OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Differentially Private Random Block Coordinate Descent

Artavazd Maranjyan HTTPS://ARTOMARANJYAN.GITHUB.IO/
Abdurakhmon Sadiev
Peter Richtárik
Center of Excellence for Generative AI, King Abdullah University of Science and Technology (KAUST) & SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI)
Thuwal, Kingdom of Saudi Arabia

Abstract

Due to their effectiveness in solving high-dimensional problems and their ability to decompose complex optimization problems, coordinate Descent (CD) methods have garnered significant interest in machine learning in the last decade. However, classical CD methods were not designed nor analyzed with any data privacy considerations in mind, while these are increasingly critical in managing sensitive information. This gap recently led to the development of differentially private CD methods, such as DP-CD proposed by [9]. Despite this progress, there remains a disparity between non-private CD and DP-CD methods. Our work proposes differentially private random block coordinate descent that allows for the selection of multiple coordinates with varying probabilities in each iteration using sketch matrices.

1. Introduction

Recently, there has been increased interest in Coordinate Descent (CD) methods because of their various applications in machine learning, where these methods are generally applicable to a variety of problems involving large or high-dimensional datasets. They naturally break down complicated optimization problems into simpler subproblems, which are easily parallelized or distributed [17, 20]. For many problems, these methods are state-of-the-art [5, 8, 11–13, 15, 16].

These coordinate descent methods are studied without considering the privacy part of the data. However, in machine learning, managing data that often contains sensitive or confidential information is a pressing challenge, as highlighted by [18]. The concept of differential privacy (DP) has become a fundamental strategy in addressing this issue. A standard and widely accepted approach for training models while controlling information leakage involves solving an empirical risk minimization (ERM) problem under DP constraints, as described by Chaudhuri et al. [2].

Recently, [9] proposed a DP version of coordinate descent called DP-CD. But there is still a gap between the non-private coordinate descent methods and their DP-CD method. In this work, we are doing the first steps in closing the gap where methods are allowed to select multiple coordinates with different probabilities in each iteration, as opposed to the method by [9], which selects a single coordinate in each iteration.

Our work is dedicated to designing a differentially private algorithm that approximates the solution to a ERM problem, as represented by the following equation:

$$x^\star \in \arg\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\} \ , \tag{1}$$

where $f_i(x) := f(x;\zeta_i) : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}$ is a loss function on ζ_i where $D = (\zeta_1, \dots, \zeta_n)$ is a dataset of n samples drawn from a universe $\mathcal{X}$.

In Mangold et al. [9], the authors consider the problem of minimizing the composite ERM objective function, i.e., a finite sum plus a nonsmooth convex regularizer term ψ. The drawback is that they assume ψ to be separable (i.e., $\psi(x) = \sum_{i=1}^{d} \psi_i(x_i)$), which we argue is not necessary. However, we leave it as a future work since it requires more advanced techniques. We discuss this in more detail in Section 4.

2. Preliminaries

In this section, we introduce key technical concepts that will be utilized throughout the paper. Throughout the paper, we will frequently use the notation $[d] := \{1, \dots, d\}$.

Norms. We begin by defining two conjugate norms, which play a vital role in our analysis, as they help in tracking coordinate-wise quantities. Let $\langle u, v\rangle = \sum_{j=1}^{d} u_i v_i$ denote the Euclidean inner product, and consider $\mathbf{M} = \mathrm{Diag}\,(M_1, \dots, M_d)$ where $M_1, \dots, M_d > 0$. We then define the norms:

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

When $\mathbf{M}$ is the identity matrix $\mathbf{I}$, the $\mathbf{I}$-norm $\|\cdot\|_{\mathbf{I}}$ becomes the standard ℓ_2-norm $\|\cdot\|_2$.

2.1. Assumptions

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

Assumption 1 (Convexity) *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $v, w \in \mathbb{R}^d$, $f(w) \geq f(v) + \langle \nabla f(v), w - v\rangle$.*

Assumption 2 (Strong convexity) *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mu_{\mathbf{M}}$-strongly-convex* w.r.t. *the norm $\|\cdot\|_{\mathbf{M}}$ if for all $v, w \in \mathbb{R}^d$, $f(w) \geq f(v) + \langle \nabla f(v), w - v\rangle + \frac{\mu_{\mathbf{M}}}{2}\|w - v\|_{\mathbf{M}}^2$. The case $M_1 = \cdots = M_d = 1$ recovers standard μ_I-strong convexity* w.r.t. *the ℓ_2-norm.*

Assumption 3 (Component Lipschitzness) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-component-Lipschitz for $L = (L_1, \dots, L_d)$ with $L_1, \dots, L_d > 0$ if for all $w \in \mathbb{R}^d$, $t \in \mathbb{R}$ and $j \in [d]$, $|f(w + te_j) - f(w)| \leq L_j |t|$. It is Λ-Lipschitz if for all $v, w \in \mathbb{R}^d$, $|f(v) - f(w)| \leq \Lambda \|v - w\|_2$.*

Assumption 4 (Component smoothness) *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathbf{M}$-component-smooth for $M_1, \dots, M_d > 0$ if for all $v, w \in \mathbb{R}^d$, $f(w) \leq f(v) + \langle \nabla f(v), w - v\rangle + \frac{1}{2}\|w - v\|_{\mathbf{M}}^2$. When $M_1 = \cdots = M_d = \beta$, f is said to be β-smooth.*

The above component-wise regularity hypotheses are not restrictive: Λ-Lipschitzness implies $(\Lambda, \ldots, \Lambda)$-component-Lipschitzness and β-smoothness implies $(\beta, \ldots, \beta)$-component-smoothness. Yet, the actual component-wise constants of a function can be much lower than what can be deduced from their global counterparts. This will be crucial for our analysis and in the performance of DP-CD.

2.2. 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 5 (Differential Privacy, [3]) *A randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{V}$ is (ϵ, δ)-differentially private if, for all neighboring datasets $D, D' \in \mathcal{D}$ and all $U \subseteq \mathcal{V}$ in the range of $\mathcal{A}$:*

$$\mathbb{P}\left(\mathcal{A}(D) \in U\right) \leq \exp(\epsilon)\mathbb{P}\left(\mathcal{A}(D') \in U\right) + \delta.$$

The value of a function $h : \mathcal{D} \to \mathbb{R}^p$ can be privately released using the Gaussian mechanism, which adds centered Gaussian noise to $h(D)$ before releasing it [4]. 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 context, we are interested in the sensitivity arising from specific coordinates. Let $U \subseteq [d]$ represent a set of coordinates. We define the sensitivity as follows:

$$\Delta_U(h) := \sup_{D \sim D'} \left\|\mathbf{I}_U \left(h(D) - h(D')\right)\right\|_2,$$

where $\mathbf{I}_U = \mathrm{Diag}\left(\delta_1, \delta_2, \ldots, \delta_d\right)$ is a diagonal matrix with

$$\delta_i = \begin{cases} 1, & \text{if} \quad i \in U, \\ 0, & \text{if} \quad i \notin U. \end{cases}$$

When U is a singleton, i.e., $U = \{i\}$, applying the L-component Lipschitz condition (Theorem 3) gives the following bound on sensitivity $\Delta_{\{i\}}(h) \leq 2L_i$ for all $i \in [d]$. Similarly, for $U = \{d\}$, $\Delta_{\{[d]\}}(h) = \Delta(h) \leq 2\Lambda$.

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].

2.3. Sketch and Sparsification

We study unbiased diagonal sketches, defined as follows:

Definition 6 (Unbiased diagonal sketch) *Let $\mathcal{S}$ be a probability distribution over the 2^n subsets of coordinates/features of the model $x \in \mathbb{R}^d$ that we wish to train. Let $\mathcal{S}$ be nonvacuous, i.e. $P(\mathcal{S} = \emptyset) = 0$ and be* proper*, meaning that for any random set $S \sim \mathcal{S}$, $p_j := \mathrm{Prob}(j \in S) > 0$ for all coordinates $j \in [d]$. For a given random set $S \sim \mathcal{S}$ we define a random diagonal matrix (sketch) $\mathbf{C} = \mathbf{C}(S) \in \mathbb{R}^{d \times d}$ via*

$$\mathbf{C} = \mathrm{Diag}\left(\mathrm{c}_1, \ldots, \mathrm{c}_d\right), \quad \mathrm{c}_j = \begin{cases} \frac{1}{p_j}, & \text{if } j \in S, \\ 0, & \text{otherwise}. \end{cases} \tag{2}$$

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

Note that given a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we have

$$(\mathbf{C}x)_j = \begin{cases} \frac{x_j}{p_j}, & \text{if} \quad j \in S, \\ 0, & \text{if} \quad j \notin S. \end{cases}$$

So, we can control the sparsity level of the product $\mathbf{C}x$ by engineering the properties of the random set S. Also note that $\mathbb{E}[\mathbf{C}x] = x$ for all x.

3. Differentially Private Sketched Gradient Descent

We propose the DP-SkGD algorithm, as shown in Algorithm 1. This approach uses sketches $\mathbf{C}$, as defined in (2). Initially, depending on the random sample S, we select random coordinates and update the model using only these chosen coordinates through a noisy gradient. The variance of the noise depends on the sampled set S. We set different step sizes for each coordinate. This leads to directionally-unbiased updates, as is common among SGD-type methods.

Algorithm 1 DP-SkGD

Input: Initial point $w^0 \in \mathbb{R}^d$, step sizes $\mathbf{\Gamma} = \text{Diag}(\gamma_1, \dots, \gamma_d)$, number of iterations T, number of inner loops K, probability distribution $\mathcal{S}$ over the subsets of $[d]$, noise scales σ_U for all $U \in \text{Range}(S)$.
for $t = 0, \dots, T-1$ **do**
  Set $\theta^0 = w^t$
  for $k = 0, \dots, K-1$ **do**
    Sample a subset $S \sim \mathcal{S}$ and let $\mathbf{C} = \mathbf{C}(S)$ (see definition 6)
    Draw $\eta \sim \mathcal{N}(0, \sigma_S \mathbf{I})$
    $\theta^{k+1} = \theta^k - \mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)$
  end for
  $w^{t+1} = \frac{1}{K}\sum_{k=1}^{K} \theta^k$
end for

In the case where $\mathcal{S}$ is a uniform distribution over single elements in $[d]$, we recover the DP-CD algorithm proposed by Mangold et al. [9]. Likewise, if $S = [d]$ with probability one, we recover DP-SGD algorithm [1].

3.1. Privacy Guarantees

For Algorithm 1 to satisfy (ϵ, δ)-DP, the noise scales σ_S should be calibrated as given by the theorem bellow.

Theorem 7 (Proof in Appendix A.1) *Let $0 < \epsilon \leq 1$, $\delta < 1/3$. If*

$$\sigma_U^2 = \Delta_U^2(\nabla l)\frac{3KT\log(1/\delta)}{n^2\epsilon^2},$$

for all $U \in \text{Range}(\mathcal{S})$, then Algorithm 1 satisfies (ϵ, δ)-DP.

3.2. Utility Guarantees

Here we state the utility results of DP-SkGD. First, we state the result for the strongly convex regime.

Theorem 8 (Proof in Appendix B) *Let f_i be convex L-component-Lipschitz loss function for all $i \in [d]$, and f be $\mathbf{M}$-component-smooth. We assume a privacy budget with $\epsilon \leq 1$ and $\delta < 1/3$. Let $w^\star$ denote a minimizer of f, with $f^\star = f(w^\star)$ representing the minimum value. Define $\mathbf{P} = \mathrm{Diag}\,(p_1, \ldots, p_d)$, where each $p_i = P(i \in \mathcal{S})$ indicates the probability of the i-th component being included in the subset $\mathcal{S}$. Consider $w_{priv} \in \mathbb{R}^p$ as the output of Algorithm 1 with step sizes $\mathbf{\Gamma} = \mathbf{P}\mathbf{M}^{-1}$, and noise scales σ_S set as in Theorem 7 (with T and K chosen below) to ensure (ϵ, δ)-DP also let $V^2 = \mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$. Then, the following holds. For f $\mu_{\mathbf{M}}$-strongly convex and for $K = 2\left(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}}\right)$ and*

$$T = \log_2\left(\frac{(f(w^0) - f(w^\star))n^2\epsilon^2}{(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}})V^2\log(1/\delta)}\right),$$

then:

$$\mathbb{E}\left[f(w^T) - f(w^\star)\right] = O\left(\frac{V^2\log(1/\delta)}{p_{\min}\mu_{\mathbf{M}}n^2\epsilon^2}\log_2\left(\frac{(f(w^0) - f(w^\star))n\epsilon p_{\min}\mu_M}{V\log(1/\delta)}\right)\right).$$

Next, this is the result for convex regime.

Theorem 9 (Proof in Appendix B) *Let f_i be convex L-component-Lipschitz loss function for all $i \in [d]$, and f be $\mathbf{M}$-component-smooth. We assume a privacy budget with $\epsilon \leq 1$ and $\delta < 1/3$. Let $w^\star$ denote a minimizer of f, with $f^\star = f(w^\star)$ representing the minimum value. Define $\mathbf{P} = \mathrm{Diag}\,(p_1, \ldots, p_d)$, where each $p_i = P(i \in \mathcal{S})$ indicates the probability of the i-th component being included in the subset $\mathcal{S}$. Consider $w_{priv} \in \mathbb{R}^p$ as the output of Algorithm 1 with step sizes $\mathbf{\Gamma} = \mathbf{P}\mathbf{M}^{-1}$, and noise scales σ_S set as in Theorem 7 (with T and K chosen below) to ensure (ϵ, δ)-DP also let $V^2 = \mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$. Then, the following holds. Under assumption f is convex and for $T = 1$ and $K = \frac{R_{\mathbf{P}^{-1}\mathbf{M}}n\varepsilon}{V\sqrt{\log(1/\delta)}}$, the utility of Algorithm 1 is defined as follows*

$$\mathbb{E}\left[f(w^{priv}) - f(w^\star)\right] = O\left(\frac{VR_{\mathbf{P}^{-1}\mathbf{M}}}{n\varepsilon}\log^{1/2}\frac{1}{\delta}\right),$$

where $R^2_{\mathbf{P}^{-1}\mathbf{M}} = \|w^0 - w^\star\|^2_{\mathbf{P}^{-1}\mathbf{M}}$.

Commentary:

- We have the flexibility to choose the probabilities p_i. In the strongly convex case, we can select them to minimize the complexity. By setting $p_i = \frac{M_i}{\sum_{i=1}^d M_i}$, we obtain the following convergence rate in the strongly convex regime.

$$\mathbb{E}\left[f(w^{\text{priv}}) - f(w^\star)\right] = O\left(\frac{V\|w^0 - w^\star\|^2}{n\varepsilon\sum_{i=1}^d M_i}\log^{1/2}\frac{1}{\delta}\right).$$

- Let $S = [d]$ with probability one, which implies $p_i = 1$ for all $i \in [d]$. In this case, we have $V^2 = \left\|\Delta_{[d]}\left(\nabla l\right)\right\|^2_{\mathbf{M}^{-1}}$ and $R^2_{\mathbf{P}^{-1}\mathbf{M}} = \|w^0 - w^\star\|^2_{\mathbf{M}}$. Therefore, we obtain the following complexity.

$$\mathbb{E}\left[f(w^{\text{priv}}) - f(w^\star)\right] = O\left(\frac{\left\|\Delta_{[d]}\left(\nabla l\right)\right\|^2 \|w^0 - w^\star\|^2}{n\varepsilon}\log^{1/2}\frac{1}{\delta}\right)$$
$$= O\left(\frac{\Lambda\left\|w^0 - w^\star\right\|^2}{n\varepsilon}\log^{1/2}\frac{1}{\delta}\right).$$

  Thus, we recover DP-SGD result by Bassily et al. [1].

- Let $\mathcal{S}$ be a uniform distribution over single coordinates, i.e. $S = \{i\}$ with probability $\frac{1}{n}$. In this case we get $V = \left\|\Delta_{\{i\}}\left(\nabla l\right)\right\|^2_{\mathbf{M}^{-1}}$. If we further assume Lipschitzness across coordinates (Assumption 3) then we get $V = \|L\|^2_{\mathbf{M}^{-1}}$, resulting the same result as in Mangold et al. [9].

There are various other possible strategies for selecting S. One option is to choose a subset of $[d]$ of size τ at each iteration. Our analysis generalizes both DP-CD by Mangold et al. [9] and DP-SGD by Bassily et al. [1].

4. Conclusion and Future Work

Our work aims to bridge the gap between non-private and private coordinate descent methods. We develop a random block coordinate descent method that allows for the selection of multiple coordinates with varying probabilities in each iteration using sketch matrices. This approach generalizes the existing method of updating a single coordinate drawn uniformly at random.

In our study, we do not address the composite Empirical Risk Minimization (ERM) problem. To tackle the composite ERM problem, a promising direction is to incorporate data-dependent sketches, as demonstrated by Safaryan et al. [14], along with other variance reduction techniques such as those outlined by Hanzely et al. [6]. We leave this exploration for future work. Another potential generalization involves considering not only diagonal matrix smoothness M but also general matrix smoothness. This would provide more detailed information about the function and lead to exploring more general descent directions beyond the coordinate direction, which would require new definitions of sensitivity and privacy analysis.

Acknowledgements

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

[1] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 464–473, 12 2014. doi: 10.1109/FOCS.2014.56.

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

[3] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[4] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[5] Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[6] Filip Hanzely, Konstantin Mishchenko, and Peter Richtárik. Sega: Variance reduction via gradient sketching. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Friedrich Liese and Igor Vajda. Convex statistical distances. *(No Title)*, 1987.

[8] Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. *Advances in Neural Information Processing Systems*, 27, 2014.

[9] Paul Mangold, Aurélien Bellet, Joseph Salmon, and Marc Tommasi. Differentially private coordinate descent for composite empirical risk minimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14948–14978. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mangold22a.html.

[10] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[11] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[12] Yurii Nesterov and Sebastian U Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

[13] Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156:433–484, 2016.

[14] Mher Safaryan, Filip Hanzely, and Peter Richtárik. Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization. *Advances in Neural Information Processing Systems*, 34:25688–25702, 2021.

[15] Shai Shalev-Shwartz and Tong Zhang. Accelerated mini-batch stochastic dual coordinate ascent. *Advances in Neural Information Processing Systems*, 26, 2013.

[16] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1), 2013.

[17] Hao-Jun Michael Shi, Shenyinying Tu, Yangyang Xu, and Wotao Yin. A primer on coordinate descent algorithms. *arXiv preprint arXiv:1610.00040*, 2016.

[18] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[19] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[20] Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

Appendix A. Privacy Analysis

The main steps of analysis are based on the concept of Rényi Differential Privacy (RDP) and the composition theorem for RDP.

Definition 10 (Rényi divergence [10]) *For two probability distributions R and T with value in the same domain $\mathcal{V}$, the Rényi divergence of order $\alpha > 1$ is*

$$D_\alpha\left(R\|T\right) = \frac{1}{\alpha-1}\log\int_{\mathcal{V}} R(x)^\alpha T(x)^{1-\alpha}dx, \tag{3}$$

where $R(x)$ and $T(x)$ are the density functions of R and T respectively at x.

Definition 11 ([10]) *A randomized algorithm $\mathcal{A}:\mathcal{D}\to\mathcal{V}$ is (α,ε)-Rényi DP (RDP) if the following inequality holds*

$$D_\alpha\left(\mathcal{A}(\mathcal{D})\|\mathcal{A}(\mathcal{D}')\right) \le \varepsilon, \tag{4}$$

for any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$.

The following statement is a closed-form expression of the Rényi divergence between a Gaussian and its offset (for a more general version see [7]), [19].

Lemma 12 (Proposition 7 [10]) $D_\alpha(N(0,\sigma^2)\|N(\mu,\sigma^2\mathbf{I})) = \frac{\alpha}{2\sigma^2}\sum_{j=1}^d \mu_j^2$.

Proof The density function of multivariate Gaussian distribution is

$$\begin{aligned}\phi(x) &= \prod_{j=1}^d \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{1}{2\sigma_j^2}\left(x_j-\mu_j\right)^2\right\}\\ &= \frac{1}{\left(2\pi\sigma_j^2\right)^{\frac{d}{2}}}\exp\left\{-\sum_{j=1}^d\frac{1}{2\sigma_j^2}\left(x_j-\mu_j\right)^2\right\}.\end{aligned}$$

By direct computation we verify that

$$\begin{aligned}D_\alpha(N(0,\mathbf{\Sigma})\|N(\mu,\mathbf{\Sigma})) &= \frac{1}{\alpha-1}\log\int_{\mathbb{R}^d}\frac{1}{\left(2\pi\sigma_j^2\right)^{\frac{d}{2}}}\exp\left\{-\alpha\sum_{j=1}^d\frac{1}{2\sigma_j^2}x_j^2\right\}\\ &\qquad\cdot\exp\left\{-(1-\alpha)\sum_{j=1}^d\frac{1}{2\sigma_j^2}\left(x_j-\mu_j\right)^2\right\}\\ &= \frac{1}{\alpha-1}\log\int_{\mathbb{R}^d}\frac{1}{\left(2\pi\sigma_j^2\right)^{\frac{d}{2}}}\exp\left\{-\sum_{j=1}^d\frac{1}{2\sigma_j^2}\left(x_j^2-2(1-\alpha)\mu_j x_j+(1-\alpha)\mu_j^2\right)\right\}\\ &= \frac{1}{\alpha-1}\log\left(\exp\left\{-\sum_{j=1}^d\frac{1}{2\sigma_j^2}\alpha\left(1-\alpha\right)\mu_j^2\right\}\right)\\ &= \sum_{j=1}^d\frac{\alpha\mu_j^2}{2\sigma_j^2}.\end{aligned}$$

■

Using this we get the following corollary.

Corollary 13 *For any nonempty $U \in [d]$, the Gaussian mechanism applied to only coordinates in U*

$$\mathcal{G}_U(\mathcal{D}) = f(\mathcal{D}) + \mathbf{I}_U \mathcal{N}\left(0, \sigma^2 \mathbf{I}\right),$$

is $\left(\alpha, \frac{\alpha \Delta_U^2(f)}{2\sigma^2}\right)$-RDP, where

$$\Delta_U(f) = \sup_{D \sim D'} \left\| \mathbf{I}_U \left(f(D) - f(D') \right) \right\|_2$$

and $\mathbf{I}_U = \operatorname{Diag}(\delta_1, \delta_2, \ldots, \delta_n)$ with

$$\delta_i = \begin{cases} 1, & \text{if} \quad i \in U, \\ 0, & \text{if} \quad i \notin U. \end{cases}$$

Now we state the composition theorem for the sequence of RDP algorithms.

Proposition 14 (Proposition 1, [10]) *Let $\mathcal{A}_1, \ldots \mathcal{A}_K : \mathcal{D} \to \mathcal{F}$ be $K > 0$ randomized algorithms, such that for each k algorithm $\mathcal{A}_k$ is $(\alpha, \varepsilon_k(\alpha))$-RDP, where these algorithms can be chosen in an adaptive way. Let $\mathcal{A} : \mathcal{D} \to \mathcal{F}^K$ be a randomized algorithm such that $\mathcal{A}(\mathcal{D}) = (\mathcal{A}_1(D), \ldots, \mathcal{A}_K(D))$. Then $\mathcal{A}$ is $\left(\alpha, \sum_{k=1}^{K} \varepsilon_k(\alpha)\right)$-RDP.*

The following Lemma gives the relationship between RDP and (ε, δ)-DP.

Lemma 15 (Proposition 3, [10]) *If $\mathcal{A}$ is (α, ε)-RDP, then it is also $\left(\varepsilon + \frac{\log \frac{1}{\delta}}{\alpha - 1}, \delta\right)$-DP for any $0 < \delta < 1$.*

Since this result holds for any α is it possible to find the minimum with respect to it. We will use the following result.

Lemma 16 (Corollary B.8, [9]) *Let $0 < \varepsilon < 1$, $0 < \delta < \frac{1}{3}$. If a randomized algorithm $\mathcal{A} : \mathcal{D} \to \mathcal{F}$ is $\left(\alpha, \frac{\gamma\alpha}{2\sigma^2}\right)$-RDP with $\gamma > 0$ and $\sigma = \frac{\sqrt{3\gamma \log \frac{1}{\delta}}}{\varepsilon}$, it is also (ε, δ)-DP.*

A.1. Proof of Theorem 7

We are now ready to prove Theorem 7. Let us first recall our Algorithm 1

Algorithm 1 DP-SkGD

Input: Initial point $w^0 \in \mathbb{R}^d$, step sizes $\mathbf{\Gamma} = \text{Diag}\,(\gamma_1, \ldots, \gamma_d)$, number of iterations T, number of inner loops K, probability distribution $\mathcal{S}$ over the subsets of $[d]$, noise scales σ_U for all $U \in \text{Range}\,(S)$.
for $t = 0, \ldots, T-1$ **do**
  Set $\theta^0 = w^t$
  for $k = 0, \ldots, K-1$ **do**
    Sample a subset $S \sim \mathcal{S}$ and let $\mathbf{C} = \mathbf{C}(S)$ (see definition 6)
    Draw $\eta \sim \mathcal{N}\,(0, \sigma_S \mathbf{I})$
    $\theta^{k+1} = \theta^k - \mathbf{\Gamma C}\left(\nabla f(\theta^k) + \eta\right)$
  end for
  $w^{t+1} = \frac{1}{K}\sum_{k=1}^{K} \theta^k$
end for

Theorem 7 *Let* $0 < \epsilon \leq 1$, $\delta < 1/3$. *If*

$$\sigma_U^2 = \Delta_U^2(\nabla l)\frac{3KT\log(1/\delta)}{n^2\epsilon^2},$$

for all $U \in \text{Range}\,(\mathcal{S})$, *then Algorithm 1 satisfies* (ϵ, δ)*-DP.*

Proof From the privacy perspective, Algorithm 1 adaptively releases and post-processes a series of gradient coordinates protected by the Gaussian mechanism. First, we focus on the inner loop of the algorithm, which can be viewed as a composition of K Gaussian mechanisms. Let $\sigma > 0$. Theorem 13 guarantees that the k-th Gaussian mechanism with noise scale $\sigma_U = \Delta_U(\nabla f)\sigma > 0$ is $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP. Then, the composition of these K mechanisms is, according to Theorem 14, $(\alpha, \frac{K\alpha}{2\sigma^2})$-RDP. This can be converted to (ϵ, δ)-DP via Theorem 16 with $\sigma = \frac{\sqrt{3\gamma\log\frac{1}{\delta}}}{\varepsilon}$ and $\gamma = K$, which gives $\sigma_U = \frac{\Delta_U(f)\sqrt{3K\log(1/\delta)}}{\epsilon}$ for $k \in [K]$. Since we repeat this inner loop T times we get

$$\sigma_U^2 = \Delta_U^2(\nabla f)\frac{3KT\log(1/\delta)}{\epsilon^2},$$

for all $U \in \text{Range}\,(\mathcal{S})$. It remains to note that $\Delta_U^2(\nabla f) = \frac{\Delta_U^2(\nabla l)}{n}$. ■

Appendix B. Proof of Utility Guarantees

Algorithm 1 DP-SkGD

Input: Initial point $w^0 \in \mathbb{R}^d$, step sizes $\mathbf{\Gamma} = \text{Diag}\,(\gamma_1, \ldots, \gamma_d)$, number of iterations T, number of inner loops K, probability distribution $\mathcal{S}$ over the subsets of $[d]$, noise scales σ_U for all $U \in \text{Range}\,(S)$.
for $t = 0, \ldots, T-1$ **do**
  Set $\theta^0 = w^t$
  for $k = 0, \ldots, K-1$ **do**
    Sample a subset $S \sim \mathcal{S}$ and let $\mathbf{C} = \mathbf{C}(S)$ (see definition 6)
    Draw $\eta \sim \mathcal{N}\,(0, \sigma_S \mathbf{I})$
    $\theta^{k+1} = \theta^k - \mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)$
  end for
  $w^{t+1} = \frac{1}{K}\sum_{k=1}^{K} \theta^k$
end for

To proof the theorem we need the following lemma.

Lemma 17 (Proof in Appendix B.3) *Under assumptions f is convex and $\mathbf{M}$-smooth and the selection of stepsize $\mathbf{\Gamma} = \mathbf{P}\mathbf{M}^{-1}$, the iterates of Algorithm 1 satisfy*

$$\mathbb{E}\left[f(w^{t+1}) - f(w^\star)\right] \leq \frac{\mathbb{E}\left[\|w^t - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right] + 2\mathbb{E}\left[f(w^t) - f(w^\star)\right]}{2K} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].$$

Remark 18 $\mathbf{M}$*-smoothness of f gives*

$$f(w^t) \leq f(w^\star) + \left\langle \nabla f(w^\star), w^t - w^\star \right\rangle + \frac{1}{2}\left\|w^t - w^\star\right\|^2_{\mathbf{M}} \leq f(w^\star) + \frac{1}{2}\left\|w^t - w^\star\right\|^2_{\mathbf{P}^{-1}\mathbf{M}}, \quad (5)$$

and the result of Lemma 17 further simplifies as:

$$\mathbb{E}\left[f(w^{t+1}) - f(w^\star)\right] \leq \frac{1}{K}\left\|w^t - w^\star\right\|^2_{\mathbf{P}^{-1}\mathbf{M}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]. \quad (6)$$

B.1. Convex Case

Theorem 9 (Convex case) *Under assumptions f is convex and $\mathbf{M}$-smooth and the selection of stepsize $\mathbf{\Gamma} = \mathbf{P}\mathbf{M}^{-1}$, for $T = 1$ and $K = \frac{R_{\mathbf{P}^{-1}\mathbf{M}} n\varepsilon}{V\sqrt{\log(1/\delta)}}$ where $V^2 = \mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$, the utility of Algorithm 1 is defined as follows*

$$\mathbb{E}\left[f(w^{priv}) - f(w^\star)\right] = O\left(\mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]\frac{R_{\mathbf{P}^{-1}\mathbf{M}}}{n\varepsilon}\log^{1/2}\frac{1}{\delta}\right). \quad (7)$$

Proof Put $T = 1$ and let $R^2_{\mathbf{M}} = \|w^0 - w^\star\|^2_{\mathbf{M}}$. From Theorem 7 we have $\sigma_U^2 = \Delta_U^2(\nabla l)\frac{3KT\log(1/\delta)}{n^2\epsilon^2}$. Putting this in Lemma 17 we get

$$\mathbb{E}\left[f(w^{\text{priv}}) - f(w^\star)\right] \leq \frac{R^2_{\mathbf{P}^{-1}\mathbf{M}}}{K} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$$
$$= \frac{R^2_{\mathbf{P}^{-1}\mathbf{M}}}{K} + \frac{3KT\log(1/\delta)}{n^2\epsilon^2}\mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].$$

To minimize the right-hand side we put $K = \frac{R_{\mathbf{P}^{-1}\mathbf{M}} n\varepsilon}{V\sqrt{\log(1/\delta)}}$, and we get

$$\mathbb{E}\left[f(w^{\text{priv}}) - f(w^\star)\right] \leq \frac{R_{\mathbf{P}^{-1}\mathbf{M}} V\sqrt{\log(1/\delta)}}{n\varepsilon} + \frac{12R_{\mathbf{P}^{-1}\mathbf{M}} V\sqrt{\log(1/\delta)}}{n\epsilon}$$
$$= O\left(\mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]\frac{R_{\mathbf{P}^{-1}\mathbf{M}}}{n\varepsilon}\log^{1/2}\frac{1}{\delta}\right).$$

■

B.2. Strongly Convex Case

Theorem 9 (Strongly Convex case) *Let f be $\mu_{\mathbf{M}}$-strongly convex w.r.t. $\|\cdot\|_{\mathbf{M}}$ and $w^\star$ be the minimizer of F. The output w^{priv} of Algorithm 1, starting from $w^0 \in \mathbb{R}^d$ with*

$$K = 2\left(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}}\right),$$

$$T = \log_2\left(\frac{(f(w^0) - f(w^\star))n^2\epsilon^2}{(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}})V^2\log(1/\delta)}\right),$$

where $V^2 = \mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$, and the σ_S's as in Theorem 7, satisfies:

$$\mathbb{E}\left[f(w^T) - f(w^\star)\right]$$
$$\leq \left(1 + \log_2\left(\frac{(f(w^0) - f(w^\star))n^2\epsilon^2}{24(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}})V^2\log(1/\delta)}\right)\right)\frac{24(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}})V^2\log(1/\delta)}{n^2\epsilon^2}$$
$$= O\left(\mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]\frac{\log(1/\delta)}{p_{\min}\mu_{\mathbf{M}}n^2\epsilon^2}\log_2\left(\frac{(f(w^0) - f(w^\star))n\epsilon p_{\min}\mu_M}{V\log(1/\delta)}\right)\right).$$

Proof As f is $\mu_{\mathbf{M}}$-strongly-convex with respect to norm $\|\cdot\|_{\mathbf{M}}$, we obtain for any $w \in \mathbb{R}^d$, that $f(w) \geq f(w^\star) + \frac{\mu_{\mathbf{M}}}{2}\|w - w^\star\|^2_{\mathbf{M}}$. Therefore, $\left\|w^t - w^\star\right\|^2_{\mathbf{M}} \leq \frac{2}{\mu_M}\left(f(w^t) - f(w^\star)\right)$ and Theorem 17 gives, for $1 \leq t \leq T-1$,

$$\mathbb{E}\left[f(w^{t+1}) - f(w^\star)\right] \leq \frac{\|w^t - w^\star\|^2_{\mathbf{P}^{-1}\mathbf{M}} + 2\left(f(w^t) - f(w^\star)\right)}{2K} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$$
$$\leq \frac{\frac{1}{p_{\min}}\|w^t - w^\star\|^2_{\mathbf{M}} + 2\left(f(w^t) - f(w^\star)\right)}{2K} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$$
$$\leq \frac{\left(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}}\right)\left(f(w^t) - f(w^\star)\right)}{K} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].$$

Setting $K = 2\left(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}}\right)$ we obtain

$$\mathbb{E}\left[f(w^{t+1}) - f(w^\star)\right] \leq \frac{f(w^t) - f(w^\star)}{2} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].$$

Recursive application of this inequality gives

$$\begin{aligned}\mathbb{E}\left[f(w^T) - f(w^\star)\right] &\leq \frac{f(w^0) - f(w^\star)}{2^T} + \sum_{t=0}^{T-1} \frac{1}{2^t}\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right] \\ &\leq \frac{f(w^0) - f(w^\star)}{2^T} + 2\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].\end{aligned}$$

Let $V^2 = \mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]$. Taking

$$T = \log_2\left(\frac{(f(w^0) - f(w^\star))n^2\epsilon^2}{\left(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}}\right)V^2\log(1/\delta)}\right)$$

gives

$$\begin{aligned}&\mathbb{E}\left[f(w^T) - f(w^\star)\right] \\ &\leq \left(1 + \log_2\left(\frac{(f(w^0) - f(w^\star))n^2\epsilon^2}{24(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}})V^2\log(1/\delta)}\right)\right)\frac{24(1 + \frac{1}{p_{\min}\mu_{\mathbf{M}}})V^2\log(1/\delta)}{n^2\epsilon^2} \\ &= O\left(\mathbb{E}\left[\|\mathbf{C}\Delta_S(\nabla l)\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]\frac{\log(1/\delta)}{p_{\min}\mu_{\mathbf{M}}n^2\epsilon^2}\log_2\left(\frac{(f(w^0) - f(w^\star))n\epsilon p_{\min}\mu_M}{V\log(1/\delta)}\right)\right),\end{aligned}$$

which is the result of our theorem. ■

B.3. Proof of Auxiliary Lemmas

Algorithm 1 DP-SkGD

Input: Initial point $w^0 \in \mathbb{R}^d$, step sizes $\mathbf{\Gamma} = \text{Diag}\,(\gamma_1, \ldots, \gamma_d)$, number of iterations T, number of inner loops K, probability distribution $\mathcal{S}$ over the subsets of $[d]$, noise scales σ_U for all $U \in \text{Range}\,(S)$.
for $t = 0, \ldots, T-1$ **do**
  Set $\theta^0 = w^t$
  for $k = 0, \ldots, K-1$ **do**
    Sample a subset $S \sim \mathcal{S}$ and let $\mathbf{C} = \mathbf{C}(S)$ (see definition 6)
    Draw $\eta \sim \mathcal{N}\,(0, \sigma_S\mathbf{I})$
    $\theta^{k+1} = \theta^k - \mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)$
  end for
  $w^{t+1} = \frac{1}{K}\sum_{k=1}^K \theta^k$
end for

Lemma 19 *Let* $\mathbf{C}$ *be an unbiased diagonal sketch defined in* (2)*, and let* $\eta \sim \mathcal{N}(0, \sigma\mathbf{I})$ *and let* $\mathbf{D}$ *be any diagonal matrix. Then*

$$\mathbb{E}\left[\mathbf{C}\left(x+\eta\right)\right] = x,$$

and

$$\mathbb{E}\left[\|\mathbf{C}\left(x+\eta\right)\|^2_{\mathbf{D}}\right] = \|x\|^2_{\mathbf{P}^{-1}\mathbf{D}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{D}}\right],$$

for any deterministic diagonal matrix D*.*

Proof We have

$$\mathbb{E}\left[\mathbf{C}(x+\eta)\right] = \mathbb{E}\left[\mathbf{C}x\right] + \mathbb{E}\left[\mathbf{C}\eta\right] = \mathbb{E}\left[\mathbf{C}\right]x + \mathbb{E}\left[\mathbf{C}\right]\mathbb{E}\left[\eta\right] = x.$$

Now we calculate the variance:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{C}(x+\eta)\|^2_{\mathbf{D}}\right] &= \mathbb{E}\left[\|\mathbf{C}x + \mathbf{C}\eta\|^2_{\mathbf{D}}\right] \\ &= \mathbb{E}\left[\|\mathbf{C}x\|^2_{\mathbf{D}}\right] + \mathbb{E}\left[2\left\langle\mathbf{D}\mathbf{C}x, \mathbf{C}\eta\right\rangle\right] + \mathbb{E}\left[\|\mathbf{C}\eta\|^2_{\mathbf{D}}\right].\end{aligned}$$

Further,

$$\mathbb{E}\left[\|\mathbf{C}x\|^2_{\mathbf{D}}\right] = \mathbb{E}\left[x^\top\mathbf{C}\mathbf{D}\mathbf{C}x\right] = \mathbb{E}\left[\sum_{j=1}^d x_j^2 c_j^2 D_j\right] = \sum_{j=1}^d x_j^2 \frac{1}{p_j} D_j = \|x\|^2_{\mathbf{P}^{-1}\mathbf{D}},$$

similarly,

$$\mathbb{E}\left[\|\mathbf{C}\eta\|^2_{\mathbf{D}}\right] = \mathbb{E}\left[\mathbb{E}\left[\|\mathbf{C}\eta\|^2_{\mathbf{D}}|S\right]\right] = \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{D}}\right].$$

It remains to note that

$$\mathbb{E}\left[2\left\langle\mathbf{D}\mathbf{C}x, \mathbf{C}\eta\right\rangle\right] = 0.$$

■

Lemma 20 (Descent Lemma) *For any* $0 < \alpha < 1$ *the following inequality holds.*

$$\begin{aligned}&\mathbb{E}\left[f(\theta^{k+1}) - f(w)\middle|\theta^k\right] - (1-\alpha)\left(f(\theta^k) - f(w)\right) \\ &\quad\leq \alpha\left(f(\theta^k) - f(w)\right) - \|\nabla f(\theta^k)\|^2_{\mathbf{\Gamma}} + \frac{1}{2}\left(\|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}^2\mathbf{M}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}^2\mathbf{M}}\right]\right).\end{aligned}$$

Proof Using the M-component smoothness of f, and lemma 19 we get:

$$\begin{aligned}\mathbb{E}\left[f(\theta^{k+1})|\theta^k\right] &= \mathbb{E}\left[f\left(\theta^k - \mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)\right)|\theta^k\right] \\ &\leq f(\theta^k) - \left\langle\nabla f(\theta^k), \mathbf{\Gamma}\mathbb{E}\left[\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)|\theta^k\right]\right\rangle \\ &\quad + \frac{1}{2}\mathbb{E}\left[\|\mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)\|^2_{\mathbf{M}}|\theta^k\right] \\ &= f(\theta^k) - \|\nabla f(\theta^k)\|^2_{\mathbf{\Gamma}} + \frac{1}{2}\mathbb{E}\left[\|\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)\|^2_{\mathbf{\Gamma}^2\mathbf{M}}|\theta^k\right] \\ &= f(\theta^k) - \|\nabla f(\theta^k)\|^2_{\mathbf{\Gamma}} + \frac{1}{2}\left(\|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}^2\mathbf{M}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}^2\mathbf{M}}\right]\right).\end{aligned}$$

For any vector w

$$\mathbb{E}\left[f(\theta^{k+1}) - f(w)\middle|\theta^k\right] \leq f(\theta^k) - f(w) - \|\nabla f(\theta^k)\|^2_{\mathbf{\Gamma}} + \frac{1}{2}\left(\|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}^2\mathbf{M}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}^2\mathbf{M}}\right]\right).$$

It remains to split $f(\theta^k) - f(w)$ into two parts. ■

Lemma 21 *Under assumptions f is convex and $\mathbf{M}$-smooth, for any $w \in \mathbb{R}^d$ and $0 < \alpha < 1$ the iterates of Algorithm 1 satisfy*

$$\left\langle \theta^k - w^\star, \nabla f(\theta^k)\right\rangle \leq \|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right] + \|\nabla f(\theta^k)\|^2_{\mathbf{M}^{-1}-\mathbf{P}^{-1}\mathbf{\Gamma}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}}\right]. \tag{8}$$

Proof

$$\begin{aligned}
\mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right] &= \mathbb{E}\left[\|\theta^k - \mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right) - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right] \\
&= \|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - 2\left\langle \theta^k - w^\star, \nabla f(\theta^k)\right\rangle + \mathbb{E}\left[\|\mathbf{\Gamma}\mathbf{C}\left(\nabla f(\theta^k) + \eta\right)\|^2_{\mathbf{\Gamma}^{-1}}\right] \\
&= \|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - 2\left\langle \theta^k - w^\star, \nabla f(\theta^k)\right\rangle + \|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}}\right] \\
&\leq \|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \left\langle \theta^k - w^\star, \nabla f(\theta^k)\right\rangle - \|\nabla f(\theta^k)\|^2_{\mathbf{M}^{-1}} \\
&\quad + \|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}}\right].
\end{aligned}$$

It remains to rearrange the terms. ■

Lemma 22 *Under assumptions f is convex and $\mathbf{M}$-smooth and the selection of stepsize $\mathbf{\Gamma} = \mathbf{P}\mathbf{M}^{-1}$ and $0 < \alpha < 1$ the iterates of Algorithm 1 satisfy*

$$\mathbb{E}\left[f(\theta^{k+1}) - f(w^\star)\middle|\theta^k\right] - (1-\alpha)\left(f(\theta^k) - f(w^\star)\right) \leq \alpha\left(\|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right]\right) + \frac{2\alpha+1}{2}\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].$$

Proof Using convexity of f and lemma 21, we get

$$\begin{aligned}
f(\theta^k) - f(w^\star) &\leq \left\langle \nabla f(\theta^k), \theta^k - w^\star\right\rangle \\
&= \|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right] + \|\nabla f(\theta^k)\|^2_{\mathbf{M}^{-1}-\mathbf{P}^{-1}\mathbf{\Gamma}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}}\right].
\end{aligned}$$

Next, we have

$$
\begin{aligned}
&\mathbb{E}\left[f(\theta^{k+1}) - f(w^\star)\middle|\theta^k\right] - (1-\alpha)\left(f(\theta^k) - f(w^\star)\right) \\
&\quad\leq \alpha\left(f(\theta^k) - f(w^\star)\right) - \|\nabla f(\theta^k)\|^2_{\mathbf{\Gamma}} + \frac{1}{2}\left(\|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}^2\mathbf{M}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}^2\mathbf{M}}\right]\right) \\
&\quad\leq \alpha\left(\|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right] + \|\nabla f(\theta^k)\|^2_{\mathbf{M}^{-1}-\mathbf{P}^{-1}\mathbf{\Gamma}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}}\right]\right) \\
&\qquad - \|\nabla f(\theta^k)\|^2_{\mathbf{\Gamma}} + \frac{1}{2}\left(\|\nabla f(\theta^k)\|^2_{\mathbf{P}^{-1}\mathbf{\Gamma}^2\mathbf{M}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}^2\mathbf{M}}\right]\right) \\
&\quad\overset{\mathbf{\Gamma}=\mathbf{P}\mathbf{M}^{-1}}{=} \alpha\left(\|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right] + \|\nabla f(\theta^k)\|^2_{\mathbf{M}^{-1}-\mathbf{P}^{-1}\mathbf{\Gamma}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}}\right]\right) \\
&\qquad - \|\nabla f(\theta^k)\|^2_{\mathbf{P}\mathbf{M}^{-1}} + \frac{1}{2}\left(\|\nabla f(\theta^k)\|^2_{\mathbf{P}\mathbf{M}^{-1}} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{\Gamma}^2\mathbf{M}}\right]\right) \\
&\quad\overset{\mathbf{\Gamma}=\mathbf{P}\mathbf{M}^{-1}}{\leq} \alpha\left(\|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right]\right) + \alpha\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right] + \frac{1}{2}\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}^2\mathbf{M}^{-1}}\right] \\
&\leq \alpha\left(\|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}} - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\middle|\theta^k\right]\right) + \frac{2\alpha+1}{2}\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].
\end{aligned}
$$

■

Lemma 17 *Under assumptions f is convex and $\mathbf{M}$-smooth and the selection of stepsize $\mathbf{\Gamma} = \mathbf{P}\mathbf{M}^{-1}$, the iterates of Algorithm 1 satisfy*

$$
\mathbb{E}\left[f(w^{t+1}) - f(w^\star)\right] \leq \frac{\mathbb{E}\left[\|w^t - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right] + 2\mathbb{E}\left[f(w^t) - f(w^\star)\right]}{2K} + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].
$$

Proof Taking $\alpha = 1/2$, the iterates of Algorithm 1 satisfy

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\left[\mathbb{E}\left[f(\theta^{k+1}) - f(w^\star)\right] - \frac{1}{2}\mathbb{E}\left[f(\theta^k) - f(w^\star)\right]\right] \\
&\quad\leq \sum_{k=0}^{K-1}\left[\frac{1}{2}\left(\mathbb{E}\left[\|\theta^k - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right] - \mathbb{E}\left[\|\theta^{k+1} - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right]\right) + \mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right]\right] \\
&\quad\leq \frac{1}{2}\left(\mathbb{E}\left[\|\theta^0 - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right] - \mathbb{E}\left[\|\theta^K - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right]\right) + K\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right] \\
&\quad\leq \frac{1}{2}\mathbb{E}\left[\|w^t - w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right] + K\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{P}\mathbf{M}^{-1}}\right].
\end{aligned}
$$

On the other hand

$$\begin{aligned}
&\sum_{k=0}^{K-1}\left[\mathbb{E}\left[f(\theta^{k+1})-f(w^\star)\right]-\frac{1}{2}\mathbb{E}\left[f(\theta^k)-f(w^\star)\right]\right]\\
&=\mathbb{E}\left[f(\theta^K)-f(w^\star)\right]-\mathbb{E}\left[f(\theta^0)-f(w^\star)\right]+\frac{1}{2}\sum_{k=0}^{K-1}\mathbb{E}\left[f(\theta^k)-f(w^\star)\right]\\
&=\mathbb{E}\left[f(\theta^K)-f(w^\star)\right]-\mathbb{E}\left[f(\theta^0)-f(w^\star)\right]\\
&\quad+\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[f(\theta^k)-f(w^\star)\right]+\frac{1}{2}\mathbb{E}\left[f(\theta^0)-f(w^\star)\right]-\frac{1}{2}\mathbb{E}\left[f(\theta^K)-f(w^\star)\right]\\
&=\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[f(\theta^k)-f(w^\star)\right]-\frac{1}{2}\mathbb{E}\left[f(w^t)-f(w^\star)\right]+\frac{1}{2}\mathbb{E}\left[f(\theta^K)-f(w^\star)\right]\\
&\geq\frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[f(\theta^k)-f(w^\star)\right]-\mathbb{E}\left[f(w^t)-f(w^\star)\right].
\end{aligned}$$

Further, using the convexity of f we have

$$f(w^{t+1})\leq\frac{1}{K}\sum_{k=1}^{K}f(\theta^k).$$

Using this, we get

$$\begin{aligned}
\mathbb{E}\left[f(w^{t+1})-f(w^\star)\right]&\leq\frac{1}{K}\sum_{k=1}^{K}f(\theta^k)-f(w^\star)\\
&\leq\frac{1}{K}\left[\frac{1}{2}\mathbb{E}\left[\|w^t-w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right]+\mathbb{E}\left[f(w^t)-f(w^\star)\right]+K\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{PM}^{-1}}\right]\right]\\
&\leq\frac{\mathbb{E}\left[\|w^t-w^\star\|^2_{\mathbf{\Gamma}^{-1}}\right]+2\mathbb{E}\left[f(w^t)-f(w^\star)\right]}{2K}+\mathbb{E}\left[\|\mathbf{C}\sigma_S\|^2_{\mathbf{PM}^{-1}}\right].
\end{aligned}$$

■