Improved Differentially Private Riemannian Optimization

Anonymous authors
Paper under double-blind review

Abstract

A common step in differentially private (DP) Riemannian optimization is sampling from the (tangent) Gaussian distribution as noise needs to be generated in the tangent space to perturb the gradient before taking the step. In this regard, existing works either use the Markov chain Monte Carlo (MCMC) sampling or explicit basis construction based sampling methods on the tangent space. This becomes a computational bottleneck in the practical use of DP Riemannian optimization, especially when performing stochastic optimization. In this paper, we discuss different sampling strategies and develop efficient sampling procedures by exploiting linear isometry between tangent spaces and show them to be orders of magnitude faster than standard sampling strategies like MCMC. We also improve utility bounds by showing them to be metric-tensor independent. Furthermore, we develop the DP Riemannian stochastic variance reduced gradient algorithm and compare it with DP Riemannian gradient descent and stochastic gradient descent algorithms on various problems.

1 Introduction

Differential privacy (DP) provides a rigorous treatment for the notion of data privacy by precisely quantifying the deviation in the model's output distribution under modification of a small number of data points (Dwork et al., 2006b). Provable guarantees of DP coupled with properties like immunity to arbitrary post-processing, and graceful composability have made it a de-facto standard of privacy with steadfast adoption in the real world (Erlingsson et al., 2014; Apple, 2017; Near, 2018; Abowd, 2018). Furthermore, it has been shown empirically that DP models resist various kinds of leakage attacks that can cause privacy violations (Rahman et al., 2018; Carlini et al., 2019; Sablayrolles et al., 2019; Zhu et al., 2019; Balle et al., 2022).

Various approaches have been explored in literature to ensure differential privacy in machine learning models. These include output perturbation (Chaudhuri & Monteleoni, 2008; Chaudhuri et al., 2011; Zhang et al., 2017) and objective perturbation (Chaudhuri & Monteleoni, 2008; Chaudhuri et al., 2011; Kifer et al., 2012; Iyengar et al., 2019; Bassily et al., 2021), in which a perturbation term is added to the output of a non-DP algorithm or the optimization objective, respectively. Another approach, gradient perturbation, involves perturbing the gradient information at every iteration of gradient based approaches and has received significant interest in context of deep learning and stochastic optimization (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016; Wang et al., 2017; Bassily et al., 2019; Wang et al., 2019a; Bassily et al., 2021).

Recently, achieving differential privacy over Riemannian manifolds has also been explored in the context of obtaining Fréchet mean (Reimherr et al., 2021) and, more generally, solving empirical risk minimization problems (Han et al., 2022). Riemannian geometry is a generalization of the Euclidean geometry (Lee, 2006; Absil et al., 2009) and includes several non-linear spaces such as set of positive definite matrices (Bhatia, 2009), set of orthogonal matrices (Edelman et al., 1998; Absil et al., 2009), and hyperbolic space (Ungar, 2008; Nickel & Kiela, 2017), among others. Several machine learning problems such as principal component analysis (Absil et al., 2007), matrix completion (Boumal & Absil, 2011), low-rank tensor learning (Nimishakavi et al., 2018), metric learning (Bhutani et al., 2018), covariance estimation, natural language processing (Jawanpuria et al., 2019), learning embeddings (Nickel & Kiela, 2017; 2018; Suzuki et al., 2019; Qi et al., 2021), etc., may be viewed as an instance of problems on Riemannian manifolds.

In differentially private Riemannian optimization (Han et al., 2022), a key step is to use tangent Gaussian sampling at every iteration to perturb the gradient direction in the tangent space. Han et al. (2022) proposed to use the Markov Chain Monte Carlo (MCMC) method (Robert & Casella, 1999), which is computationally expensive especially on matrix manifolds with large dimensions. When the underlying Riemannian metric is induced from the Euclidean metric, such as for hypersphere, Han et al. (2022) showed one can avoid MCMC via basis construction for the tangent space. For general manifolds of interest, however, a discussion on basis construction and computationally efficient sampling is missing. The sampling step is computationally prohibitive, especially when performing differential private stochastic optimization over Riemannian manifolds, where the number of sampling calls is relatively high compared to the case of deterministic optimization. It should also be noted that generalizing more sophisticated differentially private Euclidean stochastic algorithms like differentially private stochastic variance reduced gradient (Wang et al., 2017) to Riemannian geometry is non-trivial and is an active area of research. The benefits of (non-private) Riemannian stochastic variance reduction gradient (RSVRG) methods over Riemannian stochastic gradient (Bonnabel, 2013) has been studied in existing works (Zhang et al., 2016; Zhou et al., 2019; Han & Gao, 2021; Sato et al., 2019).

In this work, we propose generic fast sampling methods on the tangent space for various matrix manifolds of interest. This makes differentially private Riemannian optimization more practically appealing for real-world applications. We also propose a differentially private Riemannian stochastic variance reduction gradient (RSVRG) and illustrate its efficacy in different applications. Our main contributions are summarized below.

1. **Sampling.** We show that the computationally prohibitive MCMC sampling and explicit basis construction can be avoided in differentially private Riemannian optimization. To this end, we propose two novel sampling strategies: 1) implicit basis construction of the tangent space and 2) novel sampling procedure based on linear isometry between tangent spaces. We show that the proposed sampling strategies are remarkably fast and improve sampling time by orders of magnitude. As a side result, we improve upon the existing utility bounds (Han et al., 2022) of differentially private Riemannian gradient descent methods by removing the dependence on the Riemannian metric-tensor.

2. **DP-SVRG.** We propose a differentially private Riemannian stochastic variance reduced gradient (DP-RSVRG), expanding suite of differentially private stochastic Riemannian optimization methods. We empirically evaluate DP-RSVRG with existing differentially private Riemannian (stochastic) gradient methods and study its benefits.

Organization. The rest of the paper is organized as follows. Section 2 gives background on Riemannian geometry, Riemannian optimization, and differential privacy. We then derive various properties of tangent Gaussian distributions in Section 3 and apply them to improve the existing utility bounds (claim 2). Section 4 presents different strategies for efficient sampling. In Section 5, we develop a differentially private Riemannian stochastic variance reduction gradient algorithm (DP-RSVRG) and Section 6 discusses the empirical results. Section 7 concludes the paper.

2 Preliminaries and related work

Riemannian Geometry. A Riemannian Manifold $\mathcal{M}$ of dimension d is smooth manifold with an inner product structure $\langle ., . \rangle_w$ (i.e., having a Riemannian metric) on every tangent space $T_w\mathcal{M}$. Given a basis $\mathscr{B} = (\beta_1, \dots, \beta_d)$ for $T_w\mathcal{M}$ at $w \in \mathcal{M}$, the Riemannian metric can be represented as a symmetric positive definite matrix G_w and the inner product can be written as $\langle \nu_1, \nu_2 \rangle_w = \overrightarrow{\nu_1}^T G_w \overrightarrow{\nu_2}$, where $\overrightarrow{\nu_1}, \overrightarrow{\nu_2}$ are coordinates of the tangent vectors $\nu_1, \nu_2 \in T_w\mathcal{M}$ in the coordinate system. An induced norm is defined as $\|\nu\|_w = \sqrt{\langle \nu, \nu \rangle_w}$. A geodesic $\gamma : [0, 1] \to \mathcal{M}$ is a locally distance minimizing curve on the manifold with zero tangential acceleration. For any $v \in T_w\mathcal{M}$, the the exponential map is defined as $\mathrm{Exp}_w(v) = \gamma(1)$ where $\gamma(0) = w$ and $\gamma'(0) = v$. If, between any two points $w, w' \in \mathcal{M}$ there exists a unique geodesic connecting them, the exponential map is invertible. Transporting the vectors on the manifold requires the notion of parallel transport. In particular, parallel transport from $w_1 \in \mathcal{M}$ to $w_2 \in \mathcal{M}$ denoted as $\mathrm{PT}^{w_1 \to w_2} : T_{w_1}\mathcal{M} \to T_{w_2}\mathcal{M}$ is a linear isometry (i.e., inner product preserving). The sectional curvature at a point $w \in \mathcal{M}$ is defined as the Gauss curvature of a 2-dimensional subspace of $T_w\mathcal{M}$.

The Riemannian gradient of real valued function $f : \mathcal{M} \rightarrow \mathbb{R}$ denoted as $\operatorname{grad} f(w)$ is a tangent vector s.t for any $\nu \in T_w\mathcal{M}$, $\langle \operatorname{grad} f(w), \nu \rangle_w = \mathrm{D}f[w](\nu) = \langle \nabla f(w), \nu \rangle_2$ where $\mathrm{D}f[w](\nu)$ denotes directional derivative of f at w along ν and $\nabla f(w)$ is the Euclidean gradient. $\langle , \rangle_2$ denotes standard ℓ_2 Euclidean inner product. We refer the readers to (Do Carmo & Flaherty Francis, 1992; Lee, 2006) for a detailed exposition of Riemannian geometry and (Absil et al., 2009; Boumal, 2022) for Riemannian optimization.

Function classes on Riemannian Manifolds. We call a neighbourhood $\mathcal{W} \subseteq \mathcal{M}$ totally normal if for any two points, the exponential map is invertible. Let $\mathcal{W} \subseteq \mathcal{M}$ be a totally normal neighborhood and $D_{\mathcal{W}}$ denotes its diameter and $\kappa_{\min}$ is the lower bound on the sectional curvature of $\mathcal{W}$. A function f is called geodesic L_0-Lipschitz and geodesic L-smooth if for any $w_1, w_2 \in \mathcal{W}$, $|f(w_1) - f(w_2)| \leq L_0 \mathrm{dist}(w_1, w_2)$ and $\left\| \operatorname{grad} f(w_1) - \mathrm{PT}^{w_2 \rightarrow w_1} \operatorname{grad} f(w_2) \right\|_{w_1} \leq L \mathrm{dist}(w_1, w_2)$ respectively. A function f is called geodesic μ-strongly convex if $w, w' = \mathrm{Exp}_w(\zeta) \in \mathcal{W}$, if it satisfies $f(w') \geq f(w) + \langle \operatorname{grad} f(w), \zeta \rangle_w + \frac{\mu}{2} \mathrm{dist}^2(w, w')$. A function f is said to satisfy Riemannian Polyak–Łojasiewicz (PL)condition if there exists $\tau > 0$ $f(w) - f(w^*) \leq \tau \left\| \operatorname{grad} f(w) \right\|_w^2$ for any $w \in \mathcal{M}$ (Zhang et al., 2016; Han & Gao, 2021). The Riemannian PL condition is strictly weaker notion than geodesic strong convexity, i.e., every geodesic μ-strongly convex satisfies Riemannian PL condition (with $\tau = 1/(2\mu)$) and there exists a function that satisfies the Riemannian PL condition but not geodesic strong convexity. We also make use of the curvature constant, defined as $\zeta = \frac{\sqrt{\kappa_{\min}} D_{\mathcal{W}}}{\tanh\left(\sqrt{\kappa_{\min}} D_{\mathcal{W}}\right)}$ if $\kappa_{\min} < 0$ and $\zeta = 1$ if $\kappa_{\min} \geq 0$.

Differential privacy. Let $\mathcal{Z}$ be an input data space and two datasets of size $Z, Z' \in \mathcal{Z}^n$ of size n are called *adjacent* if they differ by at most one element. We represent adjacent datasets Z, Z' by notation $Z \sim Z'$. A manifold-valued randomized mechanism $\mathcal{R} : \mathcal{Z} \rightarrow \mathcal{M}$ is said to be (ϵ, δ)-approximately differentially private (ADP) (Dwork et al., 2006a) if for any two adjacent datasets $Z \sim Z'$ and for all measurable sets $S \subseteq \mathcal{M}$ we have $\mathbb{P}\left[\mathcal{R}(Z) \in S\right] \leq \exp(\epsilon) \mathbb{P}\left[\mathcal{R}(Z') \in S\right] + \delta$. Rényi differenital privacy (RDP) (Mironov, 2017) is a refinement of DP which gives tight privacy bounds under composition of mechanisms. λ-th moment of a mechanism $\mathcal{R}$ is defined as $\mathcal{K}_{\mathcal{R}}(\lambda) = \sup_{Z \sim Z'} \log_{o \sim \mathcal{R}(Z)} [(\frac{p(\mathcal{R}(Z)=o)}{p(\mathcal{R}(Z')=o)})^{\lambda}]$ and mechanism $\mathcal{R}$ is said to satisfy (λ, ρ)-RDP if $\frac{1}{\lambda-1} \mathcal{K}_{\mathcal{R}}(\lambda - 1) \leq \rho$. If mechanism $\mathcal{R} : \mathcal{Z} \rightarrow \mathcal{M}$ is the (adaptive) composition of k mechanism $\{\mathcal{R}_i\}_{i=1}^k$ i.e., $\mathcal{R}_i : \prod_{j=1}^{i-1} \mathcal{M}_j \times \mathcal{Z} \rightarrow \mathcal{M}_i$ then $\mathcal{K}_{\mathcal{R}}(\lambda) \leq \sum_{i=1}^k \mathcal{K}_{\mathcal{R}_i}(\lambda)$. Using moments accountant technique Abadi et al. (2016), (λ, ρ)-RDP mechanism can be given (ϵ, δ)-ADP certificate. We refer the interested readers to (Dwork et al., 2014; Vadhan, 2017) for more details.

Differential privacy on Riemannian manifolds. Reimherr et al. (2021) is the first to consider differential privacy in the Riemannian setting and derived the Riemannian Laplace mechanism based on distribution from (Hajri et al., 2016). Utpala et al. (2022) derive output perturbation for manifold of symmetric positive definite matrices (SPD) with the Log-Euclidean metric based on distribution from (Schwartzman, 2016). While (Reimherr et al., 2021; Utpala et al., 2022) focus on output perturbation, Han et al. (2022) proposes unified differentially private Riemannian optimization through gradient perturbation.

Han et al. (2022) considers the following problem (1) where the parameter of interest lies on a Riemannian manifold $\mathcal{M}$ and $z_i, i = 1, ..., n$ represent the set of data samples, i.e.,

$$\min_{w \in \mathcal{M}} \left\{ F(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \frac{1}{n} \sum_{i=1}^{n} f(w; z_i) \right\}. \tag{1}$$

The aim of differentially private Riemannian optimization is to privatize the solution from a Riemannian optimization solver by injecting noise to the Riemannian gradient, similar as in the Euclidean case. The Riemannian gradient however $\operatorname{grad} F(w)$ belongs to $(T_w\mathcal{M}, \langle , \rangle_w)$, and unlike in the Euclidean case, both the underlying space and inner product varies with the base point w. Accordingly, to perturb the Riemannian gradient, Han et al. (2022) define an intrinsic Gaussian distribution on $T_w\mathcal{M}$ with density $p(\nu) \propto \exp(-\left\| \xi - \mu \right\|_w^2 / 2\sigma^2), \nu \in T_w\mathcal{M}$, and call it tangent Gaussian. They propose differentially private Riemannian gradient and Riemannian stochastic gradient descent algorithms.

3 Properties of tangent Gaussian mechanism

In this section, we derive various properties of the tangent Gaussian distribution that are used for proposed sampling and analysis later. Proofs of the results discussed in this section are provided in Appendix C.1.

We begin with the definitions of the Lebesgue measure on tangent space and the tangent Gaussian distribution. We then show that indeed the tangent Gaussian reduces to the multivariate Gaussian when appropriate basis is constructed. Furthermore, we show how the basis construction at one point on the manifold relates to other points via the notion of isometric parallel transport. These properties allow efficient sampling without the use of MCMC. Particularly, the reduction to a multivariate Gaussian distribution allows sampling to be completed in coordinates and then transformed via a basis. The isometric transport further simplifies the basis construction process as discussed in Section 4.

Definition 1 (Lebesgue measure on tangent space)**.** *Consider a Riemannian manifold $\mathcal{M}$ with intrinsic dimension d. For $w \in \mathcal{M}$, let $\mathscr{B} = \{\beta_1, \ldots, \beta_d\}$ be an orthonormal basis of $T_w\mathcal{M}$ with respect to the Riemannian metric. Define $\phi_{\mathscr{B}} : \mathbb{R}^d \to T_p\mathcal{M}$ as $\phi_{\mathscr{B}}(c_1, \ldots, c_d) = \sum_{i=1}^d c_i\beta_i$. Let λ denote the standard Lebesgue measure on $\mathbb{R}^d$. Then, we define the Lebesgue measure on $T_p\mathcal{M}$ as the pushforward measure $\phi_*^{\mathscr{B}}$ given by $(\phi_*^{\mathscr{B}}\lambda)(S) \triangleq \lambda\left(\phi_{\mathscr{B}}^{-1}(S)\right)$.*

Remark 1. Let $\mathscr{B}_1, \mathscr{B}_2$ be two orthonormal basis of $T_p\mathcal{M}$ then $\phi_*^{\mathscr{B}_1}\lambda = \phi_*^{\mathscr{B}_2}\lambda$ because Lebesgue measure is invariant under orthogonal transformation (with respect to the Riemannian metric). Hence, in the rest of this draft, we drop the superscript $\mathscr{B}$ for clarity and denote the pushforward measure as $\phi_*\lambda$.

We now define the tangent space Gaussian distribution (Han et al., 2022) under the measure in Definition 1.

Definition 2 (Tangent Gaussian)**.** *Let $w \in M$, a tangent vector $\xi \in T_w\mathcal{M}$ follows a tangent space Gaussian distribution at w, denoted as $\xi \sim \mathcal{N}_w(\mu, \sigma^2)$ with mean $\mu \in T_w\mathcal{M}$ and standard deviation $\sigma > 0$ if its density is given by $p_w(\xi) = C_{w,\sigma} \exp\left(-\frac{\|\xi-\mu\|_w^2}{2\sigma^2}\right)$ under the pushforward measure given in Definition 1.*

Remark 1 of (Han et al., 2022) adds without any proof that if $\xi \sim \mathcal{N}_w(\mu, \sigma^2)$ and $\vec{\xi}, \vec{\mu}$ denote the coordinates of ξ, μ, respectively, then $\vec{\xi} \sim \mathcal{N}(\vec{\mu}, G_w^{-1}\sigma^2)$ with G_w being metric tensor at w. With the measure properly defined, we make use of the general change of variable theorem between measure spaces to first show that normalizing constant is independent of basepoint and is given by $C_{w,\sigma} = 1/(2\pi\sigma^2)^{d/2}$. Using this observation, we then show that infact $\vec{\xi} \sim \mathcal{N}(\vec{\mu}, \sigma^2 I)$, which is much stronger than (Han et al., 2022, Remark 1).

Claim 1. *Let $w \in \mathcal{M}$ and $\mathscr{B}$ be any orthonormal basis of $T_p\mathcal{M}$. A random tangent vector $\xi \in T_w\mathcal{M}$ follows the tangent Gaussian with mean $\mu \in T_p\mathcal{M}$ and standard deviation σ if and only if its coordinates in $\mathscr{B}$ denoted as $\vec{\xi}$ follows the d-dimensional Euclidean Gaussian distribution with mean $\vec{\mu} \in \mathbb{R}^d$ and covariance matrix $\sigma^2 I_d$. i.e., $\xi \sim \mathcal{N}_w(\mu, \sigma^2) \iff \vec{\xi} \sim \mathcal{N}(\vec{\mu}, \sigma^2 I_d)$. Hence, the density of the tangent Gaussian is given by $p_\xi(\nu) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\nu-\mu\|_w^2}{2\sigma^2}\right)$.*

A direct consequence of Claim 1 is the following claim where we improve the bounds on the variance of a tangent Gaussian sample (ξ).

Claim 2 (Metric independent utility bound)**.** *Suppose $\xi \sim \mathcal{N}_w(0, \sigma^2)$. Then, $\mathbb{E}\|\xi\|_w^2 \leq d\sigma^2$, where d is the dimension of the manifold.*

We notice in (Han et al., 2022) that the bound is $\mathbb{E}\|\xi\|_w^2 \leq d\sigma^2 c_\ell^{-1}$, where c_ℓ is the smallest eigenvalue of the metric tensor G_w. Nevertheless, from Claim 1, we show under an orthonormal basis $G_w = I_d$, which allows to improve on the variance bound and the subsequent utility bounds by removing the dependence on the metric tensor G_w.

Next, we show a result that relates the tangent Gaussian on different tangent spaces via linear isometry.

Claim 3. *Let $w_1, w_2 \in \mathcal{M}$ and let $\mathrm{I}^{w_1 \to w_2} : T_{w_1}\mathcal{M} \to T_{w_2}\mathcal{M}$ be any linear isometry (i.e., inner product preserving). If $\xi_1 \sim \mathcal{N}_{w_1}(\mu, \sigma_1^2)$ for some $\mu \in T_w\mathcal{M}$ and $\sigma > 0$, then $\mathrm{I}^{w_1 \to w_2}(\xi_1) \sim \mathcal{N}_{w_2}(\mathrm{I}^{w_1 \to w_2}(\mu), \sigma^2)$.*

Linear isometry, as defined above, is present in the form of parallel transport and more generally vector transport (Absil et al., 2009; Huang et al., 2017). We discuss relevant examples of such isometry in the

```
Algorithm 1: Sampling using explicit basis construction
  Input  : Manifold $\mathcal{M}$ of dimension $d$, base point $w \in \mathcal{M}$, Riemannian metric $\langle ., . \rangle_w$, mean $\mu \in$
           standard deviation $\sigma > 0$.
  Output: $\{\xi_1, \ldots, \xi_s\}$, s.t $\xi_i \sim \mathcal{N}_{w_i}(0, \sigma^2)$.
1 for $t = 1, \ldots, s$ do
2     construct orthonormal basis of $T_{w_i}\mathcal{M}$ wrt inner product $\langle ., . \rangle_{w_i}$ and call it $\mathscr{B}_{w_i} = \{\beta_1, \ldots, \beta_d\}$.
3     generate $d$ dimensional coordinates $\mathbf{a} \sim \mathcal{N}(0, \sigma^2 I_d)$.
4     $\xi_t = \sum_{i=1}^{d} a_i \beta_i$.
5 end
```

```
Algorithm 2: Sampling using isometric transportation
  Input  : Manifold $\mathcal{M}$ of dimension $d$, base point $w \in \mathcal{M}$, Riemannian metric $\langle ., . \rangle_p$, mean $\mu \in$ standard
           deviation $\sigma > 0$, reference point $\hat{w}$, orthonormal basis $\mathscr{B} = \{\beta_1, \ldots, \beta_d\}$ at $T_{\hat{w}}\mathcal{M}$.
  Output: $\{\xi_1, \ldots, \xi_s\}$, s.t $\xi_i \sim \mathcal{N}_{w_i}(0, \sigma^2)$.
1 for $t = 1, \ldots, s$ do
2     samples $d$ coordinates $\mathbf{a} \sim \mathcal{N}(0, \sigma^2 I_d)$.
3     Generate tangent Gaussian sample at $\hat{w}$ as $\zeta = \sum_{i=1}^{d} a_i \beta_i$.
4     Isometrically transport $\zeta$ from $T_{\hat{w}}\mathcal{M}$ to $T_{w_t}\mathcal{M}$ : $\xi_t = \mathrm{I}^{\hat{w} \to w_t}(\zeta)$.
5 end
```

context of various manifolds in Section 4. It should be noted Claim 3 is crucial to the design of efficient sampling procedures, introduced in Section 4.

4 Scaling up sampling from tangent Gaussian

As discussed earlier, efficient sampling techniques are especially useful in stochastic optimization setting. In this section, we propose novel and efficient sampling strategies from the tangent Gaussian distribution for different manifolds and different Riemannian metrics. We then show concrete implementation of the proposed sampling strategies for various manifolds of interest. We specifically discuss the SPD, Stiefel, Grassmann, and hypersphere manifolds. In addition, Appendix B.1 discusses sampling procedures for hyperbolic spaces in the Poincaré Ball and the Lorentz Hyperboloid models.

Basis construction. We begin by noting that Claim 1 in Section 3 allows to avoid the computationally expensive MCMC based sampling and and apply the (more efficient) basis construction approach for any matrix manifold. Specifically, given an orthonormal basis $\mathscr{B}$ for $T_w\mathcal{M}$ at $w \in \mathcal{M}$, the sample is generated in the basis $\mathscr{B}$ with the coordinates following the standard Gaussian distribution. However, since the underlying Riemannian metric varies from point to point, finding an orthonormal basis directly for the tangent space at a point can be difficult. In many cases, a basis that is not orthonormal can be easily obtained, which we can then orthogonalize by the Gram-Schmidt method. We term this **explicit basis construction** strategy for sampling and summarize it in Algorithm 1. Please note that (Han et al., 2022) has proposed explicit basis construction strategy limited to manifolds endowed with the Euclidean metric. On the other hand, Claim 1 allows generalizing this strategy to manifolds endowed with general Riemannian metric.

Even though the basis construction strategy is faster than MCMC, it is still computationally expensive as we need to construct basis at every iteration. For certain manifolds, however, it is possible to avoid explicit basis construction: instead of constructing $\{\beta_1, \ldots, \beta_d\}$ (Step 2 of Algorithm 1) and then producing the sample with coordinates $\mathbf{a}$ (Step 3 of Algorithm 1), we use the structure on the manifold and combine the two steps into one step efficiently. We term this as **implicit basis construction** strategy for sampling. In Sections 4.2-4.4, we detail this implicit approach for various manifolds. As discussed later in Section 6, we empirically observe that the proposed implicit strategy is much faster than the explicit strategy.

Table 1: Reference points $\hat{w}$ for Algorithm 2. $\mathbf{I} \in \mathbb{R}^{m\times m}$ denotes the identity matrix. $(\mathbf{e}_1, \dots, \mathbf{e}_r)$ denotes the standard basis vectors of $\mathbb{R}^m$ and $\tilde{\mathbf{e}}_1$ the first standard basis vector of $\mathbb{R}^{m-1}$. $\mathbf{o} \in \mathbb{R}^m$ denotes zero vector. $\langle, \rangle_{\mathrm{F}}, \langle, \rangle_2$ denote the standard Euclidean inner product on matrices and vectors respectively. We observe that at specific reference points, both the Riemannian metrics and tangent spaces can be simplified.

Manifold	Metric	Reference point $\hat{w}$	Tangent space $T_{\hat{w}}\mathcal{M}$	Metric $\langle, \rangle_{\hat{w}}$	Parallel transport
SPD	Affine-Invariant metric	$\mathbf{I} \in \mathbb{R}^{m\times m}$	$\mathrm{SYM}(m)$	$\langle, \rangle_{\mathrm{F}}$	Closed form
	Bures-Wasserstein metric	$\mathbf{I} \in \mathbb{R}^{m\times m}$	$\mathrm{SYM}(m)$	$\langle, \rangle_{\mathrm{F}}/4$	No closed form
	Log-Euclidean metric	$\mathbf{I} \in \mathbb{R}^{m\times m}$	$\mathrm{SYM}(m)$	$\langle, \rangle_{\mathrm{F}}$	Closed form
Grassmann	Grassman canonical metric	$[\mathbf{e}_1, \dots, \mathbf{e}_r] \in \mathbb{R}^{m\times r}$	$\{0\}^{r\times r} \times \mathbb{R}^{(m-r)\times r}$	$\langle, \rangle_{\mathrm{F}}$	Closed form
Stiefel	Stiefel canonical metric	$[\mathbf{e}_1, \dots, \mathbf{e}_r] \in \mathbb{R}^{m\times r}$	$\mathrm{SKEW}(r) \times \mathbb{R}^{(m-r)\times r}$	$\langle, \rangle_{\mathrm{F}}/2$	No closed form
Hypersphere	Hypersphere canonical metric	$\mathbf{e}_1 \in \mathbb{R}^m$	$\{0\} \times \mathbb{R}^{m-1}$	$\langle, \rangle_2$	Closed form
Hyperoblic	Poincaré ball metric	$\mathbf{o} \in \mathbb{R}^m$	$\mathbb{R}^m$	$\langle, \rangle_2$	Closed form
	Lorentz hyperboloid metric	$\mathbf{e}_1 \in \mathbb{R}^m$	$\{0\} \times \mathbb{R}^{m-1}$	$\langle, \rangle_2$	Closed from

Isometric transportation. We now discuss another novel sampling strategy, which altogether avoids basis construction at every iteration and is more amenable to stochastic optimization settings.

To mitigate the issue, we use Claim 3 in Section 3. In particular, Claim 3 suggests that to sample from the tangent Gaussian on $T_{w_t}\mathcal{M}$ for some $w_t \in \mathcal{M}$, one can simply sample from the tangent Gaussian from any other base point $\hat{w}$ and then transport the sample using any linear isometry from w_t to $\hat{w}$. Hence, one can fix a **reference point** $\hat{w}$ apriori and always sample from the tangent Gaussian at $T_{\hat{w}}\mathcal{M}$, and then, isometrically transport from $\hat{w}$ to required point w_t. We term this procedure as isometric transportation and summarize it in Algorithm 2.

The proposed isometric transportation based sampling approach avoids repeated basis construction and possibly orthogonalization process, when sampling from different base points. Furthermore, reference point w' can be chosen such that tangent space and/or metric has simple form that is amenable to sampling. Algorithm 2 summarizes the proposed sampling procedure. Algorithm 2 achieves significant improvement in efficiency and renders computational cost of privatizing the training process negligible, especially for high dimensional matrix manifolds. Please refer to Tables 1 and 3 for a summary of reference points and other details useful for implementing Algorithm 2 on various manifolds.

The rest of the section deals with how to implement Algorithms 1 and 2 for several important manifolds.

4.1 SPD manifold

Let $\mathrm{SPD}(m)$ denote the set of symmetric positive definite matrices of size $m \times m$ and for $\mathbf{X} \in \mathrm{SPD}(m)$, the tangent space at $\mathbf{X}$ is $T_{\mathbf{X}}\mathrm{SPD}(m) = \mathrm{SYM}(m)$, where $\mathrm{SYM}(m)$ denotes the set of symmetric matrices of size $m \times m$. We start with the standard basis for $\mathrm{SYM}(m)$, given by $\mathscr{B} = \{\mathbf{e}_i\mathbf{e}_j^T : i = 1 \dots m, j = i+1, \dots, m\}$. It can be shown that under the Euclidean metric, i.e., $\langle \mathbf{U}, \mathbf{V}\rangle^{\mathrm{E}}_{\mathbf{X}} = \mathrm{Tr}\,[\mathbf{U}\mathbf{V}]$, $\mathscr{B}$ forms an *orthogonal* basis of $(\mathrm{SYM}(k), \langle, \rangle^{\mathrm{E}}_{\mathbf{X}})$ and can be transformed into an *orthonormal* basis by scaling. $\mathbf{U}$ and $\mathbf{V}$ belong to the tangent space $T_{\mathbf{X}}\mathrm{SPD}(m)$. However, the (fixed) Euclidean metric fails to capture the geometric properties of the underlying space and alternative (varying) Riemannian metrics are usually preferred. To this end, we consider three Riemannian metrics:

- Affine-Invariant (AI) metric (Pennec, 2006; Bhatia, 2009), defined as $\langle \mathbf{U}, \mathbf{V}\rangle^{\mathrm{AI}}_{\mathbf{X}} := \mathrm{Tr}\left[\mathbf{X}^{-1}\mathbf{U}\mathbf{X}^{-1}\mathbf{V}\right]$.
- Bures-Wasserstein (BW) metric (Bhatia et al., 2019), defined as $\langle \mathbf{U}, \mathbf{V}\rangle^{\mathrm{BW}}_{\mathbf{X}} := \mathrm{Tr}[\mathcal{L}_{\mathbf{X}}[\mathbf{U}]\mathbf{V}]$, where $\mathcal{L}_{\mathbf{X}}[\mathbf{U}]$ is the solution to the matrix equation $\mathcal{L}_{\mathbf{X}}[\mathbf{U}]\mathbf{U} + \mathbf{U}\mathcal{L}_{\mathbf{X}}[\mathbf{U}] = \mathbf{U}$.
- Log-Euclidean (LE) metric (Arsigny et al., 2007), defined as $\langle \mathbf{U}, \mathbf{V}\rangle^{\mathrm{LE}}_{\mathbf{X}} := \mathrm{Tr}\,[\mathrm{DLogm}[\mathbf{X}](\mathbf{U}).\mathrm{DLogm}[\mathbf{X}](\mathbf{V})]$, where $\mathrm{DLogm}[\mathbf{X}](\mathbf{U})$ is directional derivative of matrix logarithm at $\mathbf{X}$ evaluated at $\mathbf{U}$.

For all the three metrics, $\mathscr{B}$, the standard basis of $\mathrm{SYM}(m)$ is no longer an orthogonal basis for each point. So, we use the Gram-Schmidt process on $\mathscr{B}$ to get an orthonormal basis for Algorithm 1.

For implementing Algorithm 2, $\mathbf{I}$ is used as the reference point for all the three metrics because $\langle,\rangle_{\mathbf{I}}^{\mathrm{AI}} = \langle,\rangle_{\mathbf{I}}^{\mathrm{LE}} = \langle,\rangle_{\mathrm{F}}$ and $\langle,\rangle_{\mathbf{I}}^{\mathrm{BW}} = \langle,\rangle_{\mathrm{F}}/4$ and hence $\mathscr{B}$ can be turned into orthonormal basis without the Gram-Schmidt process at $\mathbf{I}$. For the AI and LE metrics, the parallel transport expression is available in closed-form, but for BW it can be obtained by numerically solving a first-order ODE.

4.2 Stiefel manifold

The Stiefel manifold is the set of column orthonormal matrices, i.e., $\mathrm{St}(m,r) = \{\mathbf{X} \in \mathbb{R}^{m\times r} | \mathbf{X}^T\mathbf{X} = \mathbf{I}\}$ and its tangent space is $T_{\mathbf{X}}\mathrm{St}(m,r) = \{\mathbf{U} \in \mathbb{R}^{m\times r} | \mathbf{U}^T\mathbf{X} + \mathbf{X}^T\mathbf{U} = \mathbf{O}\}$. The canonical Riemannian metric is defined as $\langle \mathbf{U}, \mathbf{V}\rangle_{\mathbf{X}} = \mathrm{Tr}[\mathbf{U}^T(\mathbf{I} - \frac{1}{2}\mathbf{X}\mathbf{X}^T)\mathbf{V}]$ (Edelman et al., 1998; Absil et al., 2007; 2009).

For a point $\mathbf{X} \in \mathrm{St}(m,r)$, an explicit orthonormal basis of $T_{\mathbf{X}}\mathrm{St}(m,r)$ with respect to the canonical metric is $\mathscr{B} = \{\mathbf{X}(\mathbf{e}_i\mathbf{e}_j^T - \mathbf{e}_j\mathbf{e}_i^T) : i = 1 \dots r, j = i+1, \dots, r\} \cup \{\mathbf{X}_\perp \tilde{\mathbf{e}}_i \mathbf{e}_j^T\}$, where $(\mathbf{e}_1, \dots, \mathbf{e}_r), (\tilde{\mathbf{e}}_1, \dots, \tilde{\mathbf{e}}_{m-r})$ are the standard basis of $\mathbb{R}^r$ and $\mathbb{R}^{m-r}$ respectively and $\mathbf{X}_\perp \in \mathbb{R}^{m\times(m-r)}$ whose columns form an orthonormal basis of the orthogonal complement of column space of $\mathbf{X}$ (Huang et al., 2017, Proposition 41). Huang et al. (2017) show how to avoid explicit basis construction by representing $\mathbf{X}, \mathbf{X}_\perp$ in terms of Householder matrices (see Appendix B.2 for more details). This can be beneficial especially when $r \ll m$.

We remark that there is no closed-form expression for parallel transport under the canonical metric (for $r > 1$). However, we can construct an isometric vector transport by using the orthonormal basis, a strategy known as transportation by parallelization (Huang et al., 2015, Section 7.2).

For the Stiefel manifold with $r > 1$, Algorithm 1 with implicit basis construction and Algorithm 2 coincide. We discuss the case $r = 1$ separately below in Section 4.4.

4.3 Grassmann manifold

The Grassmann manifold $\mathrm{Gr}(m,r)$ consists of r dimensional linear subspaces of $\mathbb{R}^m$ $(r \leq m)$ and is usually represented as $\mathrm{Gr}(m,r) = \{\mathrm{colspan}(\mathbf{X}) | \mathbf{X} \in \mathbb{R}^{m\times r}, \mathbf{X}^T\mathbf{X} = \mathbf{I}_r\}$, where colspan denotes the column space. The tangent space is $T_{\mathbf{X}}\mathrm{Gr}(m,r) = \{\mathbf{U} \in \mathbb{R}^{m\times r} | \mathbf{U} \in \mathbb{R}^{m\times r}, \mathbf{X}^T\mathbf{U} = \mathbf{O}_r\}$ (Edelman et al., 1998; Absil et al., 2007; 2009). The Grassmann canonical metric coincides with the Euclidean metric, i.e., $\langle \mathbf{U}, \mathbf{V}\rangle_{\mathbf{X}} = \mathrm{Tr}\,[\mathbf{U}^T\mathbf{V}]$, for $\mathbf{U}, \mathbf{V} \in T_{\mathbf{X}}\mathrm{Gr}(m,r)$ (Edelman et al., 1998).

An explicit orthonormal basis for $T_{\mathbf{X}}\mathrm{Gr}(m,r)$ is the second part of the basis for Stiefel manifold, which is $\mathscr{B} = \{\mathbf{X}_\perp \tilde{\mathbf{e}}_i \mathbf{e}_j^T : i = 1 \dots r, j = i+1, \dots, r\}$. Similar to the Stiefel manifold, one can also perform the implicit basis construction (see Appendix B.2). For Algorithm 2 we use the reference point as $\mathbf{X} = [\mathbf{e}_1, \dots, \mathbf{e}_r] \in \mathbb{R}^{m\times r}$ where its tangent space reduces to $T_{\mathbf{X}}\mathrm{Gr}(m,r) = \{0\}^{r\times r} \times \mathbb{R}^{m-r}$. The parallel transport is also available in closed-form.

4.4 Hypersphere

The hypersphere is $\mathbb{S}^m = \{\mathbf{x} \in \mathbb{R}^m | \, \|\mathbf{x}\|_2 = 1\}$ and tangent space is given by $T_{\mathbf{x}}\mathbb{S}^m = \{\mathbf{u} \in \mathbb{R}^m | \langle \mathbf{x}, \mathbf{v}\rangle_2 = 0\}$. The Riemannian metric is the induced Euclidean metric, i.e., $\langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{x}} = \langle \mathbf{u}, \mathbf{v}\rangle_2$. This coincides with the Stiefel manifold case for $r = 1$. We follow the strategies mentiond in Section 4.2 with the additional information that the parallel transport expression is available in closed-form for the hypersphere.

Specifically, for implementing Algorithm 1, the orthonormal basis for $T_{\mathbf{x}}\mathbb{S}^{m-1}$ is $\mathscr{B} = \{\{0\} \times \tilde{\mathbf{e}}_1, \dots, \{0\} \times \tilde{\mathbf{e}}_{m-1}\}$, where $(\tilde{\mathbf{e}}_1, \dots, \tilde{\mathbf{e}}_{m-1})$ is the standard orthonormal basis for $\mathbb{R}^{m-1}$. For implementing Algorithm 2, we select the reference point as $\mathbf{x} = (1, \dots, 0)$ because $T_{\mathbf{x}}\mathbb{S}^m = \{0\} \times \mathbb{R}^{m-1}$.

5 Private Riemannian variance reduced stochastic optimization

Variance reduced stochastic optimization methods (Roux et al., 2012; Johnson & Zhang, 2013; Defazio et al., 2014; Reddi et al., 2016) employ a hybrid update rule that uses both full gradient and stochastic

Algorithm 3: DP-RSVRG

```
Input : update frequency $m$, learning rate $\eta$, number of epochs $S$, clipping parameters $\mathcal{C}_0, \mathcal{C}_1$, and
        initial iterate $w^0$.
1  initialize $\tilde{w} = w^0$.
2  for $s = 0, 1, \ldots, S-1$ do
3      $w_0^{s+1} = \tilde{w}^s$.
4      $g^{s+1} = \frac{1}{n}\sum_{i=1}^n \mathrm{clip}_{\mathcal{C}_0}(\mathrm{grad} f(\tilde{w}^s; z_i))$.
5      for $t = 0, 1, \ldots, m-1$ do
6          Randomly pick $i_t \in \{1, \ldots, n\}$
7          $v_t^{s+1} = \mathrm{clip}_{\mathcal{C}_1}(\mathrm{grad} f(w_t^{s+1}; z_{i_t})) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}}\big(\mathrm{clip}_{\mathcal{C}_1}(\mathrm{grad} f(\tilde{w}^s; z_{i_t})) - g^{s+1}\big) + \epsilon_t^{s+1}$, where
           $\epsilon_t^{s+1} \sim \mathcal{N}_{w_t^{s+1}}(0, \sigma^2)$.
8          $w_{t+1}^{s+1} = \mathrm{Exp}_{w_t^{s+1}}(-\eta v_t^{s+1})$.
9      end
10     Set $\tilde{w}_a = w_m^{s+1}$.
11 end
12 Output I : $w^{\mathrm{priv}} = \tilde{w}^S$.
13 Output II : $w^{\mathrm{priv}}$ is choosen uniformly randomly from $\{\{w_t^{s+1}\}_{t=0}^{m-1}\}_{s=0}^{S-1}$.
```

gradient simultaneously. By doing so, variance reduced methods improve the gradient complexity compared to stochastic and full gradient descent by requiring less gradient calls to achieve the same convergence rates than full gradient descent. Many variance reduction strategies that work in the Euclidean space have also been generalized to manifolds (Zhang et al., 2016; Sato et al., 2019; Zhou et al., 2019; Han & Gao, 2021).

In this section, we privatize the Riemannian stochastic variance reduced gradient (RSVRG) algorithm (Zhang et al., 2016; Sato et al., 2019) for solving (1) and develop differentially private RSVRG, henceforth denoted by DP-RSVRG. Our proposed DP-RSVRG is summarized in Algorithm 3. DP-RSVRG with restarts is presented as Algorithm 4.

DP-RSVRG takes two loops where in each inner loop, an unbiased, variance reduced stochastic gradient is constructed by correcting the Riemannian stochastic gradient with the full gradient calculated at the outer loop. We add noise from the tangent Gaussian distribution to the variance reduced gradient. The clipping operation $\mathrm{clip}_\tau : T_w\mathcal{M} \to T_w\mathcal{M}$ is defined as $\mathrm{clip}_\tau(\nu) = \max\{\frac{\tau}{\|\nu\|_w}, 1\}\nu$ and it ensures norm of ν is at most τ. The norm of gradients in full gradient are clipped with parameter $\mathcal{C}_0$ and in variance reduced gradient with parameter $\mathcal{C}_1$, respectively. PT refers to the parallel transport operation.

5.1 Privacy guarantee

In this section, we analyze the privacy guarantees of DP-RSVRG. We begin by noting that variance reduced stochastic gradient has a deterministic and a subsampled component. Hence, Step 7 of Algorithm 3 can be equivalently re-written as

$$v_t^{s+1} = \mathrm{clip}_{\mathcal{C}_1}(\mathrm{grad} f(w_t^{s+1}; z_{i_t})) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}}\big(\mathrm{clip}_{\mathcal{C}_1}(\mathrm{grad} f(\tilde{w}^s; z_{i_t})) - (g^{s+1} + \xi_{t1}^s)\big) + \xi_{t2}^s, \tag{2}$$

where $\xi_{t1}^s \sim \mathcal{N}_{\tilde{w}^s}(0, \sigma_1^2)$ and $\xi_{t2}^s \sim \mathcal{N}_{w_t^{s+1}}(0, \sigma_2^2)$. Specifically, the noise variance σ^2 is split into into σ_1^2 for the full gradient query and σ_2^2 for the variance reduced stochastic gradient query such that $\sigma_1^2 + \sigma_2^2 = \sigma^2$. Claim 3 ensures that $\mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}}\xi_{t1}^s + \xi_{t2}^{s+1} = \epsilon_t^{s+1} \sim \mathcal{N}_{w_t^{s+1}}(0, \sigma^2)$. Hence, (2) can be viewed as a composition of a full gradient tangent Gaussian mechanism $\mathcal{R}^s(Z) = r^{s+1} = \frac{1}{n}\sum_{i=1}^n \mathrm{clip}_{\mathcal{C}_0}(\mathrm{grad} f(\tilde{w}^s; z_i)) + \xi_{t1}^s; \xi_{t1}^s \sim \mathcal{N}_{\tilde{w}^s}(0, \sigma_1^2)$ and a variance reduced Gaussian mechanism $\mathcal{R}_t^{s+1}(Z) = \mathrm{clip}_{\mathcal{C}_1}(\mathrm{grad} f(w_t^{s+1}; z_{i_t})) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}}\big(\mathrm{clip}_{\mathcal{C}_1}(\mathrm{grad} f(\tilde{w}^s; z_{i_t})) - r^{s+1}\big) + \xi_{t2}^{s+1}; \xi_{t2}^{s+1} \sim \mathcal{N}_{w_t^{s+1}}(0, \sigma_2^2)$. We now prove the moments bounds on the full gradient mechanism $\mathcal{K}_{\mathcal{R}^s}$ and variance reduced mechanism $\mathcal{K}_{\mathcal{R}_t^{s+1}}$ in the following claims and the proofs are given in Section C.3.1.

Algorithm 4: DP-RSVRG with restarts

```
Input : update frequency $m$, learning rate $\eta$, number of epochs $S$, and initial iterate $w^0$.
for $k = 0, 1, \ldots, K-1$ do
    $w^{k+1} = \text{DP-RSVRG}(m, \eta, S, w^k)$ with output option II.
end
```

Claim 4. *The moments bounds satisfy* $\mathcal{K}_{\mathcal{R}^s}(\lambda) \leq \frac{2\lambda(\lambda+1)\mathcal{C}_0^2}{n^2\sigma_1^2}$ *and* $\mathcal{K}_{\mathcal{R}_t^{s+1}}(\lambda) \leq \frac{8\lambda(\lambda+1)\mathcal{C}_1^2}{\sigma_2^2}$.

Now we derive the moments bound on subsampled version of $\mathcal{R}_t^{s+1}$ using the results given in (Wang et al., 2019b;c) and the proof is given in Section C.3.2.

Claim 5. *Define* subsample : $\mathcal{Z}^n \to \mathcal{Z}$ *as the process of sampling a single data point from* n *data points uniformly randomly. Define the subsampled mechanism for* $\mathcal{R}_t^{s+1}$ *as* ${}^{\text{sub}}\mathcal{R}_t^{s+1} = \mathcal{R}_t^{s+1} \circ \text{subsample}$. *Suppose* $\sigma_2 \geq 12\mathcal{C}_1^2$ *and* $\lambda \leq 2/3\sigma_2^2 \log\big(n(\lambda+1)(1+(\sigma_2^2/16\mathcal{C}_1^2))\big)$, *we have* $\mathcal{K}_{{}^{\text{sub}}\mathcal{R}_t^{s+1}}(\lambda) \leq \frac{28\lambda(\lambda+1)\mathcal{C}_1^2}{n^2\sigma_2^2}$.

The full mechanism $\mathcal{R}$ can be seen as an adaptive composition of $\{\{\mathcal{K}_{{}^{\text{sub}}\mathcal{R}_t^{s+1}}\}_{t=0}^{m-1}\}_{s=0}^{S-1}$ and $\{\{\mathcal{K}_{\mathcal{R}^s}\}_{t=0}^{m-1}\}_{s=0}^{S-1}$. Since $\sigma_1^2+\sigma_2^2 = \sigma^2$, we can rewrite $\sigma_1^2 = \alpha\sigma^2$, $\sigma_2^2 = (1-\alpha)\sigma^2$ for some $\alpha \in (0,1)$. Using this claim, minimizing over α, and setting $\mathcal{C} = \max\{\mathcal{C}_0, \mathcal{C}_1\}$, we have

$$\begin{aligned}
&\mathcal{K}_{\mathcal{R}}(\lambda) \leq \sum_{t=0}^{m}\sum_{s=0}^{S-1} \mathcal{K}_{{}^{\text{sub}}\mathcal{R}_t^{s+1}}(\lambda) + \sum_{t=0}^{m}\sum_{s=0}^{S-1} \mathcal{K}_{\mathcal{R}^{s+1}}(\lambda) \leq \frac{2mS\lambda(\lambda+1)\mathcal{C}_0^2}{n^2\sigma_1^2} + \frac{28mS\lambda(\lambda+1)\mathcal{C}_1^2}{n^2\sigma_2^2} \\
&\Rightarrow \mathcal{K}_{\mathcal{R}}(\lambda) \leq \min_{\alpha\in(0,1)} \frac{mS\lambda(\lambda+1)\mathcal{C}^2}{n^2\sigma^2}\left[\frac{2}{\alpha} + \frac{28}{1-\alpha}\right]. \qquad (3)
\end{aligned}$$

It should be noted that for a given λ, the minimization over α has a closed-form solution.

The moments bound $\mathcal{K}_{\mathcal{R}}$ given in (3) can be converted to (ϵ, δ) guarantee using conversion rules, e.g., based on (Mironov, 2017, Proposition 3): Given $0 < \delta < 1$, $\epsilon = \min_{\lambda \geq 1} \frac{\mathcal{K}_{\mathcal{R}}(\lambda-1)+\log 1/\delta}{\lambda-1}$. Recently, however, the optimal conversion rule has been given in (Asoodeh et al., 2020, Theorem 3) for which there exists no closed-form expression but can be solved numerically to get ϵ. The solver is available in the `autodp` library (Wang et al., 2019c). The above result connecting the moment bound $\mathcal{K}_{\mathcal{R}}$ with α in (3) implies that tighter (ϵ, δ) guarantees can be obtained by optimizing over α, i.e., by exploiting the inter-play between the the noise added to the full gradient and that to the variance reduced gradient.

It should be emphasized that in the Euclidean setting, Wang et al. (2017) have not considered optimization of α as in (3). We empirically show such optimization of α obtains significant improvement in privacy in Section 6.2. We end this section with the following privacy result for Algorithms 3 and 4.

Claim 6. *Algorithms 3 and 4 are* (ϵ, δ)*-differentially private with* $\sigma^2 \geq c_1 \frac{mS\log(1/\delta)\mathcal{C}^2}{n^2\epsilon^2}$ *and* $\sigma^2 \geq c_2 \frac{mSK\log(1/\delta)\mathcal{C}^2}{n^2\epsilon^2}$, *respectively, for some positive constants* c_1, c_2 *and* $\mathcal{C} = \max\{\mathcal{C}_0, \mathcal{C}_1\}$.

5.2 Utility guarantee

In this section, we prove the utility guarantees of DP-RSVRG under various function classes on manifolds including geodesic strong convex functions, general nonconvex functions, and functions that satisfy the Riemannian Polyak–Łojasiewicz (PL) condition. In particular, geodesic strong convexity and Riemannian PL condition generalize the notion of strong convexity and PL condition (Polyak, 1963) to manifolds, allowing fast convergence (for problems satisfying these conditions) to the global optimality when optimizing on manifolds. The proofs of the results discussed in this section are discussed in Sections C.4.1-C.4.3.

Let $\mathcal{W} \subseteq \mathcal{M}$ be a totally normal neighborhood and $D_{\mathcal{W}}$ denotes its diameter and $\kappa_{\min}$ is the lower bound on curvature of $\mathcal{W}$. For more detailed introduction, please see Sections 2 and A. Following (Zhang & Sra, 2016; Han & Gao, 2021; Han et al., 2022), we make the below standard assumption.

Assumption 1. *Each f_i in (1) is L-geodesically smooth and L_0-geodesically Lipschitz over $\mathcal{W}$.*

The gradient complexity of an algorithm is measured in the number of incremental first-order oracle (IFO) calls needed. An IFO (Agarwal & Bottou, 2015) takes an index $i \in [n]$, $w \in \mathcal{W}$ and returns $(f_i(w), \operatorname{grad} f_i(w)) \in \mathbb{R} \times T_w\mathcal{M}$. Also, for readability we hide log factors in all variables through notation $\tilde{\mathcal{O}}$ in utility bounds and gradient complexities. For exact expression, see (11), (12) for μ-strongly convex, 16, 17 for non convex, and 18, 19 for the Riemannian PL functions.

Theorem 7 (Utility under geodesic strong convexity). *Suppose that Assumption 1 holds and each f_i is μ-strongly geodesic convex over $\mathcal{W}$. If we run the Algorithm 3 with learning rate $\eta = \mathcal{O}(\frac{\mu}{\zeta L^2})$, frequency $m = \tilde{\mathcal{O}}(\frac{\zeta L^2}{\mu^2})$ for $S = \mathcal{O}(\log(\frac{n\epsilon\mu}{\log(1/\delta)\zeta L_0^2 d}))$ outer loops with output* **I**, *then $\mathbb{E}[F(w^{\text{priv}}) - F(w^*)] = \tilde{\mathcal{O}}\left(\frac{d\zeta L L_0^2 \log(1/\delta)\mathbb{E}[\text{dist}^2(w_S^0, w^*)]}{\mu^2 n^2 \epsilon^2}\right)$. Furthermore, the gradient complexity is given by $\tilde{\mathcal{O}}(n + \frac{\zeta L^2}{\mu^2})$.*

Theorem 8 (Utility under nonconvex functions). *Suppose that Assumption 1 holds. If we run the Algorithm 3 with output* **II**, *learning rate $\eta = \mathcal{O}(\frac{1}{Ln^{2/3}\zeta^{1/2}})$, frequency $m = \Theta(n)$ and for $S = \sqrt{\frac{L\zeta}{d\log(1/\delta)}}\frac{n^{2/3}\epsilon}{L_0}$ outer loops, then $\mathbb{E}\|\operatorname{grad} F(w^{\text{priv}})\|^2 \leq \frac{L_0\sqrt{dL\log(1/\delta)}}{n\epsilon}$. The gradient complexity is given by $\mathcal{O}(\sqrt{\frac{L\zeta}{d\log(1/\delta)}}\frac{n^{5/3}\epsilon}{L_0})$.*

We now use Algorithm 4 obtaining utility guarantee under the Riemannian PL condition.

Theorem 9 (Utility under Riemannian PL condition). *Suppose that Assumption 1 holds and $F = \frac{1}{n}\sum_{i=1}^n f_i(w)$ satisfies the Riemannian PL condition with parameter τ. If we run Algorithm 4 with learning rate $\eta = \mathcal{O}(\frac{1}{Ln^{2/3}\zeta^{1/2}})$, frequency $m = \Theta(n)$, $S = \mathcal{O}(1)$, and $K = \log(\frac{n^2\epsilon^2}{dL\tau^2\log(1/\delta)L_0^2})$, then $\mathbb{E}[F(w^{\text{priv}}) - F(w^*)] \leq \tilde{\mathcal{O}}(\frac{dL\tau^2\log(1/\delta)L_0^2}{n^2\epsilon^2})$. Furthermore, the gradient complexity is given by $\tilde{\mathcal{O}}(L\tau\zeta^{1/2}n^{2/3})$.*

5.3 Discussion: DP-RGD vs DP-RSGD vs DP-RSVRG

In this section we compare DP-RSVRG with DP-RGD Han et al. (2022) and DP-RSGD Han et al. (2022).

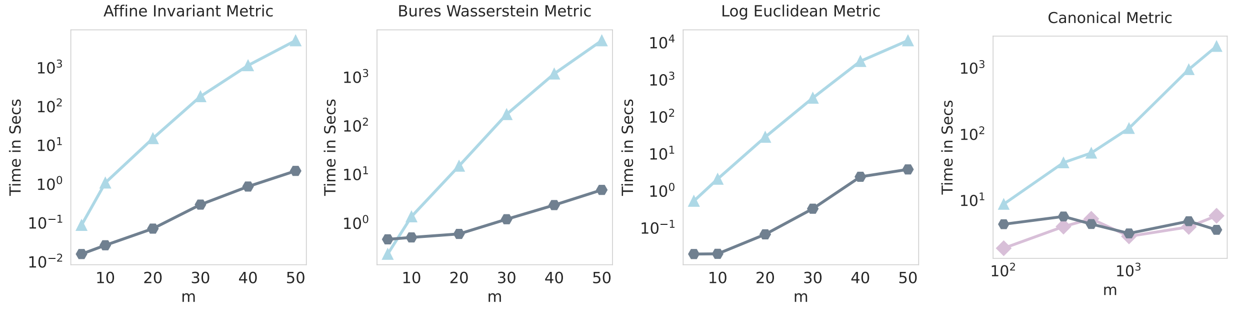
1. **Strongly geodesic convex**: For μ-strongly g-convex functions DP-RGD and DP-RSGD obtain utility bound $\mathcal{O}\left(\frac{d\zeta L_0^2 \log(1/\delta)\mathbb{E}[\text{dist}^2(w_0, w^*)]}{\mu n^2\epsilon^2}\right)$ with gradient complexities n^2 and n^3, respectively (Han et al., 2022, Theorem 3). Note that μ-stronlgy g-convex (Han et al., 2022, Theorem 3) doesn't make any assumption about L-g-smoothness for DP-RGD, DP-RSGD and while Theorem 7 make L-g-smoothness for DP-RSVRG.

   For μ-strongly g-convex functions, we see that one gets better gradient complexity by viewing it as satisfying the Riemannian PL condition with parameter $\tau = 1/(2\mu)$. For utility bounds for Riemannian PL condition one needs L-g-smoothness assumption for all DP-RGD, DP-RSGD and DP-RSVRG and hence provides ground for fair comparison.

   Now, using above observation with (Han et al., 2022, Theorem 4) DP-RGD and DP-RSGD obtain utility of $\tilde{\mathcal{O}}\left(\frac{dL_0^2\log(1/\delta)\mathbb{E}[F(w)-F(w^*)]}{n^2\epsilon^2}\right) = \tilde{\mathcal{O}}\left(\frac{dLL_0^2\log(1/\delta)\mathbb{E}[\text{dist}^2(w_0,w^*)]}{\mu n^2\epsilon^2}\right)$ (because $F(w_0) - F(w^*) \leq \frac{L}{2}\text{dist}^2(w_0, w^*)$ by definition of L-g-smoothness) with gradient complexity $\mathcal{O}\left(n\log\left(\frac{n^2\epsilon^2}{dL_0^2\log(1/\delta)}\right)\right)$ and $\mathcal{O}\left(\log\left(\frac{n^2\epsilon^2}{dL_0^2\log(1/\delta)}\right)\right)$ respectively. From Theorem 9, DP-RSVRG obtains utility of $\tilde{\mathcal{O}}\left(\frac{dL^2L_0^2\log(1/\delta)\mathbb{E}[\text{dist}^2(w_0,w^*)]}{\mu^2n^2\epsilon^2}\right)$ with gradient complexity $\tilde{\mathcal{O}}(\frac{L}{\mu}\zeta^{1/2}n^{2/3})$. Note that $\mathbb{E}[F(w^{\text{priv}}_{\text{DP-RGD}}) - F(w^*)] = \mathbb{E}[F(w^{\text{priv}}_{\text{DP-RSGD}}) - F(w^*)] \leq \frac{L}{\mu}\mathbb{E}[F(w^{\text{priv}}_{\text{DP-RSVRG}}) - F(w^*)]$. As $\frac{L}{\mu} > 1$ we can split into two cases.

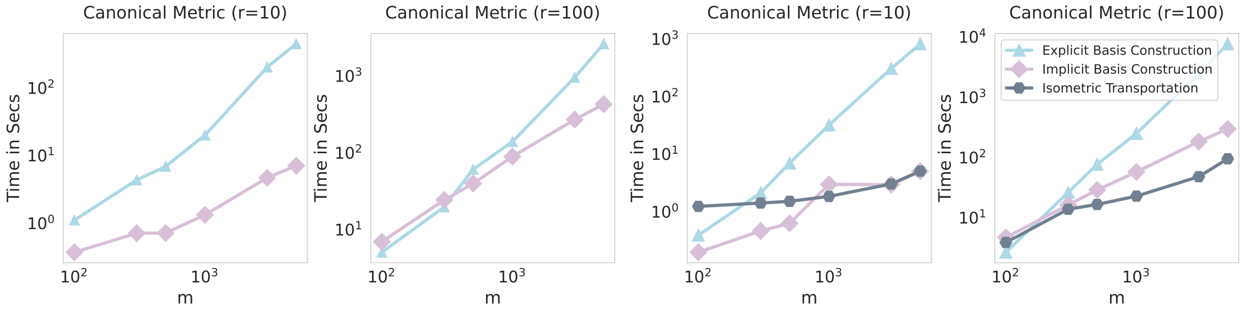
   Case (i) $\frac{L}{\mu} = \mathcal{O}(1)$: In this case, DP-RSVRG has same utility bounds compared to DP-RGD and DP-RSGD.

   Case (ii) $\frac{L}{\mu} = \Omega(n^a)$ for some $a > 0$: In this case, DP-RSVRG has worse utility bounds compared to DP-RGD and DP-RSGD.



(a) SPD manifold.

(b) Hypersphere manifold.



(c) Stiefel manifold.

(d) Grassmann manifold.

Figure 1: Benchmarking of explicit basis construction, implicit basis construction, and isometric transportation sampling strategies for the SPD, Stiefel, Grassmann, and hypersphere manifolds. It should be mentioned that implicit basis construction strategy is unavailable for the SPD manifold, while implicit basis construction and isometric transportation sampling strategy coincide for the Stiefel manifold. We consistently see the good performance of the proposed sampling strategies based on isometric transportation and implicit basis construction over the explicit basis construction sampling strategy across different manifolds.

2. **Riemannian PL condition**: Every thing exactly follows for second part of above argument.

3. **Non convex**: In the nonconvex setting, only a bound on the gradient norm can be obtained instead of a bound on the excess risk. Both DP-RGD and DP-RSGD obtain bound on gradient norm as $\mathcal{O}(\frac{L_0\sqrt{dL\log(1/\delta)}}{n\epsilon})$ in $\mathcal{O}(\frac{\sqrt{L}n^2\epsilon}{L_0\sqrt{d\log(1/\delta)}})$ and $\mathcal{O}(\frac{\sqrt{L}n\epsilon}{L_0\sqrt{d\log(1/\delta)}})$ iterations respectively (Han et al., 2022, Theorem 5). From Theorem 8, DP-RSVRG obtains bound on gradient as $\mathcal{O}(\frac{L_0\sqrt{dL\log(1/\delta)}}{n\epsilon})$ in $\mathcal{O}(\sqrt{\frac{L\zeta}{d\log(1/\delta)}}\frac{n^{5/3}\epsilon}{L_0})$ iterations. Hence in this case, DP-RGD, DP-RSGD, DP-RSVRG have same utility bounds.

Remark 2. Although the theoretically bounds are not better than DP-RGD and DP-RSGD, in practice we observe DP-RSVRG is indeed effective especially when noise added is relatively small.

.

6 Experiments

In this section, we illustrate the efficacy of the proposed sampling procedures and the proposed DP-RSVRG algorithm. We also show the benefit of α optimization (Section 5.1) in terms of the gain in privacy guarantee.

Table 2: Overhead of privatizations for DP-RSGD (with 3×10^5 epochs) for the SPD Fréchet mean and the principal eigenvector problems. Our proposed isometric sampling based sampling strategy lead to orders of magnitude improvements than those of Han et al. (2022).

Manifold	Size	Han et al. (2022)	This work
SPD	11×11	660 hrs	41 seconds ($\sim 10^4$ improvement)
Hypersphere	786	668 seconds	24 seconds (~ 10 improvement)

6.1 Benchmarking of different sampling procedures

We compare the three different sampling procedures discussed in Section 4: sampling based on explicit basis construction, implicit basis construction, and isometric transportation. We pick $s = 1000$ samples under the differentially private optimization setting (i.e., one can only see one base point at a time).

We benchmark the sampling time on various manifolds: SPD, Stiefel $\mathrm{St}(m, r)$, Grassmann $\mathrm{Gr}(m, r)$, and hypersphere $\mathbb{S}^m$. We consider $m = \{5, 10, 20, 30, 50\}$ for SPD(m) and $m \in \{100, 300, 500, 1000, 3000, 5000\}$ for $\mathbb{S}^m$. For $\mathrm{Gr}(m, r)$ and $\mathrm{St}(m, r)$, we consider $m \in \{100, 300, 500, 1000, 3000, 5000\}$ and $r \in \{10, 100\}$.

We present the sampling time plots for different manifolds (of varying dimensions) in Figure 1. We observe that the proposed isometric transportation based sampling approach significantly outperforms explicit basis construction approach in every case, especially at higher dimensions. For the Stiefel manifold, as discussed in Section 4.2, the isometric approach coincides with implicit basis construction approach and hence we show only one of them. We also observe that the performance of the proposed isometric transportation based approach is similar to that of the proposed implicit transportation based approach.

We study the benefits of the proposed sampling procedures in two problems: private estimation of SPD Fréchet mean and the principal eigenvector (discussed in Section 6.3). We use DP-RSGD algorithm for both problems and compare our sampling strategy with those developed in (Han et al., 2022). We observe that the proposed sampling strategy offers significant improvement leading to minimal overhead due to privatization.

6.2 Optimizing α in moments bound for better (ϵ, δ) guarantees

We now show better privacy guarantees can be empirically achieved by optimizing α in moments bound (Section 5.1). We use the `autodp` library (Wang et al., 2019c) and set $\sigma_1 = \sqrt{\alpha}\sigma, \sigma_2 = \sqrt{(1-\alpha)}\sigma$ instead of the standard setting $\sigma_1 = \sigma_2 = \sigma/\sqrt{2}$. We fix $\mathcal{C}_1 = 0.1, \mathcal{C}_2 = 0.01$ and frequency to $m = 10000$ and $n = 100000$. The results are shown in Figure 2 for epochs $S = \{1, 5, 10, 25, 50, 100\}$ and noise $\sigma = \{0.1, 0.05\}$. We observe that the proposed optimization over α significantly improves the privacy guarantees than the standard setting. For noise level $\sigma = 0.05$, we obtain $\epsilon = 0.47$ while the standard setting achieves $\epsilon = 0.64$, a 1.6× improvement in privacy guarantee.

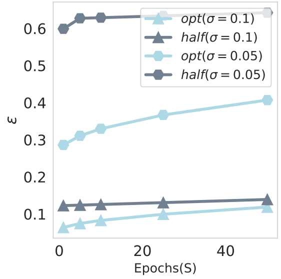
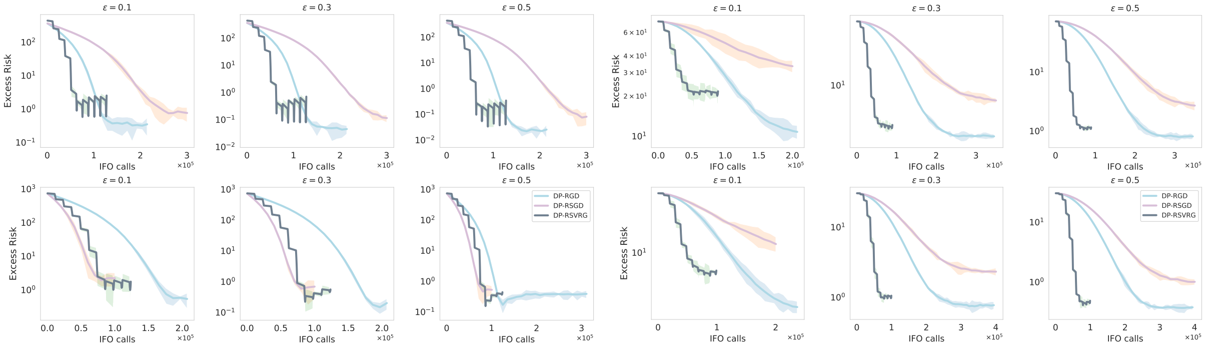


Figure 2: Improving privacy with α.

6.3 Benchmarking DP-RSVRG

In this section, we compare our proposed DP-SVRG with DP-RGD and DP-RSGD (Han et al., 2022) for the task of computing the Fréchet mean and leading eigenvector with privacy configuration $\epsilon = \{0.1, 0.3, 0.5\}$ and $\delta = 10^{-6}$. The parameter details for all the algorithms are in Section D.

Private Fréchet mean on SPD manifold. We consider the problem of privately estimating the Fréchet mean of SPD matrices under the Affine-Invariant metric. We select images from PATHMNIST medical imaging dataset (Yang et al., 2021) and pass them through the covariance descriptor pipeline to generate images, each represented as a SPD matrix of size 11×11. Please refer to Section D.1 for more details on the problem formulation and covariance descriptors. We consider the two sets consisting of 10704 and 10356 images from two different classes. For each set, we compute the optimal Fréchet mean by running



(a) Private Fréchet mean on medical imaging data.

(b) Private principal eigenvector on MNIST dataset.

Figure 3: Comparison between DP-RGD, DP-RSGD and DP-RSVRG. Each row in (a), (b) corresponds to a set consisting of images from a particular class. We see the proposed DP-SVRG achieves a comparable excess risk compared to the baselines with lower number of IFO calls.

the (non-private) RGD for 1000 epochs with learning rate set to 0.5. For both the sets, we plot excess risk against the IFO calls in Figure 3a averaged over five randomized runs.

Private principal eigenvector computation on hypersphere. We also consider the problem of computing the leading eigenvector a symmetric matrix, details in Section D.2. We take images from two classes of MNIST and generate 784 vectors to form two sets of 6903 and 7877 images. For each set, we compute the covariance matrix and compute its leading eigenvector by using eigen-decomposition of matrix $1/n \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T$ to find the optimal solution. We plot the excess risk against the IFO calls in Figure 3b averaged over five randomized runs.

Experiment results. For both the applications, we observe that the proposed DP-RSVRG obtains better or comparable excess risk against DP-GD and DP-SGD with generally fewer IFO calls. This is particularly for larger ϵ, where the level of noise injected is small.

7 Conclusion

In this work, we have improved the framework of differentially private Riemannian optimization via efficient sampling and variance reduction. We have proposed various efficient sampling procedures for tangent Gaussian distribution in order to avoid MCMC. This largely reduces the cost of privatizing Riemannian optimization. In addition, we have shown how variance reduction improves the utility and gradient complexity in practice. We believe this work allows Riemannian optimization to be privatized efficiently for large-scale applications.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016. 1, 3

John M Abowd. The US census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, 2018. 1

P-A Absil, Christopher G Baker, and Kyle A Gallivan. Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, 2007. 1, 7

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009. 1, 3, 4, 7

Alekh Agarwal and Leon Bottou. A lower bound for the optimization of finite sums. In *International conference on machine learning*, pp. 78–86. PMLR, 2015. 10

D Apple. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8), 2017. 1

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29 (1):328–347, 2007. doi: 10.1137/050637996. 6

Shahab Asoodeh, Jiachun Liao, Flavio P Calmon, Oliver Kosut, and Lalitha Sankar. A better bound gives a hundred rounds: Enhanced privacy guarantees via f-divergences. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 920–925. IEEE, 2020. 9

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. *arXiv preprint arXiv:2201.04845*, 2022. 1

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014. 1

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1

Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-Euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pp. 474–499. PMLR, 2021. 1

Rajendra Bhatia. Positive definite matrices. In *Positive Definite Matrices*. Princeton university press, 2009. 1, 6

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. 6

Mukul Bhutani, Pratik Jawanpuria, Hiroyuki Kasai, and Bamdev Mishra. Low-rank geometric mean metric learning. ICML workshop on Geometry in Machine Learning (GiMLi), 2018. 1

Silvere Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. 2

Nicolas Boumal. An introduction to optimization on smooth manifolds. To appear with Cambridge University Press, Apr 2022. URL `http://www.nicolasboumal.net/book`. 3

Nicolas Boumal and Pierre-antoine Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. *Advances in neural information processing systems*, 24, 2011. 1

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, pp. 267–284, USA, 2019. USENIX Association. ISBN 9781939133069. 1

Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, volume 21, 2008. 1

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011. 1

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014. 7

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 6. Springer, 1992. 3

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006a. 3

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b. 1

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. 3

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 1, 7

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014. 1

Hatem Hajri, Ioana Ilea, Salem Said, Lionel Bombrun, and Yannick Berthoumieu. Riemannian Laplace distribution on the space of symmetric positive definite matrices. *Entropy*, 18(3):98, 2016. 3

Andi Han and Junbin Gao. Improved variance reduction methods for Riemannian non-convex optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3, 8, 9

Andi Han, Bamdev Mishra, Pratik Jawanpuria, and Junbin Gao. Differentially private Riemannian optimization. *arXiv preprint arXiv:2205.09494*, 2022. 1, 2, 3, 4, 5, 9, 10, 11, 12, 20, 21

Wen Huang, Kyle A Gallivan, and P-A Absil. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015. 7, 18

Wen Huang, P-A Absil, and Kyle A Gallivan. Intrinsic representation of tangent vectors and vector transports on matrix manifolds. *Numerische Mathematik*, 136(2):523–543, 2017. 4, 7, 19

Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 299–316. IEEE, 2019. 1

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120, 2019. 1

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013. 7

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019. 28

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pp. 25–1. JMLR Workshop and Conference Proceedings, 2012. 1

John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006. 1, 3

Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, et al. Geomstats: a python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020. 18

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017. 3, 9

Joe Near. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*, 2018. 1

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 1

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788. PMLR, 2018. 1

Madhav Nimishakavi, Pratik Jawanpuria, and Bamdev Mishra. A dual framework for low-rank tensor completion. In *NeurIPS*, 2018. 1

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. 6

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. 9

Guodong Qi, Huimin Yu, Zhaohui Lu, and Shuzhao Li. Transductive few-shot classification on the oblique manifold. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8412–8422, 2021. 1

Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79, 2018. 1

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pp. 314–323. PMLR, 2016. 7

Matthew Reimherr, Karthik Bharath, and Carlos Soto. Differential privacy over Riemannian manifolds. *Advances in Neural Information Processing Systems*, 34:12292–12303, 2021. 1, 3

Christian P Robert and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999. 2

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25, 2012. 7

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019. 1

Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019. 2, 8

Armin Schwartzman. Lognormal distributions and geometric averages of symmetric positive definite matrices. *International Statistical Review*, 84(3):456–486, 2016. 3

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013. 1

Ryota Suzuki, Ryusuke Takahama, and Shun Onoda. Hyperbolic disk embeddings for directed acyclic graphs. In *International Conference on Machine Learning*, pp. 6066–6075. PMLR, 2019. 1

Yann Thanwerdas and Xavier Pennec. O(n)-invariant riemannian metrics on spd matrices. *arXiv preprint arXiv:2109.05768*, 2021. 18

Abraham Albert Ungar. A gyrovector space approach to hyperbolic geometry. *Synthesis Lectures on Mathematics and Statistics*, 1(1):1–194, 2008. 1

Saiteja Utpala, Praneeth Vepakomma, and Nina Miolane. Differentially private Fréchet mean on the manifold of symmetric positive definite (SPD) matrices. *arXiv preprint arXiv:2208.04245*, 2022. 3

Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pp. 347–450. Springer, 2017. 3

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2, 9

Di Wang, Changyou Chen, and Jinhui Xu. Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pp. 6526–6535, 2019a. 1

Lingxiao Wang, Bargav Jayaraman, David Evans, and Quanquan Gu. Efficient privacy-preserving stochastic nonconvex optimization. *arXiv preprint arXiv:1910.13659*, 2019b. 9, 22

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019c. 9, 12

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021. 12

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pp. 1617–1638. PMLR, 2016. 9, 22

Hongyi Zhang, Sashank J Reddi, and Suvrit Sra. Riemannian SVRG: Fast stochastic optimization on riemannian manifolds. *Advances in Neural Information Processing Systems*, 29, 2016. 2, 3, 8, 22, 24, 26, 28

Jiaqi Zhang, Kai Zheng, Wenlong Mou, and Liwei Wang. Efficient private ERM for smooth objectives. In *International Joint Conference on Artificial Intelligence*, pp. 3922–3928, 2017. 1

Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on riemannian manifolds. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 138–147. PMLR, 2019. 2, 8

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019. 1

A Function classes on manifolds

Definition 3 (Geodesic Lipschitz)**.** *A function $f : \mathcal{M} \to \mathbb{R}$ is called L_0- geodesically Lipschitz continuous if for any $w_1, w_2 \in \mathcal{M}$ $|f(w_1) - f(w_2)| \leq L_0 \mathrm{dist}(w_1, w_2)$. Under assumption of continuous gradient, function f is L_0 geodesically Lipschitz continuous if and only if $\|\mathrm{grad}\, f(w)\| \leq L_0$ for all $w \in \mathcal{M}$.*

Definition 4 (Geodesic smoothness)**.** *A differentiable function $f : \mathcal{M} \to \mathbb{R}$ geodesically L-smooth it's gradient is L-lipschitz i.e., $\|\mathrm{grad}\, f(w_1) - \mathrm{PT}^{w_2 \to w_1} \mathrm{grad}\, f(w_2)\|_{w_1} \leq L \mathrm{dist}(w_1, w_2)$. Additionally, it can be shown that if f is geodesically L-smooth following holds , $f(w_1) \leq f(w_2) + \langle \mathrm{grad}\, F(w_2), \mathrm{Exp}_{w_2}^{-1}(w_1)\rangle_{w_2} + L/2\|\mathrm{Exp}_{w_2}^{-1}(w_1)\|_{w_2}$ for all $w_1, w_2 \in \mathcal{M}$.*

Definition 5 (Geodesic convexity)**.** *A set $\mathcal{W} \subseteq \mathcal{M}$ is called geodesically convex if for any $w_1, w_2 \in \mathcal{X}$, there is geodesic γ with $\gamma(0) = 1$, $\gamma(1) = y$ and $\gamma(t) \in \mathcal{X}$ for $t \in [0, 1]$*

Definition 6 (Strong Geodesic Convexity)**.** *A function f is called geodesic μ-strongly convex if $w, w' = \mathrm{Exp}_w(\zeta) \in \mathcal{W}$, if it satisfies $f(w') \geq f(w) + \langle \mathrm{grad}\, f(w), \zeta\rangle_w + \frac{\mu}{2}\mathrm{dist}^2(w, w')$.*

Definition 7 (Riemannian Polyak–Łojasiewicz (PL)condition)**.** *A function f is said to satisfy the Riemannian Polyak–Łojasiewicz* (PL)*condition if there exists $\tau > 0$ $f(w) - f(w^*) \leq \tau \|\mathrm{grad}\, f(w)\|_w^2$ for any $w \in \mathcal{M}$.*

B Missing details about sampling

Table 3: Isometric transportation expressions useful for implementing Algorithm 2.

Manifold	Metric	Isometric transportation for Algorithm 2
SPD	Affine-Invariant metirc	$\mathrm{PT}^{\mathbf{X}\to\mathbf{Y}}(\mathbf{U}) = (\mathbf{Y}\mathbf{X}^{-1})^{\frac{1}{2}}\mathbf{U}(\mathbf{X}^{-1}\mathbf{Y})^{\frac{1}{2}}$
	Bures-Wasserstein metric	We use the implementation in `Geomstats` (Miolane et al., 2020) that solves using the method given in (Thanwerdas & Pennec, 2021).
	Log-Euclidean metric	$\mathrm{PT}^{\mathbf{X}\to\mathbf{Y}}(\mathbf{U}) = (\mathrm{DLogm}[\mathbf{Y}])^{-1}(\mathrm{DLogm}[\mathbf{X}](\mathbf{U}))$
Grasssmann	Canonical metric	$\mathrm{PT}^{\mathbf{X}\to\mathbf{Y}}(\mathbf{U}) = [-\mathbf{X}\mathbf{Q}\sin\mathbf{\Sigma}\mathbf{P}^T + \mathbf{P}\cos\mathbf{\Sigma}\mathbf{P}^T + (\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)]\mathbf{U}$ $\mathbf{P}\mathbf{\Sigma}\mathbf{Q} = \mathbf{V}$ is the compact SVD of $\mathbf{V} = \mathrm{Exp}_{\mathbf{X}}^{-1}\mathbf{Y} = \mathbf{W}\arctan\mathbf{\Theta}\mathbf{Z}^T$ where $\mathbf{W}\mathbf{\Theta}\mathbf{Z}^T = [\mathbf{X}^T\mathbf{Y}]^{-1}[\mathbf{X}^T - \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T]$
Stiefel	Canonical metric	Isometric transportation through parallelization given in (Huang et al., 2015). This approach coincides with the basis construction approach.
Hypersphere	Canonical metric	$\mathrm{PT}^{\mathbf{x}\to\mathbf{y}}(\mathbf{u}) = \left(\mathbf{I} + (\cos\Vert\mathbf{v}\Vert_2 - 1)\frac{\mathbf{v}\mathbf{v}^T}{\Vert\mathbf{v}\Vert_2} - \sin\Vert\mathbf{v}\Vert_2\frac{\mathbf{x}\mathbf{v}^T}{\Vert\mathbf{v}\Vert_2}\right)\mathbf{u}$ $\mathbf{v} = \mathrm{Exp}_{\mathbf{x}}^{-1}\mathbf{y} = \arccos\langle\mathbf{x}, \mathbf{y}\rangle_2 \frac{(\mathbf{I} - \mathbf{x}\mathbf{x}^T)(\mathbf{y} - \mathbf{x})}{\Vert(\mathbf{I} - \mathbf{x}\mathbf{x}^T)(\mathbf{y} - \mathbf{x})\Vert_2}$
Hyperbolic	Poincaré ball metric	$\mathrm{PT}^{\mathbf{x}\to\mathbf{y}}(\mathbf{u}) = \frac{1 - \Vert\mathbf{y}_2^2\Vert}{1 - \Vert\mathbf{x}_2^2\Vert}\mathrm{gyr}[\mathbf{y}, -\mathbf{x}](\mathbf{u})$, $\mathrm{gyr}[\mathbf{x}, \mathbf{y}](\mathbf{u}) = (\mathbf{o} \ominus (\mathbf{x} \oplus \mathbf{y})) \oplus (\mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{u}))$, and $\mathbf{x} \oplus \mathbf{y} = \frac{[(1 + 2\langle\mathbf{x}, \mathbf{y}\rangle_2 + \Vert\mathbf{y}\Vert_2^2)\mathbf{x} + (1 - \Vert\mathbf{x}\Vert_2^2)\mathbf{y}]}{[1 + 2\langle\mathbf{x}, \mathbf{y}\rangle_2 + \Vert\mathbf{x}\Vert_2^2\Vert\mathbf{y}\Vert_2^2]}$, $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus -\mathbf{y}$.
	Lorentz hyperboloid metric	$\mathrm{PT}^{\mathbf{x}\to\mathbf{y}}(\mathbf{u}) = \mathbf{u} - \frac{\langle\mathbf{y}, \mathbf{u}\rangle_{\mathcal{L}}}{1 - \langle\mathbf{x}, \mathbf{y}\rangle}(\mathbf{x} + \mathbf{y})$.

B.1 Sampling on hyperbolic spaces

Poincaré ball model. The Poincaré ball model consists of $\mathbb{B}(k) = \{\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 < 1\}$ with metric given by $\langle\mathbf{u}, \mathbf{v}\rangle_{\mathbf{x}}^{\mathrm{PB}} = 4\langle\mathbf{u}, \mathbf{v}\rangle_2/(1 - \|\mathbf{x}\|_2^2)$. tangent space for $\mathbf{x} \in \mathrm{PB}(k)$, $T_{\mathbf{x}}\mathrm{PB}(n) = \mathbb{R}^k$. Since Poincaré ball metric is scaled standard Euclidean inner product, standard basis $\mathscr{B} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is *orthogonal* basis of $\left(T_{\mathbf{x}}\mathrm{PB}(n), \langle\mathbf{u}, \mathbf{v}\rangle_{\mathbf{x}}^{\mathrm{PB}}\right)$. An Orthonormal basis at point $\mathbf{x}$ can simply be obtained by scaling $(1 - \|\mathbf{x}\|_2^2)/4$ and $\hat{\mathscr{B}} = \{\mathbf{e}_1[(1 - \|\mathbf{x}\|_2^2)]/4, \dots, \mathbf{e}_1[(1 - \|\mathbf{x}\|_2^2)]/4\}$.

Hence, for Algorithm 1 one can avoid the Gram-Schmidt orthogonalization process for the Poincaré ball metric . For Algorithm 2, we choose reference point as $\mathbf{o} = (0, \dots, 0)$ because $\langle\mathbf{u}, \mathbf{v}\rangle_{\mathbf{o}}^{\mathrm{PB}} = \langle\mathbf{u}, \mathbf{v}\rangle_2$. The parallel transport from $\mathbf{o}$ has expression $\mathrm{PT}^{\mathbf{o}\to\mathbf{x}}(\mathbf{v}) = (1 - \|\mathbf{x}\|_2^2)\mathbf{v}$. Hence for the Poincaré ball metric, Algorithms 1 and 2 coincide.

Lorentz hyperboloid model. The Lorentizian inner product for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is given by $\langle\mathbf{u}, \mathbf{v}\rangle_{\mathcal{L}} = -u_1v_1 + \sum_{i=2}^k u_iv_i$. The Loretnz hyperboloid model is defined as $\mathbb{H}(k) = \{\mathbf{x} \in \mathbb{R}^k | \langle\mathbf{x}, \mathbf{x}\rangle_{\mathcal{L}} = -1\}$ with the Lorentizian product as Riemannian metric. Its tangent space at point $\mathbf{x} \in \mathrm{LH}(k)$ is given by $T_{\mathbf{x}}\mathrm{LH}(k) = \{\mathbf{u} \in \mathbb{R}^k | \langle\mathbf{x}, \mathbf{u}\rangle_{\mathcal{L}} = 0\}$. Now given a point $\mathbf{x} \in \mathrm{LH}(k)$, we find basis of solution space of $\langle\mathbf{x}, \mathbf{u}\rangle_{\mathcal{L}} = 0$ and use the Gram-Schmidit orthogonalization process to construct orthonormal basis. Now for Algorithm 2 we pick

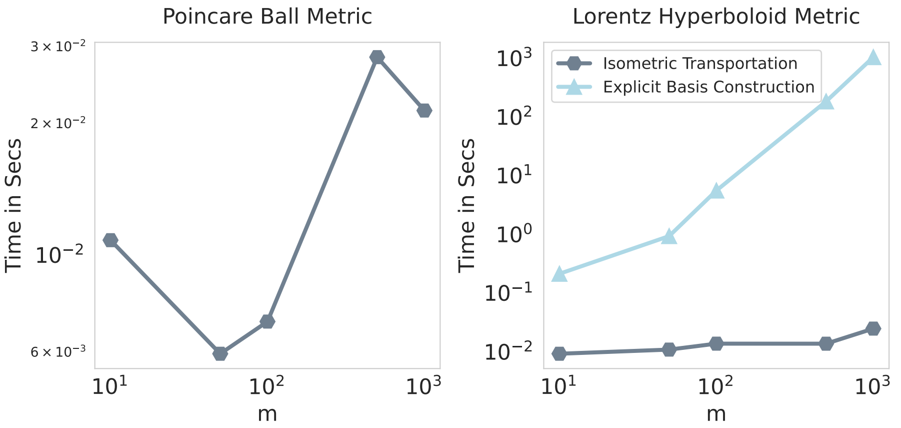


Figure 4: Hyperbolic space. For the Poincaré ball metric, the sampling strategies based on explicit and isometric transportation coincide. For the Lorentz model, we see a clear benefit of the proposed isometric transportation approach over the basis construction approach.

reference point as $\mathbf{o}$ because $T_{\mathbf{o}}\mathrm{LH}(k) = \{0\} \times \mathbb{R}^{k-1}$ and $\mathbf{u} \in T_{\mathbf{o}}\mathrm{LH}(k) \implies \|\mathbf{u}\|_{\mathcal{L}} = \|\mathbf{u}\|_2$. This implies that orthonormal basis at $\mathbf{o}$ is $\mathscr{B} = \{\{0\} \times \mathbf{e}_1, \dots, \{0\} \times \mathbf{e}_{k-1}\}$ where $\mathbf{e}_1, \mathbf{e}_{k-1}$ are standard orthnormal basis vectors of $\mathbb{R}^{k-1}$, and parallel transport from $\mathbf{o}$ is given by $\mathrm{PT}^{\mathbf{o}\to\mathbf{x}}(\mathbf{u}) = \mathbf{u} + \langle \mathbf{u}, \mathbf{x}\rangle_{\mathcal{L}}\mathbf{x}$.

B.2 Implicit basis construction for tangent space of Stiefel, Grassmann, and hypersphere

Stiefel. Let SKEW(r) denotes set of $r \times r$ skew-symmetric matrices and $\mathbf{X}_\perp \in \mathbb{R}^{m\times(m-p)}$ whose columns form an orthonormal basis of the orthogonal complement of column space of $\mathbf{X}$. For any tangent vector $\mathbf{U} \in T_{\mathbf{X}}\mathrm{St}(m,r)$, there exists a unique $\mathbf{A} \in \mathrm{SKEW}(r), \mathbf{B} \in \mathbb{R}^{(m-r)\times r}$ s.t

$$\mathbf{U} = \mathbf{X}\mathbf{A} + \mathbf{X}_\perp\mathbf{B}.$$

Note that explicitly constructing $\mathbf{X}_\perp\mathbf{B}$ would take $\mathcal{O}(m(m-r)^2)$, when $r \ll m$ this would be too expensive. Huang et al. (2017) suggested a procedure that would take $\mathcal{O}(mr^2)$ using Householder transformations.

First given a base point $\mathbf{X} \in \mathrm{St}(m,r)$, vectors corresponding to the Householder matrices $(v_1, \dots, v_r)$ and sign scalars $(s_1, \dots, s_r)$ are constructed as in (Huang et al., 2017, Algorithm 3).

Now using $(v_1, \dots, v_r)$ and $(s_1, \dots, s_r)$ and matrices $\mathbf{A}, \mathbf{B}$, tangent vector $\mathbf{U}$ can be constructed given in (Huang et al., 2017, Algorithm 5) as follows,

$$\mathbf{U} = \mathbf{I}_m - 2\mathbf{v}_1\mathbf{v}_1^T \dots \begin{bmatrix} \mathbf{I}_{r-2} & 0 \\ 0 & \mathbf{I}_{m-r+2} - 2\mathbf{v}_{r-1}\mathbf{v}_{r-1}^T \end{bmatrix} \begin{bmatrix} \mathbf{I}_{r-1} & 0 \\ 0 & \mathbf{I}_{m-r+1} - 2\mathbf{v}_r\mathbf{v}_r^T \end{bmatrix} \mathrm{diag}(s_1, s_2, \dots, s_r, \mathbf{I}_{n-r}) \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}.$$

This procedure can be shown to take $\mathcal{O}(mr^2)$.

Grassmann. It can be seen as special case of Stiefel, as any tangent $\mathbf{U} \in T_{\mathbf{X}}\mathrm{Gr}(m,r)$, there exists a $\mathbf{B} \in \mathbb{R}^{(m-r)\times r}$ s.t

$$\mathbf{U} = \mathbf{X}_\perp\mathbf{B}.$$

Hence, the rest of the procedure is similar.

Hypersphere. It can be seen as a special case of Stiefel with $r = 1$.

C Proofs

C.1 Proofs of Section 3

C.1.1 Proof of Claim 1

Theorem 10 (Change of variable formula)**.** *Let X, Y be measurable space and $\phi : X \to Y$ and $f : Y \to \mathbb{R}$ is measurable mapping and let λ be measure on X and $\phi_* \lambda$ denote the pushforward measure of λ through ϕ on Y then $\int_Y f d(\phi_* \lambda) = \int_X f \circ \phi \, d\lambda$.*

Proof. Let $\vec{\mu} \in \mathbb{R}^d$ denote coordinates of μ and consider normalizing constant,

$$\begin{aligned} C_{w,\sigma} &= \int_{T_p\mathcal{M}} \exp\left(-\frac{\|\nu-\mu\|_w^2}{2\sigma^2}\right) d(\phi_*\lambda)(\nu) \overset{(*)}{=} \int_{\mathbb{R}^d} \exp\left(-\frac{\left\|\sum_{i=1}^d c_i\beta_i - \sum_{i=1}^d \vec{\mu_i}\beta_i\right\|_w^2}{2\sigma^2}\right) d\lambda(c) \\ &= \int_{\mathbb{R}^d} \exp\left(-\frac{\sum_{i=1}^d \sum_{j=1}^d \langle (c_i-\vec{\mu_i})\beta_i, (c_j-\vec{\mu_j})\beta_j\rangle_w}{2\sigma^2}\right) d\lambda(c) \overset{(**)}{=} \int_{\mathbb{R}^d} \exp\left(-\frac{\sum_{i=1}^d (c_i-\vec{\mu_i})^2}{2\sigma^2}\right) d\lambda(c) \\ &\overset{(\dagger)}{=} (2\pi\sigma^2)^{d/2}, \end{aligned} \tag{4}$$

where we used change of variable rule (Theorem 10) under transformation ϕ in $(*)$ and that $(\beta_1, \ldots, \beta_d)$ is orthonormal tangent vectors in $(**)$ and that $\int_{\mathbb{R}^d} \exp(-\frac{\sum_{i=1}^d (c_i-\vec{\mu_i})^2}{2\sigma^2}) d\lambda$ is normalizing constant of $\mathcal{N}(\vec{\mu_i}, \sigma^2.I)$ in $(\dagger)$.

Now, let $\xi \sim \mathcal{N}_w(\mu, \sigma^2)$, we will show that $\vec{\xi} \sim \mathcal{N}(\vec{\mu}, \sigma^2 I_d)$. Let $A \subseteq \mathbb{R}^d$ be measurable set, then consider

$$\begin{aligned} \Pr[\vec{\xi} \in A] = \Pr[\xi \in \phi_{\mathscr{B}}(A)] &= \int_{\phi_{\mathscr{B}}(A)} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\nu-\mu\|_w^2}{2\sigma^2}\right) d(\phi_*\lambda)(\nu) \\ &= \int_A \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\sum_{i=1}^d (c_i-\vec{\mu_i})^2}{2\sigma^2}\right) d\lambda(c). \end{aligned}$$

The last equality is obtained similarly as in (4). Since the last expression is exactly probability that a random vector distributed as $\mathcal{N}(\vec{\mu}, \sigma^2 I_d)$ belongs to set A, we are done. The converse is shown in similar way. ☐

C.2 Proof of Claim 2

Proof. This simply follows from Claim 1 and the variance bound from the standard Gaussian distribution.

We notice that (Han et al., 2022, Remark 1) claims that $\vec{\xi} \sim \mathcal{N}(\vec{\mu}, \sigma^2 G_w^{-1})$, while not considering the fact that under an orthonormal basis, $G_w = I_d$. This implies that the normalizing constant is independent of the base point, which is $(2\pi\sigma^2)^{d/2}$ unlike the case in (Han et al., 2022). ☐

C.2.1 Proof of Claim 3

Proof. Given that $\mathrm{I}^{w_1 \to w_2}$ is linear isometric mapping, one can show that it is invertible and its inverse is again isometry, which we will denote by $\mathrm{I}^{w_2 \to w_1}$. If $\phi_*\lambda$ is Lebesuge measure on $T_{w_1}\mathcal{M}$ then ${\mathrm{I}^{w_1 \to w_2}}_* (\phi_*\lambda)$ is Lebesgue measure on $T_{w_2}\mathcal{M}$. This can be seen by observation that, if $\mathscr{B} = \{\beta_1, \ldots, \beta_d\}$ is orthonormal basis for $T_{w_1}\mathcal{M}$ then $\{\mathrm{I}^{w_1 \to w_2}\beta_1, \ldots, \mathrm{I}^{w_1 \to w_2}\beta_d\}$ is orthonormal basis for $T_{w_2}\mathcal{M}$. Let $\xi \sim \mathcal{N}_{w_1}(\mu, \sigma^2)$, we will show that $\mathrm{I}^{w_1 \to w_2}\xi \sim \mathcal{N}_{w_2}(\mathrm{I}^{w_1 \to w_2}\mu, \sigma^2)$. consider measurable set $S \subseteq T_{w_2}\mathcal{M}$

$$\begin{aligned}\Pr\left[\mathrm{I}^{w_1\to w_2}(\xi_1)\in S\right] = \Pr\left[\xi_1 \in \mathrm{I}^{w_2\to w_1}(S)\right] &= \int_{\mathrm{I}^{w_2\to w_1}(S)} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\nu-\mu\|_{w_1}^2}{2\sigma^2}\right) d(\phi_*\lambda)(\nu) \\ &\overset{(*)}{=} \int_S \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\mathrm{I}^{w_2\to w_1}(\nu)-\mu\|_{w_1}^2}{2\sigma^2}\right) d\left(\mathrm{I}_*^{w_1\to w_2}(\phi_*\lambda)\right)(\nu) \\ &\overset{(**)}{=} \int_S \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\nu-\mathrm{I}^{w_1\to w_2}(\mu)\|_{w_2}^2}{2\sigma^2}\right) d\left(\mathrm{I}_*^{w_1\to w_2}(\phi_*\lambda)\right)(\nu),\end{aligned}$$

where we used change of variables formula Theorem 10 (with $X = \mathrm{I}^{w_2\to w_1}(S), Y = S$ and $\phi = \mathrm{I}^{w_1\to w_2}$) and that I is isometry in $(**)$. Since ${\mathrm{I}^{w_1\to w_2}}_*(\phi_*\lambda)$ is the Lebesuge measure on $T_{w_2}\mathcal{M}$, we have that $\mathrm{I}^{w_1\to w_2}\xi \sim \mathcal{N}_{w_2}(\mathrm{I}^{w_1\to w_2}\mu, \sigma^2)$.

□

C.3 Proofs of Section 5

C.3.1 Proof of Claim 4

Proof. Let $\mathcal{Q}^{s+1}$ denote full gradient query given by $\mathcal{Q}^{s+1}(Z) = \frac{1}{n}\sum_{i=1}^n \operatorname{grad} f(\tilde{w}^s; z_i)$. Let $Z, Z' \in \mathcal{Z}^n$ denote adjacent datasets, consider it's sensitivity denoted at Δ^s,

$$\Delta^{s+1} = \sup_{Z\sim Z'} \|\mathcal{Q}^{s+1}(Z) - \mathcal{Q}^{s+1}(Z')\| \le \frac{1}{n}\left[\|\operatorname{grad} f(\tilde{w}^s; z_n)\|_{\tilde{w}^s} + \|\operatorname{grad} f(\tilde{w}^s; z_n')\|_{\tilde{w}^s}\right] \le \frac{2\mathcal{C}_0}{n}. \tag{5}$$

Following (Han et al., 2022, Lemma 2), the moments bound of the full gradient mechanism $\mathcal{R}^s$ is given by,

$$\mathcal{K}_{\mathcal{R}^s}(\lambda) \le \frac{\lambda(\lambda+1)}{2\sigma_1^2}(\Delta^s)^2 \overset{\text{Eq 5}}{\le} \frac{2\lambda(\lambda+1)\mathcal{C}_0^2}{n^2\sigma_1^2}.$$

Let $\mathcal{Q}_t^{s+1}$ denote variance reduced stochastic gradient query given by $\mathcal{Q}_t^{s+1}(Z) = \operatorname{grad} f(w_t^{s+1}; z) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z) - g^{s+1})$. Let $Z, Z' \in \mathcal{Z}$ denote adjacent datasets, consider it's sensitivity denoted at Δ_{t+1}^s,

$$\begin{aligned}&\Delta_t^{s+1} \\ &= \sup_{Z\sim Z'} \left\|\mathcal{Q}_{t2}^{s+1}(Z) - \mathcal{Q}_{t2}^{s+1}(Z')\right\|_{w_t^{s+1}} \\ &\overset{(*)}{\le} \sup_{Z\sim Z'} \left[\left\|\operatorname{grad} f(w_t^{s+1}; z) - \operatorname{grad} f(w_t^{s+1}; z')\right\|_{w_t^{s+1}} + \left\|\mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}\left(\operatorname{grad} f(\tilde{w}^s; z) - \operatorname{grad} f(\tilde{w}^s; z')\right)\right\|_{w_t^{s+1}}\right] \\ &\overset{(\dagger)}{=} \sup_{Z\sim Z'} \left[\left\|\operatorname{grad} f(w_t^{s+1}; z) - \operatorname{grad} f(w_t^{s+1}; z')\right\|_{w_t^{s+1}} + \|\operatorname{grad} f(\tilde{w}^s; z) - \operatorname{grad} f(\tilde{w}^s; z')\|_{\tilde{w}^s}\right] \\ &\le \sup_{Z\sim Z'} \left[\left\|\operatorname{grad} f(w_t^{s+1}; z)\right\|_{w_t^{s+1}} + \left\|\operatorname{grad} f(w_t^{s+1}; z')\right\|_{w_t^{s+1}} + \|\operatorname{grad} f(\tilde{w}^s; z)\|_{\tilde{w}^s} + \|\operatorname{grad} f(\tilde{w}^s; z')\|_{\tilde{w}^s}\right] \\ &\overset{(\ddagger)}{\le} 4\mathcal{C}_1.\end{aligned} \tag{6}$$

where we used linearity of parallel transport and triangle's inequality in $(*)$ and that parallel transport is isometric in $(\dagger)$ and triangle inequality and assumption of lipschitz in $(\ddagger)$. Now moments bound of $\mathcal{R}_t^{s+1}$ is given by,

$$\mathcal{K}_{\mathcal{R}_t^{s+1}}(\lambda) \le \frac{\lambda(\lambda+1)}{2\sigma_2^2}(\Delta_t^{s+1})^2 \overset{\text{Eq 6}}{\le} \frac{8\lambda(\lambda+1)\mathcal{C}_1^2}{\sigma_2^2}. \tag{7}$$

□

C.3.2 Proof of Claim 5

Proof. By using (Wang et al., 2019b, Lemma 3.7) and by choice of parameters σ^2, λ we have

$$\mathcal{K}_{\text{sub}_{\mathcal{R}_t^{s+1}}(\lambda)} \leq \frac{3.5}{n^2}\mathcal{K}_{\mathcal{R}_t^{s+1}}(\lambda) \overset{\text{Eq 7}}{\leq} \frac{28\lambda(\lambda+1)\mathcal{C}_1^2}{\sigma^2 n^2}.$$

$\square$

C.3.3 Proof of Claim 6

Proof. For $\mathcal{R}$ can be show (ϵ, δ)-differentially private by solving for ϵ and δ as follows,

$$\min_{\alpha\in(0,1)} \frac{mS\lambda(\lambda+1)\mathcal{C}^2}{n^2\sigma^2}\left[\frac{2}{\alpha}+\frac{28}{1-\alpha}\right] = \frac{mS\lambda(\lambda+1)\mathcal{C}^2}{n^2\sigma^2}\left[\frac{2}{\alpha^*}+\frac{28}{1-\alpha^*}\right] \leq \frac{\lambda\epsilon}{2}, \exp\left(-\frac{\lambda\epsilon}{2}\right) \leq \delta, \tag{8}$$

where where $\alpha^* = (\sqrt{14}-1)/13$ and there exists constant $c_1 > 0$ s.t $\sigma^2 \geq c_1 \frac{mS\log(1/\delta)\mathcal{C}^2}{n^2\epsilon^2}$ satisfies (8). Hence, Algorithm 3 satisfies (ϵ, δ)-DP. For Algorithm 4 using similar arguments there exists constant $c_2 > 0$ s.t $\sigma^2 \geq c_2 \frac{mSK\log(1/\delta)\mathcal{C}^2}{n^2\epsilon^2}$ guarantees (ϵ, δ)-DP . $\square$

C.4 Proofs of Section 5.2

Lemma 11 (Trigonometric distance bound (Zhang & Sra, 2016))**.** *Let $w_0, w_1, w_2 \in \mathcal{W} \subseteq \mathcal{M}$ lie in totally normal neighborhood of Riemannian manifold with curvature lower bounded by κ_{min} and $\ell_0 = \operatorname{dist}(w_0, w_1)$ and $\ell_1 = \operatorname{dist}(w_1, w_2)$ and $\ell_2 = \operatorname{dist}(w_0, w_2)$. Denote θ as the angle on $T_{w_0}\mathcal{M}$ s.t $\cos(\theta) = \frac{1}{\ell_0\ell_1}\langle \operatorname{Exp}_{w_0}^{-1}(w_1), \operatorname{Exp}_{w_0}^{-1}(w_2)\rangle_{w_0}$. Let $D_{\mathcal{W}}$ be the diameter of $\mathcal{W}$ i.e., $D_{\mathcal{W}} := \max_{w,w'} \operatorname{dist}(w, w')$. Define curvature constant $\zeta = \frac{\sqrt{\kappa_{\min}}}{\tanh\sqrt{\kappa_{\min}}}$ if $\kappa_{\min} < 0$ and $\zeta = 1$ if $\kappa_{\min} \geq 0$. Then, we have that $\ell_1^2 \leq \zeta\ell_0^2 + \ell_2^2 - 2\ell_0\ell_2\cos\theta$.*

Lemma 12.

$$\mathbb{E}_{i_t,\epsilon_t}\|v_t^{s+1}\|^2_{w_t^{s+1}} \leq \mathbb{E}_{i_t}\|\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1})\|^2_{w_t^{s+1}} + d\sigma^2. \tag{9}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{i_t,\epsilon_t}\|v_t^{s+1}\|^2_{w_t^{s+1}} &= \mathbb{E}_{i_t,\epsilon_t}\|\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1}) + \epsilon_t\|^2_{w_t^{s+1}} \\
&= \mathbb{E}_{i_t,\epsilon_t}\|\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1})\|^2_{w_t^{s+1}} + \mathbb{E}_{\epsilon_t}\|\epsilon_t\|^2_{w_t^{s+1}} \\
&\qquad + \langle \mathbb{E}_{i_t}\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1}), \mathbb{E}_{\epsilon_t}[\epsilon_t]\rangle_{w_t^{s+1}} \\
&\leq \mathbb{E}_{i_t}\|\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1})\|^2_{w_t^{s+1}} + d\sigma^2,
\end{aligned}$$

where we used that $\mathbb{E}_{\epsilon_t}[\epsilon_t] = 0$ and $\mathbb{E}_{\epsilon_t}\|\epsilon_t\|^2_{w_t^{s+1}} \leq d\sigma^2$ in last inequality. $\square$

C.4.1 Proof of Theorem 7

Proof. We bound first term $\mathbb{E}_{i_t}\|\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1})\|^2_{w_t^{s+1}}$ as in (Zhang et al., 2016)

$$
\begin{aligned}
&\mathbb{E}_{i_t} \left\| \operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}} (\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1}) \right\|_{w_t^{s+1}}^2 \\
&\leq \mathbb{E}_{i_t} \left\| \operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}} \operatorname{grad} f(\tilde{w}^s; z_{i_t}) + \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}} \left(\operatorname{grad} F(\tilde{w}^s) - \mathrm{PT}^{\tilde{w}^* \to \tilde{w}^s} \operatorname{grad} F(w^*) \right) \right\|_{w_t^{s+1}}^2 \\
&\leq 2\mathbb{E}_{i_t} \left\| \operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}} \operatorname{grad} f(\tilde{w}^s; z_{i_t}) \right\|_{w_t^{s+1}}^2 \\
&\qquad + 2\mathbb{E}_{i_t} \left\| \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}} \left(\operatorname{grad} F(\tilde{w}^s) - \mathrm{PT}^{\tilde{w}^* \to \tilde{w}^s} \operatorname{grad} F(w^*) \right) \right\|_{w_t^{s+1}}^2 \\
&= 2\mathbb{E}_{i_t} \left\| \operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s \to w_t^{s+1}} \operatorname{grad} f(\tilde{w}^s; z_{i_t}) \right\|_{w_t^{s+1}}^2 + 2\mathbb{E}_{i_t} \left\| \operatorname{grad} F(\tilde{w}^s) - \mathrm{PT}^{\tilde{w}^* \to \tilde{w}^s} \operatorname{grad} F(w^*) \right\|_{\tilde{w}^s}^2 \\
&\leq 4L^2 \| \mathrm{Exp}_{w_t^{s+1}}^{-1}(w^*) \|_{w_t^{s+1}}^2 + 6L^2 \left\| \mathrm{Exp}_{\tilde{w}^s}^{-1} w^* \right\|_{\tilde{w}^s}^2 \\
&= 4L^2 \mathrm{dist}^2(w_t^{s+1}, w^*) + 6L^2 \mathrm{dist}^2(\tilde{w}^s, w^*). \qquad (10)
\end{aligned}
$$

Using the trignometric distance bound Lemma 11 with $w_0 = x_t^{s+1}, w_1 = w_{t+1}^{s+1}, w_2 = w^*$,

$$
\begin{aligned}
\mathrm{dist}^2(w_{t+1}^{s+1}, w^*) &\leq \zeta \mathrm{dist}^2(w_{t+1}^{s+1}, w_t^{s+1}) + \mathrm{dist}^2(w_t^{s+1}, w^*) - 2\langle \mathrm{Exp}_{x_t^{s+1}}^{-1}(w_{t+1}^{s+1}), \mathrm{Exp}_{w_t^{s+1}}^{-1}(w^*) \rangle_{w_t^{s+1}} \\
&= \zeta \left\| \mathrm{Exp}_{w_t^{s+1}}^{-1} w_{t+1}^{s+1} \right\|_{w_t^{s+1}}^2 + \mathrm{dist}^2(w_t^{s+1}, w^*) - 2\langle -\eta v_t^{s+1}, \mathrm{Exp}_{w_t^{s+1}}^{-1}(w^*) \rangle_{w_t^{s+1}} \\
&= \zeta \eta^2 \left\| v_t^{s+1} \right\|_{w_t^{s+1}}^2 + \mathrm{dist}^2(w_t^{s+1}, w^*) + 2\eta \langle v_t^{s+1}, \mathrm{Exp}_{w_t^{s+1}}^{-1}(w^*) \rangle_{w_t^{s+1}}.
\end{aligned}
$$

Applying expectation we have

$$
\begin{aligned}
&\mathrm{dist}^2(w_{t+1}^{s+1}, w^*) \\
&\leq \zeta \eta^2 \mathbb{E}_{i_t, \epsilon_t} \left\| v_t^{s+1} \right\|_{w_t^{s+1}}^2 + \mathrm{dist}^2(w_t^{s+1}, w^*) + 2\eta \langle \mathbb{E}_{i_t, \epsilon_t} v_t^{s+1}, \mathrm{Exp}_{w_t^{s+1}}^{-1}(w^*) \rangle_{w_t^{s+1}} \\
&= \zeta \eta^2 L^2 \left[4\mathrm{dist}^2(w_t^{s+1}, w^*) + 6\mathrm{dist}^2(\tilde{w}^s, w^*) \right] + 2\eta \langle \operatorname{grad} F(w_t^{s+1}), \mathrm{Exp}_{w_t^{s+1}}^{-1}(w^*) \rangle_{w_t^{s+1}} + d\zeta \eta^2 \sigma^2 \\
&\leq \zeta \eta^2 L^2 \left[4\mathrm{dist}^2(w_t^{s+1}, w^*) + 6\mathrm{dist}^2(\tilde{w}^s, w^*) \right] + 2\eta [F(w^*) - F(w_t^{s+1}) - \frac{\mu}{2} \mathrm{dist}^2(w_t^{s+1}, w^*)] + d\zeta \eta^2 \sigma^2 \\
&\leq (1 + 4\zeta \eta^2 L^2 - \eta \mu) \mathrm{dist}^2(w_t^{s+1}, w^*) + 6\zeta \eta^2 L^2 \mathrm{dist}^2(\tilde{w}^s, w^*) + d\zeta \eta^2 \sigma^2.
\end{aligned}
$$

Defining $u_t = \mathrm{dist}^2(w_{t+1}^{s+1}, w^*)$, $q = (1 + 4\zeta\eta^2 L^2 - \eta\mu)$, $p = 6\zeta\eta^2 L^2, c = d\zeta\eta^2\sigma^2$ we have following recurrence $u_{t+1} - pu_0 \leq q(u_t - pu_0) + c$ from which we have that $u_m \leq (p + q^m(1-p))u_0 + \sum_{i=1}^{m-1} q^i c$. Now choosing $\eta = \frac{\mu}{17\zeta L^2}$ and $m \geq \frac{10\zeta L^2}{\mu^2}$. we get $q = 1 - \frac{\mu^2}{10\zeta L^2}$ and $p = 1/5$. Note that $0 < \frac{\mu^2}{10\zeta L^2} < 1$ ($L > \mu$, $\zeta \geq 1$) and hence $0 < q < 1$ and from which we have that $(p + q^m(1-p)) = 1/2$.

$$
\begin{aligned}
\mathbb{E}[d^2(w_m^{s+1}, w^*)] &\leq \mathbb{E}[\mathrm{dist}^2(w_m^s, w^*)] + d\zeta \frac{\mu^2 \sigma^2}{289\zeta^2 L^4} \sum_{i=1}^{m-1} \left(1 - \frac{\mu^2}{10\zeta L^2} \right)^i \\
&\leq \mathbb{E}[\mathrm{dist}^2(w_m^s, w^*)] + d\zeta \frac{\mu^2 \sigma^2}{289\zeta^2 L^4} \sum_{i=1}^{\infty} \left(1 - \frac{\mu^2}{10\zeta L^2} \right)^i \\
&= \mathbb{E}[\mathrm{dist}^2(w_m^s, w^*)] + d\zeta \frac{\mu^2 \sigma^2}{289\zeta^2 L^4} \frac{10\zeta L^2}{\mu^2} = \mathbb{E}[\mathrm{dist}^2(w_m^s, w^*)] + d\frac{10\sigma^2}{289L^2},
\end{aligned}
$$

from which we have

$$
\begin{aligned}
\mathbb{E}[\mathrm{dist}^2(w_m^S, w^*)] &= 2^{-S}\mathbb{E}[\mathrm{dist}^2(w_m^0, w^*)] + d\frac{10\sigma^2}{289L^2} \sum_{i=0}^{S} \frac{1}{2^i} \leq 2^{-S}\mathbb{E}[\mathrm{dist}^2(w_m^0, w^*)] + 2dc^{-1} \frac{10}{289L^2} \frac{mS\log(1/\delta)L_0^2}{n^2\epsilon^2} \\
&\leq 2^{-S}\mathbb{E}[\mathrm{dist}^2(w_m^0, w^*)] + d\frac{200\zeta}{289\mu^2} \frac{S\log(1/\delta)L_0^2}{n^2\epsilon^2}.
\end{aligned}
$$

$$\mathbb{E}\left[f(x^a)-f(w^*)\right] \leq \frac{1}{2}\mathbb{E}\left[Ld^2(x_a,w^*)\right] \leq 2^{-S}L\mathbb{E}[d^2(w_m^0,w^*)] + Ld\frac{\zeta}{\mu^2}\frac{S\log(1/\delta)L_0^2}{n^2\epsilon^2}.$$

Now, setting $2^{-S} = d\frac{\zeta}{\mu^2}\frac{\log(1/\delta)L_0^2}{n^2\epsilon^2\mathbb{E}[\mathrm{dist}^2(w_m^0,w^*)]} \implies 2^S = \frac{n^2\epsilon^2 289\mu^2\mathbb{E}[\mathrm{dist}^2(w_m^0,w^*)]}{d100\zeta\log(1/\delta)L_0^2} \implies S = \mathcal{O}\left(\log\left(\frac{n\epsilon\mu\mathbb{E}[\mathrm{dist}^2(w_m^0,w^*)]}{\log(1/\delta)\zeta L_0 d}\right)\right)$, substituting this we have that, and now for $S = \mathcal{O}\left(\log\left(\frac{n\epsilon\mu\mathbb{E}[\mathrm{dist}^2(w_m^0,w^*)}{\log(1/\delta)\zeta L_0^2 d}\right)\right)$

$$\mathbb{E}\left[f(x^a)-f(w^*)\right] \leq \mathcal{O}\left(\frac{d\zeta LL_0^2\log(1/\delta)\mathbb{E}[\mathrm{dist}^2(w_m^0,w^*)]}{\mu^2n^2\epsilon^2}\log\left(\frac{n\epsilon\mu}{\zeta L_0^2 d\log(1/\delta)}\right)\right). \tag{11}$$

Gradient complexity: $S\times n$ plus $m\times 2$ IFO calls $= 2nS + 2mS$,

$$\mathcal{O}\left(\left(n+\frac{\zeta L^2}{\mu^2}\right)\log\left(\frac{n\epsilon\mu\mathbb{E}[\mathrm{dist}^2(w_m^0,w^*)]}{\log(1/\delta)\zeta L_0 d}\right)\right). \tag{12}$$

This completes the proof. ☐

C.4.2 Proof of Theorem 8

Before proving Theorem 8, we state and prove following lemma that we will be using later.

Lemma 13. *Assume that each f_i is L-g-smooth, the sectional curvature in $\mathcal{X}$ is lower bounded by κ_{min} and we run Algorithm with Option II. For $c_t, c_{t+1}, \beta, \eta > 0$ and suppose we have $c_t = c_{t+1}(1+\beta\eta+2\zeta L^2\eta^2)+L^3\eta^2$ and $\delta(t) = \eta - \frac{c_{t+1}\eta}{\beta} - L\eta^2 - 2c_{t+1}\zeta\eta^2 > 0$ then the iterate w_t^{s+1} satisfies the bound*

$$\mathbb{E}\|\operatorname{grad} f(w_t^{s+1})\|^2 \leq \frac{R_t^{s+1}-R_{t+1}^{s+1}}{\delta_t} + \frac{\left(\frac{1}{2}dL\eta^2 + c_{t+1}\zeta d\eta^2\right)}{\delta_t}\sigma^2,$$

where $R_t^{s+1} := \mathbb{E}[F(w_t^{s+1}) + c_t\left\|\mathrm{Exp}_{\tilde{w}^s}w_t^{s+1}\right\|]$ for $0 \leq s \leq S-1$.

Proof. The proof is adapted from (Zhang et al., 2016, Lemma 2). Denoting $\Delta_t^{s+1} = \operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}\operatorname{grad} f(\tilde{w}^s; z_{i_t})$ it can be seen that $\mathbb{E}_{i_t|\tilde{x}^s, w_t^{s+1}}[\Delta_t^{s+1}] = \operatorname{grad} F(w_t^{s+1}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}\operatorname{grad} F(\tilde{w}^s)$

$$\begin{aligned}
\mathbb{E}_{i_t,\epsilon_t}\left\|v_t^{s+1}\right\|_{w_t^{s+1}}^2 &\overset{9}{\leq} \mathbb{E}_{i_t}\left\|\operatorname{grad} f(w_t^{s+1}; z_{i_t}) - \mathrm{PT}^{\tilde{w}^s\to w_t^{s+1}}(\operatorname{grad} f(\tilde{w}^s; z_{i_t}) - g^{s+1})\right\|_{w_t^{s+1}}^2 + d\sigma^2 \\
&= \mathbb{E}_{i_t}\left\|\Delta_t^{s+1} - \mathbb{E}_{i_t}\Delta_t^{s+1} + \operatorname{grad} F(w_t^{s+1})\right\|_{w_t^{s+1}}^2 + d\sigma^2 \\
&\overset{(*)}{\leq} 2\mathbb{E}_{i_t}\left\|\Delta_t^{s+1} - \mathbb{E}_{i_t}\Delta_t^{s+1}\right\|^2 + 2\left\|\operatorname{grad} F(w_t^{s+1})\right\|_{w_t^{s+1}}^2 + d\sigma^2 \\
&\overset{(**)}{\leq} 2\mathbb{E}_{i_t}\left\|\Delta_t^{s+1}\right\|_{w_t^{s+1}}^2 + 2\left\|\operatorname{grad} F(w_t^{s+1})\right\|_{w_t^{s+1}}^2 + d\sigma^2 \\
&\overset{(\dagger)}{\leq} 2L^2\left\|\mathrm{Exp}_{\tilde{w}^s}^{-1}(w_t^{s+1})\right\|_{\tilde{w}^s}^2 + 2\left\|\operatorname{grad} F(w_t^{s+1})\right\|_{w_t^{s+1}}^2 + d\sigma^2,
\end{aligned}$$

where $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ in $(*)$ and $\mathbb{E}_{i_t}\left\|\Delta_t^{s+1} - \mathbb{E}_{i_t}\Delta_t^{s+1}\right\|^2 = \mathbb{E}_{i_t}\left\|\Delta_t^{s+1}\right\|^2 - \left\|\mathbb{E}\Delta_t^{s+1}\right\|^2 \leq \mathbb{E}_{i_t}\left\|\Delta_t^{s+1}\right\|^2$ in $(**)$ and assumption that f_i is L-g-smooth in $(\dagger)$. Taking full expectation we have

$$\mathbb{E}\left\|v_t^{s+1}\right\|_{w_t^{s+1}}^2 \leq 2L^2\left\|\mathrm{Exp}_{\tilde{w}^s}^{-1}(w_t^{s+1})\right\|_{\tilde{w}^s}^2 + 2\left\|\operatorname{grad} F(w_t^{s+1})\right\|_{w_t^{s+1}}^2 + d\sigma^2. \tag{13}$$

For bounding the Lyapunov function $R^{s+1}_{t+1} := \mathbb{E}\left[F(w^{s+1}_{t+1}) + c_{t+1}\left\|\mathrm{Exp}_{\tilde{w}^s}(w^{s+1}_{t+1})\right\|^2\right]$ we need to bound on $\mathbb{E}[F(w^{s+1}_{t+1})]$, $\mathbb{E}[\left\|\mathrm{Exp}_{\tilde{w}^s}(w^{s+1}_{t+1})\right\|^2]$, First consider

$$
\begin{aligned}
&\mathbb{E}\left[F(w^{s+1}_{t+1})\right] \\
&\overset{(*)}{\leq} \mathbb{E}\left[F(w^{s+1}_t) + \left\langle \mathrm{grad}\, F(w^{s+1}_t), \mathrm{Exp}^{-1}_{w^{s+1}_t}(w^{s+1}_{t+1})\right\rangle_{w^{s+1}_t} + \frac{L}{2}\left\|\mathrm{Exp}^{-1}_{w^{s+1}_t}(w^{s+1}_{t+1})\right\|^2_{w^{s+1}_t}\right] \\
&\overset{(**)}{\leq} \mathbb{E}\left[F(w^{s+1}_t) - \eta\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2_{w^{s+1}_t} + \frac{L\eta^2}{2}\left\|v^{s+1}_t\right\|^2_{w^{s+1}_t}\right] \\
&\overset{13}{\leq} \mathbb{E}\left[F(w^{s+1}_t) - \eta\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2_{w^{s+1}_t} + \frac{L\eta^2}{2}\left(2L^2\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\|^2 + 2\|\,\mathrm{grad}\, F(w^{s+1}_t)\|^2 + \sigma^2 d\right)\right] \\
&= (L\eta^2 - \eta)\|\,\mathrm{grad}\, F(w^{s+1}_t)\|^2 + F(w^{s+1}_t) + L^3\eta^2\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\|^2 + \frac{1}{2}dL\eta^2\sigma^2,
\end{aligned}
\tag{14}
$$

where we used assumption f_i is L-g-smooth implies that F is L-g-smooth in $(*)$ and $\mathrm{Exp}^{-1}_{w^{s+1}_t} = v^{s+1}_t$ and $\mathbb{E}\left[v^{s+1}_t\right] = \mathrm{grad}\, F(w^{s+1}_t)$ in $(**)$. Using trignometric distance bound on $w^{s+1}_t, w^{s+1}_{t+1}, \tilde{w}^s$ we have,

$$
\begin{aligned}
\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_{t+1})\right\|^2_{\tilde{w}^s} &\leq \left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2_{\tilde{w}^s} + \zeta\left\|\mathrm{Exp}^{-1}_{w^{s+1}_t}(w^{s+1}_{t+1})\right\|^2_{w^{s+1}_t} - \left\langle \mathrm{Exp}^{-1}_{w^{s+1}_t}(w^{s+1}_{t+1}), \mathrm{Exp}^{-1}_{w^{s+1}_t}(\tilde{w}^s)\right\rangle_{w^{s+1}_t} \\
&= \left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + \zeta\eta^2\left\|v^{s+1}_t\right\|^2 + 2\eta\langle \mathrm{grad}\, F(w^{s+1}_t), \mathrm{Exp}^{-1}_{w^{s+1}_t}(\tilde{w}^s)\rangle.
\end{aligned}
$$

Taking expectation we have

$$
\begin{aligned}
&\mathbb{E}\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_{t+1})\right\|^2_{\tilde{w}^s} \\
&\leq \mathbb{E}\left[\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + \zeta\eta^2\|v^{s+1}_t\|^2 + 2\eta\langle \mathrm{grad}\, F(w^{s+1}_t), \mathrm{Exp}^{-1}_{w^{s+1}_t}(\tilde{w}^s)\rangle\right] \\
&\leq \mathbb{E}\left[\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + \zeta\eta^2\left\|v^{s+1}_t\right\|^2 + 2\eta\left[\frac{1}{2\beta}\left\|\mathrm{grad}\, f(w^{s+1}_t)\right\|^2 + \frac{\beta}{2}\left\|\mathrm{Exp}^{-1}_{w^{s+1}_t}(\tilde{w}^s)\right\|^2\right]\right] \\
&\leq \mathbb{E}\left[(1+\beta\eta)\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + \zeta\eta^2\left[2L^2\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + 2\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2 + \sigma^2 d\right]\right] \\
&\qquad + \mathbb{E}\left[\frac{\eta}{\beta}\left\|\mathrm{grad}\, f(w^{s+1}_t)\right\|^2\right] \\
&= \left(1 + 2\zeta\eta^2L^2 + \eta\beta\right)\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + \left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2 + \zeta d\eta^2\sigma^2.
\end{aligned}
\tag{15}
$$

Putting (14) and (15) into R^{s+1}_{t+1} we have

$$
\begin{aligned}
R^{s+1}_{t+1} &:= \mathbb{E}[f(w^{s+1}_{t+1}) + c_{t+1}\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_{t+1})\right\|^2] \\
&= c_{t+1}\left(1 + 2\zeta\eta^2L^2 + \eta\beta\right)\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + c_{t+1}\left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2 + c_{t+1}\zeta d\eta^2\sigma^2 \\
&\qquad + (L\eta^2 - \eta)\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2 + F(w^{s+1}_t) + L^3\eta^2\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 + \frac{1}{2}dL\eta^2\sigma^2 \\
&= F(w^{s+1}_t) + \left(c_{t+1}\left(1 + 2\zeta\eta^2L^2 + \eta\beta\right) + L^3\eta^2\right)\left\|\mathrm{Exp}^{-1}_{\tilde{w}^s}(w^{s+1}_t)\right\|^2 \\
&\qquad + \left(L\eta^2 - \eta + c_{t+1}\left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\right)\left\|\mathrm{grad}\, F(w^{s+1}_t)\right\|^2 + \left(\frac{1}{2}dL\eta^2 + c_{t+1}\zeta d\eta^2\right)\sigma^2 \\
&= R^{s+1}_t - \left(-L\eta^2 + \eta - c_{t+1}\left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\right)\|\,\mathrm{grad}\, F(w^{s+1}_t)\|^2 + \left(\frac{1}{2}dL\eta^2 + c_{t+1}\zeta d\eta^2\right)\sigma^2.
\end{aligned}
$$

Then rearranging that,

$$
\left(\eta - L\eta^2 - c_{t+1}\left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\right)\mathbb{E}\|\,\mathrm{grad}\, F(w^{s+1}_t)\|^2 \leq R^{s+1}_t - R^{s+1}_{t+1} + \left(\frac{1}{2}dL\eta^2 + c_{t+1}\zeta d\eta^2\right)\sigma^2
$$

from which we have

$$\mathbb{E}\|\operatorname{grad} F(w_t^{s+1})\|^2 \leq \frac{R_t^{s+1} - R_{t+1}^{s+1}}{\left(\eta - L\eta^2 - c_{t+1}\left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\right)} + \frac{\left(\frac{1}{2}L + c_{t+1}\zeta\right) d\eta^2}{\left(L\eta^2 - \eta - c_{t+1}\left(2\zeta\eta^2 + \frac{\eta}{\beta}\right)\right)}\sigma^2.$$

□

Now we give proof of Theorem 8.

Proof. Proof is adapted from (Zhang et al., 2016, Theorem 2, 6 and Corollary 6) Let $\delta_n = \min_t \delta_t$ and $T = mS$

$$\begin{aligned}
&\sum_{t=0}^{m-1} \mathbb{E}\left\|\operatorname{grad} f(w_t^{s+1})\right\|^2 \\
&\leq \sum_{t=0}^{m-1} \frac{R_t^{s+1} - R_{t+1}^{s+1}}{\delta_t} + \frac{\left(\frac{1}{2}L + c_{t+1}\zeta\right) d\eta^2}{\delta_t}\sigma^2 \\
&\overset{(*)}{\leq} \frac{R_0^{s+1} - R_m^{s+1}}{\delta_n} + \frac{\left(\frac{1}{2}L + c_{t+1}\zeta\right) md\eta^2}{\delta_n}\sigma^2 \\
&= \frac{\mathbb{E}\left[F(w_0^{s+1}) - F(w_m^{s+1}) + c_0\left\|\operatorname{Exp}_{w_{\tilde{s}}}(w_0^{s+1})\right\|^2 - c_m\left\|\operatorname{Exp}_{w_{\tilde{s}}}(w_m^{s+1})\right\|^2\right]}{\delta_n} + \frac{\left(\frac{1}{2}L + c_0\zeta\right) md\eta^2}{\delta_n}\sigma^2 \\
&\overset{(**)}{\leq} \frac{\mathbb{E}\left[F(\tilde{w}^s) - F(\tilde{w}^{s+1})\right]}{\delta_n} + \frac{\left(\frac{1}{2}L + c_0\zeta\right) md\eta^2}{\delta_n}\sigma^2,
\end{aligned}$$

where $\delta_t \geq \delta_n$, $c_t \leq c_0$ is used in $(*)$ and that $w_0^{s+1} = \tilde{w}^s, w_m^{s+1} = \tilde{w}^{s+1}$ and that $c_m = 0, c_0 \geq 0$ in $(**)$.

Now, summing the gradient norm square over all the epochs and using $F(w^*) \leq F(\tilde{w}^m)$, we get

$$\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1} \mathbb{E}\left\|\operatorname{grad} f(w_t^{s+1})\right\|^2 \leq \frac{\mathbb{E}\left[F(\tilde{w}^0) - F(w^*)\right]}{T\delta_n} + \frac{\left(\frac{1}{2}L + c_0\zeta\right) d\eta^2}{\delta_n}\sigma^2.$$

Choosing $\beta = L\zeta^{1-\alpha_2}/n^{\alpha_1/2}$ and solving recurrence relation c_t using η, m given by theorem as (Zhang et al., 2016, Theorem 2) one can get $c_0 = \frac{\mu_0 L}{n^{\alpha_1/2}\zeta}(e-1)$. Substituting that in $\delta_n \geq \frac{\nu}{Ln^{\alpha_1}\zeta^{\alpha_2}}$ and finally using this we have

$$\begin{aligned}
&\frac{1}{T}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1} \mathbb{E}\left\|\operatorname{grad} f(w_t^{s+1})\right\|^2 \\
&\leq c\frac{\mu_0 Ln^{\alpha_1}\zeta^{\alpha_2}}{\nu nS}\mathbb{E}\left[F(\tilde{w}^0) - F(w^*)\right] + \frac{Ln^{\alpha_1}\zeta^{\alpha_2}\left(\frac{1}{2}L + \frac{\mu_0 L}{n^{\alpha_1/2}\zeta}(e-1)\zeta\right)\frac{\mu_0^2}{L^2 n^{2\alpha_1}\zeta^{2\alpha_2}}}{\nu}d\sigma^2.
\end{aligned}$$

Finally, putting the values of $\alpha_1 = 2/3, \alpha_2 = 1/2\mu_0 = 1/10, \nu = 1/2$ and $\sigma^2 = c_2\frac{mS\log(1/\delta)L_0^2}{n^2\epsilon^2} = c_3\frac{S\log(1/\delta)L_0^2}{n\epsilon^2}$ one can get that

$$\begin{aligned}
\mathbb{E}\left\|\operatorname{grad} f(w^a)\right\|^2 &\leq c_4\left(\frac{L\zeta^{1/2}}{n^{1/3}S}\mathbb{E}\left[F(\tilde{w}^0) - F(w^*)\right] + \left[\frac{1}{n^{2/3}\zeta^{1/2}} + \frac{1}{n\zeta^{1/2}}\right]\frac{dS\log(1/\delta)L_0^2}{n\epsilon^2}\right) \\
&\leq c_4\left(\frac{L\zeta^{1/2}}{n^{1/3}S}\mathbb{E}\left[F(\tilde{w}^0) - F(w^*)\right] + \frac{dS\log(1/\delta)L_0^2}{n^{5/3}\zeta^{1/2}\epsilon^2}\right)
\end{aligned}$$

Setting $S = \sqrt{\frac{L\zeta\mathbb{E}[F(\tilde{w}^0)-F(w^*)]}{d\log(1/\delta)}\frac{n^{2/3}\epsilon}{L_0}}$ we have that,

$$\mathbb{E}\left\|\operatorname{grad} f(w^a)\right\|^2 \leq c_4 \frac{L_0\sqrt{dL\log(1/\delta)\mathbb{E}\left[F(\tilde{w}^0)-F(w^*)\right]}}{n\epsilon}. \tag{16}$$

Gradient Complexity is given by,

$$S(n+2m) = \sqrt{\frac{L\zeta\mathbb{E}\left[F(\tilde{w}^0)-F(w^*)\right]}{d\log(1/\delta)}\frac{n^{2/3}\epsilon}{L_0}}\left(n+\frac{n}{30}\right) = \sqrt{\frac{L\zeta\mathbb{E}\left[F(\tilde{w}^0)-F(w^*)\right]}{d\log(1/\delta)}\frac{n^{5/3}\epsilon}{L_0}}. \tag{17}$$

This completes the proof.

□

C.4.3 Proof of Theorem 9

Proof. With the values given in the theorem statement, $\sigma^2 = \frac{mSK\log(1/\delta)L_0^2}{n^2\epsilon^2} = \frac{Kn\lceil 6+\frac{18}{n-3}\rceil L\tau\zeta^{1/2}\frac{\mu_0}{\nu n^{1/3}}\log(1/\delta)L_0^2}{3\mu_0 n^2\epsilon^2} = \frac{K\lceil 6+\frac{18}{n-3}\rceil L\tau\zeta^{1/2}\frac{\log(1/\delta)L_0^2}{\nu n^{1/3}}}{3n\epsilon^2}$. This implies

$$\begin{aligned}
\mathbb{E}[\left\|\operatorname{grad} f(w^{k+1})\right\|^2] &\leq \frac{1}{2\tau}\mathbb{E}\left[F(\tilde{w}^0)-F(w^*)\right] + \left[\frac{1}{n^{2/3}\zeta^{1/2}}+\frac{1}{n\zeta^{1/2}}\right]\frac{dK\lceil 6+\frac{18}{n-3}\rceil L\tau\zeta^{1/2}\frac{\log(1/\delta)L_0^2}{\nu n^{1/3}}}{3n\epsilon^2} \\
&\leq \frac{1}{2\tau}\mathbb{E}\left[F(\tilde{w}^0)-F(w^*)\right] + \frac{24dKL\tau\log(1/\delta)L_0^2}{3n^2\epsilon^2}.
\end{aligned}$$

Using the Riemannian PL condition we have

$$\mathbb{E}\left[f(w^{k+1})-f(w^*)\right] \leq \tau\mathbb{E}[\left\|\operatorname{grad} f(w^{k+1})\right\|^2] \leq \frac{1}{2}\mathbb{E}\left[F(w^k)-F(w^*)\right] + \frac{24dKL\tau^2\log(1/\delta)L_0^2}{3n^2\epsilon^2}.$$

Recursively applying the above for $k=0$ to $K-1$, we have

$$\begin{aligned}
\mathbb{E}\left[f(w^K)-f(w^*)\right] &\leq \frac{1}{2^K}\mathbb{E}\left[F(w^0)-F(w^*)\right] + \frac{8dKL\tau^2\log(1/\delta)L_0^2}{n^2\epsilon^2}\sum_{i=0}^{K-1}\frac{1}{2^i} \\
&\leq \frac{1}{2^K}\mathbb{E}\left[F(w^0)-F(w^*)\right] + \frac{8dKL\tau^2\log(1/\delta)L_0^2}{n^2\epsilon^2}\sum_{i=0}^{\infty}\frac{1}{2^i} \\
&= \frac{1}{2^K}\mathbb{E}\left[F(w^0)-F(w^*)\right] + \frac{16dKL\tau^2\log(1/\delta)L_0^2}{n^2\epsilon^2}.
\end{aligned}$$

Putting $K = \log\left(\frac{n^2\epsilon^2\mathbb{E}\left[F(w^0)-F(w^*)\right]}{dL\tau^2\log(1/\delta)L_0^2}\right)$ there is a constant c s.t

$$\mathbb{E}\left[f(w^K)-f(w^*)\right] \leq c\frac{dL\tau^2\log(1/\delta)L_0^2}{n^2\epsilon^2}\log\left(\frac{n^2\epsilon^2\mathbb{E}\left[F(w^0)-F(w^*)\right]}{dL\tau^2\log(1/\delta)L_0^2}\right).$$

Ignoring the log factors,

$$\mathbb{E}\left[f(w^K)-f(w^*)\right] = \mathcal{O}\left(\frac{dL\tau^2\log(1/\delta)L_0^2}{n^2\epsilon^2}\right). \tag{18}$$

Finally, the gradient complexity is given by,

$$\begin{aligned}
KS(n+2m) &= \log\left(\frac{n^2\epsilon^2\mathbb{E}\left[F(w^0)-F(w^*)\right]}{dL\tau^2\log(1/\delta)L_0^2}\right)\left(\lceil 6+\frac{18}{n-3}\rceil L\tau\zeta^{1/2}\frac{\mu_0}{\nu n^{1/3}}\right)\left(n+\lfloor\frac{n}{3\mu}\rfloor\right) \\
&\leq L\tau\zeta^{1/2}n^{2/3}\log\left(\frac{n^2\epsilon^2\mathbb{E}\left[F(w^0)-F(w^*)\right]}{dL\tau^2\log(1/\delta)L_0^2}\right).
\end{aligned} \tag{19}$$

□

D More experimental details for Section 6

Details on the parameter configurations of DP-RSVRG, DP-RSGD, and DP-RGD. For DP-RGD, we tune the clipping parameters from the set $\mathcal{C} = \{1, 0.1, 0.01\}$ and the number of epochs from $\{10, 20, 30\}$. For DP-RSGD, clipping parameter is chosen from $\mathcal{C} = \{1, 0.1, 0.01\}$ and number of epochs from $\{100000, 200000, 300000\}$. For DP-RSVRG number of epochs is chosen from $\{5, 10\}$ and set the frequency as $m = 1000$ and full gradient clipping parameter $\mathcal{C}$ is tuned from $\{1, 0.1\}$ and variance reduced gradient clipping parameter $\mathcal{C}_2$ from $\{1, 0.1, 0.01\}$. For all three algorithms we tune the learning rate from $\eta = \{5e^{-5}, 1e^{-5}5e^{-4}, 1e^{-4}, \ldots, 5e^{-1}, 1e^{-1}, 1, 2, \ldots, 5\}$.

D.1 Details on the the Fréchet mean of SPD matrices computation and the covariance descriptors

The Riemannian distance induced by the metric is given by $\mathrm{dist}(\mathbf{Z}_1, \mathbf{Z}_2) = \|\mathrm{Logm}(\mathbf{Z}_2^{-1/2}\mathbf{Z}_1\mathbf{Z}_2^{-1/2})\|_{\mathrm{F}}$, where Logm denotes matrix logarithm. Given points $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\} \in \mathrm{SPD}(m)$, the Fréchet mean is defined as the solution to following optimization problem: $\min_{\mathbf{W} \in \mathrm{SPD}(m)} \left\{ F(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{W}; \mathbf{Z}_i) = \frac{1}{n}\sum_{i=1}^n \|\mathrm{logm}(\mathbf{W}^{-1/2}\mathbf{Z}_i\mathbf{W}^{-1/2})\|_F^2 \right\}$. Riemannian gradient of f is given in terms inverse Exponential map $\mathrm{grad}\, f(\mathbf{W}, \mathbf{X}_i) = -2\mathrm{Exp}_{\mathbf{W}}^{-1}(\mathbf{X}_i) = -2\mathbf{W}^{1/2}\mathrm{Logm}(\mathbf{W}^{-1/2}\mathbf{X}_i\mathbf{W}^{-1/2})\mathbf{W}^{1/2}$. We take first two classes from PATHMNIST (Kather et al., 2019) (ADI, adipose tissue; BACK, background).

Covariance descriptors. Let $\mathcal{I} \in \mathbb{R}^{h \times w \times 3}$ denote a RGB image with height h and width w. Let $\phi : \mathbb{R}^{h \times w \times 3} \to \mathbb{R}^{hw \times k}$ be a feature extractor of dimension k, i.e. $\phi(\mathcal{I})(\mathbf{x})$ is a k-dimensional vector at each spatial coordinate $\mathbf{x}$ in the image's domain S. Given a small $\eta > 0$, the *covariance descriptor* $R_\eta : \mathbb{R}^{h \times w \times 3} \to \mathrm{SPD}(k)$ associated with ϕ is defined as

$$R_\eta(\mathcal{I}) = \left[\frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in S} (\phi(\mathcal{I})(\mathbf{x}) - \mu)(\phi(\mathcal{I})(\mathbf{x}) - \mu)^T \right] + \eta.I,$$

where $\mu = |S|^{-1}\sum_{\mathbf{x} \in \mathcal{S}} \phi(\mathcal{I})(\mathbf{x})$, and $\eta.I$ ensures $R_\eta(\mathcal{I}) \in \mathrm{SPD}(k)$. Our experiments on the private Fréchet mean computation problem (Section 6.3) use the covariance descriptors with following feature vector:

$$\phi(\mathcal{I})(\mathbf{x}) = \left[x, y, \mathcal{I}, |\mathcal{I}_x|, |\mathcal{I}_y|, |\mathcal{I}_{xx}|, |\mathcal{I}_{yy}|, \sqrt{|\mathcal{I}_x|^2 + |\mathcal{I}_y|^2}, \arctan\left(\frac{|\mathcal{I}|_x}{|\mathcal{I}|_y}\right) \right],$$

where $\mathbf{x} = (x, y)$, intensities derivatives are denoted by $\mathcal{I}_x, \mathcal{I}_y, \mathcal{I}_{xx}, \mathcal{I}_{yy}$ and $\eta = 10^{-6}$. Let $\star$ denote convolution operation, then first and second order intensity derivatives are computed as below,

$$\mathcal{I}_x = \mathcal{I} \star \frac{1}{4}\begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +6 & 0 & -12 \end{bmatrix}, \mathcal{I}_x = \mathcal{I} \star \frac{1}{4}\begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +6 & 0 & -12 \end{bmatrix},$$

$$\mathcal{I}_{xx} = \mathcal{I} \star \frac{1}{32}\begin{bmatrix} +1 & 0 & -2 & 0 & 1 \\ +4 & 0 & -8 & 0 & 4 \\ +6 & 0 & -12 & 0 & 6 \\ +4 & 0 & -8 & 0 & 4 \\ +1 & 0 & -2 & 0 & 1 \end{bmatrix}, \mathcal{I}_{yy} = \mathcal{I} \star \frac{1}{32}\begin{bmatrix} +1 & +4 & +6 & +4 & +1 \\ 0 & 0 & 0 & 0 & 0 \\ -2 & -8 & -12 & -8 & -2 \\ 0 & 0 & 0 & 0 & 0 \\ +1 & +4 & +6 & +4 & +1 \end{bmatrix}.$$

For RGB images $\phi(\mathcal{I})(\mathbf{x})$ is a 11-dimensional vector that makes $R_\eta(\mathcal{I})$ a 11×11 SPD matrix.

D.2 Details on the private leading eigenvector computation problem

The problem of computing the leading eigenvector of sample covariance matrix is $\min_{\mathbf{w} \in \mathbb{S}^m} \left\{ F(w) = \frac{1}{n}\sum_{i=1}^n f(\mathbf{w}; \mathbf{z}_i) = -\frac{1}{n}\sum_{i=1}^n \mathbf{w}^T(\mathbf{z}_i\mathbf{z}_i^T)\mathbf{w} \right\}$. It has been shown that above problem satisfies Riemannian PL condition (Zhang et al., 2016) while the problem is nonconvex in the Euclidean setting. Riemannian gradient of f is given by $\mathrm{grad}\, f(\mathbf{w}; \mathbf{z}_i) = -2(\mathbf{I}_{d+1} - \mathbf{w}\mathbf{w}^T)\mathbf{z}_i\mathbf{z}_i^T\mathbf{w}$.