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. Differentially Private (DP) variants of optimization algorithms like Stochastic Gradient Descent (DP-SGD) have been designed to mitigate this, inducing a trade-off between privacy and utility. In this paper, we propose a new method for composite Differentially Private Empirical Risk Minimization (DP-ERM): Differentially Private proximal Coordinate Descent (DP-CD). We analyze its utility through a novel theoretical analysis of inexact coordinate descent, and highlight some regimes where DP-CD outperforms DP-SGD, thanks to the possibility of using larger step sizes. We also prove new lower bounds for composite DP-ERM under coordinate-wise regularity assumptions, that are, in some settings, nearly matched by our algorithm. In practical implementations, the coordinate-wise nature of DP-CD updates demands special care in choosing the clipping thresholds used to bound individual contributions to the gradients. A natural parameterization of these thresholds emerges from our theory, limiting the addition of unnecessarily large noise without requiring coordinate-wise hyperparameter tuning or extra computational cost.

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 $D = (d_1, \dots, d_n)$ 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 decreasing learning rates or variance reduction schemes to guarantee convergence, and they can be slow when gradients' coordinates are disparate. These flaws also hold for 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 expand this private optimization toolbox by proposing and analyzing a novel 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) ML due to their simplicity and effectiveness in high dimension (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 learning rates without variance reduction, and allow the use of larger learning rates that adapt to the coordinate-wise smoothness of the objective.

Despite these advantages, it is not obvious whether achieving a good privacy-utility trade-off is possible with CD methods. Indeed, they typically require more iterations than full gradient descent, but coordinate-wise updates do not systematically reduce the amount of perturbation needed to guarantee differential privacy. Nonetheless, through a novel and careful analysis of proximal CD with perturbed gradients, we derive upper bounds on the privacy-utility trade-off achieved by DP-CD, and show that it can outperform DP-SGD in regimes where some coordinate-wise gradients have lower sensitivity. Our analysis relies on a new recursion on distances of CD iterates to an optimal point, similarly to classical SGD analyses (Shamir and Zhang, 2013; Johnson and Zhang, 2013). This recursion allows to keep track of coordinate-wise regularity constants, which is crucial for achieving high utility by leveraging the use of large and constant learning rates. Incidentally, our result improves upon known convergence rates for inexact CD (Tappenden et al., 2016) with additive error that results from noisy gradients. We assess the optimality of DP-CD by extending known lower bounds to finer coordinate-wise Lipschitzness measures, and show that DP-CD matches those bounds in most settings. 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 suggest a simple rule for adapting these thresholds from a single hyperparameter. We provide illustrative numerical experiments which validate our theory and confirm that DP-CD is a suitable approach to 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 avoid costly coordinate-wise hyperparameter tuning, and show the relevance of DP-CD through 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 and establish utility guarantees. We compare these utility guarantees with those of DP-SGD, and propose an empirical rule for setting coordinate-wise clipping thresholds. In Section 4, we derive lower bounds under coordinate-wise regularity assumptions, and show that DP-CD can match them. Section 5 presents our numerical experiments, comparing DP-CD and DP-SGD on linear regression, LASSO and ℓ_2-regularized logistic regression. Finally, we discuss the relation to existing work in Section 6, and conclude with promising lines of future work in Section 7.

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 $M = \mathrm{diag}(M_1, \dots, M_p)$ with $M_1, \dots, M_p > 0$, and

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

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. Below and throughout the paper, we denote by $\nabla f(\cdot)$ the gradient of a differentiable function f, and by $\nabla_j f(\cdot)$ 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 $L = (L_1, \dots, L_p)$ with $L_1, \dots, L_p > 0$ 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 $M_1, \dots, M_p > 0$ 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 component-wise versions of these regularity hypotheses are not restrictive, as Λ-Lipschitzness implies $(\Lambda, \dots, \Lambda)$-component-Lipschitzness and β-smoothness implies $(\beta, \dots, \beta)$-component-smoothness. Crucially however, the actual component-wise constants of a function can much lower than what can be deduced from their global counterparts. This will play a major role in 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 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 Gaussian noise to $h(D)$ before releasing it (Dwork and Roth, 2013). 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 will prove to be central in our analysis:

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

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].

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

3 DP-CD FOR DP-ERM

In this section, we describe our main contribution: the Differentially Private proximal Coordinate Descent (DP-CD) algorithm to solve problem (1) under an (ϵ, δ)-DP constraint. We first describe our algorithm, show how to parameterize it to satisfy the desired privacy constraint, and give corresponding utility results. Then, we compare it with DP-SGD and exhibit practical strategies to set coordinate-wise clipping thresholds.

3.1 Private Proximal Coordinate Descent

Let $D = \{d_1, \dots, d_n\} \in \mathcal{X}^n$ 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 $F(w) := f(w) + \psi(w)$. Coordinate descent methods solve problem (1) by iteratively minimizing F along each of its coordinates. Exact minimization incurs a prohibitive privacy cost due to large sensitivity and/or multiple accesses to data. We thus rather minimize an upper bound of the function along one coordinate, which reveals less information and can be made private more efficiently. Such an upper bound follows from the M-component-smoothness of f and the separability of ψ. With $w \in \mathbb{R}^p$, $t \in \mathbb{R}$ and $j \in [p]$:

$$F(w + te_j) \leq f(w) + \nabla_j f(w) t + \frac{M_j}{2} t^2 + \psi_j(w_j + t) + \sum_{j' \neq j} \psi_{j'}(w_{j'}),$$

Minimizing this upper bound in t yields an update which is equivalent to a coordinate-wise proximal gradient step with learning rate $\gamma_j = \frac{1}{M_j}$. Formally, the updated j-th coordinate of w is given by

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

for $\mathrm{prox}_{\gamma_j \psi_j}(w) = \arg\min_{v \in \mathbb{R}^p} \big\{ \frac{1}{2}\|v - w\|_2^2 + \gamma_j \psi_j(v) \big\}$. We refer to Parikh and Boyd (2014) for details on proximal operators and related algorithms.

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 procedure is shown in Algorithm 1. 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 8). 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.

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

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

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

3.2 Privacy Guarantees

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

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

Sketch of Proof. We track privacy using zero concentrated differential privacy (zCDP), which gives better guarantees than classical DP for the composition of Gaussian mechanisms (Bun and Steinke, 2016). The j-th entry of ∇f has sensitivity $\Delta(\nabla_j f) = \Delta(\nabla_j \ell)/n \leq 2L_j/n$. Thus, by the Gaussian mechanism, setting $\sigma_j^2 = 4L_j^2 TK/n^2\rho$ ensures that each iteration of DP-CD is ρ/TK-zCDP. The composition theorem then guarantees ρ-zCDP of the complete procedure. We finish the proof by converting this guarantee to DP. The complete proof is provided in Appendix B. □

The dependence of the noise scales on ϵ, δ, n and on TK (the number of updates) given in Theorem 1 is standard in DP-ERM. However, the noise is calibrated to *component*-Lipschitz constants $\{L_j\}_{j=1}^p$ of f, contrary to SGD-style algorithms where the noise depends on the global Lipschitz constant of f. This is a crucial difference: component-Lipshitz constants can be much lower than the global one, and we will see that this compensates for the larger number of iterations (and thus larger noise) typically needed by CD algorithms.

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 terms. Non-asymptotic utility bounds can be found in Appendix C.

Theorem 2. *Let $\ell(\cdot; d)$ be a convex, L-component-Lipschitz, M-component-smooth loss function for all $d \in \mathcal{X}$, and $\psi : \mathbb{R}^p \to \mathbb{R}$ be a convex and separable function. Let $\epsilon < 1, \delta < 1/3$ 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 learning rates $\gamma_j = \frac{1}{M_j}$, T and K set as stated below and noise scales set as in Theorem 1 to ensure (ϵ, δ)-DP. Then, the following holds:*

1. *If F is convex, $K = O\left(R_M \sqrt{p} n\epsilon / \|L\|_{M^{-1}}\right)$ 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 $R_M = \max(\sqrt{F(w^0) - F(w^*)}, \|w^0 - w^*\|_M)$ and more simply $R_M = \|w^0 - w^*\|_M$ when $\psi = 0$.

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

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

Expectations are over the randomness of the algorithm.

Sketch of Proof. Contrary to classical analyses of CD algorithms, we prove a recursion that focuses on $\mathbb{E}\|\theta^k - w^*\|_M^2$, rather than $\mathbb{E}[F(\theta^k) - F(w^*)]$:

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

This inequality reflects the idea that coordinate-wise updates leave a fraction $\frac{p-1}{p}$ of the function "unchanged", while the remaining part decreases (up to additive noise). When summing (4) for $k = 0, \dots, K-1$, its left hand side simplifies and its right hand side is a telescoping sum with additive noise that accumulates:

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

where $\bar{w}^t$ appears since $\theta^0 = \bar{w}^t$. As F is convex, $F(\bar{w}^{t+1}) - F^* \leq \frac{1}{K}\sum_{k=1}^K F(\theta^k) - F^*$, thus the inner loop converges sublinearly (up to an additive noise term). This is in fact the result in the convex case (since $T = 1$, only one inner loop is run). For strongly convex F, we have $\mathbb{E}\|\bar{w}^t - w^*\|_M^2 \leq \frac{2}{\mu_M}\mathbb{E}[F(\bar{w}^t) - F(w^*)]$. Replacing in (5) with large enough K gives

$$\mathbb{E}[F(\bar{w}^{t+1}) - F^*] \leq \tfrac{1}{2}\mathbb{E}[F(\bar{w}^t) - F^*] + \|\sigma\|_{M^{-1}}^2,$$

and linear convergence follows (up to an additive noise term). The full proof is given in Appendix C. □

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

Our utility guarantees stated in Theorem 2 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 then show that Theorem 2 suggests values for scaling coordinate-wise clipping thresholds using a single hyperparameter, substantially easing practical implementations. We will discuss the optimality of our utility guarantees in Section 4.

3.4 Comparison with DP-SGD

We now compare more finely 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 2, 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 $\frac{\|L\|^2_{M^{-1}}}{\mu_M}$ with $\frac{\Lambda^2}{\mu_I}$ 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^*.

Balanced. When the smoothness constants M are all equal, $\|L\|_{M^{-1}} R_M = \|L\|_2 R_I$ and $\frac{\|L\|^2_{M^{-1}}}{\mu_M} = \frac{\|L\|_2^2}{\mu_I}$. The comparison thus boils down to comparing $\|L\|_2$ with Λ. As $\Lambda \leq \|L\|_2 \leq \sqrt{p}\Lambda$, DP-CD can be up to p times worse than DP-SGD. Indeed, the coordinate-wise noise is calibrated to the coordinate-wise regularity of the objective, and is thus oblivious to the potential dependences between features. As a result, when features are extremely correlated, DP-CD can end up adding overly large noise to each coordinate-wise gradient.

Unbalanced. More favorable regimes exist when smoothness constants are disparate. To illustrate this, we consider the setting where the first coordinate of ℓ dominates others. That is $M_1 =: M_{max}$ (resp. $L_1 =: L_{max}$) $\gg M_j =: M_{min}$ (resp. $L_j =: L_{min}$) for all $j \neq 1$, so that L_1^2/M_1 dominates the other terms of $\|L\|^2_{M^{-1}}$. This yields $\|L\|^2_{M^{-1}} \approx L_1^2/M_1 \approx \Lambda/M_{max}$, and $\mu_M = \mu_I M_{min}$. Moreover, assume that the first coordinate of w^* is already well estimated by w^0, so that $R_M \approx M_{min} R_I$. We obtain that $\|L\|_{M^{-1}} R_M \approx \sqrt{\frac{M_{min}}{M_{max}}} \Lambda R_I$ for convex losses and $\frac{\|L\|_{M^{-1}}}{\mu_M} \approx \frac{M_{min}}{M_{max}} \frac{\Lambda^2}{\mu_I}$ 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 learning rate to each coordinate. As such, DP-CD converges much quicker than DP-SGD on coordinates with smaller-scale gradients, requiring fewer accesses to the dataset, and in turn less noise addition overall.

We give more details in Appendix D and provide experimental evidence supporting our conclusions in a variety of settings in Section 5.

3.5 Coordinate-wise Gradient Clipping

In this section, we go back to inequality (2) and discuss gradient sensitivities directly. In practice, Lipschitz constants can indeed give pessimistic estimates of sensitivities, making gradients overly noisy, thereby hurting convergence. To address this issue, common practice in DP-SGD is to clip per-sample gradients at a fixed threshold $C > 0$ on their ℓ_2-norm (Abadi et al., 2016):

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

This effectively ensures that the sensitivity of the clipped gradient is bounded, *i.e.,* $\Delta(\mathrm{clip}(\nabla \ell, C)) \leq 2C$. In coordinate descent, gradients are released one coordinate at a time. Clipping in DP-CD must thus be done coordinate-wise, which requires setting p coordinate-wise thresholds $\{C_j\}_{j=1}^p$. Uniform thresholds $C_1 = \cdots = C_p = C$ would be oblivious to coordinate-wise gradient sensitivities, whereas tuning them individually is impractical. Instead, we can leverage our utility results from Theorem 2 to adapt them using a single hyperparameter. We first remark that these are invariant to the scale of the matrix M. We thus

Table 1: Utility guarantees for DP-CD and DP-SGD for L-component-Lipschitz, Λ-Lipschitz loss.

	Convex	Strongly-convex
DP-CD (this paper)	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\|L\|_{M^{-1}} R_M\right)$	$\widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\|L\|_{M^{-1}}^2}{\mu_M}\right)$
DP-SGD (Bassily et al., 2014) DP-SVRG (Wang et al., 2017)	$\widetilde{O}\left(\frac{\sqrt{p\log(1/\delta)}}{n\epsilon}\Lambda R_I\right)$	$\widetilde{O}\left(\frac{p\log(1/\delta)}{n^2\epsilon^2}\frac{\Lambda^2}{\mu_I}\right)$

rescale M as proposed by Richtárik and Takáč (2014), using matrix $\widetilde{M} = \frac{p}{\mathrm{tr}(M)}M$ so that $\mathrm{tr}(\widetilde{M}) = \mathrm{tr}(I) = p$. The key quantity that appears in our utility bounds with the rescaled matrix $\widetilde{M}$ is now $\Delta_{\widetilde{M}^{-1}}(\nabla\ell)$. This suggests choosing $C_j = \sqrt{\frac{M_j}{\mathrm{tr}(M)}}C$, for some parameter $C > 0$, which gives

$$\Delta_{\widetilde{M}^{-1}}(\{\mathrm{clip}(\nabla_j\ell, C_j)\}_{j=1}^p) \leq 2C.$$

The parameter C thus controls the overall sensitivity for DP-CD, allowing it to perform p iterations for the same privacy budget as one update of clipped DP-SGD.

4 LOWER BOUNDS

We now prove a new lower bound on the error achievable for composite DP-ERM with L-component-Lipschitz loss functions, thereby extending the results of Bassily et al. (2014) for Λ-Lipschitz losses. Deriving these lower bounds requires to adapt the worst-case objectives used in the proof of Bassily et al. (2014) to our unconstrained setting. It also involves revisiting 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 (see Appendix E for details).

Theorem 3. *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 $\lceil \frac{p}{75} \rceil$, $\sum_{j\in\mathcal{J}} L_j^2 = \Omega(\|L\|_2^2)$. 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:*

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

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

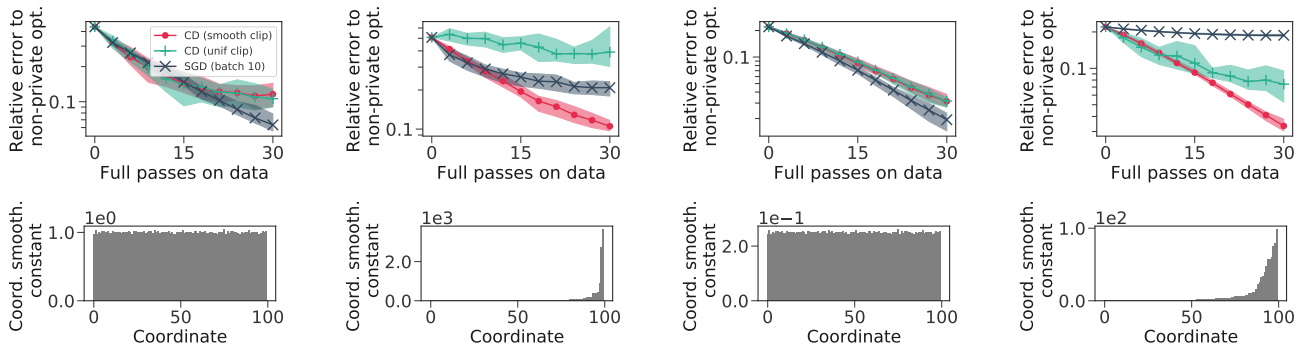
$$\mathbb{E}\big[F(w^{priv}; D) - F(w^*)\big] = \Omega\Big(\frac{p\,\|L\|_2^2}{\mu_I n^2\epsilon^2}\Big).$$

We recover the lower bounds of Bassily et al. (2014) for Λ-Lipschitz losses as a special case of ours by setting $L_1 = \cdots = L_p = \Lambda/\sqrt{p}$. In this case, the loss function used in the proof is indeed $(\sum_{j=1}^p L_j^2)^{1/2} = \Lambda$-Lipschitz. To relate these lower bounds with DP-CD, we consider a suboptimal version of our algorithm, where we set the learning rates to $\gamma_1 = \cdots = \gamma_p = (\max_j M_j)^{-1}$. In this setting, results from Theorem 2 still hold, and match the lower bounds from Theorem 3 up to logarithmic factors. We leave open the question of 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 3 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 much looser. Although the above assumption may seem solely technical, we conjecture that better utility is possible when a small number of coordinate-wise Lipschitz constants dominate the others. We discuss this perspective further in Section 7.

5 NUMERICAL EXPERIMENTS

In this section, we illustrate the practical performance of our DP-CD algorithm by comparing it with DP-SGD (with minibatches of 10 records) on LASSO (Tibshirani, 1996) and ℓ_2-regularized logistic regression. To ensure privacy, both algorithms use the Gaussian mechanism with appropriate noise, as described in Section 3.2. For DP-SGD, we also use amplification by subsampling. The privacy budget is fixed to $\epsilon = 1$, $\delta = 1/n^2$, which is generally regarded as providing good privacy guarantees. We report the loss for each iterate without periodic averaging, as we observe that this performs slightly better empirically. Hyperparameters are tuned by grid search, choosing the values that yield the lowest average loss on the training set. As we aim to compare both algorithms at their best, we do not account for tuning in our privacy budget. We also stress the fact that while smoothness constants could leak information, they can be estimated using global, non-confidential bounds/statistics on the data features. Experiments are run on a computer running Debian 11, with 16GB RAM and Intel Core i7-10610U CPU. Algorithms are implemented in C++, with Python bindings.



(a) LASSO, $\lambda = 50$, balanced smoothness.

(b) LASSO, $\lambda = 100$, unbalanced smoothness.

(c) Logistic with ℓ_2-reg, balanced smoothness.

(d) Logistic with ℓ_2-reg, unbalanced smoothness.

Figure 1: Comparison of DP-CD and DP-SGD on different problems and datasets. The top row reports the evolution of the relative error (with respect to the optimal non-private loss) across iterations, with average, minimum and maximum values over 10 runs. The bottom row reports coordinate-wise smoothness constants.

LASSO. Let $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}$, $\ell(w;(x,y)) = \frac{1}{2}(x^\top w - y)^2$ and $\psi = \lambda \|w\|_1 = \lambda \sum_{j=1}^p |w_j|$, where λ is high enough to enforce sparsity of the optimal solution (values of λ are given in Figure 1). For DP-SGD, we use standard gradient clipping, as described in (6). For DP-CD we use our coordinate-wise rule described in Section 3.5 with $C_j = \sqrt{M_j / \operatorname{tr}(M)}C$ (referred to as `CD (smooth clip)`), as well as the uniform clipping rule $C_j = \frac{C}{\sqrt{p}}$ (`CD (unif clip)`), with $C > 0$ to be tuned. Each algorithm is run for 30 passes over the full dataset and the experiment is repeated over 10 different random seeds. We report the mean, minimum and maximum of the relative error to the optimal (non-private) solution over the 10 random runs with chosen hyperparameters.

We generate two datasets of $n = 10,000$ records with $p = 100$ features independently drawn from a standard Gaussian distribution. We then generate noisy labels from a linear predictor perturbed with small Gaussian noise. In the first dataset, we keep the features balanced, so that the loss has balanced smoothness constants. In the second one, we rescale features so that smoothness constants follow a lognormal distribution, which we expect to resemble real-world datasets (Chmiel et al., 2021) while remaining close to the case discussed in Section 3.4. Figure 1a shows results on the first dataset. Here, the two clipping rules for DP-CD are equivalent and perform slightly worse than DP-SGD. Figure 1b reports results on the second dataset: there, DP-CD with uniform clipping suffers from its large learning rates, as these also amplify the final amount of noise. Remarkably, our clipping rule from Section 3.5 avoids this pitfall and improves over DP-SGD.

Logistic regression. We now set $\ell(w;(x,y)) = \log(1+\exp(-yx^\top w))$ and $\psi = \lambda \|w\|_2^2$. The design matrix X is generated as in the LASSO case. Labels y are computed from a linear predictor, and assigned to ± 1 depending on a fixed threshold. They are then independently flipped with probability 0.2. We set λ to $\frac{1}{n}$, following classical machine learning guidelines. Results are shown in Figures 1c and 1d, and are similar to the ones of the LASSO: in the balanced setting, the two clipping rules for DP-CD are equivalent, and a little worse than DP-SGD. In the unbalanced setting, choosing the appropriate clipping rule improves convergence, and DP-CD largely outperforms DP-SGD.

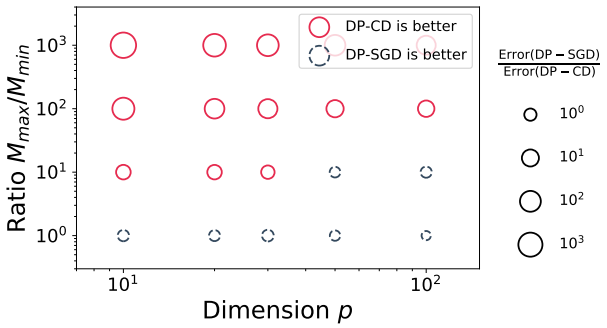


Figure 2: Comparison of DP-CD (with clipping rule $C_j = \sqrt{M_j / \operatorname{tr}(M)}$) and DP-SGD on linear regression in function of the dimension and the ratio M_{max}/M_{min}, with tuned hyperparameters. Circle sizes give the ratio of the optimization error of DP-SGD over the one of DP-CD, averaged over 10 random runs.

Dimension and conditioning. In this experiment, we aim to numerically reproduce the results from Section 3.4, in the setting where one parameter dominates. To this end, we generate multiple datasets of $n = 10,000$ records, with one larger smoothness constant, where we vary the dimension p and the ratio $\frac{M_{max}}{M_{min}}$. We consider a linear regression problem

(*i.e.,* quadratic loss and $\psi = 0$) on each dataset and run DP-SGD and DP-CD on each of them, tuning hyperparameters and reporting the mean optimization error for the last iterate over 10 runs. In Figure 2, we plot the ratio of the error of DP-SGD over that of DP-CD. As predicted by our theoretical analysis, the advantage of DP-CD increases as $\frac{M_{max}}{M_{min}}$ increases. When the latter is large, DP-SGD suffers from small learning rates and is unable to get close to the optimal. This effect is mitigated as the dimension increases, which is due to the fact that coordinate-wise clipping is oblivious to correlated gradient coordinates, which makes the amount of noise added by DP-CD increase faster with the dimension than the one added by DP-SGD.

6 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 for non-smooth objectives. Wang et al. (2017) obtained faster convergence rate using a differentially private version of the SVRG algorithm (Johnson and Zhang, 2013; Xiao and Zhang, 2014). Owing to the simplicity and popularity of SGD in machine learning, 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 our DP-CD algorithm, 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) was the first to prove a global non-asymptotic $1/k$ convergence rate for CD with random choice of coordinates for a convex, smooth objective. Parallel, proximal variants of CD were developed by Richtárik and Takáč (2014); Fercoq and Richtárik (2015), while Hanzely et al. (2018) further considered non-separable non-smooth parts. See Wright (2015) or 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 not easily compatible with DP. We give tighter results for inexact CD with noisy gradients (see Remark 1). Dual CD algorithms were introduced in (Shalev-Shwartz and Zhang, 2013) for smooth ERM, with performance similar to SVRG.

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. Our work shows 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 specific (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.

7 CONCLUSION AND DISCUSSION

We presented the first differentially private proximal coordinate descent algorithm for composite DP-ERM. We showed theoretically and experimentally that DP-CD strongly outperforms DP-SGD when gradients coordinates are disparate. Notably, the choice of coordinate-wise clipping thresholds is crucial for DP-CD to achieve good utility, and we provided a simple rule to set them. We believe that adaptive clipping techniques (Pichapati et al., 2019; Thakkar et al., 2021) may help to further improve the performance of DP-CD, for instance when gradients coordinates are more balanced.

We also derived novel lower bounds under a component-Lipschitzness assumption, and showed that DP-CD matches these bounds in some settings. Although DP-CD already achieves good utility when most coordinates have small sensitivity, our lower bounds suggest that even better utility could be achieved in such cases by dynamically allocating more privacy budget to the coordinates with largest sensitivities. This is in line with recent work on DP-SGD with a subspace assumption (Zhou et al., 2021; Kairouz et al., 2021). A promising direction for DP-CD is to 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).

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

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, Oct. 2016. Association for Computing Machinery. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318.

H. Asi, V. Feldman, T. Koren, and K. Talwar. Private stochastic convex optimization: Optimal rates in ℓ_1 geometry. In *ICML*, volume 139, pages 393–403, 2021.

R. Bassily, A. Smith, and A. 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, Oct. 2014. IEEE. ISBN 978-1-4799-6517-5. doi: 10.1109/FOCS.2014.56.

R. Bassily, C. Guzman, and A. Nandi. Non-Euclidean Differentially Private Stochastic Convex Optimization. In *COLT*, pages 474–499. PMLR, July 2021.

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

M. Bun and T. Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In M. Hirt and A. Smith, editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 635–658, Berlin, Heidelberg, 2016. Springer. ISBN 978-3-662-53641-4. doi: 10.1007/978-3-662-53641-4_24.

M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, page 10, 2014.

K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. Coordinate Descent Method for Large-scale L2-loss Linear Support Vector Machines. *Journal of Machine Learning Research*, 9:1369–1398, June 2008. ISSN 1532-4435.

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011. ISSN 1533-7928.

B. Chmiel, L. Ben-Uri, M. Shkolnik, E. Hoffer, R. Banner, and D. Soudry. Neural gradients are near-lognormal: improved quantized and sparse training. In *ICLR*, 2021.

G. Damaskinos, C. Mendler-Dünner, R. Guerraoui, N. Papandreou, and T. Parnell. Differentially private stochastic coordinate descent. In *AAAI*, pages 7176–7184. AAAI Press, 2021.

M. De Santis, S. Lucidi, and F. Rinaldi. A fast active set block coordinate descent algorithm for ℓ_1-regularized least squares. *SIAM J. Optim.*, 26(1):781–809, Jan. 2016. ISSN 1052-6234. doi: 10.1137/141000737.

C. Dwork. Differential Privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 1–12, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-35908-1. doi: 10.1007/11787006_1.

C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4): 211–407, 2013. ISSN 1551-305X, 1551-3068. doi: 10.1561/0400000042.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-32732-5. doi: 10.1007/11681878_14.

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

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1): 1–22, 2010. ISSN 1548-7660.

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

F. Hanzely, D. Kovalev, and P. 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.

R. Johnson and T. 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.

P. Kairouz, M. R. Diaz, K. Rush, and A. Thakurta. (Nearly) Dimension Independent Private ERM with AdaGrad Rates via Publicly Estimated Subspaces. In M. Belkin and S. Kpotufe, editors, *COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 2717–2746. PMLR, 2021.

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

A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1):501–546, July 2016. ISSN 1436-4646. doi: 10.1007/s10107-015-0943-9.

H. Liu, M. Palatucci, and J. 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. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553458.

Z.-Q. Luo and P. 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.

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

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

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

N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, Jan. 2014. ISSN 2167-3888. doi: 10.1561/2400000003.

V. Pichapati, A. T. Suresh, F. X. Yu, S. J. Reddi, and S. Kumar. AdaCliP: Adaptive Clipping for Private SGD. *arXiv:1908.07643 [cs, stat]*, Oct. 2019.

P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, Apr. 2014. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-012-0614-z.

S. Sardy, A. G. Bruce, and P. Tseng. Block Coordinate Relaxation Methods for Nonparametric Wavelet Denoising. *Journal of Computational and Graphical Statistics*, 9(2):361–379, June 2000. ISSN 1061-8600. doi: 10.1080/10618600.2000.10474885.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, Feb. 2013. ISSN 1532-4435.

O. Shamir and T. Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes. In *ICML*, pages 71–79. PMLR, Feb. 2013.

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

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, May 2017. doi: 10.1109/SP.2017.41.

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

R. Tappenden, P. Richtárik, and J. Gondzio. Inexact Coordinate Descent: Complexity and Preconditioning. *J. Optim. Theory Appl.*, 170(1):144–176, July 2016. ISSN 0022-3239, 1573-2878. doi: 10.1007/s10957-016-0867-4.

O. Thakkar, G. Andrew, and H. B. McMahan. Differentially Private Learning with Adaptive Clipping. In *Advances in Neural Information Processing Systems*, 2021.

R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1996.tb02080.x.

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

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

T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2014.2320500.

D. Wang, M. Ye, and J. 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.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, June 2015. ISSN 1436-4646. doi: 10.1007/s10107-015-0892-3.

L. Xiao and T. Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. *SIAM J. Optim.*, 24(4):2057–2075, Jan. 2014. ISSN 1052-6234. doi: 10.1137/140961791.

G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin. A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification. *Journal of Machine Learning Research*, 11:3183–3234, Dec. 2010. ISSN 1532-4435.

Y. Zhou, S. Wu, and A. 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 1 that (coordinate) gradients are bounded by (coordinate-wise-)Lipschitz constants. We then link this upper bound with gradients' sensitivities in Lemma 2.

Lemma 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 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 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 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 1. *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 1

To track the privacy loss of an adaptive composition of K Gaussian mechanisms, we use zero Concentrated Differential Privacy (zCDP). This flavor of differential privacy, tailored for the Gaussian mechanism, 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, zCDP 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 Concentrated Differential Privacy

We will use the Rényi divergence (Definition 2), which gives a distribution-oriented vision of privacy.

Definition 2 (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 $\alpha > 1$,*

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

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

Definition 3 (Zero-Concentrated Differential Privacy, Bun and Steinke 2016)**.** *A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is (ξ, ρ)-zero-concentrated differentially private (zCDP) if, for all $\alpha > 1$ and all datasets $D, D' \in \mathcal{D}$ differing on at most on element,*

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

If $\xi = 0$, the algorithm is said ρ-zero-concentrated differentially private or simply ρ-zCDP.

Lemma 3 (Bun and Steinke 2016, Proposition 1.3)**.** *If a randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is ρ-zCDP, then it is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-differentially private.*

Corollary 2. *Let $0 < \epsilon \leq 1, 0 < \delta < \frac{1}{3}$. A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ that satisfies $\frac{\epsilon^2}{6\log(1/\delta)}$-zCDP also satisfies (ϵ, δ)-DP.*

Proof. From Lemma 3 with $\rho = \frac{\epsilon^2}{6\log(1/\delta)}$ it holds that $\mathcal{A}$ is (ϵ', δ)-DP with

$$\epsilon' = \frac{\epsilon^2}{6\log(1/\delta)} + 2\sqrt{\frac{\epsilon^2}{6}} \leq (1/6 + \sqrt{2/3})\epsilon \leq \epsilon, \tag{16}$$

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 $1/6 + \sqrt{2/3} \approx 0.983 < 1$. □

We can now restate the composition theorem of zCDP (Theorem 4), which is key in designing private iterative algorithms.

Theorem 4 (Bun and Steinke 2016, Lemma 2.3)**.** *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 (ξ_k, ρ_k)-zCDP, 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 $\left(\sum_{k=1}^K \xi_k, \sum_{k=1}^K \rho_k\right)$-zCDP.*

Finally, we define the Gaussian mechanism (Definition 4), as used in Algorithm 1, and restate in Lemma 4 the privacy guarantees that it satisfies in terms of zCDP.

Definition 4 (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 4 (Bun and Steinke 2016, Lemma 2.5)**.** *For $\sigma^2 = \frac{\Delta(f)^2}{2\rho}$, the Gaussian mechanism is ρ-zCDP.*

B.2 Proof of Theorem 1

We are now ready to prove Theorem 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 5, which gives an (ϵ, δ)-differential privacy guarantee for the adaptive composition of K Gaussian mechanisms.

Lemma 5. *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. Lemma 4 guarantees that the k-th Gaussian mechanism with noise $\sigma_k^2 = \frac{\Delta(f_k)^2 K}{2\rho}$ is $\frac{\rho}{K}$-zCDP. Then, the composition of K such mechanisms is, according to Theorem 4, ρ-zCDP.

This can be converted to DP using Corollary 2: the composition of these mechanisms is (ϵ, δ)-DP for $\rho = \frac{\epsilon^2}{6\log(1/\delta)}$, which gives, after replacing ρ by this value, for $k \in [K]$,

$$\sigma_k^2 = \frac{\Delta(f_k)^2 K}{2\rho} = \frac{6\Delta(f_k)^2 K \log(1/\delta)}{2\epsilon^2} = \frac{3\Delta(f_k)^2 K \log(1/\delta)}{\epsilon^2}.$$

Thus, to ensure (ϵ, δ)-DP, one needs to set noise scale $\sigma_k = \frac{\Delta(f_k)\sqrt{3K\log(1/\delta)}}{\epsilon}$. □

We now restate Theorem 1 and prove it.

Theorem 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 5 and the post-processing property of DP, Algorithm 1 is (ϵ, δ)-differentially private. □

C Proof of Utility (Theorem 2)

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:

$$w^* \in \operatorname*{arg\,min}_{w \in \mathbb{R}^p} \left\{ F(w; D) := \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(w; d_i)}_{=: f(w; D)} + \psi(w) \right\}, \tag{18}$$

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 2

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 6, 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 (18).

Lemma 6. *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 $\Gamma = \mathrm{diag}(\gamma_1, \dots, \gamma_p)$. It holds that*

$$\begin{aligned}&\mathbb{E}_j[F(\theta - \gamma_j g_j e_j) - F(w)|\theta] - \frac{p-1}{p}(F(\theta) - F(w)) \\ &\quad\leq \frac{1}{p}\left(f(\theta) - f(w) + \langle \nabla f(\theta), -\Gamma g\rangle + \frac{1}{2}\|\Gamma g\|_M^2 + \psi(\theta - \Gamma g) - \psi(w)\right). \end{aligned} \tag{19}$$

Remark 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:

$$\mathbb{E}_j[F(\theta - \gamma_j g_j e_j) - F(w)|\theta] = \sum_{j=1}^{p} \frac{1}{p}(F(\theta - \gamma_j g_j) - F(w)) \tag{20}$$

$$\overset{F=f+\psi}{=} \frac{1}{p}\sum_{j=1}^{p} f(\theta - \gamma_j g_j) - f(w) + \psi(\theta - \gamma_j g_j) - \psi(w) \tag{21}$$

$$\overset{f \text{ smooth}}{\leq} \frac{1}{p}\sum_{j=1}^{p}\left(f(\theta) + \langle \nabla f(\theta), -\gamma_j g_j\rangle + \frac{1}{2}\|\gamma_j g_j\|_M^2 - f(w) + \psi(\theta - \gamma_j g_j) - \psi(w)\right) \tag{22}$$

$$= f(\theta) - f(w) + \frac{1}{p}\sum_{j=1}^{p}\left(\langle \nabla f(\theta), -\gamma_j g_j\rangle + \frac{1}{2}\|\gamma_j g_j\|_M^2 + (\psi(\theta - \gamma_j g_j) - \psi(w))\right) \tag{23}$$

$$= f(\theta) - f(w) + \frac{1}{p}\langle \nabla f(\theta), -\Gamma g\rangle + \frac{1}{2p}\|\Gamma g\|_M^2 + \frac{1}{p}\sum_{j=1}^{p}(\psi(\theta - \gamma_j g_j) - \psi(w)). \tag{24}$$

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

$$\sum_{j=1}^{p}(\psi(\theta - \gamma_j g_j) - \psi(w)) = \sum_{j=1}^{p}\left(\psi_j(\theta_j - \gamma_j g_j) - \psi_j(w_j) + \sum_{j' \neq j}\psi_{j'}(\theta_{j'}) - \psi(w_{j'})\right) \tag{25}$$

$$= \psi(\theta - \Gamma g) - \psi(w) + (p-1)(\psi(\theta) - \psi(w)). \tag{26}$$

Plugging (26) in (24) results in the following:

$$\begin{aligned}\mathbb{E}_j[F(\theta-\gamma_j g_j e_j)-F(w)|\theta] &\leq f(\theta)-f(w)+\frac{1}{p}\langle\nabla f(\theta),-\Gamma g\rangle+\frac{1}{2p}\|\Gamma g\|_M^2 \\ &\quad+\frac{1}{p}(\psi(\theta-\Gamma g)-\psi(w))+\frac{p-1}{p}(\psi(\theta)-\psi(w)) \qquad (27)\\ &=\frac{1}{p}\left(f(\theta)-f(w)+\langle\nabla f(\theta),-\Gamma g\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\psi(\theta-\Gamma g)-\psi(w)\right)\\ &\quad+\frac{p-1}{p}(f(\theta)+\psi(\theta)-f(w)-\psi(w)) \qquad (28)\end{aligned}$$

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

To exploit this result, we need to upper bound the right hand side of (19) 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 (19) so as to obtain telescopic terms plus a bias term resulting from the addition of noise, as shown in Lemma 7.

Lemma 7. *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 learning rates (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*

$$\mathbb{E}_{j,\eta}\big[F(\theta^{k+1})-F(w)\big|\theta^k\big]-\frac{p-1}{p}(F(\theta^k)-F(w))\leq\tfrac{1}{2}\left\|\theta^k-w\right\|_{\Gamma^{-1}}^2-\tfrac{1}{2}\mathbb{E}_{j,\eta}\Big[\left\|\theta^{k+1}-w\right\|_{\Gamma^{-1}}^2\Big|\theta^k\Big]+\tfrac{1}{p}\|\sigma\|_\Gamma^2, \quad (29)$$

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 $\Gamma=\mathrm{diag}(\gamma_1,\dots,\gamma_p)$ the diagonal matrix having the learning rates as its coefficients.

From Lemma 6 with $\theta=\theta^k$, $w=w$ and $g=g$ as defined above we obtain

$$\begin{aligned}&\mathbb{E}_j\big[F(\theta^k-\gamma_j g_j e_j)-F(w)\big|\theta^k\big]-\frac{p-1}{p}(F(\theta^k)-F(w))\\ &\quad\leq\frac{1}{p}\left(f(\theta^k)-f(w)+\big\langle\nabla f(\theta^k),-\Gamma g\big\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\psi(\theta^k-\Gamma g)-\psi(w)\right). \qquad (30)\end{aligned}$$

We can upper bound the right hand term of (30) using the convexity of f and ψ:

$$\begin{aligned}&f(\theta^k)-f(w)+\big\langle\nabla f(\theta^k),-\Gamma g\big\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\psi(\theta^k-\Gamma g)-\psi(w) \qquad (31)\\ &\quad\leq\big\langle\nabla f(\theta^k),\theta^k-w\big\rangle+\big\langle\nabla f(\theta^k),-\Gamma g\big\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\big\langle\partial\psi(\theta^k-\Gamma g),\theta^k-\Gamma g-w\big\rangle \qquad (32)\\ &\quad=\big\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g),\theta^k-\Gamma g-w\big\rangle+\frac{1}{2}\|\Gamma g\|_M^2, \qquad (33)\end{aligned}$$

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:

$$\begin{aligned}&\big\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g),\theta^k-\Gamma g-w\big\rangle+\frac{1}{2}\|\Gamma g\|_M^2 \qquad (34)\\ &\quad=\big\langle g,\theta^k-\Gamma g-w\big\rangle+\frac{1}{2}\|\Gamma g\|_M^2+\big\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g)-g,\theta^k-\Gamma g-w\big\rangle \qquad (35)\\ &\quad=\underbrace{\big\langle g,\theta^k-w\big\rangle-\|g\|_\Gamma^2+\frac{1}{2}\|g\|_{\Gamma^2M}^2}_{\text{“descent” term}}+\underbrace{\big\langle\nabla f(\theta^k)+\partial\psi(\theta^k-\Gamma g)-g,\theta^k-\Gamma g-w\big\rangle}_{\text{“noise” term}}, \qquad (36)\end{aligned}$$

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 (36) 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 learning rates Γ^{-1}:

$$\begin{aligned}
\text{"descent" term} &= \langle g, \theta^k - w\rangle - \frac{1}{2}\|g\|_\Gamma^2 && (37)\\
&= \langle \Gamma g, \theta^k - w\rangle_{\Gamma^{-1}} - \frac{1}{2}\|\Gamma g\|_{\Gamma^{-1}}^2 && (38)\\
&= \frac{1}{2}\|\theta^k - w\|_{\Gamma^{-1}}^2 - \frac{1}{2}\|\theta^k - w\|_{\Gamma^{-1}}^2 + \langle \Gamma g, \theta^k - w\rangle_{\Gamma^{-1}} - \frac{1}{2}\|\Gamma g\|_{\Gamma^{-1}}^2 && (39)\\
&= \frac{1}{2}\|\theta^k - w\|_{\Gamma^{-1}}^2 - \frac{1}{2}\|\theta^k - \Gamma g - w\|_{\Gamma^{-1}}^2, && (40)
\end{aligned}$$

where we factorized the norm to obtain the last inequality. We can rewrite (40) 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):

$$\begin{aligned}
\frac{1}{2}\|\theta^k - w\|_{\Gamma^{-1}}^2 - \frac{1}{2}\|\theta^k - \Gamma g - w\|_{\Gamma^{-1}}^2 &= \frac{p}{2} \times \left(\frac{1}{p}\sum_{j=1}^p \gamma_j^{-1}\left|\theta_j^k - w_j\right|^2 - \gamma_j^{-1}\left|\theta_j^k - \gamma_j g_j - w_j\right|^2\right) && (41)\\
&= \frac{p}{2} \times \mathbb{E}_j\Big[\gamma_j^{-1}\left|\theta_j^k - w_j\right|^2 - \gamma_j^{-1}\left|\theta_j^k - \gamma_j g_j - w_j\right|^2 \Big| \theta^k, \eta\Big]. && (42)
\end{aligned}$$

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

$$\begin{aligned}
\text{"descent" term} &= \mathbb{E}_j\Big[\frac{p}{2}\|\theta^k - w\|_{\Gamma^{-1}}^2 - \frac{p}{2}\|\theta^k - \gamma_j g_j - w\|_{\Gamma^{-1}}^2 \Big| \theta^k, \eta\Big] && (43)\\
&= \frac{p}{2}\|\theta^k - w\|_{\Gamma^{-1}}^2 - \frac{p}{2}\mathbb{E}_j\Big[\|\theta^{k+1} - w\|_{\Gamma^{-1}}^2 \Big| \theta^k, \eta\Big]. && (44)
\end{aligned}$$

Upper bounding the "noise" term. We now upper bound the "noise" term in (36). We first recall the definition of the noisy proximal update g_j (line 7 of Algorithm 1), and define its non-noisy counterpart $\tilde{g}_j$:

$$\begin{aligned}
g_j &= \gamma_j^{-1}\Big(\operatorname{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) - \theta_j^k\Big) && (45)\\
\tilde{g}_j &= \gamma_j^{-1}\Big(\operatorname{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k)) - \theta_j^k\Big). && (46)
\end{aligned}$$

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:

$$\begin{aligned}
0 &\in \theta_j^{k+1} - \theta_j^k + \gamma_j(\nabla_j f(\theta^k) + \eta_j)) + \frac{1}{M_j}\partial\psi_j(\theta_j^k - \gamma_j g_j) && (47)\\
&= \gamma_j \times \left(\frac{1}{\gamma_j}(\theta_j^{k+1} - \theta_j^k) + \nabla_j f(\theta^k) + \eta_j + \partial\psi_j(\theta_j^k - \gamma_j g_j)\right). && (48)
\end{aligned}$$

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 (36) can be thus be rewritten using v:

$$\text{"noise" term} = \langle \nabla f(\theta^k) + v - g, \theta^k - \Gamma g - w\rangle = \langle \eta, \theta^k - \Gamma g - w\rangle, \tag{49}$$

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

$$\text{"noise" term} = \sum_{j=1}^p \eta_j(\theta_j^k - \gamma_j g_j - w_j) = \sum_{j=1}^p \eta_j(\theta_j^k - \gamma_j \widetilde{g}_j - w_j) + \sum_{j=1}^p \eta_j(\gamma_j \widetilde{g}_j - \gamma_j g_j). \tag{50}$$

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

$$\mathbb{E}_\eta\Big[\sum_{j=1}^p \eta_j(\theta_j^k - \gamma_j\widetilde{g}_j - w_j) \mid \theta^k\Big] = \sum_{j=1}^p \mathbb{E}_\eta[\eta_j]\,(\theta_j^k - \gamma_j\widetilde{g}_j - w_j) = 0. \tag{51}$$

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

$$\mathbb{E}_\eta\Big[\sum_{j=1}^p \eta_j(\gamma_j\widetilde{g}_j - \gamma_j g_j) \mid \theta^k\Big] \le \mathbb{E}_\eta\Big[\gamma_j\Big|\sum_{j=1}^p \eta_j(\widetilde{g}_j - g_j)\Big| \mid \theta^k\Big] \le \sum_{j=1}^p \gamma_j\mathbb{E}_\eta\big[\ |\eta_j|\,|\widetilde{g}_j - g_j|\ \big|\ \theta^k\big]\,, \tag{52}$$

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

$$|\widetilde{g}_j - g_j| = \gamma_j^{-1}\left|\mathrm{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k))) - \mathrm{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j))\right| \le |\eta_j|\,, \tag{53}$$

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

$$\sum_{j=1}^p \gamma_j\mathbb{E}_\eta\big[|\eta_j|\,|\widetilde{g}_j - g_j|\big|\theta^k\big] \le \sum_{j=1}^p \gamma_j\mathbb{E}_\eta[|\eta_j|\,|\eta_j|] = \sum_{j=1}^p \gamma_j\mathbb{E}_\eta[\eta_j^2] = \|\sigma\|_\Gamma^2\,. \tag{54}$$

We now have everything to prove the lemma by plugging (54) and (51) into expected value of (50), and then (50) and (40) back into (36) to obtain, after using the Tower property of conditional expectations:

$$\frac{1}{p}\mathbb{E}_{j,\eta}\Big[f(\theta^k) - f(w) + \big\langle\nabla f(\theta^k), -\Gamma g\big\rangle + \frac{1}{2}\|\Gamma g\|_M^2 + \psi(\theta^k - \Gamma g) - \psi(w)\Big|\theta^k\Big] \tag{55}$$

$$\le \frac{1}{p}\,(\text{“descent” term} + \text{“noise” term}) \tag{56}$$

$$\le \frac{1}{2}\big\|\theta^k - w\big\|_{\Gamma^{-1}}^2 - \frac{1}{2}\mathbb{E}_{j,\eta}\Big[\big\|\theta^{k+1} - w\big\|_{\Gamma^{-1}}^2\Big|\theta^k\Big] + \frac{1}{p}\|\sigma\|_\Gamma^2\,, \tag{57}$$

which is the result of the lemma. □

C.2.2 Convergence Lemma

Lemma 7 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 8, which is the key lemma of our proof.

Lemma 8. *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 $\theta^1, \dots, \theta^K$ computed from the inner loop of Algorithm 1 starting from the point $\bar{w}^t$, with learning rates $\gamma_j = \frac{1}{M_j}$ and noise scales σ_j. Letting $\bar{w}^{t+1} = \frac{1}{K}\sum_{k=1}^K \theta^k$, it holds that*

$$\mathbb{E}\big[F(\bar{w}^{t+1}) - F(w^*)\big] \le \frac{p(\|\bar{w}^t - w^*\|_M^2 + 2(F(\bar{w}^t) - F(w^*)))}{2K} + \|\sigma\|_{M^{-1}}^2\,. \tag{58}$$

Remark 3. *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*

$$f(\bar{w}^t) \le f(w^*) + \big\langle\nabla f(w^*), \bar{w}^t - w^*\big\rangle + \frac{1}{2}\big\|\bar{w}^t - w^*\big\|_M^2 = f(w^*) + \frac{1}{2}\big\|\bar{w}^t - w^*\big\|_M^2\,, \tag{59}$$

and the result of Lemma 8 further simplifies as:

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

Proof. Summing Lemma 7 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:

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\big[F(\theta^{k+1}) - F(w^*)\big] - \frac{p-1}{p} \sum_{k=0}^{K-1} \mathbb{E}\big[(F(\theta^k) - F(w^*))\big] \\
&\quad \leq \sum_{k=0}^{K-1} \frac{1}{2}\mathbb{E}\Big[\big\|\theta^k - w^*\big\|^2_{\Gamma^{-1}}\Big] - \frac{1}{2}\mathbb{E}\Big[\big\|\theta^{k+1} - w^*\big\|^2_{\Gamma^{-1}}\Big] + \frac{1}{p}\left\|\sigma\right\|^2_{\Gamma} \qquad (61) \\
&\quad = \frac{1}{2}\mathbb{E}\Big[\big\|\bar{w}^0 - w^*\big\|^2_{\Gamma^{-1}}\Big] - \frac{1}{2}\mathbb{E}\Big[\big\|\theta^K - w^*\big\|^2_{\Gamma^{-1}}\Big] + \frac{K}{p}\left\|\sigma\right\|^2_{\Gamma}. \qquad (62)
\end{aligned}$$

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 (62):

$$\sum_{k=0}^{K-1} \mathbb{E}\big[F(\theta^{k+1}) - F(w^*)\big] - \tfrac{p-1}{p} \sum_{k=0}^{K-1} \mathbb{E}\big[(F(\theta^k) - F(w^*))\big] \geq \tfrac{1}{p} \sum_{k=1}^{K} \mathbb{E}\big[F(\theta^k) - F(w^*)\big] - (F(\bar{w}^0) - F(w^*)). \quad (63)$$

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 (63) and combining the result with (62) gives

$$F(\bar{w}^{t+1}) - F(w^*) \leq \frac{p(\frac{1}{2}\left\|\bar{w}^0 - w^*\right\|^2_{\Gamma^{-1}} + F(\bar{w}^0) - F(w^*))}{K} + \left\|\sigma\right\|^2_{\Gamma}. \qquad (64)$$

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

C.2.3 Convex Case

Theorem 2 (Convex case)**.** *Let w^* be a minimizer of F and $R_M^2 = \max(\left\|\bar{w}^0 - w^*\right\|^2_M, F(\bar{w}^0) - F(w^*))$. 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 1, satisfies:*

$$F(w^{priv}) - F(w^*) \leq \frac{3pR_M^2}{2K} + \frac{12\left\|L\right\|^2_{M^{-1}} K \log(1/\delta)}{n^2\epsilon^2}. \qquad (65)$$

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

$$F(w^{priv}) - F(w^*) \leq \frac{9\sqrt{p}\left\|L\right\|_{M^{-1}} R_M\sqrt{\log(1/\delta)}}{n\epsilon} = \widetilde{O}\left(\frac{\sqrt{p}R_M\left\|L\right\|_{M^{-1}}}{n\epsilon}\right). \qquad (66)$$

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 8 with $\bar{w}^{t+1} = \bar{w}^1$, $w^t = \bar{w}^0$ and σ_j chosen as in Theorem 1 gives the result. Taking $K = \frac{R_M\sqrt{p}n\epsilon}{\left\|L\right\|_{M^{-1}}\sqrt{8\log(1/\delta)}}$ then gives

$$F(\bar{w}_1^{t+1}) - F(w^*) \leq \frac{2\sqrt{8p\log(1/\delta)}\left\|L\right\|_{M^{-1}} R_M}{n\epsilon} + \frac{12\sqrt{p\log(1/\delta)}\left\|L\right\|_{M^{-1}} R_M}{\sqrt{8}n\epsilon}, \qquad (67)$$

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

C.2.4 Strongly Convex Case

Theorem 2 (Strongly-convex case)**.** *Let F be μ_M-strongly convex w.r.t. $\left\|\cdot\right\|_M$ 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 1, satisfies:*

$$F(w^{priv}) - F(w^*) \leq \frac{F(\bar{w}^0) - F(w^*)}{2^T} + \frac{24p(1 + 1/\mu_M)T\left\|L\right\|^2_{M^{-1}}\log(1/\delta)}{n^2\epsilon^2}. \qquad (68)$$

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

$$\mathbb{E}\big[F(w^{priv}) - F(w^*)\big] \leq \left(1 + \log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n^2\epsilon^2}{24p(1+1/\mu_M)\|L\|_{M^{-1}}^2\log(1/\delta)}\right)\right)\frac{24p(1+1/\mu_M)\|L\|_{M^{-1}}^2\log(1/\delta)}{n^2\epsilon^2} \tag{69}$$

$$= O\left(\frac{p\|L\|_{M^{-1}}^2\log(1/\delta)}{\mu_M n^2\epsilon^2}\log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n\epsilon\mu_M}{p\|L\|_{M^{-1}}\log(1/\delta)}\right)\right) \tag{70}$$

Proof. As F is μ_M-strongly-convex with respect to norm $\|\cdot\|_M$, we obtain for any $w \in \mathbb{R}^p$, that $F(w) \geq F(w^*) + \frac{\mu_M}{2}\|w - w^*\|_M^2$. Therefore, $F(\bar{w}^0) - F(w^*) \leq \frac{2}{\mu_M}\left\|\bar{w}^0 - w^*\right\|_M^2$ and Lemma 8 gives, for $1 \leq t \leq T-1$,

$$F(\bar{w}^{t+1}) - F(w^*) \leq \frac{(1+1/\mu_M)p(F(\bar{w}^t) - F(w^*))}{K} + \|\sigma\|_M^2\,. \tag{71}$$

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

$$F(\bar{w}^{t+1}) - F(w^*) \leq \frac{F(\bar{w}^t) - F(w^*)}{2} + \|\sigma\|_M^2\,. \tag{72}$$

Recursive application of this inequality gives

$$\mathbb{E}\big[F(\bar{w}^T) - F(w^*)\big] \leq \frac{F(\bar{w}^0) - F(w^*)}{2^T} + \sum_{t=0}^{T-1}\frac{1}{2^t}\|\sigma\|_M^2 \leq \frac{F(\bar{w}^0) - F(w^*)}{2^T} + 2\|\sigma\|_M^2\,, \tag{73}$$

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

$$\mathbb{E}\big[F(\bar{w}^T) - F(w^*)\big] \leq \left(1 + \log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n^2\epsilon^2}{24p(1+1/\mu_M)\|L\|_{M^{-1}}^2\log(1/\delta)}\right)\right)\frac{24p(1+1/\mu_M)\|L\|_{M^{-1}}^2\log(1/\delta)}{n^2\epsilon^2} \tag{74}$$

$$= O\left(\frac{p\|L\|_{M^{-1}}^2\log(1/\delta)}{\mu_M n^2\epsilon^2}\log_2\left(\frac{(F(\bar{w}^0) - F(w^*))n\epsilon\mu_M}{p\|L\|_{M^{-1}}\log(1/\delta)}\right)\right), \tag{75}$$

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 $V_j(\theta, t) = \nabla_j(\theta)t + \frac{M_j}{2}|t|^2 + \psi_j(\theta_j^k + t)$. For $\eta \in \mathbb{R}$, we also define its noisy counterpart, $V_j^\eta(\theta, t) = (\nabla_j(\theta) + \eta)t + \frac{M_j}{2}|t|^2 + \psi_j(\theta_j^k + t)$. We aim at finding δ_j such that for any $\theta^k \in \mathbb{R}^p$ used in the inner loop of Algorithm 1:

$$\mathbb{E}_{\eta_j}\big[V_j(\theta^k, -\gamma_j g_j)\big] \leq \min_{\widetilde{g}\in\mathbb{R}} V_j(\theta^k, -\gamma_j\widetilde{g}) + \delta_j, \tag{76}$$

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

$$\mathrm{prox}_{\gamma_j\psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) = \arg\min_{v\in\mathbb{R}}\frac{1}{2}\left\|v - \theta_j^k + \gamma_j(\nabla_j f(\theta^k) + \eta_j)\right\|_2^2 \tag{77}$$

$$= \arg\min_{v\in\mathbb{R}}\left\langle\gamma_j(\nabla_j f(\theta_j^k) + \eta_j), v - \theta_j^k\right\rangle + \frac{1}{2}\left\|v - \theta_j^k\right\|_2^2 + \gamma_j\psi_j(v) \tag{78}$$

$$= \arg\min_{v\in\mathbb{R}}\left\langle\nabla_j f(\theta^k) + \eta_j, v - \theta_j^k\right\rangle + \frac{M_j}{2}\left\|v - \theta_j^k\right\|_2^2 + \psi_j(v) \tag{79}$$

$$= \theta_j^k + \arg\min_{t\in\mathbb{R}}\left\langle\nabla_j f(\theta^k) + \eta_j, t\right\rangle + \frac{M_j}{2}\|t\|_2^2 + \psi_j(\theta_j^k + t). \tag{80}$$

Which means that $-\gamma_j g_j = \mathrm{prox}_{\gamma_j \psi_j}(\theta_j^k - \gamma_j(\nabla_j f(\theta^k) + \eta_j)) - \theta_j^k \in \arg\min_{t\in\mathbb{R}} V_j^{\eta_j}(\theta^k, t)$. Let $-\gamma_j g_j^* = \mathrm{prox}_{\gamma_j \psi_j}(\theta_j^k - \gamma_j \nabla_j(\theta^k)) - \theta_j^k$ 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

$$V_j^{\eta_j}(\theta^k, -\gamma_j g_j) \leq \left\langle \nabla_j f(\theta^k) + \eta_j, -\gamma_j g_j^* \right\rangle + \frac{M_j}{2} \left\| -\gamma_j g_j^* \right\|_2^2 + \psi_j(\theta_j^k + -\gamma_j g_j^*) \tag{81}$$

$$= \min_t V_j(\theta^k, t) + \left\langle \eta_j, -\gamma_j g_j^* \right\rangle, \tag{82}$$

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

$$\mathbb{E}_{\eta_j}\big[V_j(\theta^k, -\gamma_j g_j)\big] \leq \min_t V_j(\theta^k, t) + \mathbb{E}_{\eta_j}\big[\left\langle \eta_j, \gamma_j(g_j - g_j^*) \right\rangle\big]. \tag{83}$$

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

$$\mathbb{E}_{\eta_j}\big[V_j(\theta^k, -\gamma_j g_j)\big] \leq \min_t V_j(\theta^k, t) + \gamma_j \sigma_j^2, \tag{84}$$

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 $\beta = \frac{1}{p}\left\|\sigma\right\|_{M^{-1}}^2$.

Convex functions. When the objective function F is convex, we use Lemma 8 to obtain, since $\left\|\sigma\right\|_{M^{-1}}^2 = \beta p$,

$$F(w^1) - F(w^*) \leq \frac{2pR_M^2}{K} + \left\|\sigma\right\|_{M^{-1}}^2 = \frac{2pR_M^2}{K} + \beta p. \tag{85}$$

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 $\left\|\cdot\right\|_M$, then from (73) we obtain, as long as $K \geq 4/\mu_M$, that

$$\mathbb{E}\big[F(w^T) - F(w^*)\big] \leq \frac{F(w^0) - F(w^*)}{2^T} + 2\beta p. \tag{86}$$

This proves that $\mathbb{E}\big[F(w^T) - F(w^*)\big] \leq \xi$ 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 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

$$\left\|L\right\|_{M^{-1}} = \sqrt{\sum_{j=1}^p \frac{1}{M_j} L_j^2} = \sqrt{\frac{1}{M}\sum_{j=1}^p L_j^2} = \frac{1}{\sqrt{M}} \left\|L\right\|_2. \tag{87}$$

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 $x, y \in \mathbb{R}^p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_M}{2} \|y - x\|_M^2 = f(x) + \langle \nabla f(x), y - x \rangle + \frac{M\mu_M}{2} \|y - x\|_2^2 , \tag{88}$$

  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

$$\|L\|_{M^{-1}}^2 = \sum_{j=1}^p \frac{L_j^2}{M_j} \approx \frac{L_1^2}{M_1} \approx \frac{\Lambda^2}{M_{max}}. \tag{89}$$

- *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,

$$R_M^2 \approx \sum_{j \neq 1} M_j \left|w_j^0 - w_j^*\right|^2 = M_{min} \sum_{j \neq 1} \left|w_j^0 - w_j^*\right|^2 \approx M_{min} R_I^2. \tag{90}$$

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 $x, y \in \mathbb{R}^p$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_M}{2} \|y - x\|_M^2 \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M_{min}\mu_M}{2} \|y - x\|_2^2 , \tag{91}$$

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 $\frac{M_{max}}{M_{min}}$ increases.

The two hypotheses we describe above are of course very restrictive. However, it gives the intuition as to when and why DP-CD outperforms DP-SGD. Our numerical experiments in Section 5 confirm this analysis, even in some less favorable cases. In practice, our approximation $\|L\|_{M^{-1}}^2 \approx \frac{\Lambda^2}{M_{max}}$ can be quite loose, particularly as the dimension grows. Indeed, the accumulation of smaller scaled coordinates, where DP-CD uses the large learning rate $1/M_{min}$, causes it to add large amounts of noise on many potentially insignificant coordinates. In contrast, DP-SGD almost ignores these coordinates, due to its small learning rate, which can improve utility when these coordinates have a negligible impact on the objective. We show numerically in Section 5 that the rule we proposed to set coordinate-wise clipping thresholds for DP-CD helps to mitigate this effect.

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 $\left\|\sum_{i=1}^{n} d_i\right\|_2 = \Omega(\min(n\left\|L\right\|_2, \sqrt{p}\left\|L\right\|_2/\epsilon))$, 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 5) and accuracy (Definition 6), 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 5 (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 $\mathcal{D}$:*

$$q(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} q(d_i). \tag{92}$$

Definition 6 (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 $D \in \mathcal{X}^n$,*

$$\Pr\left[\mathcal{A}(D) \text{ is } (\alpha,\beta)\text{-accurate for } \mathcal{Q}\right] \geq 2/3. \tag{93}$$

In our proof, we will use a specific class of queries: one-way marginals (Definition 7), that compute the arithmetic mean of a dataset along one of its column.

Definition 7 (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 9 (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 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 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*

$$q_j(D_L^{aug}) = \frac{2L_j}{3} q_j(D) + \frac{L_j}{3}, \tag{94}$$

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

$$q_j(D^{aug}) = \frac{1}{3n} \sum_{i=1}^{3n} D_{i,j}^{aug} = \frac{1}{3} \left(\frac{1}{n} \sum_{i=1}^{n} D_{i,j}^{aug} \right) + \frac{1}{3n} \sum_{i=n+1}^{3n} 1 = \frac{1}{3} q_j(D) + \frac{2}{3} \in [0,1]. \tag{95}$$

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

$$q_j(D_L^{aug}) = \frac{1}{3n} \sum_{i=1}^{3n} (D_L^{aug})_{i,j} = \frac{1}{3n} \sum_{i=1}^{3n} 2L_j((D^{aug})_{i,j} - 1/2) = 2L_j(q_j(D^{aug}) - 1/2) \in [-L_j, L_j], \tag{96}$$

combining (95) and (96) gives the result. □

Theorem 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 $\mathcal{A}$:*

$$\|\mathcal{A}(D) - q(D)\|_2 = \Omega\left(\min\left(\|L\|_2, \frac{\sqrt{p}\,\|L\|_2}{n\epsilon}\right)\right). \tag{97}$$

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 $n \leq n^*$. 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

$$\Pr\left[\exists \mathcal{J} \subseteq [p] \text{ such that } |\mathcal{J}| \geq (1-\beta)p \text{ and } \forall j \in \mathcal{J},\ |\mathcal{A}_j(D') - q_j(D')| < \frac{2L_j}{3}\alpha\right] \geq 2/3. \tag{98}$$

Importantly, for all $D \in (\{0,1\})^p)^n$, the randomized algorithm $\mathcal{A}$ satisfies (98) 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

$$\left|\widetilde{\mathcal{A}}(D) - q_j(D)\right| = \left|\frac{3}{2L_j}\mathcal{A}_j(D_L^{aug}) - \frac{L_j}{3} - \frac{3}{2L_j}q_j(D_L^{aug}) + \frac{L_j}{3}\right| = \frac{3}{2L_j}\left|\mathcal{A}_j(D_L^{aug}) - q_j(D_L^{aug})\right|. \tag{99}$$

Therefore, if $\mathcal{A}$ satisfies (98) and (99), then $\widetilde{\mathcal{A}} : (\{0,1\}^p)^n \to [0,1]^p$ satisfies, for all $D \in (\{0,1\}^p)^n$,

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

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 $n \leq n^*$, which contradicts the result of Lemma 9 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 $|\mathcal{J}| \geq \lceil \beta p \rceil$ such that

$$\left\|\mathcal{A}(D_L^{aug}) - q(D_L^{aug})\right\|_2 \overset{(98)}{\geq} \sqrt{\sum_{j\in\mathcal{J}} \frac{4L_j^2}{9}} \geq \Omega(\|L\|_2), \tag{101}$$

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

$$\left\|\mathcal{A}(D_L^{aug}) - q(D_L^{aug})\right\|_2 \overset{(98)}{\geq} \sqrt{\sum_{j\in\mathcal{J}} \frac{4L_j^2}{9}} \geq \sqrt{\frac{4}{9\times 75}}\|L\|_2 \geq \frac{2}{27}, \tag{102}$$

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 4. *Without the assumption on the distribution of the L_j's, we can still get an inequality that resembles* (101)*: $\left\|\mathcal{A}(D_L^{aug}) - q(D_L^{aug})\right\|_2 \overset{(98)}{\geq} \sqrt{\sum_{j\in\mathcal{J}} \frac{4L_j^2}{9}} \geq \frac{2}{27}\frac{L_{min}}{L_{max}}\|L\|_2$, with probability at least* $1/3$*, and we get a result similar to Theorem 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 $\psi(w) = \frac{\|\sum_{i=1}^n d_i\|_2}{\beta n}\|w\|_2^2$:

$$w^* = \operatorname*{arg\,min}_{w\in\mathbb{R}^p} \left\{ F(w;D) := -\frac{1}{n}\langle w, \textstyle\sum_{i=1}^n d_i\rangle + \frac{\|\sum_{i=1}^n d_i\|_2}{\beta n}\|w\|_2^2 \right\}, \tag{103}$$

To find the solution of (103), we look for w^* so that the objective's gradient is zero, that is

$$w^* = \frac{\beta}{\|\sum_{i=1}^n d_i\|_2} \sum_{i=1}^n d_i, \tag{104}$$

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

$$F(w;D) - F(w^*;D) = -\frac{1}{n}\left\langle w - w^*, \sum_{i=1}^n d_i\right\rangle + \frac{\left\|\sum_{i=1}^n d_i\right\|}{2\beta n}(\|w\|_2^2 - \|w^*\|_2^2) \tag{105}$$

$$= -\frac{1}{n}\left\langle w - w^*, \frac{\left\|\sum_{i=1}^n d_i\right\|}{\beta} w^*\right\rangle + \frac{\left\|\sum_{i=1}^n d_i\right\|}{2\beta n}(\|w\|_2^2 - \|w^*\|_2^2) \tag{106}$$

$$= \frac{\left\|\sum_{i=1}^n d_i\right\|}{\beta n}\left(\langle w^* - w, w^*\rangle + \frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|w^*\|_2^2\right) \tag{107}$$

$$= \frac{\left\|\sum_{i=1}^n d_i\right\|}{\beta n}\left(-\langle w, w^*\rangle + \frac{1}{2}\|w\|_2^2 + \frac{1}{2}\|w^*\|_2^2\right) \tag{108}$$

$$= \frac{\left\|\sum_{i=1}^n d_i\right\|}{2\beta n}\|w - w^*\|_2^2 . \tag{109}$$

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 (103). Suppose, for the sake of contradiction, that for every dataset D with $\|\sum_{i=1}^n d_i\|_2 \in [M-1; M+1]$,

$$\left\|w^{priv} - w^*\right\| \neq \Omega(\beta), \text{ with probability at least } 2/3. \tag{110}$$

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,

$$\left\|\widetilde{\mathcal{A}}(D) - q(D)\right\|_2 = \frac{M}{n\beta}\left\|w^{priv} - \frac{\beta}{M}\sum_{i=1}^n d_i\right\|_2 \neq \Omega\left(\frac{M}{n}\right) = \Omega\left(\min\left(\|L\|_2, \frac{\|L\|_2\sqrt{p}}{n\epsilon}\right)\right). \tag{111}$$

where $q(D) = \frac{1}{n}\sum_{i=1}^n d_i$. This is in contradiction with Theorem 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,

$$F(w^{priv};D) - F(w^*;D) = \frac{\left\|\sum_{i=1}^n d_i\right\|}{2\beta n}\left\|w^{priv} - w^*\right\|_2^2 = \Omega\left(\min\left(\|L\|_2\beta, \frac{\beta\|L\|_2\sqrt{p}}{n\epsilon}\right)\right), \tag{112}$$

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 9 (that is the dataset from Theorem 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 $\pm nL_j$ (for some j's), but no lower bound on the sum of the remaining columns can be derived. Thus, assuming (110) holds, then (111) can be rewritten as

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

with probability at least 1/3, which is in contradiction with Remark 4. We thus get an additional factor of L_{min}/L_{max} in the lower bound:

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

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:

$$w^* = \operatorname*{arg\,min}_{w\in\mathbb{R}^p}\left\{F(w;D) := \frac{\mu_I}{2n}\sum_{i=1}^n\|w - d_i\|_2^2 + i_{\mathcal{W}}(w)\right\} \tag{115}$$

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:

$$|\nabla_j \ell(w; d_i)| \leq \mu_I(|w_j| + |d_{i,j}|) \leq \mu_I \left(\frac{L_j}{2\mu_I} + \frac{L_j}{2\mu_I} \right) \leq L_j. \tag{116}$$

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

$$\ell(w; d_i) = \frac{\mu_I}{2} \|w - d_i\|_2^2 = \frac{\mu_I}{2} \|w' - d_i + w - w'\|_2^2 = \frac{\mu_I}{2} \left(\|w' - d_i\|_2^2 + 2 \langle w' - d_i, w - w' \rangle + \|w - w'\|_2^2 \right), \tag{117}$$

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 (115) 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

$$\begin{aligned}
F(w; D) - F(w^*) &= \frac{\mu_I}{2n} \sum_{i=1}^n \|w - d_i\|_2^2 - \|w^* - d_i\|_2^2 && (118)\\
&= \frac{\mu_I}{2n} \sum_{i=1}^n \|w\|^2 - \|w^*\|^2 + 2 \langle d_i, w^* - w \rangle && (119)\\
&= \frac{\mu_I}{2} \|w\|^2 - \frac{1}{2} \|w^*\|^2 + \langle w^*, w^* - w \rangle && (120)\\
&= \frac{\mu_I}{2} \|w - q(D)\|_2^2 . && (121)
\end{aligned}$$

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

$$F(w^{priv}; D) - F(w^*) = \Omega \left(\min \left(\frac{\|L\|_2^2}{\mu_I}, \frac{\|L\|_2^2 p}{\mu_I n^2 \epsilon^2} \right) \right), \tag{122}$$

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 4 directly gives the result with an additional multiplicative factor $(L_{min}/L_{max})^2$:

$$F(w^{priv}; D) - F(w^*) = \Omega \left(\min \left(\frac{L_{min}^2}{L_{max}^2} \frac{\|L\|_2^2}{\mu_I}, \frac{L_{min}^2}{L_{max}^2} \frac{\|L\|_2^2 p}{\mu_I n^2 \epsilon^2} \right) \right), \tag{123}$$

with probability at least 1/3.