On Riemannian Stochastic Approximation Schemes with Fixed Step-Size

Alain Durmus
Université Paris-Saclay
ENS Paris-Saclay, CNRS
Centre Borelli, F-91190 Gif-sur-Yvette, France
alain.durmus@ens-paris-saclay.fr

Pablo Jiménez
CMAP, École Polytechnique
Institut Polytechnique de Paris
pablo.jimenez-moreno@polytechnique.edu

Éric Moulines
CMAP, École Polytechnique, CNRS
Institut Polytechnique de Paris
eric.moulines@polytechnique.edu

Salem Said
Laboratoire IMS, Université de Bordeaux
salem.said@u-bordeaux.fr

Abstract

This paper studies fixed step-size stochastic approximation (SA) schemes, including stochastic gradient schemes, in a Riemannian framework. It is motivated by several applications, where geodesics can be computed explicitly, and their use accelerates crude Euclidean methods. A fixed step-size scheme defines a family of time-homogeneous Markov chains, parametrized by the step-size. Here, using this formulation, non-asymptotic performance bounds are derived, under Lyapunov conditions. Then, for any step-size, the corresponding Markov chain is proved to admit a unique stationary distribution, and to be geometrically ergodic. This result gives rise to a family of stationary distributions indexed by the step-size, which is further shown to converge to a Dirac measure, concentrated at the solution of the problem at hand, as the step-size goes to 0. Finally, the asymptotic rate of this convergence is established, through an asymptotic expansion of the bias, and a central limit theorem.

1 INTRODUCTION

This paper deals with the study of fixed step-size Stochastic Approximation (SA) algorithms (Robbins and Monro, 1951; Kushner and Yin, 2003; Polyak and Juditsky, 1992), defined on a Riemannian manifold Θ with metric $\mathfrak{g}$. Specifically, consider the problem

$$\begin{aligned} &\text{find } \theta \in \Theta \text{ satisfying } h(\theta) = 0 \,, \\ &\text{for a vector field } h : \Theta \to \mathrm{T}\Theta \,, \end{aligned} \tag{1}$$

where $\mathrm{T}\Theta$ denotes the tangent bundle of Θ, and h is only accessible through an oracle returning noisy estimates. The setting where $h = -\mathrm{grad}\, f$ is of particular interest for minimizing a smooth function $f : \Theta \to \mathbb{R}$. In the Euclidean setting, Stochastic Gradient Descent (SGD) and its variants are now common methods for solving this problem (Bottou, 2010; Bottou and Bousquet, 2008). However, it should be stressed that (1) encompasses several other applications in stochastic optimization, reinforcement learning or maximum likelihood estimation, such as online Expectation Maximization algorithms (Cappé and Moulines, 2009), policy gradient (Baxter and Bartlett, 2001) or Q-learning (Jaakkola et al., 1993). Minimization over a Riemannian manifold or its general formulation (1) arises in many applications: Principal Component Analysis (Edelman et al., 1998), dictionary recovery (Sun et al., 2017), matrix completion (Boumal and Absil, 2011), smooth semidefinite programs (Boumal et al., 2016), tensor factorization (Ishteva et al., 2011), and Riemannian barycenter estimation (Said and Manton, 2019; Arnaudon et al., 2012). This has motivated the development of a comprehensive framework for stochastic optimization problems on Riemannian manifolds. One of the first contributions in this field is Bonnabel (2013), which derives asymptotic convergence results for SA on Riemannian manifolds. Non-asymptotic results are obtained by Zhang and Sra (2016) for a geodesically convex function f. This study has been followed and completed by Zhang et al. (2016); Sato et al. (2019)

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

which introduce and analyze a Riemannian counterpart of the Stochastic Variance Reduced Gradient (SVRG) algorithm. Since then, many existing methods or results from the Euclidean case have been considered in a Riemannian setting. For example, Khuzani and Li (2017) suggest a Riemannian stochastic primal-dual algorithm and most recently Tripuraneni et al. (2018) study an averaged version of Riemannian SGD.

In this paper, we are interested in the study of fixed step-size SA methods of the form

$$\begin{aligned} &\theta_{n+1} = \mathrm{proj}_{\mathsf{S}}\left[\mathrm{Exp}_{\theta_n}\left\{\eta H_{\theta_n}(X_{n+1})\right\}\right] , \\ &\text{where } H_{\theta_n}(X_{n+1}) = h(\theta_n) + e_{\theta_n}(X_{n+1}) . \end{aligned} \tag{2}$$

In (2), $\eta > 0$ is a step-size, $(X_n)_{n\in\mathbb{N}^*}$ is an $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted process, defined on a filtered probability space, with values in a measurable space $(\mathsf{X}, \mathcal{X})$, and $e : \Theta \times \mathsf{X} \to \mathrm{T}\Theta$ is a measurable function, such that $\theta \mapsto e_\theta(x)$ is a vector field over Θ, for any $x \in \mathsf{X}$. In addition, $\mathrm{Exp}_\theta : \mathrm{T}_\theta\Theta \to \Theta$ is the Riemannian exponential mapping and $\mathrm{proj}_{\mathsf{S}} : \Theta \to \mathsf{S}$ is a projection-like operator onto a subset $\mathsf{S} \subset \Theta$. This recursion is a natural extension of Euclidean SA, akin to the Robbins-Monroe algorithm, in a Riemannian setting.

In the Euclidean setting, the study of fixed step-size SA, and in particular SGD, has recently attracted much attention, see e.g. Ma et al. (2018); Vaswani et al. (2019); Dieuleveut et al. (2017); Bach (2020); Bach and Moulines (2011). Indeed, first of all, the step-size η is the only parameter to tune, in contrast to the case where a decreasing sequence of step-sizes is used in (2). Furthermore, the forgetting of the initial condition is exponentially fast (Nedić and Bertsekas, 2001; Needell et al., 2014).

We aim to show, in a general Riemannian framework, that the use of (2) provides a good solution for (1). To this end, we establish non-asymptotic and asymptotic properties of $(\theta_n)_{n\in\mathbb{N}}$, in the limit $\eta \to 0$. Our contributions can be summarized as follows.

(1) We derive non-asymptotic bounds, for the convergence of $(\theta_n)_{n\in\mathbb{N}}$ to approximate solutions of (1), under general Lyapunov assumptions and mild assumptions on the manifold Θ and the subset S.
(2) Under additional regularity conditions, we show that $(\theta_n)_{n\in\mathbb{N}}$, as a Markov chain, admits a unique stationary distribution μ^η and is geometrically ergodic, *i.e.* converges to μ^η exponentially fast.
(3) We study the limiting behavior of the family $(\mu^\eta)_{\eta>0}$ as $\eta \to 0$. In particular, we show that if (1) admits a unique solution $\theta^\star$ and other suitable conditions hold, this family converges to the Dirac measure at $\theta^\star$. In addition, we asymptotically quantify this convergence, through a central limit theorem. Precisely, we prove that after a $\eta^{-1/2}$-rescaling, this family of stationary distributions converges weakly to a normal distribution as $\eta \to 0$. These results illustrate the exponential forgetting of initial condition of the scheme and that, at stationarity, the iterates $(\theta_n)_{n\in\mathbb{N}}$ stay in a $\mathcal{O}(\eta^{1/2})$-neighborhood of $\theta^\star$. In addition, they can be understood as generalizations to Riemannian spaces of Pflug (1986, Theorem 1) and Dieuleveut et al. (2017, Theorem 4).
(4) We apply our results to SGD. In particular, we establish the first non-asymptotic convergence bounds for strongly geodesically convex functions, without boundedness assumptions on the manifold Θ.
(5) Finally, we introduce and prove the convergence of an SGD scheme to compute the Riemannian barycenter, also known as the Karcher mean, of distributions on Hadamard manifolds. To the authors' knowledge, our contribution on this topic is one of the few without boundedness assumptions on the distribution.

For ease of reading, all the assumptions are gathered in the supplement Section S1. In the derivation of our results, we use crucially the fact that $(\theta_n)_{n\in\mathbb{N}}$ defines a Markov chain in Θ, under mild conditions. This interpretation has been successfully used in several papers dealing with the convergence of SA or SGD in Euclidean spaces; see e.g. Benveniste et al. (1990); Kushner and Huang (1981); Fort and Pagès (1999); Pflug (1986).

We consider a more general setting and milder conditions in comparison with most other studies in the field. Indeed, most papers do not consider the general SA framework, but only the case $h = -\mathrm{grad}\, f$, dealing with SGD and its variants. To the authors' knowledge, only Bonnabel (2013); Durmus et al. (2020) tackle the general SA problem (1). Our main contribution, compared to these two works, is to deal with the fixed step-size setting. Besides, our study considers general geodesically complete Riemannian manifolds which encompass Hadamard spaces, which have been the primary focus for Zhang and Sra (2016); Zhang et al. (2016); Tripuraneni et al. (2018).

Furthermore, a majority of the previous studies on SGD in a Riemannian space (see e.g. Zhang and Sra (2016); Zhang et al. (2016); Tripuraneni et al. (2018); Alimisis et al. (2020); Han and Gao (2020)), are purely local in nature, because of the assumption that $(\theta_n)_{n\in\mathbb{N}}$ stays almost surely in a (fixed and deterministic) compact and geodesically convex subset of Θ. For example, note that all the convergence results derived in Zhang and Sra (2016) depend on the diameter of the compact in which $(\theta_n)_{n\in\mathbb{N}}$ is assumed to stay. This assumption rarely holds in practice, and is quite difficult to verify in theory. It strongly limits the applicability of many results in the literature over the past few years. On the contrary, our results do not suffer from this problem,

and can all be applied either on a compact or non-compact Riemannian manifold. As a result, we consider a new SA method to estimate the Karcher mean of a distribution π on Θ, see Arnaudon et al. (2012); Le (2004); Zhang and Sra (2016); Iannazzo and Porcelli (2018), for which we derive non-asymptotic convergence bounds without boundedness conditions on the support of π.

Notations For any $\theta \in \Theta$ and $v, w \in \mathrm{T}_\theta\Theta$, denote by $\mathfrak{g}_\theta(v,w) = \langle v, w\rangle_\theta$ and its corresponding norm by $\mathfrak{g}_\theta(v,v) = \|v\|_\theta^2$. $\rho_\Theta : \Theta \times \Theta \to \mathbb{R}_+$ denotes the distance associated with the Riemannian metric $\mathfrak{g}$. For any $\theta_0 \in \Theta, r > 0$, set $\mathrm{B}(\theta_0, r) = \{\theta_1 \in \Theta : \rho_\Theta(\theta_0, \theta_1) < r\}$, the open ball centered at θ_0 with radius r. Similarly, we define closed balls in Θ by $\overline{\mathrm{B}}(\theta_0, r) = \{\theta_1 \in \Theta : \rho_\Theta(\theta_0, \theta_1) \le r\}$.

For a smooth function $g : \Theta \to \mathbb{R}$, we denote by $\operatorname{grad} g$ its Riemannian gradient (Lee, 2019, p. 27) and by $\operatorname{Hess} g$ its Riemannian, or covariant, Hessian (Lee, 2019, Example 4.22). For a curve $\gamma : I \to \Theta, \mathrm{T}^\gamma_{t_0,t_1} : \mathrm{T}_{\gamma(t_0)}\Theta \to \mathrm{T}_{\gamma(t_1)}\Theta$ stands for the parallel transport map associated to the Levi-Civita connection along γ from $\gamma(t_0)$ to $\gamma(t_1)$ (Lee, 2019, Equation 4.22). Moreover, for any $\theta \in \Theta$, under the assumption that Θ is complete, consider the Riemannian exponential map $\mathrm{Exp}_\theta : \mathrm{T}_\theta\Theta \to \Theta$, see Lee (2019, Proposition 5.19). This map projects a vector from the tangent space $\mathrm{T}_\theta\Theta$ onto the manifold Θ, following a geodesic curve.

2 CONSTANT STEPSIZE ANALYSIS FOR A CONSTRAINED SCHEME

2.1 Main Results

In this section, we study the Stochastic Approximation scheme (2), which is constrained on a subset $\mathsf{S} \subset \Theta$. The following assumption on the manifold Θ and S is considered all along this paper and allows us to rigorously define proj_S.

A 1. *Assume one of the following conditions.*

(i) Θ is a Hadamard manifold, i.e. a complete, simply connected Riemannian manifold with non-positive sectional curvature. In addition, S is a closed geodesically convex subset of Θ with non-empty interior.
(ii) Θ is a complete, connected Riemannian manifold and $\mathsf{S} = \Theta$.

Note that under **A**1, the exponential map $\mathrm{Exp} : \mathrm{T}\Theta \to \Theta$ is well-defined, see Lee (2019, Theorem 6.19). Under **A**1-(i), Sturm (2003, Proposition 2.6) shows that there exists $\mathrm{proj}_\mathsf{S} : \Theta \to \mathsf{S}$ which is the Riemannian counterpart of the Euclidean projection onto a closed convex subset. More precisely, proj_S is the unique mapping from Θ to S such that for any $\theta \in \Theta$, $\rho_\Theta(\mathrm{proj}_\mathsf{S}(\theta), \theta) = \inf_{\theta' \in \mathsf{S}} \rho_\Theta(\theta', \theta)$. Under **A**1-(ii), we simply set $\mathrm{proj}_\mathsf{S} = \mathrm{Id}$.

Recall that the recursion (2) only uses a noisy estimate H_θ of the mean field $h(\theta)$, for any $\theta \in \Theta$. We assume the following conditions on the noise to ensure convergence.

MD 1. *The sequence $(X_n)_{n\in\mathbb{N}^*}$ is independent and identically distributed (i.i.d.). In addition, for any $\theta \in \Theta$, $\mathbb{E}\left[e_\theta(X_1)\right] = 0$ and there exist $\sigma_0^2, \sigma_1^2 > 0$ such that for any $\theta \in \mathsf{S}$, $\mathbb{E}[\|e_\theta(X_1)\|_\theta^2] \le \sigma_0^2 + \sigma_1^2 \|h(\theta)\|_\theta^2$.*

MD 1 is referred to as the martingale difference setting which implies that $(\theta_n)_{n\in\mathbb{N}}$ is a time-homogeneous $(\mathcal{F}_n)_{n\in\mathbb{N}}$-Markov chain, for which we denote by Q_η its corresponding Markov kernel. The Euclidean counterpart of this assumption consists in replacing the Riemannian norm with the Euclidean one, which is usual in Stochastic Approximation (see Duflo, 1997).

MD 2. *(i) $\mathbb{P}$-almost surely, the vector field $\theta \mapsto e_\theta(X_1)$ is continuous on Θ.*
(ii) For any $\theta \in \Theta$, Leb_θ and the distribution of $e_\theta(X_1)$ are mutually absolutely continuous, where Leb_θ stands for the Lebesgue measure on $\mathrm{T}_\theta\Theta$.

MD2 ensures topological and aperiodicity properties of the Markov chain under consideration. This condition is used in the study of the limiting behaviour of $(\theta_n)_{n\in\mathbb{N}}$. Note that the condition **MD** 2-(ii) is automatically satisfied adding some Gaussian noise, *i.e.*, when $e_\theta(X_i)$ is replaced by $e_\theta(X_i) + \mathrm{p}_\theta(Z_i)$ where for any $\theta \in \Theta$, p_θ is any invertible linear application from $\mathbb{R}^d$ to $\mathrm{T}_\theta\Theta$ and $(Z_i)_{i\in\mathbb{N}^*}$ is a sequence of i.i.d. d-dimensional Gaussian random variables with zero-mean and covariance matrix identity. The same assumption is considered in the Euclidean case in Pflug (1986, Assumption A(v)-A(vi)).

To ensure recurrence of $(\theta_n)_{n\in\mathbb{N}}$, we assume the existence of a Lyapunov function $V : \Theta \to \mathbb{R}_+$ for the mean vector field h.

H 1. *(i) For any $\theta \in \Theta$, $V \circ \mathrm{proj}_\mathsf{S}(\theta) \le V(\theta)$.*
(ii) V is continuously differentiable on Θ and its Riemannian gradient $\operatorname{grad} V$ is geodesically L-Lipschitz, i.e., there exists $L \ge 0$ such that for any $\theta_0, \theta_1 \in \Theta$, and geodesic curve $\gamma : [0,1] \to \Theta$ such that $\gamma(0) = \theta_0$ and $\gamma(1) = \theta_1$,

$$\|\operatorname{grad} V(\theta_1) - \mathrm{T}^\gamma_{01} \operatorname{grad} V(\theta_0)\|_{\theta_1} \le L\ell(\gamma)\,, \qquad (3)$$

where $\ell(\gamma) = \|\dot\gamma(0)\|_{\theta_0}$ is the length of the geodesic.
(iii) V is proper on S, i.e., for any $M \ge 0$, there exists a compact set $\mathsf{K} \subset \mathsf{S}$ such that for any $\theta \in \mathsf{S} \setminus \mathsf{K}$, $V(\theta) > M$.

H 2. *There exist $C_1 \ge 0$ and $C_2 > 0$ such that for any $\theta \in \mathsf{S}$, $\|h(\theta)\|_\theta^2 + C_2 \langle \operatorname{grad} V(\theta), h(\theta)\rangle_\theta \le C_1$.*

In addition, to quantify the convergence of $(\theta_n)_{n\in\mathbb{N}}$ in

a neighborhood of a solution of (1), we consider the following condition for some compact set $\mathsf{K}^\star \subset \mathsf{S}$.

H3 ($\mathsf{K}^\star$). *There exists $\lambda > 0$ such that for any $\theta \in \mathsf{S}$, $\langle \operatorname{grad} V(\theta), h(\theta) \rangle_\theta \leq -\lambda V(\theta) \mathbb{1}_{\mathsf{S} \setminus \mathsf{K}^\star}(\theta)$.*

Note that under **H**3($\emptyset$), if $h(\theta) = 0$, then $V(\theta) = 0$ since V is a nonnegative function.

It is relevant to recognize that **H**1, **H**2 and **H**3 boil down to standard stability and recurrence conditions; see e.g. Benveniste et al. (1990); Duflo (1997). In the Euclidean case when we assume the uniqueness of a solution $\theta^\star$, a common choice for V is $\theta \mapsto \|\theta - \theta^\star\|^2$ because **H**1 is always satisfied. In this context, **H**2 and **H**3 would read $\|h(\theta)\| \leq A + B\|\theta - \theta^\star\|$ and $\inf_{\theta \in \mathsf{S} \setminus \mathsf{K}} \langle \theta - \theta^\star, h(\theta^\star) \rangle / \|\theta - \theta^\star\|^2 \geq \lambda$ respectively. They are therefore the Riemannian counterparts of Pflug (1986, Assumption A(i)-A(ii)). However, the square distance is no longer a suitable candidate in non-compact Riemannian settings, and therefore selecting a Lyapunov function adapted to the manifold Θ and the geometry of the mean field h is all the more important. Note that **H**1-(iii) is automatically satisfied if S is compact. In addition, in most cases $\mathsf{K}^\star$ and V are chosen such that $\mathsf{K}^\star = \emptyset$ or $\{\theta \in \mathsf{S} : \|h(\theta)\|_\theta \leq \varepsilon\}$ for some $\varepsilon \geq 0$, $-C_2 \langle h(\theta), \operatorname{grad} V(\theta) \rangle_\theta \geq \|h(\theta)\|_\theta^2$ for some $C_2 > 0$ and any $\theta \in \Theta$, and therefore **H**2 is satisfied with $C_1 = 0$.

The use of Lyapunov functions is really common and widespread to analyze stochastic approximation schemes, see Kushner and Yin (2003); Kushner and Huang (1981); Duflo (1997). However, compared to the Euclidean setting, the square distance cannot be used in many situations because it does not satisfy **H** 1-(ii). This brought us to consider a different Lyapunov function and therefore develop an adapted framework for the Riemannian case; see Section 2.2 hereafter for more details.

We start with our first result which is established along with all the other statements of this section in the supplement Section S3.

Theorem 1. *Assume **A**1, **MD**1, **H**1-(i)-(ii), **H**2.*

(a) Suppose in addition that for any $\theta \in \mathsf{S}, \langle \operatorname{grad} V(\theta), h(\theta) \rangle_\theta \leq 0$. Then, for any $\eta \in (0, \overline{\eta}]$, $\theta_0 \in \mathsf{S}$, and $n \in \mathbb{N}^*$,

$$n^{-1} \sum_{k=0}^{n-1} \mathbb{E}\left[- \langle \operatorname{grad} V(\theta_k), h(\theta_k) \rangle_{\theta_k} \right] \leq 2V(\theta_0)/(n\eta) + \eta b \,, \tag{4}$$

where $(\theta_n)_{n \in \mathbb{N}}$ is defined by (2) starting from θ_0, $\overline{\eta} = [2C_2 L(1 + \sigma_1^2)]^{-1}$, $b = 2L\{\sigma_0^2 + C_1(1 + \sigma_1^2)\}$.

*Suppose in addition that **H**3($\mathsf{K}^\star$) holds for some compact set $\mathsf{K}^\star \subset \mathsf{S}$.*

(b) Then for any $\eta \in (0, \overline{\eta}]$, $\theta_0 \in \mathsf{S}$, and $n \in \mathbb{N}^*$,

$$n^{-1} \sum_{k=0}^{n-1} \mathbb{E}[\mathbb{1}_{\mathsf{S} \setminus \mathsf{K}^\star}(\theta_k) V(\theta_k)] \leq V(\theta_0)/(an\eta) + \eta b/(2a) \,, \tag{5}$$

where $a = \lambda/2$.

(c) Define $\|V\|_{\mathsf{K}^\star} = \sup\{V(\theta) : \theta \in \mathsf{K}^\star\}$ if $\mathsf{K}^\star \neq \emptyset$ and $\|V\|_{\mathsf{K}^\star} = 0$ otherwise. Then for any $\eta \in (0, \overline{\eta}]$, $\theta_0 \in \mathsf{S}$, and any $n \in \mathbb{N}^*$,

$$\mathbb{E}[V(\theta_n)] \leq \{1 - \eta a\}^n V(\theta_0) + \|V\|_{\mathsf{K}^\star} + \eta b/(2a) \,. \tag{6}$$

Note that Theorem 1 gives, in the case $\mathsf{K}^\star = \emptyset$, non-asymptotic bounds of order η on $n^{-1} \sum_{k=0}^{n-1} \mathbb{E}[- \langle \operatorname{grad} V(\theta_k), h(\theta_k) \rangle_{\theta_k}]$, $n^{-1} \sum_{k=0}^{n-1} \mathbb{E}[V(\theta_k)]$ and $\mathbb{E}[V(\theta_n)]$ as $n \to +\infty$. In addition, the forgetting of the initial condition in (4) and (5) is linear w.r.t. n, contrary to (6) where it is exponential. A statement similar to Theorem 1-(b) holds only assuming **H**1-(i)-(ii) and replacing **H**3($\mathsf{K}^\star$) by the condition that there exists $\lambda > 0$ such that for any $\theta \in \mathsf{S}$, $\langle \operatorname{grad} V(\theta), h(\theta) \rangle_\theta \leq -\lambda \|h(\theta)\|_\theta^2 \mathbb{1}_{\mathsf{S} \setminus \mathsf{K}^\star}(\theta)$. This result is postponed to the supplement Theorem S2-Section S3.2. Theorem 1-(a) is a generalization of Hosseini and Sra (2019, Lemma 7) for SGD under a general Lyapunov condition and milder assumptions. Theorem 1-(a) is also stated with very similar assumptions in Durmus et al. (2020, Theorem 2) – except that $\langle \operatorname{grad} V(\theta), h(\theta) \rangle_\theta \leq 0$ in our setting is replaced with $\|\operatorname{grad} V(\theta)\|_\theta \leq \overline{c} \|h(\theta)\|_\theta$ – where the result is not restricted to constant step-size settings. However, Theorem 1-(c) can only be obtained for constant step-size schemes, and the following results use it as a stepping stone. We show in Section 4, how this generalization can be applied to SGD to obtain better convergence guarantees. Finally, in the same Section, we show that Theorem 1-(b)-(c) can be used to derive non-asymptotic convergence bounds for SGD applied to a geodesically strongly convex function, without any boundedness assumptions on Θ.

The study of the asymptotic behavior of $(\theta_n)_{n \in \mathbb{N}}$ is the second step towards understanding the quality of the approximation to the solution of (1). We now show, under suitable assumptions and for $\eta \leq \overline{\eta}$ given in Theorem 1, first, that the chain is ergodic and admits a unique invariant distribution, and second, that this measure converges weakly to the Dirac measure at some point $\theta^\star$, as the stepsize of the scheme goes to zero. In other words, the family of stationary distributions $(\mu^\eta)_{\eta \in (0, \overline{\eta}]}$ concentrates around $\theta^\star$ as $\eta \to 0$. Possible approximations of $\theta^\star$ are therefore derived from sampling from μ^η or taking its Riemannian barycenter, for a small enough η. If the sequence $(\theta_n)_{n \in \mathbb{N}}$ is ergodic, then as $n \to +\infty$ the marginal distributions of this Markov chain converge to μ^η and can be used in turn as proxy to solve (1). A remaining question is

to provide an estimate of the approximation error as a function of the step-size η. This is tackled in Section 3.

Theorem 2. *Assume* ***A*** *1,* ***MD*** *1,* ***MD*** *2,* ***H*** *1,* ***H*** *2 and* ***H*** *3($\mathsf{K}^\star$) for some compact set $\mathsf{K}^\star \subset \mathsf{S}$. Let $\eta \in (0, \overline{\eta}]$ where $\overline{\eta} = [2C_2L(1+\sigma_1^2)]^{-1}$. Then, $(\theta_n)_{n\in\mathbb{N}}$ defined by (2) admits a unique stationary distribution μ^η and is Harris-recurrent. In addition, there exist $\rho \in [0,1)$ and $C \geq 0$ such that for any $\theta_0 \in \mathsf{S}$ and $k \in \mathbb{N}$, $\left|\mathbb{E}[g(\theta_n)] - \int_\Theta g(\theta)\mathrm{d}\mu^\eta(\theta)\right| \leq C\rho^n(1+V(\theta_0))$, for any measurable function $g : \Theta \to \mathbb{R}$ satisfying $\sup_{\theta\in\Theta}\{|g|/V\} \leq 1$.*

Taking $n \to +\infty$ in Theorem 1-(c), we obtain by Theorem 2 that

$$\left|\int_\Theta g(\theta)\mathrm{d}\mu^\eta(\theta)\right| \leq \|V\|_{\mathsf{K}^\star} + \eta b/(2a)\,, \tag{7}$$

for any measurable function $g : \Theta \to \mathbb{R}$ satisfying $\sup_{\theta\in\Theta}\{|g|/V\} \leq 1$. In the case $\|V\|_{\mathsf{K}^\star} = 0$ (then $V(\theta) = 0$ for any $\theta \in \mathsf{K}^\star$), we get $\int_\Theta V(\theta)\mathrm{d}\mu^\eta(\theta) \leq \eta b/(2a)$. Therefore, this result indicates that the family $\{\mu^\eta : \eta \in (0,\overline{\eta}]\}$ concentrates in a $\mathcal{O}(\eta)$-neighborhood of $\mathsf{K}^\star$ as $\eta \to 0$. In particular, if V admits a unique zero $\theta^\star$ which corresponds in many applications to a solution of (1), then we can expect that $\{\mu^\eta : \eta \in (0,\overline{\eta}]\}$ converges in distribution to $\delta_{\theta^\star}$, the Dirac measure at $\theta^\star$, as $\eta \to 0$. The specific additional conditions to obtain such a result are the following.

H4. *There exists $\theta^\star \in \mathsf{S}$ such that for any $r > 0$,* ***H*** *3($\overline{\mathrm{B}}(\theta^\star, r)$) holds and that there exists $c_r > 0$ satisfying for any $\theta \in \mathsf{S} \setminus \overline{\mathrm{B}}(\theta^\star, r)$, $c_r \leq V(\theta)$.*

Note that assuming **H**4 is weaker than assuming **H** 3($\{\theta^\star\}$) since in the first case the constant $\lambda > 0$ in **H** 3($\overline{\mathrm{B}}(\theta^\star, r)$) may depend on r. Pflug (1986) considers a similar assumption in the Euclidean case (see Pflug, 1986, Assumption A(i)).

As announced previously, we obtain the convergence in distribution of $\{\mu^\eta : \eta \in (0,\overline{\eta}]\}$.

Theorem 3. *Assume* ***A*** *1,* ***MD*** *1,* ***MD*** *2,* ***H*** *1 and* ***H*** *2 and let $\overline{\eta} = [2C_2L(1+\sigma_1^2)]^{-1}$.*

(a) In addition suppose ***H*** *3($\mathsf{K}^\star$) holds for some compact set $\mathsf{K}^\star \subset \mathsf{S}$ and that there exists $c > 0$ such that for any $\theta \in \mathsf{S} \setminus \mathsf{K}^\star$, $c \leq V(\theta)$. Then $\lim_{\eta\to 0} \mu^\eta\{\mathsf{K}^\star\} = 1$, where μ^η is the stationary distribution of Q_η for $\eta \in (0,\overline{\eta}]$.*
(b) In addition suppose ***H*** *4 holds. Then $(\mu^\eta)_{\eta\in(0,\overline{\eta}]}$ converges weakly to $\delta_{\theta^\star}$, as $\eta \to 0$.*

2.2 Two Examples of Lyapunov Functions

Having stated the main results of this section, we give two examples of Lyapunov functions V under the following setting for Θ.

A 2. *Θ is a Hadamard manifold. In addition, there exists $\kappa > 0$ such that the sectional curvature of Θ is bounded below by $-\kappa^2$.*

A classical choice of Lyapunov function on Euclidean spaces is $\theta \mapsto \rho_\Theta^2(\theta, \theta^\star)$, being both strongly convex and Lipschitz-gradient. However, this function does not satisfy **H**1-(ii) as soon as Θ has non-zero curvature and is non-compact. In an effort to show the capital impact of curvature and in order to obtain a valid Lyapunov function satisfying the conditions **H**1 and **H**3($\overline{\mathrm{B}}(\theta^\star, r)$) for $r > 0$, we now introduce the necessary assumptions and consider a truncated version of $\theta \mapsto \rho_\Theta^2(\theta, \theta^\star)$.

Let $\mathsf{H} = \{\mathrm{Exp}_\theta(tH_\theta(x)) : \theta \in \mathsf{S}, x \in \mathsf{X}, t \in [0,\eta]\}$ be the set of all points reached from geodesics $\gamma : [0,1] \to \Theta$ of the form $\gamma(0) \in \mathsf{S}$ and $\dot\gamma(0) = \eta H_{\gamma(0)}(x)$, for any $x \in \mathsf{X}$. We assume in our next result that the closure of H is compact which is implied for example in the case where S is compact and $(\theta, x) \mapsto H_\theta(x)$ is bounded on $\mathsf{S} \times \mathsf{X}$.

Proposition 4. *Assume* ***A*** *2 and that the closure $\overline{\mathsf{H}}$ of H is compact, denote $\mathrm{D}_\mathsf{H} = \mathrm{diam}(\overline{\mathsf{H}})$. Consider a smooth function $\chi_\mathsf{H} : \Theta \to [0,1]$ with compact support satisfying $\chi_\mathsf{H}(\theta) = 1$ for any $\theta \in \overline{\mathsf{H}}$ and for any $\theta \in \Theta$ such that $\inf_{\theta'\in\overline{\mathsf{H}}} \rho_\Theta(\theta', \theta) \geq 1$, it holds $\chi_\mathsf{H}(\theta) = 0$. Consider now $V_2 : \Theta \to \mathbb{R}_+$ defined for any $\theta \in \Theta$ by*

$$V_2(\theta) = \chi_\mathsf{H}(\theta)\rho_\Theta^2(\theta^\star, \theta) + (1 - \chi_\mathsf{H}(\theta))\mathrm{D}_\mathsf{H}^2\,.$$

Then, ***H*** *1-(i)-(ii) holds with $V \leftarrow V_2$ and $L \leftarrow C_\chi(\mathrm{D}_\mathsf{H}+1)(1+\kappa\coth(\kappa\mathrm{D}_\mathsf{H}))$ where $C_\chi \geq 0$ is a constant only depending on χ_H. Suppose in addition that there exist $r > 0, \lambda_\rho > 0$ such that for any $\theta \in \mathsf{S}$,*

$$-\left\langle \mathrm{Exp}_\theta^{-1}(\theta^\star), h(\theta)\right\rangle_\theta \leq -\lambda_\rho\rho_\Theta^2(\theta^\star, \theta)\mathbb{1}_{\mathsf{S}\setminus\overline{\mathrm{B}}(\theta^\star, r)}(\theta)\,. \tag{8}$$

Then ***H*** *3($\overline{\mathrm{B}}(\theta^\star, r)$) holds with $\lambda \leftarrow \lambda_\rho$.*

Note that under the setting of Proposition 4, $V_2(\theta) \geq c$ for any $\theta \in \mathsf{S} \setminus \overline{\mathrm{B}}(\theta^\star, r)$ by definition, since it is continuous. Clearly, **H**1-(iii) does not hold for V_2 if S is non-compact, since V_2 is constant outside of the support of χ_H. For this reason, and to weaken the assumptions of Proposition 4, we introduce a "Huberized" version of the distance to $\theta^\star$.

Proposition 5. *Assume* ***A*** *2. Let $\delta > 0$ and consider $V_1 : \Theta \to \mathbb{R}_+$ defined for any $\theta \in \Theta$ by*

$$V_1(\theta) = \delta^2\left\{(\rho_\Theta(\theta^\star, \theta)/\delta)^2 + 1\right\}^{1/2} - \delta^2\,. \tag{9}$$

Then, ***H*** *1 holds with $V \leftarrow V_1$ and $L \leftarrow 1+\kappa\delta$. Suppose in addition that there exist $r > 0, \lambda_\rho > 0$ such that for any $\theta \in \mathsf{S}$, (8) holds. Then,* ***H*** *3($\overline{\mathrm{B}}(\theta^\star, r)$) holds and $\lambda \leftarrow \lambda_\rho$.*

The proof is an extension of Durmus et al. (2020, Lemma 17) and relies heavily on this result. This Lyapunov function is still constructed upon the distance function, but as Proposition 5 shows, it is better

suited for non-positive curvature spaces. expansion of V, which is only possible because V is smooth. Note that under the setting of Proposition 5, $V_1(\theta) \geq c$ for any $\theta \in \mathsf{S} \setminus \overline{\mathrm{B}}(\theta^\star, r)$ by definition since it is continuous.

It is worth mentioning that if either Proposition 4 or Proposition 5 can be applied, in order to use Theorem 1 and Theorem 2 (resp. Theorem 3-(b)) the only condition to verify (relative to the Lyapunov function) is **H**2 (resp. are **H**2 and **H**4).

3 ASYMPTOTIC EXPANSION AND LAW IN THE UNCONSTRAINED CASE

The purpose of this section is to quantify the convergence derived in Theorem 3-(b). First, we establish an asymptotic expansion for the bias $\int_\Theta \mathrm{g}(\theta)\mathrm{d}\mu^\eta(\theta) - \mathrm{g}(\theta^\star)$ w.r.t. the step size η for g belonging to a certain class of smooth functions from Θ to $\mathbb{R}$. Our result can be applied to SGD ($h = -\mathrm{grad}\, f$) and implies then an asymptotic expansion of $\int_\Theta \|\mathrm{grad}\, f(\theta)\|_\theta^2 \mathrm{d}\mu^\eta(\theta)$. Secondly, we establish that the convergence derived in Theorem 3-(b) occurs at a rate $\eta^{1/2}$, through a central limit theorem for $(\mu^\eta)_{\eta\in(0,\overline{\eta}]}$. These two results can be understood as a bias-variance decomposition in which both terms are of order η and are therefore weak counterparts of Dieuleveut et al. (2017, Proposition 3, Theorem 5), Pflug (1986, Theorem 1) in a Riemannian setting. The related proofs are postponed to the supplement Section S4.

3.1 Asymptotic Expansion as $\eta \to 0$

Here, we assume that **A**1-(ii) holds, Θ is compact and the conditions of Theorem 3-(b) hold. In addition, define the covariance tensor field Σ on Θ, for any $\theta \in \Theta$ by,

$$\Sigma(\theta) = \mathbb{E}\left[e_\theta(X_1) \otimes e_\theta(X_1)\right] \ . \tag{10}$$

Under appropriate conditions, letting $n \to +\infty$ in Theorem 1-(b), and Theorem S2 in the supplement, show respectively that $\int_\Theta V(\theta)\mathrm{d}\mu^\eta(\theta)$ and $\int_\Theta \|h(\theta)\|_\theta^2 \mathrm{d}\mu^\eta(\theta)$ are bounded by a term of order η. We specify this result in the case where $h = -\mathrm{grad}\, f$ for a smooth objective function $f : \Theta \to \mathbb{R}$. More precisely, we establish in what follows a weak asymptotic expansion for $\int_\Theta \|\mathrm{grad}\, f(\theta)\|_\theta^2 \mathrm{d}\mu^\eta(\theta)$, as $\eta \to 0$ based on the following result for which we assume:

MD3. *Σ is a continuous $(2,0)$-tensor field on Θ.*

Denote the contraction of a covariant 2-tensor F with a contravariant 2-tensor G on Θ by $[F : G]$; see Section S6.2 (S69)-(S70) in the supplementary for more details. For two matrices A, B, $[A : B]$ just corresponds to $\mathrm{Tr}(AB^\top)$, where $^\top$ denotes the transpose.

Theorem 6. *Assume **A** 1-(ii), h is continuous, $h(\theta^\star) = 0$ and Θ is compact. Assume also **MD**1, **MD** 2, **MD**3, **H**1, **H**2 and **H**4. Let $\overline{\eta} = [2C_2L(1+\sigma_1^2)]^{-1}$. Then for any $\eta \in (0, \overline{\eta}]$ and smooth function $g : \Theta \to \mathbb{R}$, we have*

$$-\int_\Theta \langle \mathrm{grad}\, g(\theta), h(\theta)\rangle_\theta \mathrm{d}\mu^\eta(\theta) = \frac{\eta}{2}\left[\mathrm{Hess}\, g : \Sigma\right](\theta^\star) + \mathscr{R}_{g,\eta} \ ,$$

where $\lim_{\eta\to 0}\{|\mathscr{R}_{g,\eta}|/\eta\} = 0$.

Applying this result to SGD, *i.e.* $h = -\mathrm{grad}\, f$ and $g = f$, we obtain that $\int_\Theta \|\mathrm{grad}\, f(\theta)\|_\theta^2 \mathrm{d}\mu^\eta(\theta) = (\eta/2)\left[\mathrm{Hess}\, f : \Sigma\right](\theta^\star) + \mathscr{R}_{f,\eta}$ with $\lim_{\eta\to 0}\{|\mathscr{R}_{f,\eta}|/\eta\} = 0$.

3.2 A Central Limit Theorem on $(\mu^\eta)_{\eta\in(0,\overline{\eta}]}$

Now, we assume both **A** 1-(i) and **A** 1-(ii), meaning $\mathsf{S} = \Theta$ and Θ is a Hadamard manifold. Note that under this setting $\mathrm{Exp}_\theta^{-1} : \Theta \to \mathrm{T}_\theta\Theta$ is a well defined diffeomorphism for any $\theta \in \Theta$ by (Lee, 2019, Proposition 12.9). In addition, we assume that the other conditions of Theorem 3-(b) hold. Following the approach of Pflug (1986) in Euclidean SA, to find the asymptotic rate of convergence of the family $(\mu^\eta)_{\eta\in(0,\overline{\eta}]}$ defined in Section 2, we establish a central limit theorem in $\mathrm{T}_{\theta^\star}\Theta$, for the family of pushforward measures $(\overline{\nu}^\eta)_{\eta\in(0,\overline{\eta}]}$ defined for any $\mathsf{A} \in \mathcal{B}(\mathrm{T}_{\theta^\star}\Theta)$ by

$$\overline{\nu}^\eta(\mathsf{A}) = \mu^\eta\left(\mathrm{Exp}_{\theta^\star}(\eta^{1/2}\mathsf{A})\right) \ . \tag{11}$$

It is shown in Section S4 that for any $\eta \in (0, \overline{\eta}]$, $\overline{\nu}^\eta$ is the stationary distribution of the rescaled and projected Markov chain $(\overline{U}_n)_{n\in\mathbb{N}}$ defined for any $n \in \mathbb{N}$ by $\overline{U}_n = \eta^{-1/2}\mathrm{Exp}_{\theta^\star}^{-1}(\theta_n)$. Therefore, since under **A**1-(i)-(ii), for any $u \in \mathrm{T}_{\theta^\star}\Theta$, $\rho_\Theta(\theta^\star, \mathrm{Exp}_{\theta^\star}(u)) = \|u\|_{\theta^\star}$ by (Lee, 2019, Corollary 6.12,Proposition 12.9), showing a central limit theorem for the family $(\overline{\nu}^\eta)_{\eta\in(0,\overline{\eta}]}$ as $\eta \to 0$ shows that asymptotically $(\mu^\eta)_{\eta\in(0,\overline{\eta}]}$ concentrates in regions of diameter $\mathcal{O}(\eta^{1/2})$ around $\theta^\star$ for the Riemannian distance.

We consider the following assumptions.

MD4. *There exist $\varepsilon_e > 0$, $\tilde{\sigma}_0^2, \tilde{\sigma}_1^2 \geq 0$ such that for any $\theta \in \Theta$, $\mathbb{E}[\|e_\theta(X_1)\|_\theta^{2+\varepsilon_e}] \leq \tilde{\sigma}_0^2 + \tilde{\sigma}_1^2 V(\theta)$.*

H5. *There exist a linear mapping $\mathbf{A} : \mathrm{T}_{\theta^\star}\Theta \to \mathrm{T}_{\theta^\star}\Theta$ and a map $\mathscr{H} : \Theta \to \mathrm{T}_{\theta^\star}\Theta$, such that for any $\theta \in \Theta$,*

$$h(\theta) = \mathrm{T}_{01}^\gamma\left(\mathbf{A}\mathrm{Exp}_{\theta^\star}^{-1}(\theta) + \mathscr{H}(\theta)\right) \ , \tag{12}$$

*where $\theta^\star$ is defined in **H**4, T_{01}^γ denotes parallel transport along the geodesic $\gamma : [0,1] \to \Theta$ with $\gamma(0) = \theta^\star$ and $\gamma(1) = \theta$, and $\lim_{\theta\to\theta^\star}\{\|\mathscr{H}(\theta)\|_{\theta^\star}/\rho_\Theta(\theta^\star, \theta)\} = 0$. In addition, the eigenvalues of the matrix $\mathbf{A}$ all have strictly negative real parts. Finally, there exists $C_3 > 0$ such that for any $\theta \in \Theta$, $\|h(\theta)\|_\theta \leq C_3\rho_\Theta(\theta^\star, \theta)$.*

MD4 is a strengthened version of **MD** 1. We show in Theorem S16 that (12) holds in the case h is twice continuously differentiable on Θ with $\mathbf{A} = \nabla h(\theta^\star)$. Note

that when $\Theta = \mathbb{R}^d$ is equipped with its Euclidean metric, (12) translates into $h(\theta) = \mathbf{A}(\theta - \theta^\star) + \mathscr{H}(\theta)$ with $\lim_{\theta \to \theta^\star} \{\|\mathscr{H}(\theta)\| / \|\theta - \theta^\star\|\} = 0$, which is exactly the assumption in Pflug (1986, Assumption B(ii)). For ease of notation, we also denote by $\mathbf{A}$ and $\Sigma(\theta^\star)$ the matrices associated with these two linear applications in some orthonormal basis of $\mathrm{T}_{\theta^\star}\Theta$. **H**5 guarantees the existence and uniqueness of the solution $\mathbf{V} \in \mathbb{R}^{d \times d}$ of the Lyapunov equation $\mathbf{A}\mathbf{V} + \mathbf{V}\mathbf{A}^\top = \Sigma(\theta^\star)$, see (Horn and Johnson, 1994, Theorem 2.2.1).

We also assume that V can be compared to a function of the distance on Θ which leads to the strengthening of **H**4.

H 6. *There exists $\theta^\star$ such that **H** 3($\{\theta^\star\}$) holds and there exists $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any $\theta \in \Theta$, $V(\theta) \geq \phi(\rho_\Theta(\theta^\star, \theta))$ and for any $r > 0$, $\inf_{[r,+\infty)} \phi > 0$. In addition, there exists $\overline{a} > 0$, such that $\lim_{r \to +\infty} \sup_{a \leq \overline{a}} a/\phi(a^{1/2} r) = 0$.*

Note that the assumption on the growth rate of the Lyapunov function is verified when $V = V_1$, considered in Proposition 5. In this case, we can take $\phi(r) = \delta^2[1 + (r/\delta)^2]^{1/2} - \delta^2$. Pflug (1986) only considers the Euclidean case with $V : \theta \mapsto \|\theta - \theta^\star\|^2$, in which case **H**6 boils down to $\langle \theta - \theta^\star, h(\theta) \rangle \geq \|\theta - \theta^\star\|^2 \lambda$. In addition, for this particular choice of Lyapunov function, it is equivalent to **H**3($\emptyset$), as the other conditions result from the existence and uniqueness of the solution $\theta^\star$ and the Euclidean properties of the squared distance function (see Pflug, 1986, Assumption A(i)-B(iii)). However, for our study, we have to generalize this assumption to any well-suited Lyapunov function, since, as we have already discussed, taking $V : \theta \mapsto \|\theta - \theta^\star\|^2$ is no longer an option in the Riemannian case.

Theorem 7. *Assume **A**1-(i)-(ii), **MD**1, **MD**2, **MD**3, **MD**4, **H**1, **H**2, **H**5 and **H**6 hold. Suppose in addition that $h(\theta^\star) = 0$, h is continuous and let $\overline{\eta} = [2C_2 L(1 + \sigma_1^2)]^{-1} \wedge (4C_3)^{-1}$. Then, the family of distributions $(\overline{\nu}^\eta)_{\eta \in (0,\overline{\eta}]}$, defined by (11), converges weakly to $\mathrm{N}(0, \mathbf{V})$ as $\eta \to 0$ on $\mathrm{T}_{\theta^\star}\Theta$, where $\mathbf{V}$ is the unique solution to the Lyapunov equation $\mathbf{A}\mathbf{V} + \mathbf{V}\mathbf{A}^\top = \Sigma(\theta^\star)$.*

Even though $\mathrm{N}(0, \mathbf{V})$ is a distribution on $\mathbb{R}^d$, we identify $\mathbb{R}^d$ with $\mathrm{T}_{\theta^\star}\Theta$ using the same orthonormal basis as before. As mentioned in Section 2, Theorem 7 complements Theorem 3 because it proves that the asymptotic rate of convergence of $(\mu^\eta)_{\eta \in (0,\overline{\eta}]}$ to $\delta_{\theta^\star}$ is $\eta^{1/2}$, since $(\overline{U}_n)_{n \in \mathbb{N}}$ is rescaled by this factor with respect to the actual SA scheme $(\theta_n)_{n \in \mathbb{N}}$. Finally Theorem 7 can be seen as a Riemannian counterpart of Pflug (1986, Theorem 1). In the following section, we illustrate our results on SGD.

4 APPLICATION TO SGD

We assume throughout this section that **A** 1-(i)-(ii) holds. We apply the results of Section 2 and Section 3, to the unconstrained stochastic gradient scheme, *i.e.* $(\theta_n)_{n \in \mathbb{N}}$ defined by (2) with $h = -\mathrm{grad}\, f$ and $\mathsf{S} = \Theta$. Proofs are postponed to the supplement, Section S5.

Geodesically Strongly Convex and Smooth Function First, the objective function $f : \Theta \to \mathbb{R}$ is subject to the following assumptions.

F 1. *$f : \Theta \to \mathbb{R}$ is twice continuously differentiable and $\mathrm{grad}\, f$ is geodesically L_f-Lipschitz, see (3).*

F 2. *f is continuously differentiable on Θ and λ_f-strongly geodesically convex, for some $\lambda_f > 0$, i.e. for any $\theta_1, \theta_2 \in \Theta$, $f(\theta_2) \geq f(\theta_1) + \langle \mathrm{Exp}_{\theta_1}^{-1}(\theta_2), \mathrm{grad}\, f(\theta_1) \rangle_{\theta_1} + \lambda_f \rho_\Theta^2(\theta_1, \theta_2)$.*

Under **F**2, f admits a unique minimizer denoted by $\theta^\star$. In addition, we have the following inequalities.

Lemma 8. *Assume **A**1-(i)-(ii) and **F**2. Then for any $\theta \in \Theta$, we have*

$$\begin{aligned} &\|\mathrm{grad}\, f(\theta)\|_\theta^2 \geq \lambda_f (f(\theta) - f(\theta^\star)) \quad \text{and}\,, \\ &f(\theta) - f(\theta^\star) \geq \lambda_f \rho_\Theta^2(\theta, \theta^\star)\,. \end{aligned} \tag{13}$$

Under **F** 1 and **F** 2, Lemma 8 implies that $V(\theta) = f(\theta) - f(\theta^\star)$ and $h = -\mathrm{grad}\, f$ satisfy **H**1 with $L \leftarrow L_f$, **H**2 with $C_1 \leftarrow 0, C_2 \leftarrow 1$ and **H**3($\emptyset$) with $\lambda \leftarrow \lambda_f$. A direct application of Theorem 1-(c) leads to the following result.

Corollary 9. *Assume **A** 1-(i)-(ii), **MD** 1, **F** 1, **F** 2. Consider $(\theta_n)_{n \in \mathbb{N}}$ defined by (2) with $h = -\mathrm{grad}\, f$. Let $\overline{\eta} = [2L_f(1 + \sigma_1^2)]^{-1}$ and $\eta \in (0, \overline{\eta}]$. For any $\theta_0 \in \Theta$, and $n \in \mathbb{N}$,*

$$\mathbb{E}[f(\theta_n) - f(\theta^\star)] \leq (1 - \eta\lambda_f/2)^n (f(\theta_0) - f(\theta^\star)) + 2\eta\, L_f \sigma_0^2/\lambda_f\,.$$

Then, setting $\eta = \overline{\eta} \wedge [\varepsilon \lambda_f / \{4\sigma_0^2 L_f\}]$, for $\varepsilon \in (0,1)$, and $n = \lceil [\log(1/\varepsilon) - \log(f(\theta_0) - f(\theta^\star))] / \log(1 - \eta\lambda_f/2) \rceil$, we get $\mathbb{E}[f(\theta_n) - f(\theta^\star)] \leq \varepsilon$.

Corollary 9 shows that (2) has a computational complexity of order $\mathcal{O}(\log(1/\varepsilon)\varepsilon^{-1})$ to minimize f, without any boundedness assumptions on Θ, contrary to Zhang and Sra (2016). In addition, Lemma 8 also implies that **H**5 and **H**6 hold if f is three times continuously differentiable and therefore Theorem 7 can be applied.

Geodesically Quasi-Convex Function with Bounded Gradient Consider the following assumption.

F 3. *f is twice continuously differentiable. Further, there exists $\tilde{\lambda}_f > 0$ such that for any $\theta \in \Theta$, $-\langle \mathrm{Exp}_\theta^{-1}(\theta^\star), \mathrm{grad}\, f(\theta) \rangle_\theta \geq \tilde{\lambda}_f V_1(\theta)$, where V_1 is defined by (9) with $\delta = 1$. In addition, there exists*

$C_f > 0$ *such that for any* $\theta \in \Theta$, $\|\mathrm{grad}\, f(\theta)\|_\theta^2 \leq C_f(\rho_\Theta^2(\theta^\star, \theta) \wedge 1)$.

In the Euclidean case, this assumption corresponds to weak quasi-convexity as considered in Hardt et al. (2019, Definition 2.1), $\langle \mathrm{grad}\, f(\theta), \theta - \theta^\star \rangle \geq \tau[f(\theta) - f(\theta^\star)]$ where the Lyapunov function is $\theta \mapsto f(\theta) - f(\theta^\star)$. Note that the geodesical quasi-convexity condition **F**3 is a weaker version of the usual geodesical convexity in the SGD setting (see Zhang and Sra, 2016). By introducing **F**3, we can relax the condition **F**1 using the following result.

Lemma 10. *Assume* ***A***2 *and* ***F***2. *Suppose in addition that* f *is twice continuously differentiable and there exists* $M_f > 0$ *such that for any* $\theta \in \Theta$, $\|\mathrm{grad}\, f(\theta)\|_\theta^2 \leq M_f \rho_\Theta^2(\theta^\star, \theta)$. *Let* $\tilde{f} = \{f - f(\theta^\star) + 1\}^{1/2}$. *Then* $\tilde{f}$ *satisfies* ***F***3 *with* $C_f \leftarrow (M_f/4)[1 \wedge \lambda_f]$ *and* $\tilde{\lambda}_f \leftarrow \lambda_f/(2M_f^{1/2})$.

Note that the condition introduced in Lemma 10 is a relaxation of the condition that $\mathrm{grad}\, f$ is geodesically Lipschitz. Indeed, by Jost (2005, Theorem 5.6.1), for $\theta^\star \in \Theta$, $\theta \mapsto \rho_\Theta^2(\theta^\star, \theta)$ satisfies the conditions of Lemma 10 but its gradient is not geodesically Lipschitz. A non-asymptotic bound is now given in terms of the distance-like function, defined for any $\theta_1, \theta_2 \in \Theta$ by

$$D_\Theta^2(\theta_1, \theta_2) = \rho_\Theta^2(\theta_1, \theta_2)/(1 + \rho_\Theta^2(\theta_1, \theta_2)) \,. \tag{14}$$

Proposition 11. *Assume that* ***A***2, ***MD***1, ***F***3 *hold. Let* $\overline{\eta} = [(8C_f/\tilde{\lambda}_f)(1+\kappa)(1+\sigma_1^2)]^{-1}$ *and* $\eta \in (0, \overline{\eta}]$. *Consider* $(\theta_n)_{n\in\mathbb{N}}$ *defined by* (2) *with* $h = -\mathrm{grad}\, f$ *and* $\mathsf{S} = \Theta$. *Then, for any* $\theta_0 \in \Theta$ *and* $n \in \mathbb{N}^*$,

$$n^{-1} \sum_{k=0}^{n-1} \mathbb{E}\left[D_\Theta^2(\theta^\star, \theta_n)\right] \leq 4V_1(\theta_0)/(n\eta\tilde{\lambda}_f) + 4\eta(1+\kappa)\sigma_0^2/\tilde{\lambda}_f \,,$$

where κ *is given in* ***A***2, *and* V_1 *is defined by* (9) *with* $\delta = 1$.

To the authors' knowledge, such a bound is novel even in a deterministic setting.

Application to the Riemannian Barycenter Problem To conclude our study, we consider the problem of computing the Riemannian barycenter $\theta^\star$ of a probability distribution π on a Hadamard manifold Θ. First, we look at the discrete case:

$$\pi = M_\pi^{-1} \textstyle\sum_{i=1}^{M_\pi} \delta_{\overline{\theta}_i} \,, \tag{15}$$

where $M_\pi \in \mathbb{N}^*$ and $\{\overline{\theta}_i\}_{i=1}^{M_\pi} \in \Theta^{M_\pi}$. The Riemannian barycenter $\theta^\star$ or Karcher mean of π (Arnaudon et al., 2012) is the unique global minimum of the function $f_\pi : \theta \mapsto \sum_{i=1}^{M_\pi} \rho_\Theta^2(\theta, \overline{\theta}_i)/(2M_\pi)$. By Jost (2005, Theorem 5.6.1), $\mathrm{grad}\, f_\pi(\theta) = -M_\pi^{-1} \sum_{i=1}^{M_\pi} \mathrm{Exp}_\theta^{-1}(\overline{\theta}_i)$ for any $\theta \in \Theta$ and f_π satisfies **F**2 with $\lambda_f = 1/2$ using Durmus et al. (2020, Lemma 10). Therefore, by Lemma 10, Proposition 11 can be applied. In addition, we get the following result, as an application of Proposition 4 and Theorem 1-(c).

Proposition 12. *Assume* ***A***2. *Let* $\theta_\pi^\star$ *be the Riemannian barycenter of the probability measure* π *in* (15) *on the Hadamard manifold* Θ, *and let* $(\theta_n)_{n\in\mathbb{N}}$ *be given by* $\theta_{n+1} = \mathrm{Exp}_{\theta_n}(\eta \mathrm{Exp}_{\theta_n}^{-1}(X_{n+1}))$, *where* $(X_n)_{n\in\mathbb{N}^*}$ *is a sequence of i.i.d. random variables with distribution* π. *Then, for any* $\eta \in (0, 1/(CL_\pi^3)]$, $\theta_0 \in \Theta$ *and* $n \in \mathbb{N}$,

$$\mathbb{E}[\rho_\Theta^2(\theta_n, \theta_\pi^\star)] \leq (1 - \eta/4)^n \rho_\Theta^2(\theta_0, \theta_\pi^\star) + C\eta\, L_\pi \mathrm{D}^2 \,,$$

where $L_\pi = (1 + \mathrm{D})(1 + \kappa \coth(\kappa \mathrm{D}))$, C *is a universal constant, and* $\mathrm{D} = \max_{i=1,\ldots,M_\pi} \rho_\Theta(\theta_0, \overline{\theta}_i)$.

Secondly, we tackle the general case where π is not required to be discrete or compactly supported. In this case, the mapping that we are looking to minimize is

$$f_\pi : \theta \mapsto (1/2) \int_\Theta \rho_\Theta^2(\theta, \nu)\, \pi(\mathrm{d}\nu) \,. \tag{16}$$

The function f_π is well-defined and finite under the following assumption.

MD 5. *There exists* $\theta \in \Theta$ *such that* $\int_\Theta \rho_\Theta^2(\theta, \nu)\pi(\mathrm{d}\nu) < +\infty$.

Note that by the triangle inequality, **MD**5 is equivalent to for any $\theta \in \Theta$ such that $\int_\Theta \rho_\Theta^2(\theta, \nu)\pi(\mathrm{d}\nu) < +\infty$ and therefore f_π is finite. Using the Lebesgue's dominated convergence theorem and Jost (2005, Theorem 5.6.1), we can compute its Riemannian gradient given for any $\theta \in \Theta$ by, $\mathrm{grad}\, f_\pi(\theta) = -\int_\Theta \mathrm{Exp}_\theta^{-1}(\nu)\, \pi(\mathrm{d}\nu)$. Then, f_π satisfies **F**2 with $\lambda_f = 1/2$ and admits a unique minimizer $\theta_\pi^\star$. However, $\mathrm{grad}\, f$ does not satisfy **F**1 in general. More precisely, it fails to be geodesically Lipschitz, see Jost (2005, Theorem 5.6.1). In the Euclidean setting, several modifications of SGD have been suggested to rescale the gradient such as RMSProp, AdaGrad and Adam (Geoffrey, 2014; Duchi et al., 2011; Kingma and Ba, 2017). Inspired by these methods, we consider the stochastic approximation scheme (2) with $\mathsf{S} = \Theta$ and

$$H_\theta(X_{n+1}) = (1/2)\mathrm{Exp}_\theta^{-1}\left(X_{n+1}^{(1)}\right)\{\rho_\Theta^2(\theta, X_{n+1}^{(2)})/2 + 1\}^{-1/2} \,, \tag{17}$$

where $X_{n+1} = (X_{n+1}^{(1)}, X_{n+1}^{(2)})$ and $(X_k^{(1)}, X_k^{(2)})_{k\in\mathbb{N}^*}$ is an i.i.d. sequence of pairs of independent random variables with distribution π. The following result establishes non-asymptotic convergence bounds for the resulting recursion.

Theorem 13. *Assume* ***A***2 *and* ***MD***5. *Let* $\theta_\pi^\star$ *be the Riemannian barycenter of the probability measure* π. *Let* $(\theta_n)_{n\in\mathbb{N}}$ *be given by* (2) *with* $\mathsf{S} = \Theta$ *and* H *defined by* (17). *Then, for any* $n \in \mathbb{N}$,

$$n^{-1} \sum_{k=0}^{n-1} \mathbb{E}\left[D_\Theta^2(\theta_k, \theta_\pi^\star)\right] \leq 4V_1(\theta_0) C_\pi^{1/2} \big/ (\eta n) + 4\eta B_\pi \,,$$

where V_1 *is defined by* (9) *with* $\delta \leftarrow 1$, $\theta^\star \leftarrow \theta_\pi^\star$, $C_\pi = 1 + 2f_\pi(\theta_\pi^\star)$, $B_\pi = (1+\kappa)(f_\pi(\theta_\pi^\star)+1)(f_\pi(\theta_\pi^\star)+2)C_\pi^{-1/2}$ *and* D_Θ^2 *is defined in* (14).

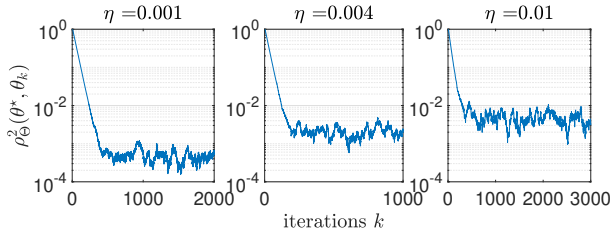

Figure 1: Paths of the algorithm in Proposition 12

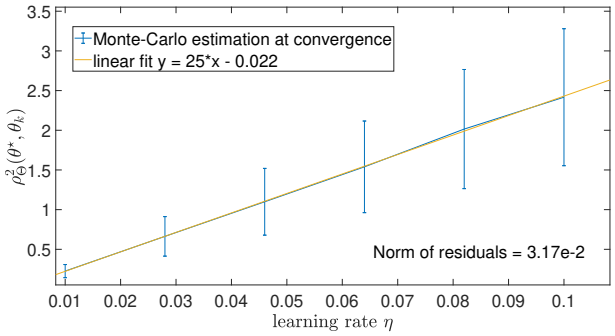

Figure 2

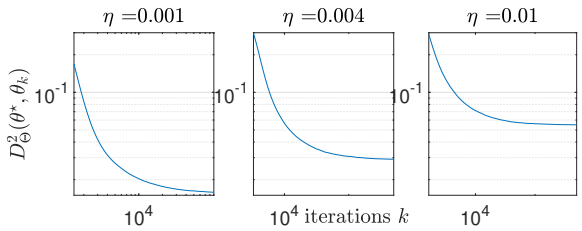

Figure 3: Paths of the algorithm in Theorem 13

5 NUMERICAL EXPERIMENTS

We consider in our experiments the Karcher mean estimation problem on $\Theta = \mathrm{Sym}^+_{50}(\mathbb{R}) \subset \mathbb{R}^{50\times 50}$, the symmetric definite positive matrix manifold (SPD) equipped with its affine-invariant metric, see Pennec et al. (2006). Note that the dimension of Θ is 1275.

We first consider the case where $\pi = (15)^{-1}\sum_{i=1}^{15}\delta_{x_i}$ is a discrete distribution, where $\{x_i\}_{i=1}^{15}$ are random samples from the Wishart distribution $\mathbf{W}(50, \mathrm{Id})$ *i.e.* with 50 degrees of freedom and scale matrix identity. The Karcher mean $\theta^\star_\pi$ associated with π is estimated using the Matrix Means Toolbox (Bini and Iannazzo, 2013).

Figure 1 represents the behavior of the squared distance to the barycenter $\theta^\star_\pi$ for a single path and three step-sizes $\eta \in \{10^{-3}, 4\times 10^{-3}, 10^{-2}\}$. As expected from Proposition 12, two regimes can be observed. At first, the squared-distance to the barycenter exponentially decreases and then the iterates oscillate in a $\mathcal{O}(\eta^{1/2})$-neighborhood of $\theta^\star_\pi$. In addition, the rate of convergence in the exponential decay depends on the step-size.

In Figure 2, we aim at illustrating (7), Theorem 6 and Theorem 7. To this end, 1000 replications of the previous experiment are performed to obtain $\{(\theta_n^{(i)}) : i \in \{1,\dots,1000\}\}$ for $n = \lceil 10/\eta\rceil$ and $\eta \in \{1, 2.8, 4.6, 6.4, 8.2, 10\}\times 10^{-2}$. These samples are used to estimate the mean and the variance of $\rho^2_\Theta(\theta, \theta^\star_\pi)$, for θ following the stationary distribution μ^η. As expected, the mean and the variance are both linear w.r.t. the step-size η, further confirming that the iterates remain in a neighborhood of diameter $\mathcal{O}(\eta^{1/2})$ to the ground truth.

Secondly, we examine the barycenter problem for $\pi = \mathbf{W}(50, \mathrm{Id})$, following the scheme introduced in (17). The estimation of $\theta^\star_\pi$, relative to the new distribution π, is now done with a 100-batch-size version of our methodology, with 10^6 iterations and $\eta = 10^{-4}$.

As a counterpart to Figure 1, in Figure 3 we are interested in the mean values of $(D^2_\Theta(\theta_n, \theta^\star_\pi))_{n\in\mathbb{N}}$ along a single path for three step-sizes $\eta \in \{10^{-3}, 4\times 10^{-3}, 10^{-2}\}$, with respective burn-ins $\{13, 3.3, 1.645\}\times 10^3$. As predicted by Theorem 13, an initial decrease in $\mathcal{O}(n^{-1})$ is followed by a plateau in $\mathcal{O}(\eta)$. We can observe that compared to Figure 1, averaging smoothes oscillations.

Finally, we also perform the experiment corresponding to Figure 2 for the discrete setting to illustrate numerically that the conclusions of (7), Theorem 6 and Theorem 7 still hold. However, due to space constraints and since the conclusions are the same than for Figure 2, the corresponding figure is postponed to the supplement Figure S1.

Acknowledgments

AD and EM acknowledge support of the Lagrange Mathematical and Computing Research Center.

References

Alimisis, F., Orvieto, A., Becigneul, G., and Lucchi, A. (2020). A Continuous-time Perspective for Modeling Acceleration in Riemannian Optimization. volume 108 of *Proceedings of Machine Learning Research*, pages 1297–1307, Online. PMLR.

Arnaudon, M., Dombry, C., Phan, A., and Yang, L. (2012). Stochastic algorithms for computing means of probability measures. *Stochastic Processes and their Applications*, 122(4):1437 – 1455.

Bach, F. (2020). On the effectiveness of richardson extrapolation in machine learning.

Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 451–459.

Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. Translated from the French by Stephen S. Wilson.

Bini, D. A. and Iannazzo, B. (2013). Computing the Karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4):1700 – 1710. 16th ILAS Conference Proceedings, Pisa 2010.

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

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France. Springer.

Bottou, L. and Bousquet, O. (2008). The tradeoffs of large scale learning. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 161–168. NIPS Foundation (http://books.nips.cc).

Boumal, N. and Absil, P.-A. (2011). RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in neural information processing systems*, pages 406–414.

Boumal, N., Voroninski, V., and Bandeira, A. (2016). The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765.

Cappé, O. and Moulines, E. (2009). On-line expectation-maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):593–613.

Dieuleveut, A., Durmus, A., and Bach, F. (2017). Bridging the gap between Constant Step Size Stochastic Gradient Descent and Markov Chains.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61):2121–2159.

Duflo, M. (1997). *Random Iterative Models*. Springer-Verlag, Berlin, Heidelberg, 1st edition.

Durmus, A., Jiménez, P., Moulines, E., Said, S., and Wai, H. (2020). Convergence analysis of Riemannian stochastic approximation schemes. *arXiv preprint arXiv:2005.13284*.

Edelman, A., Arias, T., and Smith, S. (1998). The geometry of Algorithms with Orthogonality Constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

Fort, J.-C. and Pagès, G. (1999). Asymptotic behavior of a Markovian Stochastic Algorithm with Constant Step. *SIAM Journal on Control and Optimization*, 37(5):1456–1482.

Geoffrey, H. (2014). Lecture 6e RMSprop: Divide the gradient by a running average of its recent magnitude.

Han, A. and Gao, J. (2020). Variance reduction for Riemannian non-convex optimization with batch size adaptation.

Hardt, M., Ma, T., and Recht, B. (2019). Gradient descent learns linear dynamical systems.

Horn, R. A. and Johnson, C. R. (1994). *Topics in matrix analysis*. Cambridge university press.

Hosseini, R. and Sra, S. (2019). An alternative to em for gaussian mixture models: Batch and stochastic riemannian optimization. *Mathematical Programming*, pages 1–37.

Iannazzo, B. and Porcelli, M. (2018). The riemannian barzilai–borwein method with nonmonotone line search and the matrix geometric mean computation. *Ima Journal of Numerical Analysis*, 38:495–517.

Ishteva, M., Absil, P.-A., Van Huffel, S., and De Lathauwer, L. (2011). Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1993). Convergence of stochastic iterative dynamic programming algorithms. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 703–710, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jost, J. (2005). *Riemannian Geometry and Geometric Analysis*. Springer Universitat texts. Springer.

Khuzani, M. B. and Li, N. (2017). Stochastic primal-dual method on riemannian manifolds with bounded sectional curvature. *arXiv preprint arXiv:1703.08167*.

Kingma, D. P. and Ba, J. (2017). Adam: A Method for Stochastic Optimization.

Kushner, H. J. and Huang, H. (1981). Asymptotic Properties of Stochastic Approximations with Constant Coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition. Stochastic Modelling and Applied Probability.

Le, H. (2004). Estimation of riemannian barycentres. *Lms Journal of Computation and Mathematics*, 7:193–200.

Lee, J. (2019). *Introduction to Riemannian Manifolds*. Springer International Publishing.

Ma, S., Bassily, R., and Belkin, M. (2018). The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334.

Nedić, A. and Bertsekas, D. (2001). *Convergence Rate of Incremental Subgradient Algorithms*, pages 223–264. Springer US, Boston, MA.

Needell, D., Ward, R., and Srebro, N. (2014). Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025.

Pennec, X., Fillard, P., and Ayache, N. (2006). A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66.

Pflug, G. (1986). Stochastic Minimization with Constant Step-Size: Asymptotic Laws. *SIAM Journal on Control and Optimization*, 24(4):655–666.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855.

Robbins, H. and Monro, S. (1951). A stochastic approxiation method. *The Annals of mathematical Statistics*, 22(3):400–407.

Said, S. and Manton, J. (2019). The riemannian barycentre as a proxy for global optimisation. In *Geometric Science of Information, GSI*, pages 657–664.

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

Sturm, K. T. (2003). Probability Measures on Metric Spaces of Nonpositive Curvature. *Contemporary Mathematics*, 338.

Sun, J., Qu, Q., and Wright, J. (2017). Complete Dictionary Recovery Over the Sphere II: Recovery by Riemannian Trust-Region Method. *IEEE Transactions on Information Theory*, 63(2):885–914.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. (2018). Averaging Stochastic Gradient Descent on Riemannian Manifolds. In *Conference On Learning Theory, COLT*, pages 650–687.

Vaswani, S., Bach, F., and Schmidt, M. (2019). Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204.

Zhang, H., Reddi, S. J., and Sra, S. (2016). Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600.

Zhang, H. and Sra, S. (2016). First-order Methods for Geodesically Convex Optimization. In *Conference on Learning Theory, COLT*, pages 1617–1638.