tends to reach these results in up to 10 times fewer passes on the data than DP-SGD (see Appendix G.1 for detailed results). Additionally, when accounting for running time, DP-CD significantly outperforms DP-SGD: we refer to Appendix G.2 for the counterparts of Figure 1 and 2 as a function of the running time instead of the number of passes.

7 Related Work

DP-ERM. Differentially Private Empirical Risk Minimization was first studied by Chaudhuri et al. [2011], using output perturbation (adding noise to the solution of the non-private ERM problem) and objective perturbation (adding noise to the ERM objective itself). Bassily et al. [2014] then proposed DP-SGD and proved its near-optimality. Wang et al. [2017] obtained faster convergence rates using a DP version of the SVRG algorithm [Johnson and Zhang, 2013, Xiao and Zhang, 2014]. DP-SGD has become the standard approach to DP-ERM. In our work, we show that coordinate-wise updates can have lower sensitivity than DP-SGD updates and propose a DP-CD algorithm achieving competitive results.

Private versions of Frank-Wolfe algorithms (DP-FW) were also proposed to solve *constrained* DP-ERM problems [Talwar et al., 2015, Asi et al., 2021, Bassily et al., 2021]. Although these algorithms achieve a good privacy-utility trade-off in theory, we are not aware of any empirical evaluation. DP-FW algorithms access gradients indirectly through a linear optimization oracle over a constrained set. Restricting to a constrained set is not necessary in DP-CD, allowing its use for a different family of problems.

Coordinate descent. Coordinate descent (CD) algorithms have a long history in optimization. Luo and Tseng [1992], Tseng [2001], Tseng and Yun [2009] have shown convergence results for (block) CD algorithms for nonsmooth optimization. Nesterov [2012] later proved a global non-asymptotic $1/k$ convergence rate for CD with random choice of coordinates for a convex, smooth objective. Parallel, proximal variants were developed by Richtárik and Takáč [2014], Fercoq and Richtárik [2015], while Hanzely et al. [2018] further considered non-separable non-smooth parts. Shalev-Shwartz and Zhang [2013] introduced Dual CD algorithms for smooth ERM, showing performance similar to SVRG. We refer to Wright [2015] and Shi et al. [2017] for detailed reviews on CD. Inexact CD was studied by Tappenden et al. [2016], but their analysis requires updates not to increase the objective, which is hardly compatible with DP. We obtain tighter results for inexact CD with noisy gradients (see Remark 3.4).

Private coordinate descent. Damaskinos et al. [2021] introduced a CD method to privately solve the dual problem associated with generalized linear models with ℓ_2 regularization. Dual CD is tightly related to SGD, as each coordinate in the dual is associated with one data point. The authors briefly mention the possibility of performing primal coordinate descent but discard it on account of the seemingly large sensitivity of its updates. We show that primal DP-CD is in fact quite effective, and can be used to solve more general problems than considered by Damaskinos et al. [2021]. Primal CD was successfully used by Bellet et al. [2018] to privately learn personalized models from decentralized datasets. For the smooth objective they consider, each coordinate depends only on a subset of the full dataset, which directly yields low coordinate-wise sensitivity updates. In contrast, we introduce a general algorithm for composite DP-ERM, for which a novel utility analysis was required.

8 Conclusion and Discussion

We presented the first differentially private proximal coordinate descent algorithm for composite DP-ERM. Using an original approach to analyze proximal CD with perturbed gradients, we derived optimal upper bounds on the privacy-utility trade-off achieved by DP-CD. We also prove new lower bounds under a component-Lipschitzness assumption, and showed that DP-CD matches these bounds. Our results demonstrate that DP-CD strongly outperforms DP-SGD when gradients' coordinates are imbalanced. Numerical experiments show that DP-CD also performs very well in balanced regimes. The choice of coordinate-wise clipping thresholds is crucial for DP-CD to achieve good utility in practice, and we provided a simple rule to set them.

Although DP-CD already achieves good utility when most coordinates have small sensitivity, our lower bounds suggest that even better utility could be achieved by dynamically allocating more privacy budget to coordinates with largest sensitivities. A promising direction is to design DP-CD algorithms that leverage active set methods [Yuan et al., 2010, Lewis and Wright, 2016, Nutini et al., 2017, De Santis et al., 2016, Massias et al., 2018], which could provide practical alternatives to recent DP-SGD approaches that use a subspace assumption [Zhou et al., 2021, Kairouz et al., 2021]. Finally, we believe that adaptive clipping techniques [Pichapati et al., 2019, Thakkar et al., 2021] may help to further improve the practical performance of DP-CD when coordinate-wise smoothness constants are more balanced.

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

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *ICML*, volume 139, pages 393–403, 2021.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy Amplification by Subsampling: Tight Analyses via Couplings and Divergences. In *NeurIPS*, 2018.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, Philadelphia, PA, USA, October 2014. IEEE.

Raef Bassily, Cristobal Guzman, and Anupama Nandi. Non-Euclidean Differentially Private Stochastic Convex Optimization. In *COLT*, pages 474–499. PMLR, 2021.

Aurélien Bellet, Rachid Guerraoui, Mahsa Taziki, and Marc Tommasi. Personalized and Private Peer-to-Peer Machine Learning. In *International Conference on Artificial Intelligence and Statistics*, pages 473–481. PMLR, March 2018.

Mark Bun and Thomas Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In Martin Hirt and Adam Smith, editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 635–658, Berlin, Heidelberg, 2016. Springer.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, page 10, 2014.

E. J. Candès, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *J. Fourier Anal. Applicat.*, 14(5-6):877–905, 2008.

Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. Coordinate Descent Method for Large-scale L2-loss Linear Support Vector Machines. *J. Mach. Learn. Res.*, 9:1369–1398, June 2008.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially Private Empirical Risk Minimization. *J. Mach. Learn. Res.*, 12(29):1069–1109, 2011.

Georgios Damaskinos, Celestine Mendler-Dünner, Rachid Guerraoui, Nikolaos Papandreou, and Thomas Parnell. Differentially private stochastic coordinate descent. In *AAAI*, pages 7176–7184. AAAI Press, 2021.

Marianna De Santis, Stefano Lucidi, and Francesco Rinaldi. A fast active set block coordinate descent algorithm for ℓ_1-regularized least squares. *SIAM J. Optim.*, 26(1):781–809, January 2016.

Cynthia Dwork. Differential Privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 1–12, Berlin, Heidelberg, 2006. Springer.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284, Berlin, Heidelberg, 2006. Springer.

Electricity. Electricity Dataset. URL `https://www.openml.org/d/151`.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel and proximal coordinate descent. *SIAM J. Optim.*, 25(3): 1997–2013, 2015.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660.

Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, NIPS'18, pages 2086–2097, Red Hook, NY, USA, December 2018. Curran Associates Inc.

Filip Hanzely, Dmitry Kovalev, and Peter Richtárik. Variance reduced coordinate descent with acceleration: New method with a surprising application to finite-sum problems. In *ICML*, volume 119, pages 4039–4048. PMLR, 2020.

Paul W. Holland and Roy E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827, January 1977.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Peter Kairouz, Mónica Ribero Diaz, Keith Rush, and Abhradeep Thakurta. (Nearly) Dimension Independent Private ERM with AdaGrad Rates via Publicly Estimated Subspaces. In Mikhail Belkin and Samory Kpotufe, editors, *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 2717–2746. PMLR, 2021.

Sai Praneeth Karimireddy, Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Efficient Greedy Coordinate Descent for Composite Problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2887–2896. PMLR, April 2019.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33, May 1997.

A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1): 501–546, July 2016.

Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*, pages 649–656, New York, NY, USA, June 2009. Association for Computing Machinery.

Zlhi-Quau Luo and Paul Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72(1):7–35, 1992.

M. Massias, A. Gramfort, and J. Salmon. Celer: a Fast Solver for the Lasso with Dual Extrapolation. In *ICML*, volume 80, pages 3315–3324, 2018.

Ilya Mironov. Renyi Differential Privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, August 2017. arXiv: 1702.07476.

Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv:1908.10530 [cs, stat]*, August 2019.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.*, 22(2): 341–362, 2012.

Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection. In *ICML*, pages 1632–1641. PMLR, June 2015.

Julie Nutini, Issam Laradji, and Mark Schmidt. Let's Make Block Coordinate Descent Go Fast: Faster Greedy Rules, Message-Passing, Active-Set Complexity, and Superlinear Convergence. *arXiv:1712.08859 [math]*, December 2017.

Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, January 2014.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. AdaCliP: Adaptive Clipping for Private SGD. *arXiv:1908.07643 [cs, stat]*, October 2019.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, April 2014.

Sylvain Sardy, Andrew G. Bruce, and Paul Tseng. Block Coordinate Relaxation Methods for Nonparametric Wavelet Denoising. *Journal of Computational and Graphical Statistics*, 9(2):361–379, June 2000.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.*, 14(1):567–599, February 2013.

Hao-Jun Michael Shi, Shenyinying Tu, Yangyang Xu, and Wotao Yin. A Primer on Coordinate Descent Algorithms. *arXiv:1610.00040 [math, stat]*, January 2017.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, May 2017.

Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly Optimal Private LASSO. *Advances in Neural Information Processing Systems*, 28, 2015.

Rachael Tappenden, Peter Richtárik, and Jacek Gondzio. Inexact Coordinate Descent: Complexity and Preconditioning. *J. Optim. Theory Appl.*, 170(1):144–176, July 2016.

Om Thakkar, Galen Andrew, and H. Brendan McMahan. Differentially Private Learning with Adaptive Clipping. In *Advances in Neural Information Processing Systems*, 2021.

P. Tseng and S. Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *J. Optim. Theory Appl.*, 140(3):513, 2009.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, June 2015.

Lin Xiao and Tong Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. *SIAM J. Optim.*, 24(4):2057–2075, January 2014.

Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification. *J. Mach. Learn. Res.*, 11:3183–3234, December 2010.

Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *ICLR*, 2021.

A Lemmas on Sensitivity

In this section, we let $\mathcal{X}$ be the universe where the data is drawn from. To upper bound the sensitivities of a function's gradient, we start by recalling in Lemma A.1 that (coordinate) gradients are bounded by (coordinate-wise-)Lipschitz constants. We then link this upper bound with gradients' sensitivities in Lemma A.2.

Lemma A.1. *Let $\ell : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be convex and differentiable in its first argument, $\Lambda > 0$ and $L_1, \dots, L_p > 0$.*

1. *If $\ell(\cdot; d)$ is Λ-Lipschitz for all $d \in \mathcal{X}$, then $\|\nabla \ell(w; d)\|_2 \leq \Lambda$ for all $w \in \mathbb{R}^p$ and $d \in \mathcal{X}$.*
2. *If $\ell(\cdot; d)$ is L-component-Lipschitz for all $d \in \mathcal{X}$, then $|\nabla_j \ell(w; d)| \leq L_j$ for all $w \in \mathbb{R}^p$, $d \in \mathcal{X}$ and $j \in [p]$.*

Proof. Let $d \in \mathcal{X}$. We start by proving the first statement. First, if $\nabla \ell(w; d) = 0$, $\|\nabla \ell(w; d)\|_2 = 0 \leq \Lambda$ and the result holds. Second, we focus on the case where $\nabla \ell(w; d) \neq 0$. The convexity of ℓ gives, for $w \in \mathbb{R}^p$, $d \in \mathcal{X}$:

$$\ell(w + \nabla \ell(w; d); d) \geq \ell(w; d) + \langle \nabla \ell(w; d), \nabla \ell(w; d) \rangle = \ell(w; d) + \|\nabla \ell(w; d)\|_2^2 , \tag{7}$$

then, reorganizing the terms and using Λ-Lipschitzness of ℓ yields

$$\|\nabla \ell(w; d)\|_2^2 \leq \ell(w + \nabla \ell(w; d); d) - \ell(w; d) \leq |\ell(w + \nabla \ell(w; d); d) - \ell(w; d)| \leq \Lambda \|\nabla \ell(w; d)\|_2 , \tag{8}$$

and the result follows after dividing by $\|\nabla \ell(w; d)\|_2$. To prove the second statement, we set $j \in [p]$, and $w \in \mathbb{R}^p$, and remark that if $\nabla_j \ell(w; d) = 0$, then $|\nabla_j \ell(w; d)| \leq L_j$. When $\nabla_j \ell(w; d) \neq 0$, the convexity of ℓ yields

$$\ell(w + \nabla_j \ell(w; d) e_j; d) \geq \ell(w; d) + \langle \nabla \ell(w; d), \nabla_j \ell(w; d) e_j \rangle = \ell(w; d) + \nabla_j \ell(w; d)^2 . \tag{9}$$

Reorganizing the terms and using L-component-Lipschitzness of ℓ gives

$$\nabla_j \ell(w; d)^2 \leq \ell(w + \nabla_j \ell(w; d) e_j; d) - \ell(w; d) \leq |\ell(w + \nabla_j \ell(w; d) e_j; d) - \ell(w; d)| \leq L_j |\nabla_j \ell(w; d)| , \tag{10}$$

and we get the result after dividing by $|\nabla_j \ell(w; d)|$. □

Lemma A.2. *Let $\ell : \mathbb{R}^p \times \mathcal{X} \to \mathbb{R}$ be convex and differentiable in its 1st argument, $\Lambda > 0$ and $L_1, \dots, L_p > 0$.*

1. *If $\ell(\cdot; d)$ is Λ-Lipschitz for all $d \in \mathcal{X}$, then $\Delta(\nabla \ell) \leq 2\Lambda$.*
2. *If $\ell(\cdot; d)$ is L-component-Lipschitz for all $d \in \mathcal{X}$, then $\Delta(\nabla_j \ell) \leq L_j$ for all $j \in [p]$.*

Proof. We start by proving the first statement. Let $w, w' \in \mathbb{R}^p$, $d, d' \in \mathcal{X}$. From the triangle inequality and Lemma A.1, we get the following upper bounds:

$$\|\nabla \ell(w; d) - \nabla \ell(w'; d')\|_2 \leq |\nabla \ell(w; d)| + |\nabla \ell(w'; d')| \leq 2\Lambda, \tag{11}$$

which is the claim of the first statement. To prove the second statement, we proceed similarly: the triangle inequality and Lemma A.1 give the following upper bounds:

$$|\nabla_j \ell(w; d) - \nabla_j \ell(w'; d')| \leq |\nabla_j \ell(w; d)| + |\nabla_j \ell(w'; d')| \leq 2L_j , \tag{12}$$

which is the desired result. □

We obtain the inequality (2) stated in Section 2 as a corollary.

Corollary A.3. *Let $L_1, \dots, L_p > 0$. Let $\ell(\cdot; d) : \mathbb{R}^p \to \mathbb{R}$ be a convex, L-component-Lipschitz function for all $d \in \mathcal{X}$. Then*

$$\Delta_{M^{-1}}(\nabla \ell) = \Big(\sum_{j=1}^{p} \frac{1}{M_j} \Delta(\nabla_j \ell)^2 \Big)^{\frac{1}{2}} \leq \Big(\sum_{j=1}^{p} \frac{4}{M_j} L_j^2 \Big)^{\frac{1}{2}} = 2 \|L\|_{M^{-1}} . \tag{13}$$

B Proof of Theorem 3.1

To track the privacy loss of an adaptive composition of K Gaussian mechanisms, we use Rényi Differential Privacy [Mironov, 2017, RDP]. We note that similar results are obtained with zero Concentrated Differential Privacy [Bun and Steinke, 2016]. This flavor of differential privacy, gives tighter privacy guarantees in that setting, as it reduces the noise variance by a multiplicative factor of $\log(K/\delta)$ in comparison to the usual advanced composition theorem of differential privacy [Dwork et al., 2006]. Importantly, RDP can be translated back to differential privacy.

In this section, we recall the definition and main properties of zCDP. We denote by $\mathcal{D}$ the set of all datasets over a universe $\mathcal{X}$ and by $\mathcal{F}$ the set of possible outcomes of the randomized algorithms we consider.

B.1 Rényi Differential Privacy

We will use the Rényi divergence (Definition B.1), which gives a distribution-oriented vision of privacy.

Definition B.1 (Rényi divergence, van Erven and Harremoës 2014)**.** For two random variables Y and Z with values in the same domain $\mathcal{C}$, the Rényi divergence is, for $\alpha > 1$,

$$D_\alpha(Y||Z) = \frac{1}{\alpha - 1} \log \int_{\mathcal{C}} \Pr[Y = z]^\alpha \Pr[Z = z]^{1-\alpha} \, dz. \tag{14}$$

We now define RDP in Definition B.2. RDP provides a strong privacy guarantee that can be converted to classical differential privacy (Lemma B.3 and Corollary B.8).

Definition B.2 (Rényi Differential Privacy, Mironov 2017)**.** A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is (α, ϵ)-Rényi-differentially private (RDP) if, for all all datasets $D, D' \in \mathcal{D}$ differing on at most one element,

$$D_\alpha(\mathcal{A}(D)||\mathcal{A}(D')) \leq \epsilon. \tag{15}$$

Lemma B.3 (Mironov 2017, Proposition 3)**.** *If a randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is (α, ϵ)-RDP, then it is $(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-differentially private for all $0 < \delta < 1$.*

Remark B.4. The above (α, ϵ)-RDP guarantees hold for multiple values of α, ϵ. As such, $\epsilon = \epsilon(\alpha)$ can be seen as a function of α, and Lemma B.3 ensures that the algorithm is (ϵ', δ)-DP for

$$\epsilon' = \min_{\alpha > 1} \left\{ \epsilon(\alpha) + \frac{\log(1/\delta)}{\alpha - 1} \right\}. \tag{16}$$

We can now restate in Theorem B.5 the composition theorem of RDP, which is key in designing private iterative algorithms.

Theorem B.5 (Mironov 2017, Proposition 1)**.** *Let $\mathcal{A}_1, \ldots, \mathcal{A}_K : \mathcal{D} \to \mathcal{F}$ be $K > 0$ randomized algorithms, such that for $1 \leq k \leq K$, $\mathcal{A}_k$ is $(\alpha, \epsilon_k(\alpha))$-RDP, where these algorithms can be chosen adaptively (i.e., $\mathcal{A}_k$ can use to the output of $\mathcal{A}_{k'}$ for all $k' < k$). Let $\mathcal{A} : \mathcal{D} \to \mathcal{F}^K$ such that for $D \in \mathcal{D}$, $\mathcal{A}(D) = (\mathcal{A}_1(D), \ldots, \mathcal{A}_K(D))$. Then $\mathcal{A}$ is $\left(\alpha, \sum_{k=1}^K \epsilon_k(\alpha)\right)$-RDP.*

Finally, we define the Gaussian mechanism (Definition B.6), as used in Algorithm 1, and restate in Lemma B.7 the privacy guarantees that it satisfies in terms of RDP.

Definition B.6 (Gaussian mechanism)**.** Let $f : \mathcal{D} \to \mathbb{R}^p$, $\sigma > 0$, and $D \in \mathcal{D}$. The Gaussian mechanism for answering the query f is defined as:

$$\mathcal{M}_f^{Gauss}(D; \sigma) = f(D) + \mathcal{N}\left(0, \sigma^2 I_p\right). \tag{17}$$

Lemma B.7 (Mironov 2017, Corollary 3)**.** *The Gaussian mechanism with noise σ^2 is $(\alpha, \frac{\Delta(f)^2 \alpha}{2\sigma^2})$-RDP, where $\Delta(f) = \sup_{D,D'} \|f(D) - f(D')\|_2$ (for neighboring D, D') is the sensitivity of f.*

Proof. The function $h = \frac{f}{\Delta(f)}$ has sensitivity 1, thus for any $s > 0$, the Gaussian mechanism $\mathcal{M}_h^{Gauss}(\cdot; s)$ is $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP [Mironov, 2017, Corollary 1]. As $f = \Delta(f) \times h$, we have $\mathcal{M}_f^{Gauss}(\cdot; \sigma) = \Delta(f) \times \mathcal{M}_h^{Gauss}(\cdot; \frac{\sigma}{\Delta(f)})$. This mechanism is thus $(\alpha, \frac{\Delta(f)^2 \alpha}{2\sigma^2})$-RDP □

Corollary B.8. *Let $0 < \epsilon \leq 1, 0 < \delta < \frac{1}{3}$. If a randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is $(\alpha, \frac{\gamma\alpha}{2\sigma^2})$-RDP with $\gamma > 0$ and $\sigma = \frac{\sqrt{3\gamma \log(1/\delta)}}{\epsilon}$ for all $\alpha > 1$, it is also (ϵ, δ)-DP.*

Proof. From Remark B.4 it holds that $\mathcal{A}$ is (ϵ', δ)-DP with $\epsilon' = \min_{\alpha > 1} \left\{ \frac{\gamma\alpha}{2\sigma^2} + \frac{\log(1/\delta)}{\alpha - 1} \right\}$. This minimum is attained when the derivative of the objective is zero, which is the case when $\frac{\gamma}{2\sigma^2} = \frac{\log(1/\delta)}{(\alpha-1)^2}$, resulting in $\alpha = 1 + \sqrt{\frac{2\log(1/\delta)\sigma^2}{\gamma}}$. $\mathcal{A}$ is thus (ϵ', δ)-DP with

$$\epsilon' = \frac{\gamma}{2\sigma^2} + \frac{\sqrt{\gamma \log(1/\delta)}}{\sqrt{2}\sigma} + \frac{\sqrt{\gamma \log(1/\delta)}}{\sqrt{2}\sigma} = \frac{\gamma}{2\sigma^2} + \frac{\sqrt{2\gamma \log(1/\delta)}}{\sigma}. \tag{18}$$

Choosing $\sigma = \frac{\sqrt{3\gamma \log(1/\delta)}}{\epsilon}$ now gives

$$\epsilon' = \frac{\epsilon^2}{6\log(1/\delta)} + \sqrt{2/3}\epsilon \leq (1/6 + \sqrt{2/3})\epsilon \leq \epsilon, \tag{19}$$

where the first inequality comes from $\epsilon \leq 1$, thus $\epsilon^2 \leq \epsilon$ and $\delta < 1/3$ thus $\frac{1}{\log(1/\delta)} \leq 1$. The second inequality follows from $1/6 + \sqrt{2/3} \approx 0.983 < 1$. □

B.2 Proof of Theorem 3.1

We are now ready to prove Theorem 3.1. From the privacy perspective, Algorithm 1 adaptively releases and post-processes a series of gradient coordinates protected by the Gaussian mechanism. We thus start by proving Lemma B.9, which gives an (ϵ, δ)-differential privacy guarantee for the adaptive composition of K Gaussian mechanisms.

Lemma B.9. *Let $0 < \epsilon \leq 1$, $\delta < 1/3$, $K > 0$, $p > 0$, and $\{f_k : \mathbb{R}^p \to \mathbb{R}\}_{k=1}^{k=K}$ a family of K functions. The adaptive composition of K Gaussian mechanisms, with the k-th mechanism releasing f_k with noise scale $\sigma_k = \frac{\Delta(f_k)\sqrt{3K\log(1/\delta)}}{\epsilon}$ is (ϵ, δ)-differentially private.*

Proof. Let $\sigma > 0$. Lemma B.7 guarantees that the k-th Gaussian mechanism with noise scale $\sigma_k = \Delta(f_k)\sigma > 0$ is $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP. Then, the composition of these K mechanisms is, according to Theorem B.5, $(\alpha, \frac{k\alpha}{2\sigma^2})$-RDP. This can be converted to (ϵ, δ)-DP via Corollary B.8 with $\gamma = K$, which gives $\sigma_k = \frac{\Delta(f_k)\sqrt{3k\log(1/\delta)}}{\epsilon}$ for $k \in [K]$. □

We now restate Theorem 3.1 and prove it.

Theorem 3.1. *Assume $\ell(\cdot; d)$ is L-component-Lipschitz $\forall d \in \mathcal{X}$. Let $\epsilon < 1$ and $\delta < 1/3$. If $\sigma_j^2 = \frac{12L_j^2 TK\log(1/\delta)}{n^2\epsilon^2}$ for all $j \in [p]$, then Algorithm 1 satisfies (ϵ, δ)-DP.*

Proof. For $j \in [1, p]$, $\nabla_j f$ in Algorithm 1 is released using the Gaussian mechanism with noise variance σ_j^2. The sensitivity of $\nabla_j f$ is $\Delta(\nabla_j f) = \frac{\Delta(\nabla_j \ell)}{n} \leq \frac{2L_j}{n}$. Note that TK gradients are released, and

$$\sigma_j^2 = \frac{12L_j^2 TK\log(1/\delta)}{n^2\epsilon^2} \text{ for } j \in [1, p],$$

thus by Lemma B.9 and the post-processing property of DP, Algorithm 1 is (ϵ, δ)-differentially private. □

C Proof of Utility (Theorem 3.3)

C.1 Problem Statement

Let $D \in \mathcal{X}^n$ be a dataset of n elements drawn from a universe $\mathcal{X}$. Recall that we consider the following composite empirical risk minimization problem:

$$w^* \in \underset{w \in \mathbb{R}^p}{\arg\min} \left\{ F(w; D) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(w; d_i)}_{=: f(w; D)} + \psi(w) \right\}, \tag{20}$$

where $\ell(\cdot, d)$ is convex, L-component-Lipschitz, and M-component-smooth for all $d \in \mathcal{X}$, and $\psi(w) = \sum_{j=1}^{p} \psi_j(w_j)$ is convex and separable. We denote by F the complete objective function, and by f its smooth part. For readability, we omit the dependence on their second argument (*i.e.,* the data) in the rest of this section.

C.2 Proof of Theorem 3.3

In this section, we prove our central theorem that guarantees the utility of the DP-CD algorithm. To this end, we start by proving a lemma that upper bounds the expected value of $F(\theta^{k+1})$ in Algorithm 1. Using this lemma, we prove sub-linear convergence for the inner loop of DP-CD. This gives the sub-linear convergence of our algorithm for convex losses. Under the additional hypothesis that F is strongly convex, we show that iterates of the outer loop of DP-CD converge linearly towards the (unique) minimum of F.

We recall that in Algorithm 1, iterates of the inner loop are denoted by $\theta_1, \dots, \theta_K$, and those of the outer loop by $\bar{w}_1, \dots, \bar{w}_T$, with $\bar{w}_t = \frac{1}{K}\sum_{k=1}^{K} \theta^k$ for $t > 0$. Algorithm 1 is randomized in two ways: when choosing the coordinate to update and when drawing noise. For convenience, we denote by $\mathbb{E}_j[\cdot]$ the expectation *w.r.t.* the choice of coordinate, by $\mathbb{E}_\eta[\cdot]$ the one *w.r.t.* the noise, and by $\mathbb{E}_{j,\eta}[\cdot]$ the expectation *w.r.t.* both. When no subscript is used, the expectation is taken over all random variables. We will also use the notation $\mathbb{E}_{j,\eta}[\cdot|\theta_k]$ for the conditional expectation of a random variable, given a realization of θ_k.

C.2.1 Descent Lemma

We begin by proving Lemma C.1, which decomposes the change of a function F when updating its argument $\theta \in \mathbb{R}^p$, in relation to a vector $w \in \mathbb{R}^p$, into two parts: one that remains fixed, corresponding to the unchanged entries of θ, and a second part corresponding to the objective decrease due to the update. At this point, the vector w is arbitrary, but we will later choose w to be a minimizer of F, that is a solution to (20).

Lemma C.1. *Let ℓ, f, ψ, and F be defined as in Section C.1. Take a random variable $\theta \in \mathbb{R}^p$ and two arbitrary vectors $w, g \in \mathbb{R}^p$. Let a random variable j, taking its values uniformly randomly in $[p]$, Choose $\gamma_1, \dots, \gamma_p > 0$ and $\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_p)$. It holds that*

$$\begin{aligned}
&\mathbb{E}_j[F(\theta - \gamma_j g_j e_j) - F(w)|\theta] - \frac{p-1}{p}(F(\theta) - F(w)) \\
&\quad \leq \frac{1}{p}\left(f(\theta) - f(w) + \langle \nabla f(\theta), -\Gamma g\rangle + \frac{1}{2}\|\Gamma g\|_M^2 + \psi(\theta - \Gamma g) - \psi(w)\right). \qquad (21)
\end{aligned}$$

Remark C.2. To avoid notational clutter, we will write $\gamma_j g_j$ instead of $\gamma_j g_j e_j$ throughout this section.

Proof. We start the proof by finding an upper bound on $\mathbb{E}_j[F(\theta - \gamma_j g_j e_j) - F(w)|\theta]$, using the *$M$-component-smoothness* of f:

$$\begin{aligned}
\mathbb{E}_j[F(\theta - \gamma_j g_j e_j) - F(w)|\theta] &= \sum_{j=1}^{p} \frac{1}{p}(F(\theta - \gamma_j g_j) - F(w)) && (22)\\
&\overset{F=f+\psi}{=} \frac{1}{p}\sum_{j=1}^{p} f(\theta - \gamma_j g_j) - f(w) + \psi(\theta - \gamma_j g_j) - \psi(w) && (23)\\
&\overset{f \text{ smooth}}{\leq} \frac{1}{p}\sum_{j=1}^{p}\left(f(\theta) + \langle \nabla f(\theta), -\gamma_j g_j\rangle + \frac{1}{2}\|\gamma_j g_j\|_M^2 - f(w) + \psi(\theta - \gamma_j g_j) - \psi(w)\right) && (24)\\
&= f(\theta) - f(w) + \frac{1}{p}\sum_{j=1}^{p}\left(\langle \nabla f(\theta), -\gamma_j g_j\rangle + \frac{1}{2}\|\gamma_j g_j\|_M^2 + (\psi(\theta - \gamma_j g_j) - \psi(w))\right) && (25)\\
&= f(\theta) - f(w) + \frac{1}{p}\langle \nabla f(\theta), -\Gamma g\rangle + \frac{1}{2p}\|\Gamma g\|_M^2 + \frac{1}{p}\sum_{j=1}^{p}(\psi(\theta - \gamma_j g_j) - \psi(w)). && (26)
\end{aligned}$$

The regularization terms can now be reorganized using the separability of ψ, as done by Richtárik and Takáč [2014]. Indeed, we notice that

$$\begin{aligned}
\sum_{j=1}^{p}(\psi(\theta - \gamma_j g_j) - \psi(w)) &= \sum_{j=1}^{p}\left(\psi_j(\theta_j - \gamma_j g_j) - \psi_j(w_j) + \sum_{j' \neq j}\psi_{j'}(\theta_{j'}) - \psi(w_{j'})\right) && (27)\\
&= \psi(\theta - \Gamma g) - \psi(w) + (p-1)(\psi(\theta) - \psi(w)). && (28)
\end{aligned}$$

Plugging (28) in (26) results in the following:

$$\mathbb{E}_j[F(\theta-\gamma_j g_j e_j)-F(w)|\theta] \le f(\theta)-f(w)+\frac{1}{p}\langle\nabla f(\theta),-\Gamma g\rangle+\frac{1}{2p}\|\Gamma g\|_M^2 + \frac{1}{p}(\psi(\theta-\Gamma g)-\psi(w))+\frac{p-1}{p}(\psi(\theta)-\psi(w)) \tag{29}$$

$$= \frac{1}{p}\left(f(\theta)-f(w)+\langle\nabla f(\theta),-\Gamma g\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\psi(\theta-\Gamma g)-\psi(w)\right) + \frac{p-1}{p}(f(\theta)+\psi(\theta)-f(w)-\psi(w)) \tag{30}$$

which gives the lemma since $F=f+\psi$. □

To exploit this result, we need to upper bound the right hand side of (21) for the realizations of θ^k in Algorithm 1. This is where our proof differs from classical convergence proofs for coordinate descent methods. Namely, we rewrite the right hand side of (21) so as to obtain telescopic terms plus a bias term resulting from the addition of noise, as shown in Lemma C.3.

Lemma C.3. *Let ℓ, f, ψ, and F defined as in Section C.1. For $k>0$, let θ^k and θ^{k+1} be two consecutive iterates of the inner loop of Algorithm 1, $\gamma_1=\frac{1}{M_1},\dots,\gamma_p=\frac{1}{M_p}>0$ the coordinate-wise step sizes (where M_j are the coordinate-wise smoothness constants of f), and $g_j=\frac{1}{\gamma_j}(\theta_j^{k+1}-\theta_j^k)$. Let $w\in\mathbb{R}^p$ an arbitrary vector and $\sigma_1,\dots,\sigma_p>0$ the coordinate-wise noise scales given as input to Algorithm 1. It holds that*

$$\mathbb{E}_{j,\eta}\left[F(\theta^{k+1})-F(w)\middle|\theta^k\right]-\frac{p-1}{p}(F(\theta^k)-F(w)) \le \tfrac{1}{2}\left\|\theta^k-w\right\|_{\Gamma^{-1}}^2-\tfrac{1}{2}\mathbb{E}_{j,\eta}\left[\left\|\theta^{k+1}-w\right\|_{\Gamma^{-1}}^2\middle|\theta^k\right]+\tfrac{1}{p}\|\sigma\|_\Gamma^2, \tag{31}$$

where $\|\sigma\|_\Gamma^2=\sum_{j=1}^p\gamma_j\sigma_j^2$ and the expectations are taken over the random choice of j and η, conditioned upon the realization of θ^k.

Proof. We define g the vector $(g_1,\dots,g_p)\in\mathbb{R}^p$ with $g_j=\frac{1}{\gamma_j}(\theta_j^{k+1}-\theta_j^k)$ when coordinate j is chosen in Algorithm 1. We also denote by $\Gamma=\operatorname{diag}(\gamma_1,\dots,\gamma_p)$ the diagonal matrix having the step sizes as its coefficients.

From Lemma C.1 with $\theta=\theta^k$, $w=w$ and $g=g$ as defined above we obtain

$$\mathbb{E}_j\left[F(\theta^k-\gamma_j g_j e_j)-F(w)\middle|\theta^k\right]-\frac{p-1}{p}(F(\theta^k)-F(w)) \le \frac{1}{p}\left(f(\theta^k)-f(w)+\left\langle\nabla f(\theta^k),-\Gamma g\right\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\psi(\theta^k-\Gamma g)-\psi(w)\right). \tag{32}$$

We can upper bound the right hand term of (32) using the convexity of f and ψ:

$$f(\theta^k)-f(w)+\left\langle\nabla f(\theta^k),-\Gamma g\right\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\psi(\theta^k-\Gamma g)-\psi(w) \tag{33}$$

$$\le \left\langle\nabla f(\theta^k),\theta^k-w\right\rangle+\left\langle\nabla f(\theta^k),-\Gamma g\right\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\left\langle\partial\psi(\theta^k-\Gamma g),\theta^k-\Gamma g-w\right\rangle \tag{34}$$

$$= \left\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g),\theta^k-\Gamma g-w\right\rangle+\frac{1}{2}\|\Gamma g\|_M^2, \tag{35}$$

where we use the slight abuse of notation $\partial\psi(\theta^k-\Gamma g)$ to denote any vector in the subdifferential of ψ at the point $\theta^k-\Gamma g$. We now rewrite the dot product:

$$\left\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g),\theta^k-\Gamma g-w\right\rangle+\frac{1}{2}\|\Gamma g\|_M^2 \tag{36}$$

$$= \left\langle g,\theta^k-\Gamma g-w\right\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\left\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g)-g,\theta^k-\Gamma g-w\right\rangle \tag{37}$$

$$= \underbrace{\left\langle g,\theta^k-w\right\rangle-\|g\|_\Gamma^2+\frac{1}{2}\|g\|_{\Gamma^2 M}^2}_{\text{“descent” term}}+\underbrace{\left\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g)-g,\theta^k-\Gamma g-w\right\rangle}_{\text{“noise” term}}, \tag{38}$$

where the second equality follows from $\langle g, -\Gamma g\rangle = -\|g\|_\Gamma^2$ and $\|\Gamma g\|_M^2 = \|g\|_{\Gamma^2 M}^2$. We split (38) into two terms: a "descent" term and a "noise" term.

Rewriting the "descent" term. We first focus on the "descent" term. As $\gamma_j = \frac{1}{M_j}$ for all $j \in [p]$, it holds that $\gamma_j^2 M_j = \gamma_j$ which gives $-\|g\|_\Gamma^2 + \frac{1}{2}\|g\|_{\Gamma^2 M}^2 = -\|g\|_\Gamma^2 + \frac{1}{2}\|g\|_\Gamma^2 = -\frac{1}{2}\|g\|_\Gamma^2$. We can now rewrite the "descent" term as a difference of two norms, materializing the distance to w, weighted by the inverse of the step sizes Γ^{-1}:

$$\text{"descent" term} = \left\langle g, \theta^k - w\right\rangle - \frac{1}{2}\|g\|_\Gamma^2 \tag{39}$$

$$= \left\langle \Gamma g, \theta^k - w\right\rangle_{\Gamma^{-1}} - \frac{1}{2}\|\Gamma g\|_{\Gamma^{-1}}^2 \tag{40}$$

$$= \frac{1}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \frac{1}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 + \left\langle \Gamma g, \theta^k - w\right\rangle_{\Gamma^{-1}} - \frac{1}{2}\|\Gamma g\|_{\Gamma^{-1}}^2 \tag{41}$$

$$= \frac{1}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \frac{1}{2}\left\|\theta^k - \Gamma g - w\right\|_{\Gamma^{-1}}^2, \tag{42}$$

where we factorized the norm to obtain the last inequality. We can rewrite (42) as an expectation over the random choice of the coordinate j (drawn uniformly in $[p]$), given the realizations of θ^k and of the noise η (which determines g):

$$\frac{1}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \frac{1}{2}\left\|\theta^k - \Gamma g - w\right\|_{\Gamma^{-1}}^2 = \frac{p}{2} \times \left(\frac{1}{p}\sum_{j=1}^p \gamma_j^{-1}\left|\theta_j^k - w_j\right|^2 - \gamma_j^{-1}\left|\theta_j^k - \gamma_j g_j - w_j\right|^2\right) \tag{43}$$

$$= \frac{p}{2} \times \mathbb{E}_j\Big[\gamma_j^{-1}\left|\theta_j^k - w_j\right|^2 - \gamma_j^{-1}\left|\theta_j^k - \gamma_j g_j - w_j\right|^2 \Big| \theta^k, \eta\Big]. \tag{44}$$

Finally, we remark that $\gamma_j^{-1}\left|\theta_j^k - w_j\right|^2 - \gamma_j^{-1}\left|\theta_j^k - \gamma_j g_j - w_j\right|^2 = \left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \left\|\theta^k - \gamma_j g_j - w\right\|_{\Gamma^{-1}}^2$, as only one coordinate changes between the two vectors, and the squared norm $\|\cdot\|_{\Gamma^{-1}}^2$ is separable. We thus obtain

$$\text{"descent" term} = \mathbb{E}_j\Big[\frac{p}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \frac{p}{2}\left\|\theta^k - \gamma_j g_j - w\right\|_{\Gamma^{-1}}^2 \Big| \theta^k, \eta\Big] \tag{45}$$

$$= \frac{p}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \frac{p}{2}\mathbb{E}_j\Big[\left\|\theta^{k+1} - w\right\|_{\Gamma^{-1}}^2 \Big| \theta^k, \eta\Big]. \tag{46}$$

Upper bounding the "noise" term. We now upper bound the "noise" term in (38). We first recall the definition of the noisy proximal update g_j (line 7 of Algorithm 1), and define its non-noisy counterpart $\tilde{g}_j$:

$$g_j = \gamma_j^{-1}\Big(\operatorname{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) - \theta_j^k\Big) \tag{47}$$

$$\tilde{g}_j = \gamma_j^{-1}\Big(\operatorname{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k)) - \theta_j^k\Big). \tag{48}$$

For an update of the coordinate $j \in [p]$, the optimality condition of the proximal operator gives, for η_j the realization of the noise drawn at the current iteration when coordinate j is chosen:

$$0 \in \theta_j^{k+1} - \theta_j^k + \gamma_j(\nabla_j f(\theta^k) + \eta_j)) + \frac{1}{M_j}\partial\psi_j(\theta_j^k - \gamma_j g_j) \tag{49}$$

$$= \gamma_j \times \left(\frac{1}{\gamma_j}(\theta_j^{k+1} - \theta_j^k) + \nabla_j f(\theta^k) + \eta_j + \partial\psi_j(\theta_j^k - \gamma_j g_j)\right). \tag{50}$$

As such, there exists a real number $v_j \in \partial\psi_j(\theta_j^k - \gamma_j g_j)$ such that $g_j = -\frac{1}{\gamma_j}(\theta_j^{k+1} - \theta_j^k) = \nabla_j f(\theta^k) + \eta_j + v_j$. We denote by $v \in \mathbb{R}^p$ the vector having this v_j as j-th coordinate. Recall that ψ is separable, therefore $v \in \partial\psi(\theta^k - \Gamma g)$. The "noise" term of (38) can be thus be rewritten using v:

$$\text{"noise" term} = \left\langle \nabla f(\theta^k) + v - g, \theta^k - \Gamma g - w\right\rangle = \left\langle \eta, \theta^k - \Gamma g - w\right\rangle, \tag{51}$$

and we now separate this term in two using $\widetilde{g}$:

$$\text{"noise" term} = \sum_{j=1}^p \eta_j(\theta_j^k - \gamma_j g_j - w_j) = \sum_{j=1}^p \eta_j(\theta_j^k - \gamma_j\widetilde{g}_j - w_j) + \sum_{j=1}^p \eta_j(\gamma_j\widetilde{g}_j - \gamma_j g_j). \tag{52}$$

It is now time to consider the expectation with respect to the noise of these terms. First, as $\widetilde{g}_j$ is not dependent on the noise anymore, it simply holds that

$$\mathbb{E}_\eta\Big[\sum_{j=1}^p \eta_j(\theta_j^k - \gamma_j\widetilde{g}_j - w_j) \mid \theta^k\Big] = \sum_{j=1}^p \mathbb{E}_\eta[\eta_j]\,(\theta_j^k - \gamma_j\widetilde{g}_j - w_j) = 0. \tag{53}$$

The last step of our proof now takes care of the following term:

$$\mathbb{E}_\eta\Big[\sum_{j=1}^p \eta_j(\gamma_j\widetilde{g}_j - \gamma_j g_j) \mid \theta^k\Big] \leq \mathbb{E}_\eta\Big[\gamma_j\Big|\sum_{j=1}^p \eta_j(\widetilde{g}_j - g_j)\Big| \mid \theta^k\Big] \leq \sum_{j=1}^p \gamma_j\mathbb{E}_\eta\big[\ |\eta_j|\,|\widetilde{g}_j - g_j|\ \big|\ \theta^k\big]\,, \tag{54}$$

where each inequality comes from the triangle inequality. The non-expansiveness property of the proximal operator (see Parikh and Boyd [2014], Section 2.3) is now key to our result, as it yields

$$|\widetilde{g}_j - g_j| = \gamma_j^{-1}\left|\mathrm{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k))) - \mathrm{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j))\right| \leq |\eta_j|\,, \tag{55}$$

which directly gives, as $\mathbb{E}_\eta\big[\eta_j^2\big] = \sigma_j^2$ (and $\|\sigma\|_\Gamma^2 = \sum_{j=1}^p \gamma_j\sigma_j^2$),

$$\sum_{j=1}^p \gamma_j\mathbb{E}_\eta\big[|\eta_j|\,|\widetilde{g}_j - g_j|\big|\theta^k\big] \leq \sum_{j=1}^p \gamma_j\mathbb{E}_\eta[|\eta_j|\,|\eta_j|] = \sum_{j=1}^p \gamma_j\mathbb{E}_\eta\big[\eta_j^2\big] = \|\sigma\|_\Gamma^2\,. \tag{56}$$

We now have everything to prove the lemma by plugging (56) and (53) into expected value of (52), and then (52) and (42) back into (38) to obtain, after using the Tower property of conditional expectations:

$$\frac{1}{p}\mathbb{E}_{j,\eta}\Big[f(\theta^k) - f(w) + \big\langle\nabla f(\theta^k), -\Gamma g\big\rangle + \frac{1}{2}\|\Gamma g\|_M^2 + \psi(\theta^k - \Gamma g) - \psi(w)\Big|\theta^k\Big] \tag{57}$$

$$\leq \frac{1}{p}\,(\text{“descent” term} + \text{“noise” term}) \tag{58}$$

$$\leq \frac{1}{2}\left\|\theta^k - w\right\|_{\Gamma^{-1}}^2 - \frac{1}{2}\mathbb{E}_{j,\eta}\Big[\left\|\theta^{k+1} - w\right\|_{\Gamma^{-1}}^2\Big|\theta^k\Big] + \frac{1}{p}\|\sigma\|_\Gamma^2\,, \tag{59}$$

which is the result of the lemma. □

C.2.2 Convergence Lemma

Lemma C.3 allows us to prove a result on the mean of K consecutive noisy coordinate-wise gradient updates, by simply summing it and rewriting the terms. This gives Lemma C.4, which is the key lemma of our proof.

Lemma C.4. *Assume $\ell(\cdot, d)$ is convex, L-component-Lipschitz and M-component-smooth for all $d \in \mathcal{X}$, ψ is convex and separable, such that $F = f + \psi$ and w^* is a minimizer of F. For $t \in [T]$, consider the K successive iterates $\theta^1, \ldots, \theta^K$ computed from the inner loop of Algorithm 1 starting from the point $\bar{w}^t$, with step sizes $\gamma_j = \frac{1}{M_j}$ and noise scales σ_j. Letting $\bar{w}^{t+1} = \frac{1}{K}\sum_{k=1}^K \theta^k$, it holds that*

$$\mathbb{E}\big[F(\bar{w}^{t+1}) - F(w^*)\big] \leq \frac{p(\|\bar{w}^t - w^*\|_M^2 + 2(F(\bar{w}^t) - F(w^*)))}{2K} + \|\sigma\|_{M^{-1}}^2\,. \tag{60}$$

Remark C.5. The term $F(\bar{w}^t) - F(w^*)$ essentially remains in the inequality due to the composite nature of F. When $\psi = 0$, M-component-smoothness of $f(\cdot; d)$ (for $d \in \mathcal{X}$) gives

$$f(\bar{w}^t) \leq f(w^*) + \big\langle\nabla f(w^*), \bar{w}^t - w^*\big\rangle + \frac{1}{2}\left\|\bar{w}^t - w^*\right\|_M^2 = f(w^*) + \frac{1}{2}\left\|\bar{w}^t - w^*\right\|_M^2\,, \tag{61}$$

and the result of Lemma C.4 further simplifies as:

$$\mathbb{E}\big[F(\bar{w}^{t+1}) - F(w^*)\big] \leq \frac{p\,\|\bar{w}^t - w^*\|_M^2}{K} + \|\sigma\|_{M^{-1}}^2\,. \tag{62}$$

Proof. Summing Lemma C.3 for $k = 0$ to $k = K$ and $w = w^*$, taking expectation with respect to all choices of coordinate and random noise and using the tower property gives:

$$\sum_{k=0}^{K-1} \mathbb{E}\big[F(\theta^{k+1}) - F(w^*)\big] - \frac{p-1}{p} \sum_{k=0}^{K-1} \mathbb{E}\big[(F(\theta^k) - F(w^*))\big]$$

$$\leq \sum_{k=0}^{K-1} \frac{1}{2}\mathbb{E}\Big[\big\|\theta^k - w^*\big\|_{\Gamma^{-1}}^2\Big] - \frac{1}{2}\mathbb{E}\Big[\big\|\theta^{k+1} - w^*\big\|_{\Gamma^{-1}}^2\Big] + \frac{1}{p}\|\sigma\|_\Gamma^2 \tag{63}$$

$$= \frac{1}{2}\mathbb{E}\Big[\big\|\bar{w}^0 - w^*\big\|_{\Gamma^{-1}}^2\Big] - \frac{1}{2}\mathbb{E}\Big[\big\|\theta^K - w^*\big\|_{\Gamma^{-1}}^2\Big] + \frac{K}{p}\|\sigma\|_\Gamma^2 . \tag{64}$$

Remark that $\sum_{k=0}^{K-1} \mathbb{E}\big[F(\theta^k) - F(w^*)\big] = \sum_{k=1}^{K} \mathbb{E}\big[F(\theta^k) - F(w^*)\big] + (F(\bar{w}^0) - F(w^*)) - \mathbb{E}\big[F(\theta^K) - F(w^*)\big]$, then as $\mathbb{E}\big[F(\theta^K) - F(w^*)\big] \geq 0$, we obtain a lower bound on the left hand side of (64):

$$\sum_{k=0}^{K-1} \mathbb{E}\big[F(\theta^{k+1}) - F(w^*)\big] - \tfrac{p-1}{p} \sum_{k=0}^{K-1} \mathbb{E}\big[(F(\theta^k) - F(w^*))\big] \geq \tfrac{1}{p} \sum_{k=1}^{K} \mathbb{E}\big[F(\theta^k) - F(w^*)\big] - (F(\bar{w}^0) - F(w^*)). \tag{65}$$

As $\bar{w}^{t+1} = \frac{1}{K}\sum_{k=1}^{K} \theta^k$, the convexity of F gives $F(\bar{w}^{t+1}) \leq \frac{1}{K}\sum_{k=1}^{K} F(\theta^k) - F(w^*)$. Plugging this inequality into (65) and combining the result with (64) gives

$$F(\bar{w}^{t+1}) - F(w^*) \leq \frac{p(\frac{1}{2}\big\|\bar{w}^0 - w^*\big\|_{\Gamma^{-1}}^2 + F(\bar{w}^0) - F(w^*))}{K} + \|\sigma\|_\Gamma^2 . \tag{66}$$

We conclude the proof by using the fact that $\Gamma_j = M_j^{-1}$ for all $j \in [p]$, thus $\|\cdot\|_\Gamma = \|\cdot\|_{M^{-1}}$ and $\|\cdot\|_{\Gamma^{-1}} = \|\cdot\|_M$. □

C.2.3 Convex Case

Theorem 3.3 (Convex case)**.** *Let w^* be a minimizer of F and $R_M^2 = \max(\big\|\bar{w}^0 - w^*\big\|_M^2, F(\bar{w}^0) - F(w^*))$. The output w^{priv} of DP-CD (Algorithm 1), starting from $\bar{w}^0 \in \mathbb{R}^p$ with $T = 1$, $K > 0$ and the σ_j's as in Theorem 3.1, satisfies:*

$$F(w^{priv}) - F(w^*) \leq \frac{3pR_M^2}{2K} + \frac{12\|L\|_{M^{-1}}^2 K \log(1/\delta)}{n^2\epsilon^2}. \tag{67}$$

Setting $K = \frac{R_M\sqrt{p}n\epsilon}{\|L\|_{M^{-1}}\sqrt{8\log(1/\delta)}}$ yields:

$$F(w^{priv}) - F(w^*) \leq \frac{9\sqrt{p}\,\|L\|_{M^{-1}} R_M\sqrt{\log(1/\delta)}}{n\epsilon} = \widetilde{O}\left(\frac{\sqrt{p}R_M\|L\|_{M^{-1}}}{n\epsilon}\right). \tag{68}$$

Proof. In the convex case, we iterate only once in the inner loop (since $T = 1$). As such, $w^{priv} = \bar{w}^1$, and applying Lemma C.4 with $\bar{w}^{t+1} = \bar{w}^1$, $w^t = \bar{w}^0$ and σ_j chosen as in Theorem 3.1 gives the result. Taking $K = \frac{R_M\sqrt{p}n\epsilon}{\|L\|_{M^{-1}}\sqrt{8\log(1/\delta)}}$ then gives

$$F(\bar{w}_1^{t+1}) - F(w^*) \leq \frac{2\sqrt{8p\log(1/\delta)}\,\|L\|_{M^{-1}} R_M}{n\epsilon} + \frac{12\sqrt{p\log(1/\delta)}\,\|L\|_{M^{-1}} R_M}{\sqrt{8}n\epsilon}, \tag{69}$$

and the result follows from $2\sqrt{8} + \frac{12}{\sqrt{8}} \approx 8.48 < 9$. □

C.2.4 Strongly Convex Case

Theorem 3.3 (Strongly-convex case)**.** *Let F be μ_M-strongly convex w.r.t. $\|\cdot\|_M$ and w^* be the minimizer of F. The output w^{priv} of DP-CD (Algorithm 1), starting from $\bar{w}^0 \in \mathbb{R}^p$ with $T > 0$, $K = 2p(1 + 1/\mu_M)$ and the σ_j's as in Theorem 3.1, satisfies:*

$$F(w^{priv}) - F(w^*) \leq \frac{F(\bar{w}^0) - F(w^*)}{2^T} + \frac{24p(1 + 1/\mu_M)T\|L\|_{M^{-1}}^2 \log(1/\delta)}{n^2\epsilon^2}. \tag{70}$$

Setting $T = \log_2\left(\frac{32n^2\epsilon^2(F(\bar{w}^0)-F(w^*))}{p(1+1/\mu_M)\|L\|^2_{M^{-1}}\log(1/\delta)}\right)$ *yields:*

$$\mathbb{E}\big[F(w^{priv}) - F(w^*)\big] \leq \left(1 + \log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n^2\epsilon^2}{24p(1+1/\mu_M)\|L\|^2_{M^{-1}}\log(1/\delta)}\right)\right)\frac{24p(1+1/\mu_M)\|L\|^2_{M^{-1}}\log(1/\delta)}{n^2\epsilon^2} \tag{71}$$

$$= O\left(\frac{p\|L\|^2_{M^{-1}}\log(1/\delta)}{\mu_M n^2\epsilon^2}\log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n\epsilon\mu_M}{p\|L\|_{M^{-1}}\log(1/\delta)}\right)\right) \tag{72}$$

Proof. As F is μ_M-strongly-convex with respect to norm $\|\cdot\|_M$, we obtain for any $w \in \mathbb{R}^p$, that $F(w) \geq F(w^*) + \frac{\mu_M}{2}\|w - w^*\|^2_M$. Therefore, $F(\bar{w}^0) - F(w^*) \leq \frac{2}{\mu_M}\left\|\bar{w}^0 - w^*\right\|^2_M$ and Lemma C.4 gives, for $1 \leq t \leq T-1$,

$$F(\bar{w}^{t+1}) - F(w^*) \leq \frac{(1+1/\mu_M)p(F(\bar{w}^t) - F(w^*))}{K} + \|\sigma\|^2_M\,. \tag{73}$$

It remains to set $K = 2p(1+1/\mu_M)$ to obtain

$$F(\bar{w}^{t+1}) - F(w^*) \leq \frac{F(\bar{w}^t) - F(w^*)}{2} + \|\sigma\|^2_M\,. \tag{74}$$

Recursive application of this inequality gives

$$\mathbb{E}\big[F(\bar{w}^T) - F(w^*)\big] \leq \frac{F(\bar{w}^0) - F(w^*)}{2^T} + \sum_{t=0}^{T-1}\frac{1}{2^t}\|\sigma\|^2_M \leq \frac{F(\bar{w}^0) - F(w^*)}{2^T} + 2\|\sigma\|^2_M\,, \tag{75}$$

where we upper bound the sum by the value of the complete series. It remains to replace $\|\sigma\|^2_M$ by its value to obtain the result. Taking $T = \log_2\left(\frac{(F(\bar{w}^0)-F(w^*))n^2\epsilon^2}{24p(1+1/\mu_M)\|L\|^2_{M^{-1}}\log(1/\delta)}\right)$ then gives

$$\mathbb{E}\big[F(\bar{w}^T) - F(w^*)\big] \leq \left(1 + \log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n^2\epsilon^2}{24p(1+1/\mu_M)\|L\|^2_{M^{-1}}\log(1/\delta)}\right)\right)\frac{24p(1+1/\mu_M)\|L\|^2_{M^{-1}}\log(1/\delta)}{n^2\epsilon^2} \tag{76}$$

$$= O\left(\frac{p\|L\|^2_{M^{-1}}\log(1/\delta)}{\mu_M n^2\epsilon^2}\log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n\epsilon\mu_M}{p\|L\|_{M^{-1}}\log(1/\delta)}\right)\right), \tag{77}$$

which is the result of our theorem. □

C.3 Proof of Remark 1

We recall the notations of Tappenden et al. [2016]. For $\theta \in \mathbb{R}^p$, $t \in \mathbb{R}$ and $j \in [p]$, let $V_j(\theta, t) = \nabla_j(\theta)t + \frac{M_j}{2}|t|^2 + \psi_j(\theta^k_j + t)$. For $\eta \in \mathbb{R}$, we also define its noisy counterpart, $V^\eta_j(\theta, t) = (\nabla_j(\theta) + \eta)t + \frac{M_j}{2}|t|^2 + \psi_j(\theta^k_j + t)$. We aim at finding δ_j such that for any $\theta^k \in \mathbb{R}^p$ used in the inner loop of Algorithm 1:

$$\mathbb{E}_{\eta_j}\big[V_j(\theta^k, -\gamma_j g_j)\big] \leq \min_{\widetilde{g}\in\mathbb{R}} V_j(\theta^k, -\gamma_j\widetilde{g}) + \delta_j, \tag{78}$$

where the expectation is taken over the random noise η_j, and $-\gamma_j g_j = \mathrm{prox}_{\gamma_j\psi_j}(\theta^k_j - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) - \theta^k_j$ as defined in the analysis of Algorithm 1. We need to link the proximal operator we use in DP-CD with the quantity $V^{\eta_j}_j$ that we just defined:

$$\mathrm{prox}_{\gamma_j\psi_j}(\theta^k_j - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) = \arg\min_{v\in\mathbb{R}}\frac{1}{2}\left\|v - \theta^k_j + \gamma_j(\nabla_j f(\theta^k) + \eta_j)\right\|^2_2 \tag{79}$$

$$= \arg\min_{v\in\mathbb{R}}\left\langle\gamma_j(\nabla_j f(\theta^k_j) + \eta_j), v - \theta^k_j\right\rangle + \frac{1}{2}\left\|v - \theta^k_j\right\|^2_2 + \gamma_j\psi_j(v) \tag{80}$$

$$= \arg\min_{v\in\mathbb{R}}\left\langle\nabla_j f(\theta^k) + \eta_j, v - \theta^k_j\right\rangle + \frac{M_j}{2}\left\|v - \theta^k_j\right\|^2_2 + \psi_j(v) \tag{81}$$

$$= \theta^k_j + \arg\min_{t\in\mathbb{R}}\left\langle\nabla_j f(\theta^k) + \eta_j, t\right\rangle + \frac{M_j}{2}\|t\|^2_2 + \psi_j(\theta^k_j + t). \tag{82}$$

Which means that $-\gamma_j g_j = \operatorname{prox}_{\gamma_j \psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) - \theta_j^k \in \arg\min_{t\in\mathbb{R}} V_j^{\eta_j}(\theta^k, t)$. Let $-\gamma_j g_j^* = \operatorname{prox}_{\gamma_j \psi_j}(\theta_j^k - \gamma_j \nabla_j(\theta^k)) - \theta_j^k$ be the non-noisy counterpart of $-\gamma_j g_j$. Since $-\gamma_j g_j$ is a minimizer of $V_j^{\eta_j}(\theta^k, \cdot)$, it holds that

$$V_j^{\eta_j}(\theta^k, -\gamma_j g_j) \leq \left\langle \nabla_j f(\theta^k) + \eta_j, -\gamma_j g_j^* \right\rangle + \frac{M_j}{2} \left\| -\gamma_j g_j^* \right\|_2^2 + \psi_j(\theta_j^k + -\gamma_j g_j^*) \tag{83}$$

$$= \min_t V_j(\theta^k, t) + \left\langle \eta_j, -\gamma_j g_j^* \right\rangle, \tag{84}$$

which can be rewritten as $V_j(\theta^k, -\gamma_j g_j) \leq \min_t V_j(\theta^k, t) + \left\langle \eta_j, \gamma_j(g_j - g_j^*) \right\rangle$. Taking the expectation yields

$$\mathbb{E}_{\eta_j}\left[V_j(\theta^k, -\gamma_j g_j)\right] \leq \min_t V_j(\theta^k, t) + \mathbb{E}_{\eta_j}\left[\left\langle \eta_j, \gamma_j(g_j - g_j^*) \right\rangle\right]. \tag{85}$$

Finally, we remark that $\left|g_j - g_j^*\right| \leq |\gamma_j \eta_j|$ and the non-expansiveness of the proximal operator gives

$$\mathbb{E}_{\eta_j}\left[V_j(\theta^k, -\gamma_j g_j)\right] \leq \min_t V_j(\theta^k, t) + \gamma_j \sigma_j^2, \tag{86}$$

which implies an upper bound on the expectation of δ_j: $\mathbb{E}_{j,\eta_j}[\delta_j] = \frac{1}{p}\sum_{j=1}^p \mathbb{E}_{\eta_j}[\delta_j] \leq \frac{1}{p}\sum_{j=1}^p \gamma_j \sigma_j^2 = \frac{1}{p}\sum_{j=1}^p \sigma_j^2/M_j$, when $\gamma_j = 1/M_j$. In the formalism of Tappenden et al. [2016], this amounts to setting $\alpha = 0$ and $\beta = \frac{1}{p}\|\sigma\|_{M^{-1}}^2$.

Convex functions. When the objective function F is convex, we use Lemma C.4 to obtain, since $\|\sigma\|_{M^{-1}}^2 = \beta p$,

$$F(w^1) - F(w^*) \leq \frac{2pR_M^2}{K} + \|\sigma\|_{M^{-1}}^2 = \frac{2pR_M^2}{K} + \beta p. \tag{87}$$

Therefore, when F is convex, we get $F(w^1) - F(w^*) \leq \xi$, for $\xi > \beta p$, as long as $\frac{2pR_M^2}{K} \leq \xi - \beta p$, that is $K \geq \frac{2pR_M^2}{\xi - \beta p}$.

In comparison, Tappenden et al. [2016, Theorem 5.1 therein] gives convergence to $\xi > \sqrt{2pR_M^2\beta}$ when $K \geq \frac{2pR_M^2}{\xi - \sqrt{2pR_M^2\beta}}$. We thus gain a factor $\sqrt{\beta p / 2R_M^2}$ in utility. Importantly, our utility upper bound does not depend on initialization in that setting, whereas the one of Tappenden et al. [2016] does.

Strongly-convex functions. When the objective function F is μ_M-strongly-convex *w.r.t.* to $\|\cdot\|_M$, then from (75) we obtain, as long as $K \geq 4/\mu_M$, that

$$\mathbb{E}\left[F(w^T) - F(w^*)\right] \leq \frac{F(w^0) - F(w^*)}{2^T} + 2\beta p. \tag{88}$$

This proves that $\mathbb{E}\left[F(w^T) - F(w^*)\right] \leq \xi$ for $\xi > 2\beta p$ when $\frac{F(w^0)-F(w^*)}{2^T} \leq \xi - 2\beta p$ that is $T \geq \log \frac{F(w^0)-F(w^*)}{\xi - 2\beta p}$ and $TK \geq \frac{4p}{\mu_M} \log \frac{F(w^0)-F(w^*)}{\xi - 2\beta p}$. In comparison, Tappenden et al. [2016, Theorem 5.2 therein] shows convergence to $\xi > \frac{\beta p}{\mu_M}$ for $K \geq \frac{p}{\mu_M} \log \frac{F(w^0)-F(w^*)-\frac{\beta p}{\mu_M}}{\xi - \frac{\beta p}{\mu_M}}$. We thus gain a factor $\mu_M/2$ in utility.

D Comparison with DP-SGD

In this section, we provide more details on the arguments of Section 3.4, where we suppose that ℓ is L-component-Lipschitz and Λ-Lipschitz. To ease the comparison, we assume that $R_M = \left\|w^0 - w^*\right\|_M$, which is notably the case in the smooth setting with $\psi = 0$ (see Remark C.2).

Balanced. We start by the scenario where coordinate-wise smoothness constants are balanced and all equal to $M = M_1 = \cdots = M_p$. We observe that

$$\|L\|_{M^{-1}} = \sqrt{\sum_{j=1}^p \frac{1}{M_j} L_j^2} = \sqrt{\frac{1}{M}\sum_{j=1}^p L_j^2} = \frac{1}{\sqrt{M}} \|L\|_2. \tag{89}$$

We then consider the convex and strongly-convex functions separately:

- *Convex functions:* it holds that $R_M = \sqrt{M} R_I$, which yields the equality $\|L\|_{M^{-1}} R_M = \|L\|_2 R_I$.

- *Strongly convex functions:* if f is μ_M-strongly-convex with respect to $\|\cdot\|_M$, then for any $x, y \in \mathbb{R}^p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_M}{2} \|y - x\|_M^2 = f(x) + \langle \nabla f(x), y - x \rangle + \frac{M\mu_M}{2} \|y - x\|_2^2, \quad (90)$$

which means that f is $M\mu_M$-strongly-convex with respect to $\|\cdot\|_2$. This gives $\frac{\|L\|_{M^{-1}}^2}{\mu_M} = \frac{\|L\|_2^2/M}{\mu_I/M} = \frac{\|L\|_2^2}{\mu_I}$.

In light of the results summarized in Table 1, it remains to compare $\|L\|_2 = \sqrt{\sum_{j=1}^p L_j^2}$ with Λ, for which it holds that $\Lambda \leq \sqrt{\sum_{j=1}^p L_j^2} \leq \sqrt{p}\Lambda$, which is our result.

Unbalanced. When smoothness constants are disparate, we discuss the case where

- *one coordinate of the gradient dominates the others:* we assume without loss of generality that the dominating coordinate is the first one. It holds that $M_1 =: M_{\max} \gg M_{\min} =: M_j$, for all $j \neq 1$ and $L_1 =: L_{\max} \gg L_{\min} =: L_j$, for all $j \neq 1$ such that $\frac{L_1^2}{M_1} \gg \sum_{j \neq 1} \frac{L_j^2}{M_j}$. As L_1 dominates the other component-Lipschitz constants, most of the variation of the loss comes from its first coordinate. This implies that L_1 is close to the global Lipschitz constant Λ of ℓ. As such, it holds that

$$\|L\|_{M^{-1}}^2 = \sum_{j=1}^p \frac{L_j^2}{M_j} \approx \frac{L_1^2}{M_1} \approx \frac{\Lambda^2}{M_{\max}}. \quad (91)$$

- *the first coordinate of $\bar{w}^0$ is already very close to its optimal value* so that $M_1 \left|\bar{w}_1^0 - w_1^*\right| \ll \sum_{j \neq 1} M_j \left|\bar{w}_j^0 - w_j^*\right|$. Under this hypothesis,

$$R_M^2 \approx \sum_{j \neq 1} M_j \left|w_j^0 - w_j^*\right|^2 = M_{\min} \sum_{j \neq 1} \left|w_j^0 - w_j^*\right|^2 \approx M_{\min} R_I^2. \quad (92)$$

We can now easily compare DP-CD with DP-SGD in this scenario. First, if ℓ is convex, then $\|L\|_{M^{-1}} R_M \approx \sqrt{\frac{M_{\min}}{M_{\max}}} \Lambda R_I$. Second, when ℓ is strongly-convex, we observe that for $x, y \in \mathbb{R}^p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_M}{2} \|y - x\|_M^2 \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_{\min}\mu_M}{2} \|y - x\|_2^2, \quad (93)$$

which implies that when f is μ_M strongly-convex with respect to $\|\cdot\|_M$, it is $M_{\min}\mu_M$ strongly-convex with respect to $\|\cdot\|_2$. This yields, under our hypotheses, $\frac{\|L\|_{M^{-1}}^2}{\mu_M} \approx \frac{\Lambda^2/M_{\max}}{\mu_I/M_{\min}} = \frac{M_{\min}}{M_{\max}} \frac{\Lambda^2}{\mu_I}$. In both cases, DP-CD can get arbitrarily better than DP-SGD, and gets better as the ratio $M_{\max}/M_{\min}$ increases.

The two hypotheses we describe above are of course very restrictive. However, it gives some insight about when and why DP-CD can outperform DP-SGD. Our numerical experiments in Section 6 confirm this analysis, even in less favorable cases.

E Proof of Lower Bounds

To prove lower bounds on the utility of L-component-Lipschitz functions, we extend the proof of Bassily et al. [2014] to our setting (that is, L-component-Lipschitz functions and unconstrained composite optimization). There are three main difficulties in adapting their proof:

- First, the optimization problem (1) is not constrained. We stress that while convex constraints can be enforced using the regularizer ψ (using the characteristic function of a convex set), its separable nature only allows box constraints. In contrast, Bassily et al. [2014] rely on an ℓ_2-norm constraint to obtain their lower bounds.
- Second, Lemma 5.1 of Bassily et al. [2014] must be extended to our L-component-Lipschitz setting. To do so, we consider datasets with points in $\prod_{j=1}^p \{-L_j, L_j\}$ rather than $\{-1/\sqrt{p}, 1/\sqrt{p}\}^p$, and carefully adapt the construction of the dataset D so that $\|\sum_{i=1}^n d_i\|_2 = \Omega(\min(n \|L\|_2, \sqrt{p} \|L\|_2/\epsilon))$, which is essential to prove our lower bounds.

- Third, the lower bounds of Bassily et al. [2014] rely on fingerprinting codes, and in particular on the result of Bun et al. [2014] which uses such codes to prove that (when n is smaller than some n^* we describe later) differential privacy is incompatible with precisely and simultaneously estimating *all* p counting queries defined over the columns of the dataset D. In our construction, since all columns of D now have different scales, we need an additional hypothesis on the repartition of the L_j's, (*i.e.,* that $\sum_{j\in\mathcal{J}} L_j^2 = \Omega(\|L\|_2)$ for all $\mathcal{J} \subseteq [p]$ of a given size), which is not required in existing lower bounds (where all columns have equal scale).

E.1 Counting Queries and Accuracy

We start our proof by recalling and extending to our setting the notions of counting queries (Definition E.1) and accuracy (Definition E.2), as described by Bun et al. [2014]. The main feature of our definitions is that we allow the set $\mathcal{X}$ to have different scales for each of its coordinates, and that we account for this scale in the definition of accuracy. We denote by $\mathrm{conv}(\mathcal{X})$ the convex hull of a set $\mathcal{X}$.

Definition E.1 (Counting query)**.** Let $n > 0$. A counting query on $\mathcal{X}$ is a function $q : \mathcal{X}^n \to \mathrm{conv}(\mathcal{X})$ defined using a predicate $q : \mathcal{X} \to \mathcal{X}$. The evaluation of the query q over a dataset $\mathcal{D} \in \mathcal{X}^n$ is defined as the arithmetic mean of q on $\mathcal{D}$:

$$q(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} q(d_i). \tag{94}$$

Definition E.2 (Accuracy)**.** Let $n, p \in \mathbb{N}$, $\alpha, \beta \in [0,1]$, $L_1, \dots, L_p > 0$, and $\mathcal{X} = \prod_{j=1}^{p}\{-L_j; L_j\}$ or $\mathcal{X} = \{0, L_j\}^p$. Let $\mathcal{Q} = \{q_1, \dots, q_p\}$ be a set of p counting queries on $\mathcal{X}$ and $D \in \mathcal{X}^n$ a dataset of n elements. A sequence of answers $a = (a_1, \dots, a_p)$ is said (α, β)-accurate for $\mathcal{Q}$ if $|q_j(D) - a_j| \leq L_j\alpha$ for at least a $1-\beta$ fraction of indices $j \in [p]$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^{|\mathcal{Q}|}$ is said (α, β)-accurate for $\mathcal{Q}$ on $\mathcal{X}$ if for every $D \in \mathcal{X}^n$,

$$\Pr\left[\mathcal{A}(D) \text{ is } (\alpha, \beta)\text{-accurate for } \mathcal{Q}\right] \geq 2/3. \tag{95}$$

In our proof, we will use a specific class of queries: one-way marginals (Definition E.3), that compute the arithmetic mean of a dataset along one of its column.

Definition E.3 (One-way marginals)**.** Let $\mathcal{X} = \prod_{j=1}^{p}\{-L_j; L_j\}$ or $\mathcal{X} = \{0, L_j\}^p$. The family of one-way marginals on $\mathcal{X}$ is defined by queries with predicates $q_j(x) = x_j$ for $x \in \mathcal{X}$. For a dataset $D \in \mathcal{X}^n$ of size n, we thus have $q_j(D) = \frac{1}{n}\sum_{i=1}^{n} d_{i,j}$.

E.2 Lower Bound for One-Way Marginals

We can now restate a key result from Bun et al. [2014], which shows that there exists a minimal number n^* of records needed in a dataset to allow achieving both accuracy and privacy on the estimation of one-way marginals on $\mathcal{X} = (\{0,1\}^p)^n$. This lemma relies on the construction of re-identifiable distribution (see Bun et al. 2014, Definition 2.10). One can then use this distribution to find a dataset on which a private algorithm can not be accurate (see Bun et al. 2014, Lemma 2.11).

Lemma E.4 (Bun et al. 2014, Corollary 3.6)**.** *For $\epsilon > 0$ and $p > 0$, there exists a number $n^* = \Omega(\frac{\sqrt{p}}{\epsilon})$ such that for all $n \leq n^*$, there exists no algorithm that is both $(1/3, 1/75)$-accurate and $(\epsilon, o\left(\frac{1}{n}\right))$-differentially private for the estimation of one-way marginals on $(\{0,1\}^p)^n$.*

To leverage this result in our setting of private empirical risk minimization, we start by extending it to queries on $\mathcal{X} = \prod_{j=1}^{p}\{-L_j; L_j\}$. Before stating the main theorem of this section (Theorem E.5), we describe a procedure $\chi_L : (\{0,1\}^p)^n \to \mathcal{X}^{3n}$ (with $L_1, \dots, L_p > 0$), that takes as input a dataset $D \in (\{0,1\}^p)^n$ and outputs an augmented and rescaled version. This procedure is crucial to our proof and is defined as follows. First, it adds $2n$ rows filled with 1's to D, which ensures that the sum of each column of D is $\Theta(n)$ (which gives the lower bound on M in Theorem E.5). Then it rescales each of these columns by subtracting $1/2$ to each coefficient and multiplying the j-th column of D ($j \in [p]$) by $2L_j$. The resulting dataset $D_L^{aug} = \chi_L(D)$ is a set of $3n$ points with values in $\mathcal{X} = \prod_{j=1}^{p}\{-L_j, L_j\}$, with the property that, for all $j \in [p]$, $3nL_j \geq \sum_{i=1}^{n}(D_L^{aug})_{i,j} \geq nL_j$. For $D \in (\{0,1\}^p)^n$, we show how to reconstruct $q_j(\chi_L(D))$ from $q_j(D)$ in Claim 1.

Claim 1. Let $n \in \mathbb{N}$, $j \in [p]$, $L_j > 0$ and q_j the j-th one-way marginal on datasets D with p columns such that for $d_i \in D$, $q_j(d_i) = d_{i,j}$. Let $D_L^{aug} = \chi_L(D)$. It holds that

$$q_j(D_L^{aug}) = \frac{2L_j}{3}q_j(D) + \frac{L_j}{3}, \tag{96}$$

where we use the slight abuse of notation by denoting the one-way marginals $q_j : \mathcal{X}^{3n} \to \mathrm{conv}(\mathcal{X})$ and $q_j : (\{0,1\}^p)^n \to [0,1]^p$ in the same way.

Proof. Let $D \in (\{0,1\}^p)^n$, and let $D^{aug} \in (\{0,1\}^p)^{3n}$ constructed by adding $2n$ rows of 1's at the end of D. Let $D_L^{aug} = \chi_L(D)$. We remark that

$$q_j(D^{aug}) = \frac{1}{3n}\sum_{i=1}^{3n} D_{i,j}^{aug} = \frac{1}{3}\left(\frac{1}{n}\sum_{i=1}^{n} D_{i,j}^{aug}\right) + \frac{1}{3n}\sum_{i=n+1}^{3n} 1 = \frac{1}{3}q_j(D) + \frac{2}{3} \in [0,1]. \tag{97}$$

Then, we link $q_j(D^{aug})$ with $q_j(D_L^{aug})$:

$$q_j(D_L^{aug}) = \frac{1}{3n}\sum_{i=1}^{3n}(D_L^{aug})_{i,j} = \frac{1}{3n}\sum_{i=1}^{3n} 2L_j((D^{aug})_{i,j} - 1/2) = 2L_j(q_j(D^{aug}) - 1/2) \in [-L_j, L_j], \tag{98}$$

combining (97) and (98) gives the result. □

Theorem E.5. *Let $n, p \in \mathbb{N}$, and $L_1, \dots, L_p > 0$. Assume that for all subsets $\mathcal{J} \subseteq [p]$ of size at least $\lceil \frac{p}{75} \rceil$, $\sqrt{\sum_{j\in\mathcal{J}} L_j^2} = \Omega(\|L\|_2)$. Define $\mathcal{X} = \prod_{j=1}^p \{-L_j; +L_j\}$, and let $q_j : \mathcal{X} \to \{-L_j, L_j\}$ be the predicate of the j-th one-way marginal on $\mathcal{X}$. Take $\epsilon > 0$ and $\delta = o(\frac{1}{n})$. There exists a number $M = \Omega\left(\min\left(n\|L\|_2, \frac{\sqrt{p}\|L\|_2}{\epsilon}\right)\right)$ such that for every (ϵ,δ)-differentially private algorithm $\mathcal{A}$, there exists a dataset $D = \{d_1, \dots, d_n\} \in \mathcal{X}^n$ with $\|\sum_{i=1}^n d_i\|_2 \in [M-1, M+1]$ such that, with probability at least $1/3$ over the randomness of $\mathcal{A}$:*

$$\|\mathcal{A}(D) - q(D)\|_2 = \Omega\left(\min\left(\|L\|_2, \frac{\sqrt{p}\,\|L\|_2}{n\epsilon}\right)\right). \tag{99}$$

Proof. Let $M = \Omega\left(\min\left(n\|L\|_2, \frac{\sqrt{p}\|L\|_2}{\epsilon}\right)\right)$, and define the set of queries $\mathcal{Q}$ composed of p queries $q_j(D) = \frac{1}{n}\sum_{i=1}^n d_{i,j}$ for $j \in [p]$. Let $\mathcal{A}$ be a (ϵ,δ)-differentially-private randomized algorithm. Let $\alpha, \beta \in [0,1]$. We will show that there exists a dataset D such that $\|\sum_{i=1}^n d_i\|_2 \in [M-1, M+1]$ for which $\mathcal{A}(D)$ is not (α,β)-accurate.

When $n \le n^*$. Assume, for the sake of contradiction, that $\mathcal{A} : \mathcal{X}^{3n} \to \mathrm{conv}(\mathcal{X})$ is $(\frac{1}{3}\alpha, \beta)$-accurate for $\mathcal{Q}$. Then, for each dataset $D' \in \mathcal{X}^{3n}$, we have

$$\Pr\left[\exists \mathcal{J} \subseteq [p] \text{ such that } |\mathcal{J}| \ge (1-\beta)p \text{ and } \forall j \in \mathcal{J},\ |\mathcal{A}_j(D') - q_j(D')| < \frac{2L_j}{3}\alpha\right] \ge 2/3. \tag{100}$$

Importantly, for all $D \in (\{0,1\})^p)^n$, the randomized algorithm $\mathcal{A}$ satisfies (100) for the dataset $D_L^{aug} = \chi_L(D) \in \mathcal{X}^{3n}$. We now construct the mechanism $\widetilde{\mathcal{A}} : (\{0,1\}^p)^n \to [0,1]^p$ that takes a dataset $D \in (\{0,1\}^p)^n$, constructs $D_L^{aug} = \chi_L(D)$ and runs $\mathcal{A}$ on it. It then outputs $\widetilde{\mathcal{A}}(D)$ such that, for $j \in [p]$, $\widetilde{\mathcal{A}}_j(D) = \frac{3}{2L_j}\mathcal{A}_j(D_L^{aug}) - \frac{L_j}{3}$. Using Claim 1, the results of $\widetilde{\mathcal{A}}$ and be linked to the ones of $\mathcal{A}$, as

$$\left|\widetilde{\mathcal{A}}(D) - q_j(D)\right| = \left|\frac{3}{2L_j}\mathcal{A}_j(D_L^{aug}) - \frac{L_j}{3} - \frac{3}{2L_j}q_j(D_L^{aug}) + \frac{L_j}{3}\right| = \frac{3}{2L_j}\left|\mathcal{A}_j(D_L^{aug}) - q_j(D_L^{aug})\right|. \tag{101}$$

Therefore, if $\mathcal{A}$ satisfies (100) and (101), then $\widetilde{\mathcal{A}} : (\{0,1\}^p)^n \to [0,1]^p$ satisfies, for all $D \in (\{0,1\}^p)^n$,

$$\Pr\left[\exists \mathcal{J} \subseteq [p] \text{ such that } |\mathcal{J}| \ge (1-\beta)p \text{ and } \forall j \in \mathcal{J},\ \left|\widetilde{\mathcal{A}}_j(D) - q_j(D)\right| < \alpha\right] \ge 2/3, \tag{102}$$

which is exactly the definition of (α,β)-accuracy for $\widetilde{\mathcal{A}}$. Remark that since $\widetilde{\mathcal{A}}$ is only a post-processing of $\mathcal{A}$, without additional access to the dataset itself, $\widetilde{\mathcal{A}}$ is itself (ϵ,δ)-differentially-private. We have thus constructed an algorithm that is both accurate and private for $n \le n^*$, which contradicts the result of Lemma E.4 when $\beta = \frac{1}{75}$. This proves the existence of a dataset $D \in (\{0,1\}^p)^n$ such that for $D_L^{aug} = \chi_L(D)$, $\mathcal{A}(D_L^{aug})$ is not $(\frac{1}{3}\alpha,\beta)$-accurate on $\mathcal{Q}$, which means that with probability at least $1/3$, there exists a subset $\mathcal{J} \subseteq [p]$ of cardinal $|\mathcal{J}| \ge \lceil \beta p \rceil$ such that

$$\|\mathcal{A}(D_L^{aug}) - q(D_L^{aug})\|_2 \overset{(100)}{\ge} \sqrt{\sum_{j\in\mathcal{J}} \frac{4L_j^2}{9}} \ge \Omega(\|L\|_2), \tag{103}$$

where the second inequality comes from the fact that $|\mathcal{J}| \geq \lceil \beta p \rceil = \lceil \frac{p}{75} \rceil$ and our hypothesis on $\sum_{j\in\mathcal{J}} L_j^2$. Notice that when $L_1 = \cdots = L_p = \frac{1}{\sqrt{p}}$, we recover the result of Bassily et al. [2014], since $\|L\|_2 = 1$ it holds with probability at least $1/3$ that

$$\|\mathcal{A}(D_L^{aug}) - q(D_L^{aug})\|_2 \overset{(100)}{\geq} \sqrt{\sum_{j\in\mathcal{J}} \frac{4L_j^2}{9}} \geq \sqrt{\frac{4}{9\times 75}} \|L\|_2 \geq \frac{2}{27}, \tag{104}$$

and in that case, since all L_j's are equal, it indeed holds that $\sqrt{\sum_{j\in\mathcal{J}} L_j^2} = \Omega(\|L\|_2)$. Finally, we remark that the sum of each column of D_L^{aug} is $\sum_{i=1}^n d_{i,j} \geq nL_j$, and as such, we have $\|\sum_{i=1}^n d_i\|_2 = \sqrt{\sum_{j=1}^p (\sum_{i=1}^n d_{i,j})^2} \geq \sqrt{\sum_{j=1}^p n^2 L_j^2} = n\|L\|_2$.

When $n > n^*$. We get the result in that case by augmenting the dataset D^* that we constructed in the first part of this proof. To do so, we follow the steps described by Bassily et al. [2014] in the proof of their Lemma 5.1. The construction consists in choosing a vector $c \in \mathcal{X}$, and adding $\lceil \frac{n-n^*}{2} \rceil$ rows with c, and $\lfloor \frac{n-n^*}{2} \rfloor$ rows with $-c$ to the dataset D^*. This results in a dataset D' such that $\|\sum_{i=1}^n d_i\| = \Omega(n^*\|L\|_2) = \Omega(\frac{\sqrt{p}\|L\|_2}{\epsilon})$, since the contributions of rows $-c$ and c (almost) cancel out. The theorem follows from observing that $(\frac{n^*}{n}\alpha, \beta)$-accuracy on this augmented dataset implies (α, β)-accuracy on the original dataset. As such, if an algorithm is both private and $(\frac{n^*}{n}\alpha, \beta)$-accurate on the dataset D', we get a contradiction, which gives the theorem as $\frac{n^*}{n} = \frac{\sqrt{p}}{n\epsilon}$. □

Remark E.6. Without the assumption on the distribution of the L_j's, we can still get an inequality that resembles (103): $\|\mathcal{A}(D_L^{aug}) - q(D_L^{aug})\|_2 \overset{(100)}{\geq} \sqrt{\sum_{j\in\mathcal{J}} \frac{4L_j^2}{9}} \geq \frac{2}{27}\frac{L_{\min}}{L_{\max}}\|L\|_2$, with probability at least $1/3$, and we get a result similar to Theorem E.5, except with an additional multiplicative factor $L_{\min}/L_{\max}$.

E.3 Lower Bound for Convex Functions

To prove a lower bound for our problem in the convex case, we let $L_1, \cdots, L_p > 0$ and define a dataset $D = \{d_1, \ldots, d_n\}$ taking its values in a set $\mathcal{X} = \prod_{j=1}^p \{\pm L_j\}$. For $\beta > 0$, we consider the problem (1) with the convex, smooth and L-component-Lipschitz loss function $\ell(w; d) = -\langle w, d\rangle$ and the convex, separable regularizer $\psi(w) = \frac{\|\sum_{i=1}^n d_i\|_2}{\beta n}\|w\|_2^2$:

$$w^* = \operatorname*{arg\,min}_{w\in\mathbb{R}^p} \left\{ F(w; D) = -\frac{1}{n}\langle w, \textstyle\sum_{i=1}^n d_i\rangle + \frac{\|\sum_{i=1}^n d_i\|_2}{\beta n}\|w\|_2^2 \right\}, \tag{105}$$

To find the solution of (105), we look for w^* so that the objective's gradient is zero, that is

$$w^* = \frac{\beta}{\|\sum_{i=1}^n d_i\|_2} \sum_{i=1}^n d_i, \tag{106}$$

so that $\|w^*\|_2 = \frac{\beta}{\|\sum_{i=1}^n d_i\|_2}\|\sum_{i=1}^n d_i\|_2 = \beta$. To prove the lower bound, we remark that

$$\begin{aligned}
F(w; D) - F(w^*; D) &= -\frac{1}{n}\langle w - w^*, \textstyle\sum_{i=1}^n d_i\rangle + \frac{\|\sum_{i=1}^n d_i\|}{2\beta n}(\|w\|_2^2 - \|w^*\|_2^2) && (107)\\
&= -\frac{1}{n}\left\langle w - w^*, \frac{\|\sum_{i=1}^n d_i\|}{\beta} w^* \right\rangle + \frac{\|\sum_{i=1}^n d_i\|}{2\beta n}(\|w\|_2^2 - \|w^*\|_2^2) && (108)\\
&= \frac{\|\sum_{i=1}^n d_i\|}{\beta n}\left(\langle w^* - w, w^*\rangle + \frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|w^*\|_2^2\right) && (109)\\
&= \frac{\|\sum_{i=1}^n d_i\|}{\beta n}\left(-\langle w, w^*\rangle + \frac{1}{2}\|w\|_2^2 + \frac{1}{2}\|w^*\|_2^2\right) && (110)\\
&= \frac{\|\sum_{i=1}^n d_i\|}{2\beta n}\|w - w^*\|_2^2. && (111)
\end{aligned}$$

At this point, we can proceed similarly to Bassily et al. [2014] to relate this quantity to private estimation of one-way marginals. We let $M = \Omega(\min(n \|L\|_2, \|L\|_2 \sqrt{p}/\epsilon))$ and $\mathcal{A}$ be an (ϵ, δ)-differentially private mechanism that outputs a private solution w^{priv} to (105). Suppose, for the sake of contradiction, that for every dataset D with $\|\sum_{i=1}^n d_i\|_2 \in [M-1; M+1]$,

$$\left\|w^{priv} - w^*\right\| \neq \Omega(\beta), \text{ with probability at least } 2/3. \tag{112}$$

We now derive from $\mathcal{A}$ a mechanism $\widetilde{\mathcal{A}}$ to estimate one-way marginals. To do this, $\widetilde{\mathcal{A}}$ runs $\mathcal{A}$ to obtain w^{priv} and outputs $\frac{M}{n\beta} w^{priv}$. We obtain that with probability at least $2/3$,

$$\left\|\widetilde{\mathcal{A}}(D) - q(D)\right\|_2 = \frac{M}{n\beta} \left\|w^{priv} - \frac{\beta}{M} \sum_{i=1}^n d_i\right\|_2 \neq \Omega\left(\frac{M}{n}\right) = \Omega\left(\min\left(\|L\|_2, \frac{\|L\|_2 \sqrt{p}}{n\epsilon}\right)\right). \tag{113}$$

where $q(D) = \frac{1}{n}\sum_{i=1}^n d_i$. This is in contradiction with Theorem E.5. We thus proved that $\left\|w^{priv} - w^*\right\| = \Omega(\beta)$, with probability at least $1/3$. As a consequence, we now obtain that with probability at least $1/3$,

$$F(w^{priv}; D) - F(w^*; D) = \frac{\|\sum_{i=1}^n d_i\|}{2\beta n} \left\|w^{priv} - w^*\right\|_2^2 = \Omega\left(\min\left(\|L\|_2 \beta, \frac{\beta \|L\|_2 \sqrt{p}}{n\epsilon}\right)\right), \tag{114}$$

which gives the desired result on the expectation of $F(w^{priv}; D) - F(w^*; D)$.

Finally, if we do not make any hypothesis on the L_j's distribution, we can directly use the non-augmented dataset constructed by Bun et al. [2014] to prove Lemma E.4 (that is the dataset from Theorem E.5, rescaled but not augmented). The ℓ_2-norm of the sum of this dataset is $\|\sum_{i=1}^n d_j\|_2 = [M'-1, M'+1]$ with $M' = \Omega\left(\min\left(\frac{L_{\min}}{L_{\max}} n \|L\|_2, \frac{L_{\min}}{L_{\max}} \frac{\sqrt{p}\|L\|_2}{\epsilon}\right)\right)$. This holds since four columns of this dataset out of five have sum of $\pm n L_j$ (for some j's), but no lower bound on the sum of the remaining columns can be derived. Thus, assuming (112) holds, then (113) can be rewritten as

$$\left\|\widetilde{\mathcal{A}}(D) - q(D)\right\|_2 = \frac{M'}{n\beta} \left\|w^{priv} - \frac{\beta}{M} \sum_{i=1}^n d_i\right\|_2 \neq \Omega\left(\frac{M'}{n}\right) = \Omega\left(\min\left(\frac{L_{\min}}{L_{\max}} \|L\|_2, \frac{L_{\min}}{L_{\max}} \frac{\|L\|_2 \sqrt{p}}{n\epsilon}\right)\right), \tag{115}$$

with probability at least $1/3$, which is in contradiction with Remark E.6. We thus get an additional factor of $L_{\min}/L_{\max}$ in the lower bound:

$$F(w^{priv}; D) - F(w^*; D) = \frac{\|\sum_{i=1}^n d_i\|}{2\beta n} \left\|w^{priv} - w^*\right\|_2^2 = \Omega\left(\min\left(\frac{L_{\min}}{L_{\max}} \|L\|_2 \beta, \frac{L_{\min}}{L_{\max}} \frac{\beta \|L\|_2 \sqrt{p}}{n\epsilon}\right)\right). \tag{116}$$

E.4 Lower Bound for Strongly-Convex Functions

To prove a lower bound for strongly-convex functions, we let $\mu_I > 0$, $L_1, \ldots, L_p > 0$, $\mathcal{W} = \prod_{j=1}^p [-\frac{L_j}{2\mu_I}, +\frac{L_j}{2\mu_I}]$ and $D = \{d_1, \ldots, d_n\} \in \prod_{j=1}^p \{\pm \frac{L_j}{2\mu_I}\}$. We consider the following problem, which fits in our setting:

$$w^* = \operatorname*{arg\,min}_{w \in \mathbb{R}^p} \left\{ F(w; D) = \frac{\mu_I}{2n} \sum_{i=1}^n \|w - d_i\|_2^2 + i_{\mathcal{W}}(w) \right\} \tag{117}$$

where $i_{\mathcal{W}}$ is the (separable) characteristic function of the set $\mathcal{W}$. The associated loss function $\ell(w; d_i) = \frac{\mu_I}{2} \|w - d_i\|_2^2$ is L-component-Lipschitz as, for $w \in \mathcal{W}$ and $j \in [p]$, the triangle inequality gives:

$$|\nabla_j \ell(w; d_i)| \leq \mu_I (|w_j| + |d_{i,j}|) \leq \mu_I \left(\frac{L_j}{2\mu_I} + \frac{L_j}{2\mu_I}\right) \leq L_j. \tag{118}$$

This loss is also μ_I-strongly convex *w.r.t.* ℓ_2-norm since for $w, w' \in \mathcal{W}$,

$$\ell(w; d_i) = \frac{\mu_I}{2} \|w - d_i\|_2^2 = \frac{\mu_I}{2} \|w' - d_i + w - w'\|_2^2 = \frac{\mu_I}{2} \left(\|w' - d_i\|_2^2 + 2\langle w' - d_i, w - w'\rangle + \|w - w'\|_2^2\right), \tag{119}$$

which is exactly μ_I-strong convexity since $\ell(w'; d_i) = \frac{\mu_I}{2} \|w' - d_i\|_2^2$ and $\nabla \ell(w'; d_i) = \mu_I(w - d_i)$. The minimum of the objective function in (117) is attained at $w^* = \frac{1}{n} \sum_{i=1}^n d_i = q(D) \in \mathcal{W}$. The excess risk of F is thus

$$F(w; D) - F(w^*) = \frac{\mu_I}{2n} \sum_{i=1}^n \|w - d_i\|_2^2 - \|w^* - d_i\|_2^2 \tag{120}$$

$$= \frac{\mu_I}{2n} \sum_{i=1}^n \|w\|^2 - \|w^*\|^2 + 2 \langle d_i, w^* - w \rangle \tag{121}$$

$$= \frac{\mu_I}{2} \|w\|^2 - \frac{1}{2} \|w^*\|^2 + \langle w^*, w^* - w \rangle \tag{122}$$

$$= \frac{\mu_I}{2} \|w - q(D)\|_2^2 . \tag{123}$$

It remains to apply Theorem E.5 to obtain that, with probability at least $1/3$,

$$F(w^{priv}; D) - F(w^*) = \Omega \left(\min \left(\frac{\|L\|_2^2}{\mu_I}, \frac{\|L\|_2^2 p}{\mu_I n^2 \epsilon^2} \right) \right), \tag{124}$$

which gives the lower bound on the expected value of $F(w^{priv}; D) - F(w^*)$. Note that without the additional assumption on the distribution of the L_j's, Remark E.6 directly gives the result with an additional multiplicative factor $(L_{\min}/L_{\max})^2$:

$$F(w^{priv}; D) - F(w^*) = \Omega \left(\min \left(\frac{L_{\min}^2}{L_{\max}^2} \frac{\|L\|_2^2}{\mu_I}, \frac{L_{\min}^2}{L_{\max}^2} \frac{\|L\|_2^2 p}{\mu_I n^2 \epsilon^2} \right) \right), \tag{125}$$

with probability at least $1/3$.

F Private Estimation of Smoothness Constants

In this section, we explain how a fraction ϵ' of the ϵ budget of DP can be used to estimate the coordinate-wise smoothness constants, which are essential to the good performance of DP-CD on imbalanced problems. Let f be defined as the average loss over the dataset D as in problem (1). We denote by $M_j^{(i)}$ the j-th component-smoothness constant of $\ell(\cdot, d_i)$, where d_i is the i-th point in D. The j-th smoothness constant of the function f is thus the average of all these constants: $M_j = \frac{1}{n} \sum_{i=1}^n M_j^{(i)}$.

Assuming that the practitioner knows an approximate upper bound b_j over the $M_j^{(i)}$'s, they can enforce it by clipping $M_j^{(i)}$ to b_j for each $i \in [n]$. The sensitivity of the average of the clipped $M_j^{(i)}$'s is thus $2b_j/n$. One can then compute an estimate of $M_1, \ldots, M_p$ under ϵ-DP using the Laplace mechanism as follows:

$$M_j^{priv} = \frac{1}{n} \sum_{i=1}^n \text{clip}(M_j^{(i)}, b_j) + \text{Lap} \left(\frac{2b_j p}{n\epsilon'} \right), \quad \text{for each } j \in [p], \tag{126}$$

where the factor p in noise scale comes from using the simple composition theorem Dwork and Roth [2014], and $\text{Lap}(\lambda)$ is a sample drawn in a Laplace distribution of mean zero and scale λ. The computed constant can then directly be used in DP-CD, allocating the remaining budget $\epsilon - \epsilon'$ to the optimization procedure.

G Additional Experimental Details and Results

G.1 Hyperparameter Tuning

DP-SGD and DP-CD both depend on three hyperparameters: step size, clipping threshold and number of passes on data. For DP-CD, step sizes are adapted from a parameter as described in Section 6, and clipping thresholds as well (see Section 5.1). For DP-SGD, the step size is given by γ/β, where γ is the hyperparameter and β is the problem's global smoothness constant (which we consider given), and the clipping threshold is used directly to clip gradients along their ℓ_2-norm.

We simultaneously tune these three hyperparameters for each algorithm across the following grid:

Table 2: Relative error to non-private optimal value of the objective function for different number of passes on the data. Results are reported for each dataset and for DP-CD and DP-SGD, after tuning step size and clipping hyperparameters. A star indicates the lowest error in each row.

	Passes on data	2	5	10	20	50
Electricity (imbalanced)	DP-CD	$0.1458 \pm$ 6e-04	$0.0842 \pm$ 1e-03	$0.0436 \pm$ 2e-03	$0.0147 \pm$ 2e-03	$0.0020 \pm$ 1e-03*
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.2047 \pm$ 2e-02	$0.1804 \pm$ 2e-02	$0.1766 \pm$ 2e-02	$0.1644 \pm$ 2e-02	$0.1484 \pm$ 1e-02*
Electricity (balanced)	DP-CD	$0.0186 \pm$ 4e-04	$0.0023 \pm$ 4e-04	$0.0013 \pm$ 6e-04*	$0.0013 \pm$ 4e-04	$0.0019 \pm$ 8e-04
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.0391 \pm$ 1e-02	$0.0189 \pm$ 5e-03	$0.0123 \pm$ 4e-03	$0.0106 \pm$ 3e-03	$0.0040 \pm$ 2e-03*
California (imbalanced)	DP-CD	$0.1708 \pm$ 7e-03	$0.1232 \pm$ 1e-02	$0.0598 \pm$ 1e-02	$0.0287 \pm$ 5e-03	$0.0124 \pm$ 7e-03*
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.2799 \pm$ 9e-02	$0.1863 \pm$ 2e-02	$0.1476 \pm$ 2e-02	$0.1094 \pm$ 2e-02	$0.1068 \pm$ 2e-02*
California (balanced)	DP-CD	$0.0007 \pm$ 3e-04*	$0.0011 \pm$ 6e-04	$0.0012 \pm$ 5e-04	$0.0010 \pm$ 1e-04	$0.0017 \pm$ 1e-03
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.0351 \pm$ 2e-02	$0.0226 \pm$ 8e-03	$0.0125 \pm$ 3e-03	$0.0087 \pm$ 2e-03	$0.0042 \pm$ 1e-03*
Sparse LASSO	DP-CD	$0.2498 \pm$ 4e-02*	$0.4702 \pm$ 9e-02	$0.5982 \pm$ 4e-02	$0.7160 \pm$ 2e-02	$0.7551 \pm$ 0e+00
$\epsilon = 10, \delta = 1/n^2$	DP-SGD	$0.7551 \pm$ 0e+00	$0.7551 \pm$ 3e-09*	$0.7551 \pm$ 0e+00	$0.7551 \pm$ 0e+00	$0.7551 \pm$ 0e+00

- step size: 10 logarithmically-spaced values between 10^{-6} and 1 for DP-SGD, and between 10^{-2} and 10 for DP-CD.[6]
- clipping threshold: 100 logarithmically-spaced values, between 10^{-3} and 10^{6}.
- number of passes: 5 values (2, 5, 10, 20 and 50).

We run each algorithm on each dataset 5 times on each combination of hyperparameter values. We then keep the set of hyperparameters that yield the lowest value of the objective at the last iterate, averaged across the 5 runs.

In Table 2, we report the best relative error (in comparison to optimal objective value) at the last iterate, averaged over five runs, for each dataset, algorithm, and total number of passes on the data. As such, each cell of this table corresponds to the best value obtained after tuning the step size and clipping hyperparameters for a given number of passes.

G.2 Running Time

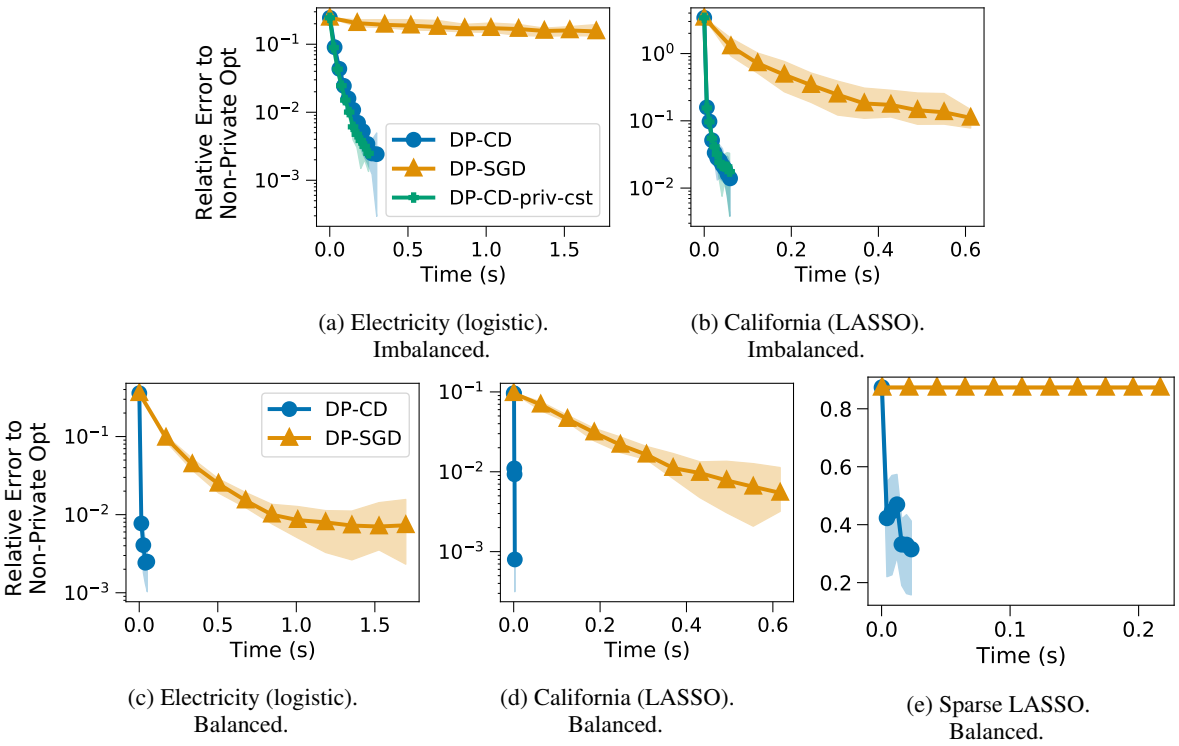
In this section, we report the running times of DP-CD and DP-SGD. We implemented DP-CD and DP-SGD in C++, with Python bindings. The design matrix and the labels are kept in memory as dense matrices of the Eigen library. No special code optimization nor tricks is applied to the algorithms, except for the update of residuals at each iteration of DP-CD, which prevents from accessing the complete dataset at each step.

Figure 3 shows the same experiments as in Figure 1 and Figure 2, but as a function of the running time. In our implementation, DP-CD runs about 4 times as fast as DP-SGD for a given number of iterations (see Figure 3a and Figure 3b for 50 iterations). On the three other plots, Figure 3c, Figure 3d and Figure 3e, DP-CD yields better results in less iterations. DP-CD is thus particularly valuable in these scenarios: combined with its faster running time, it provides accurate results extremely fast. For completeness, we provide in Table 3 the full table of running time, corresponding to Table 2 and Figure 3. These results show that, for a given number of passes on the data, DP-CD consistently runs about 5 times faster than DP-SGD.

Table 3: Time of execution (in seconds) for different number of passes on the data (averaged over 10 runs). Results are reported for each dataset and for DP-CD and DP-SGD, after tuning step size and clipping hyperparameters.

	Passes on data	2	5	10	20	50
Electricity (imbalanced)	DP-CD	$0.0128 \pm$ 1e-03	$0.0274 \pm$ 1e-03	$0.0500 \pm$ 1e-03	$0.0980 \pm$ 7e-04	$0.2457 \pm$ 2e-03
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.0663 \pm$ 2e-03	$0.1722 \pm$ 1e-02	$0.3321 \pm$ 1e-02	$0.6729 \pm$ 1e-02	$1.8588 \pm$ 2e-01
Electricity (balanced)	DP-CD	$0.0121 \pm$ 7e-04	$0.0281 \pm$ 3e-03	$0.0529 \pm$ 2e-03	$0.1062 \pm$ 6e-03	$0.2577 \pm$ 2e-03
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.0686 \pm$ 4e-03	$0.1768 \pm$ 1e-02	$0.3578 \pm$ 2e-02	$0.6787 \pm$ 2e-02	$1.6766 \pm$ 2e-02
California (imbalanced)	DP-CD	$0.0029 \pm$ 9e-05	$0.0065 \pm$ 8e-05	$0.0130 \pm$ 1e-04	$0.0258 \pm$ 1e-04	$0.0647 \pm$ 2e-04
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.0269 \pm$ 1e-03	$0.0665 \pm$ 1e-03	$0.1318 \pm$ 2e-03	$0.2628 \pm$ 3e-03	$0.6476 \pm$ 8e-03
California (balanced)	DP-CD	$0.0031 \pm$ 2e-04	$0.0065 \pm$ 2e-04	$0.0132 \pm$ 1e-04	$0.0262 \pm$ 2e-04	$0.0649 \pm$ 3e-04
$\epsilon = 1, \delta = 1/n^2$	DP-SGD	$0.0261 \pm$ 7e-04	$0.0641 \pm$ 5e-04	$0.1295 \pm$ 2e-03	$0.2592 \pm$ 4e-03	$0.6469 \pm$ 7e-03
Sparse LASSO	DP-CD	$0.0244 \pm$ 6e-04	$0.0760 \pm$ 6e-04	$0.1614 \pm$ 4e-03	$0.3213 \pm$ 5e-04	$0.6598 \pm$ 1e-02
$\epsilon = 10, \delta = 1/n^2$	DP-SGD	$0.0718 \pm$ 3e-03	$0.1788 \pm$ 4e-03	$0.3654 \pm$ 7e-03	$0.7292 \pm$ 2e-02	$1.8110 \pm$ 3e-02

[6] Recall that step sizes for CD algorithms are coordinate-wise, and thus larger than in SGD algorithms. We empirically verify that the best step size always lies strictly inside the considered interval for both DP-CD and DP-SGD.



(a) Electricity (logistic). Imbalanced.

(b) California (LASSO). Imbalanced.

(c) Electricity (logistic). Balanced.

(d) California (LASSO). Balanced.

(e) Sparse LASSO. Balanced.

Figure 3: Relative error to non-private optimal for DP-CD (blue, round marks), DP-CD with privately estimated coordinate-wise smoothness constants (green, + marks) and DP-SGD (orange, triangle marks) on five problems. We report average, minimum and maximum values over 10 runs for each algorithm, as a function of the algorithm running time (in seconds).