Averaging Stochastic Gradient Descent on Riemannian Manifolds

Nilesh Tripuraneni   NILESH_TRIPURANENI@BERKELEY.EDU
University of California, Berkeley

Nicolas Flammarion   FLAMMARION@BERKELEY.EDU
University of California, Berkeley

Francis Bach   FRANCIS.BACH@INRIA.FR
INRIA, Ecole Normale Supérieure and PSL Research University

Michael I. Jordan   JORDAN@CS.BERKELEY.EDU
University of California, Berkeley

Abstract

We consider the minimization of a function defined on a Riemannian manifold $\mathcal{M}$ accessible only through unbiased estimates of its gradients. We develop a geometric framework to transform a sequence of slowly converging iterates generated from stochastic gradient descent (SGD) on $\mathcal{M}$ to an averaged iterate sequence with a robust and fast $O(1/n)$ convergence rate. We then present an application of our framework to geodesically-strongly-convex (and possibly Euclidean non-convex) problems. Finally, we demonstrate how these ideas apply to the case of streaming k-PCA, where we show how to accelerate the slow rate of the randomized power method (without requiring knowledge of the eigengap) into a robust algorithm achieving the optimal rate of convergence.

Keywords: Optimization, Riemannian Manifold, Stochastic Approximation, k-PCA.

1. Introduction

We consider stochastic optimization of a (potentially non-convex) function f defined on a Riemannian manifold $\mathcal{M}$, and accessible only through unbiased estimates of its gradients. The framework is broad—encompassing fundamental problems such as principal components analysis (PCA) (Edelman et al., 1998), dictionary learning (Sun et al., 2017), low-rank matrix completion (Boumal and Absil, 2011) and tensor factorization (Ishteva et al., 2011).

The classical setting of stochastic approximation in $\mathbb{R}^d$, first appearing in the work of Robbins and Monro (1951), has been thoroughly explored in both the optimization and machine learning communities. A key step in the development of this theory was the discovery of Polyak-Ruppert averaging—a technique in which the iterates are averaged along the optimization path. Such averaging provably reduces the impact of noise on the problem solution and improves convergence rates in certain important settings (Polyak, 1990; Ruppert, 1988).

By contrast, the general setting of stochastic approximation on Riemannian manifolds has been far less studied. There are many open questions regarding achievable rates and the possibility of accelerating these rates with techniques such as averaging. The problems are twofold: it is not always clear how to extend a gradient-based algorithm to a setting which is missing the global vector-space structure of Euclidean space, and, equally importantly, classical analysis techniques often rely on the Euclidean structure and do not always carry over. In particular, Polyak-Ruppert

averaging relies critically on the Euclidean structure of the configuration space, both in its design and its analysis. We therefore ask: *Can the classical technique of Polyak-Ruppert iterate averaging be adapted to the Riemannian setting?* Moreover, do the advantages of iterate averaging, in terms of rate and robustness, carry over from the Euclidean setting to the Riemannian setting?

Indeed, in the traditional setting of Euclidean stochastic optimization, averaged optimization algorithms not only improve convergence rates, but also have the advantage of adapting to the hardness of the problem (Bach and Moulines, 2011). They provide a single, robust algorithm that achieves optimal rates with and without strong convexity, and achieve the statistically optimal asymptotic variance. In the presence of strong convexity, setting the step size proportional to $\gamma_n = \frac{1}{\mu n}$ is sufficient to achieve the optimal $O(\frac{1}{n})$ rate. However, as highlighted by Nemirovski et al. (2008) and Bach and Moulines (2011), the convergence of such a scheme is highly sensitive to the choice of constant prefactor C in the step size; an improper choice can lead to an arbitrarily slow convergence *rate*. Moreover, since μ is often never known, properly calibrating C is often impossible (unless an explicit regularizer is added to the objective, which adds an extra hyperparameter).

In this paper, we provide a practical iterate-averaging scheme that enhances the robustness and speed of stochastic, gradient-based optimization algorithms, applicable to a wide range of Riemannian optimization problems—including those that are (Euclidean) non-convex. Principally, our framework extends the classical Polyak-Ruppert iterate-averaging scheme (and its inherent benefits) to the Riemannian setting. Moreover, our results hold in the general stochastic approximation setting and do not rely on any finite-sum structure of the objective.

Our main contributions are:

- The development of a geometric framework to transform a sequence of slowly converging iterates, produced from SGD, to an iterate-averaged sequence with a robust, fast $O(\frac{1}{n})$ rate.
- A general formulation of geometric iterate averaging for a class of locally smooth and geodesically-strongly-convex optimization problems.
- An application of our framework to the (non-convex) problem of streaming PCA, where we show how to transform the slow rate of the randomized power method (with no knowledge of the unknown eigengap) into an algorithm that achieves the optimal rate of convergence and which empirically outperforms existing algorithms.

1.1. Related Work

Stochastic Optimization: The literature on (Euclidean) stochastic optimization is vast, having been studied through the lens of machine learning (Bottou, 1998; Shalev-Shwartz et al., 2009), optimization (Nesterov and Vial, 2008), and stochastic approximation (Kushner and Yin, 2003). Polyak-Ruppert averaging first appeared in the works of Polyak (1990) and Ruppert (1988); Polyak and Juditsky (1992) then provided asymptotic normality results for the distribution of the averaged iterate sequence. Bach and Moulines (2011) later generalized these results, providing non-asymptotic guarantees for the convergence of the averaged iterates. An important contribution of Bach and Moulines (2011) was to present a unified analysis showing that averaging coupled with sufficiently slow step-sizes achieves optimal convergence in all settings (with and without strong convexity).

Riemannian Optimization: Riemannian optimization has not been explored in the machine learning community until relatively recently. Udriste (1994) and Absil et al. (2009) provide comprehensive background on the topic. Most existing work has primarily focused on providing asymptotic convergence guarantees for non-stochastic algorithms (see, e.g., Absil et al., 2007; Ring and Wirth, 2012, who analyze the convergence of Riemannian trust-region and Riemannian L-BFGS methods).

Bonnabel (2013) provided the first asymptotic convergence proof of stochastic gradient descent (SGD) on Riemannian manifolds while highlighting diverse applications of the Riemannian framework to problems such as PCA. The first global complexity results for first-order Riemannian optimization, utilizing the notion of *functional g-convexity*, were obtained in the foundational work of Zhang and Sra (2016). The finite-sum, stochastic setting has been further investigated by Zhang et al. (2016) and Sato et al. (2017), who developed Riemannian SVRG methods. However, the potential utility of Polyak-Ruppert averaging in the Riemannian setting has been unexplored.

2. Results

We consider the optimization of a function f over a compact, connected subset $\mathcal{X} \subset \mathcal{M}$,

$$\min_{x \in \mathcal{X} \subset \mathcal{M}} f(x),$$

with access to a (noisy) first-order oracle $\{\nabla f_n(x)\}_{n \geq 1}$. Given a sequence of iterates $\{x_n\}_{n \geq 0}$ in $\mathcal{M}$ produced from the first-order optimization of f,

$$x_n = R_{x_{n-1}}\left(-\gamma_n \nabla f_n(x_{n-1})\right), \tag{1}$$

that are converging to a *strict* local minimum of f, denoted by $x_\star$, we consider (and analyze the convergence of) a streaming average of iterates:

$$\tilde{x}_n = R_{\tilde{x}_{n-1}}\left(\frac{1}{n} R_{\tilde{x}_{n-1}}^{-1}(x_n)\right). \tag{2}$$

Here we use R_x to denote a retraction mapping (defined formally in Section 3), which provides a natural means of moving along a vector (such as the gradient) while restricting movement to the manifold. As an example, when $\mathcal{M} = \mathbb{R}^d$ we can take R_x as vector addition by x. In this setting, Eq. (1) reduces to the standard gradient update $x_n = x_{n-1} - \gamma_n \nabla f_n(x_{n-1})$ and Eq. (2) reduces to the ordinary average $\tilde{x}_n = \tilde{x}_{n-1} + \frac{1}{n}(x_n - \tilde{x}_{n-1})$. In the update in Eq. (1), we will always consider step-size sequences of the form $\gamma_n = \frac{C}{n^\alpha}$ for $C > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$, which satisfy the usual stochastic approximation step-size rules $\sum_{i=1}^{\infty} \gamma_i = \infty$ and $\sum_{i=1}^{\infty} \gamma_i^2 < \infty$ (see, e.g., Benveniste et al., 1990).

Intuitively, our main result states that if the iterates x_n converge to $x_\star$ at a slow $O(\gamma_n)$ rate, their streaming Riemannian average will converge to $x_\star$ at the the optimal $O(\frac{1}{n})$ rate. This result requires several technical assumptions, which are standard generalizations of those appearing in the Riemannian optimization and stochastic approximation literatures (detailed in Section 4). The critical assumption we make is that all iterates remain bounded in $\mathcal{X}$—where the manifold behaves well and the algorithm is well-defined (Assumption 2). The notion of slow convergence to an optimum is formalized in the following assumption:

Assumption 1 *If $\Delta_n = R_{x_\star}^{-1}(x_n)$ for a sequence of iterates evolving in Eq. (1), then*

$$\mathbb{E}[\|\Delta_n\|^2] = O(\gamma_n).$$

Assumption 1 can be verified in a variety of optimization problems, and we provide such examples in Section 6. As $x_\star$ is unknown, Δ_n is not computable but is primarily a tool for our analysis. Importantly, Δ_n is a tangent vector in $T_{x_\star}\mathcal{M}$. Note also that the norm $\|\Delta_n\|$ is locally equivalent to the geodesic distance $d(x_n, x_\star)$ on $\mathcal{M}$ (see Section 3).

We use Σ to denote the covariance of the noisy gradients at the optima $x_\star$. Formally, our main convergence result regarding Polyak-Ruppert averaging in the manifold setting is as follows (where Assumptions 2 through 6 will be presented later):

Theorem 1 *Let Assumptions 1, 2, 3, 4, 5, and 6 hold for the iterates evolving according to Eq. (1) and Eq. (2). Then $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$ satisfies:*

$$\sqrt{n}\tilde{\Delta}_n \xrightarrow{D} \mathcal{N}(0, \nabla^2 f(x_\star)^{-1}\Sigma\nabla^2 f(x_\star)^{-1}).$$

If we additionally assume a bound on the fourth moment of the iterates—of the form $\mathbb{E}[\|\Delta_n\|^4] = O(\gamma_n^2)$—then a non-asymptotic result holds:

$$\mathbb{E}[\|\tilde{\Delta}_n\|^2] \leq \frac{1}{n}\operatorname{tr}\big[\nabla^2 f(x_\star)^{-1}\Sigma\nabla^2 f(x_\star)^{-1}\big] + O(n^{-2\alpha}) + O(n^{\alpha-2}).$$

We make several remarks regarding this theorem:

- The asymptotic result in Theorem 1 is a generalization of the classical asymptotic result of Polyak and Juditsky (1992). In particular, the leading term has variance $O(\frac{1}{n})$ *independently* of the step-size choice γ_n. In the presence of strong convexity, SGD can achieve the $O(\frac{1}{n})$ rate with a carefully chosen step size, $\gamma_n = \frac{C}{\mu n}$ (for $C = 1$). However, the result is fragile: too small a value of C can lead to an arbitrarily slow convergence rate, while too large a C can lead to an "exploding," non-convergent sequence (Nemirovski et al., 2008). In practice determining μ is often as difficult as the problem itself.

- Theorem 1 implies that the distance (measured in $T_{x_\star}\mathcal{M}$) of the streaming average $\tilde{x}_n$ to the optimum, asymptotically saturates the Cramer-Rao bound on the manifold $\mathcal{M}$ (Smith, 2005; Boumal, 2013)—asymptotically achieving the statistically optimal covariance[1]. SGD, even with the carefully calibrated step-size choice of $\gamma_n = \frac{1}{\mu n}$, does not achieve this optimal asymptotic variance (Nevelson and Hasminski, 1973).

We exhibit two applications of this general result in Section 6. Next, we introduce the relevant background and assumptions that are necessary to prove our theorem.

3. Preliminaries

We recall some important concepts from Riemannian geometry. Do Carmo (2016) provides more a thorough review, with Absil et al. (2009) providing a perspective particularly relevant for Riemannian optimization.

As a base space we consider a Riemannian manifold $(\mathcal{M}, \mathfrak{g})$–a smooth manifold equipped with a Riemannian metric $\mathfrak{g}$ containing a compact, connected subset $\mathcal{X}$. At all $x \in \mathcal{M}$, the metric $\mathfrak{g}$ induces a natural inner product on the tangent space $T_x\mathcal{M}$, denoted by $\langle\cdot,\cdot\rangle$–this inner product induces a norm on each tangent space denoted by $\|\cdot\|$. The metric $\mathfrak{g}$ also provides a means of measuring the length of a parametrized curve from $\mathbb{R}$ to the manifold; a *geodesic* is a constant speed curve $\gamma\colon[0, 1] \to \mathcal{M}$ that is locally distance-minimizing with respect to the distance d induced by $\mathfrak{g}$.

When considering functions $f : \mathcal{M} \to \mathbb{R}$ we will use $\nabla f(x) \in T_x M$ to denote the *Riemannian gradient* of f at $x \in \mathcal{M}$, and $\nabla^2 f(x) : T_x M \to T_x M$, the *Riemannian Hessian* of f at x. When considering functions between manifolds $F : \mathcal{M} \to \mathcal{M}$, we will use $DF(x) : T_x\mathcal{M} \to T_{F(x)}\mathcal{M}$ to denote the *differential* of the mapping at x (its linearization) (see Absil et al., 2009, for more formal definitions of these objects). The *exponential map* $\operatorname{Exp}_x(v) : T_x\mathcal{M} \to \mathcal{M}$ maps $v \in T_x\mathcal{M}$

---

1. Note the estimator $\tilde{\Delta}_n$ is only asymptotically unbiased, and hence the Cramer-Rao bound is only meaningful in the asymptotic limit. However, this result can also be understood as saturating the Hàjek-Le Cam local asymptotic minimax lower bound (Van der Vaart, 1998, Ch. 8).
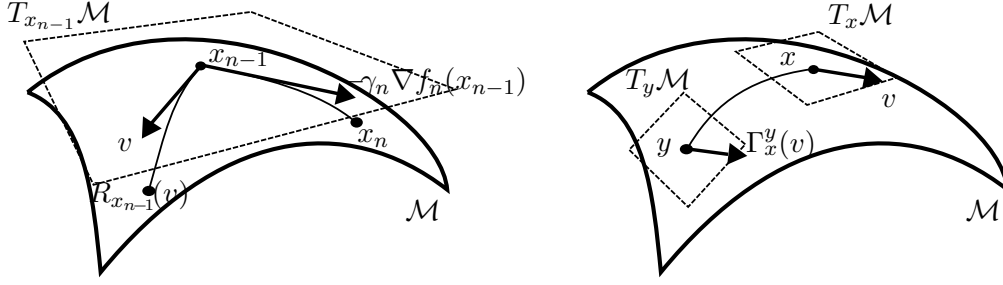
Figure 1: Left: In the tangent space at iterate x_{n-1}, a retraction along an arbitrary vector v generating a curve pointing in the "direction" of v, and a retraction along the gradient update to x_n. Right: The parallel transport of a different v along the same path.

to $y \in \mathcal{M}$ such that there is a geodesic with $\gamma(0) = x$, $\gamma(1) = y$, and $\frac{d}{dt}\gamma(0) = v$; although it may not be defined on the whole tangent space. If there is a unique geodesic connecting $x, y \in \mathcal{X}$, the exponential map will have a well-defined inverse $\mathrm{Exp}_x^{-1}(y) : \mathcal{M} \to T_x\mathcal{M}$, such that the length of the connecting geodesic is $d(x, y) = \|\mathrm{Exp}_x^{-1}(y)\|$. We also use $R_x : T_x\mathcal{M} \to \mathcal{M}$ and $R_x^{-1} : \mathcal{M} \to T_x\mathcal{M}$ to denote a *retraction mapping* and its inverse (when well defined), which is an approximation to the exponential map (and its inverse). R_x is often computationally cheaper to compute then the entire exponential map Exp_x. Formally, the map R_x is defined as a first-order retraction if $R_x(0) = x$ and $DR_x(0) = id_{T_x\mathcal{M}}$—so locally $R_x(\xi)$ must move in the "direction" of ξ. The map R_x is a second-order retraction if R_x also satisfies $\frac{D^2}{dt^2}R_x(t\xi)|_0 = 0$ for all $\xi \in T_x\mathcal{M}$, where $\frac{D^2}{dt^2}\gamma = \frac{D}{dt}\dot{\gamma}$ denotes the acceleration vector field (Absil et al., 2009, Sec. 5.4). This condition ensures R_x satisfies a "zero-acceleration" initial condition. Note that locally, for x close to y, the retraction satisfies $\|R_x^{-1}(y)\| = d(x, y) + o(d(x, y))$.

If we consider the example where the manifold is a sphere (i.e., $\mathcal{M} = S^{d-1}$ with the round metric $\mathfrak{g}$), the exponential map along a vector will generate a curve that is a great circle on the underlying sphere. A nontrivial example of a retraction R on the sphere can be defined by first moving along the tangent vector in the embedded Euclidean space $\mathbb{R}^d$, and then projecting this point to the closest point on the sphere.

We further define the *parallel translation* $\Gamma_x^y : T_x\mathcal{M} \to T_y\mathcal{M}$ as the map transporting a vector $v \in T_x\mathcal{M}$ to $\Gamma_x^y v$, along a path $R_x(\xi)$ connecting x to $y = R_x(\xi)$, such that the vector stays "constant" by satisfying a zero-acceleration condition. This is illustrated in Figure 1. The map Γ_x^y is an isometry. We also consider, a different vector transport map $\Lambda_x^y : T_x\mathcal{M} \to T_y\mathcal{M}$ which is the differential $DR_x(R_x^{-1}(y))$ of the retraction R (Absil et al., 2009, Sec. 8.1).

Following Huang et al. (2015), we will call a function f on $\mathcal{X}$ *retraction convex* on $\mathcal{X}$ (with respect to R) if for all $x \in \mathcal{X}$ and all $\eta \in T_x\mathcal{M}$ satisfying $\|\eta\| = 1$, $t \mapsto f(R_x(t\eta))$ is convex for all t such that $R_x(t\eta) \in \mathcal{X}$; similarly f is *retraction strongly convex* on $\mathcal{X}$ if $t \mapsto f(R_x(t\eta))$ is strongly convex under the same conditions. If R_x is the exponential map, this reduces to the definition of geodesic convexity (see the work of Zhang and Sra, 2016, for further details).

4. Assumptions

We introduce several assumptions on the manifold $\mathcal{M}$, function f, and the noise process $\{\nabla f_n\}_{n\geq 1}$ that will be relevant throughout the paper.

4.1. Assumptions on $\mathcal{M}$

First, we assume the iterates of the algorithm in Eq. (1) and Eq. (2) remain in $\mathcal{X}$ where the manifold "behaves well." Formally,

Assumption 2 *For a sequence of iterates $\{x_n\}_{n\geq 0}$ defined in Eq. (1), there exists a compact, connected subset $\mathcal{X}$ such that $x_n \in \mathcal{X}$ for all $n \geq 0$, and $x_\star \in \mathcal{X}$. Furthermore, $\mathcal{X}$ is totally retractive (with respect to the retraction R) and the function $x \mapsto \|\frac{1}{2} R_y^{-1}(x)\|^2$ is retraction strongly convex on $\mathcal{X}$ for all $y \in \mathcal{X}$. Also, R is a second-order retraction at $x_\star$.*

Assumption 2 is restrictive, but standard in prior work on stochastic approximation on manifolds (e.g., Bonnabel, 2013; Zhang et al., 2016; Sato et al., 2017). As further detailed by Huang et al. (2015), a totally retractive neighborhood $\mathcal{X}$ is such that for all $x \in \mathcal{X}$ there exists $r > 0$ such that $\mathcal{X} \subset R_x(\mathbb{B}_r(0))$ where R_x is a diffeomorphism on $\mathbb{B}_r(0)$. A totally retractive neighborhood is analogous to the concept of a totally normal neighborhood (see, e.g., Do Carmo, 2016, Chap. 3, Sec. 3). Principally, Assumption 2 ensures that the retraction map (and its respective inverse) are well-defined when applied to the iterates of our algorithm.

If $\mathcal{M}$ is a Hadamard manifold, the exponential map (and its inverse) is defined everywhere on $\mathcal{M}$, although this globally may not be true for a retraction R. Similarly, if $\mathcal{M}$ is a compact manifold the first statement of Assumption 2 is always satisfied. Moreover, in the case of the exponential map, $x \mapsto \frac{1}{2}\|\mathrm{Exp}_y^{-1}(x)\|^2$ is strongly convex in a ball around y whose radius depends on the curvature, as explained by Afsari (2011) and Sakai (1996, Ch. IV, Sec. 2 Lemma 2.9). For our present purpose, we also assume the retraction R agrees with the Riemannian exponential map up to second order near $x_\star$. This assumption, that R is a second-order retraction, is fairly general and is satisfied by projection-like retraction maps on matrix manifolds (see Absil and Malick, 2012).

4.2. Assumptions on f

We now introduce some regularity assumptions on the function f ensuring sufficient differentiability and strong convexity at $x_\star$:

Assumption 3 *The function f is twice-continuously differentiable on $\mathcal{X}$. Further the Hessian of the function f at $x_\star$, $\nabla^2 f(x_\star)$, satisfies, for all $v \in T_{x_\star}\mathcal{M}$ and $\mu > 0$,*

$$\langle v, \nabla^2 f(x_\star)v \rangle \geq \mu \|v\|^2 > 0.$$

Continuity of the Hessian also ensures local retraction strong convexity in a neighborhood of $x_\star$ (Absil et al., 2009, Prop. 5.5.6). Moreover, since the function f is twice-continuously differentiable on $\mathcal{X}$ its Hessian is Lipschitz on this compact set. We formalize this as follows:

Assumption 4 *There exists $M > 0$ such that the Hessian of the function f, $\nabla^2 f$, is M-Lipschitz at $x_\star$. That is, for all $y \in \mathcal{X}$ and $v \in T_y\mathcal{M}$,*

$$\|\Gamma_y^{x_\star} \circ \nabla^2 f(y) \circ \Gamma_{x_\star}^y - \nabla^2 f(x_\star)\|_{op} \leq M \|R_{x_\star}^{-1}(y)\|.$$

Note that $\|R_{x_\star}^{-1}(y)\|$ is not necessarily symmetric under the exchange of $x_\star$ and y. This term could also be replaced with $d(x_\star, y)$, since these expressions will be locally equivalent in a neighborhood of $x_\star$, but would come at the cost of a less transparent analysis.

4.3. Assumptions on the noise

We state several assumptions on the noise process that will be relevant throughout our discussion. Let $(\mathcal{F}_n)_{n \geq 0}$ be an increasing sequence of sigma-fields. We will assume access to a sequence $\{\nabla f_n\}_{n \geq 1}$ of noisy estimates of the true gradient ∇f of the function f,

Assumption 5 *The sequence of (random) vector fields $\{\nabla f_n\}_{n \geq 1} : \mathcal{M} \to T\mathcal{M}$ is $\mathcal{F}_n$-measurable, square-integrable and unbiased:*

$$\forall x \in \mathcal{X}, \ \forall n \geq 1, \ \mathbb{E}[\nabla f_n(x)|\mathcal{F}_{n-1}] = \nabla f(x).$$

This general framework subsumes two situations of interest.
- Statistical Learning (on Manifolds): minimizing a loss function $\ell : \mathcal{M} \times \mathcal{Z} \to \mathbb{R}$ over $x \in \mathcal{X}$, given a sequence of i.i.d. observations in $\mathcal{Z}$, with access only to noisy, unbiased estimates of the gradient $\nabla f_n = \nabla \ell(\cdot, z_n)$ (Aswani et al., 2011).
- Stochastic Approximation (on Manifolds): minimizing a function $f(x)$ over $x \in \mathcal{X}$, with access only to the (random) vector field $\nabla f(x) + \epsilon_n(x)$ at each iteration. Here the gradient vector field is perturbed by a square-integrable martingale-difference sequence (for all $x \in \mathcal{M}, \mathbb{E}[\epsilon_n(x)|\mathcal{F}_{n-1}] = 0$) (Bonnabel, 2013).

Lastly, we will assume the vector fields $\{\nabla f_n\}_{n \geq 1}$ are individually Lipschitz and have bounded covariance at the optimum $x_\star$:

Assumption 6 *There exists $L > 0$ such that for all $x \in \mathcal{X}$ and $n \geq 1$, the vector field ∇f_n satisfies*

$$\mathbb{E}[\|\Gamma_x^{x_\star} \nabla f_n(x) - \nabla f_n(x_\star)\|^2 |\mathcal{F}_{n-1}] \leq L^2 \|R_{x_\star}^{-1}(x)\|^2,$$

there exists $\tau > 0$ such that $\mathbb{E}[\|\nabla f_n(x)\|^4 |\mathcal{F}_{n-1}] \leq \tau^4$ for all $x \in \mathcal{X}$, and a symmetric positive-definite matrix Σ such that,

$$\mathbb{E}[\nabla f_n(x_\star) \otimes \nabla f_n(x_\star)|\mathcal{F}_{n-1}] = \Sigma \ a.s.$$

These are natural generalizations of standard assumptions in the optimization literature (Fabian, 1968) to the setting of Riemannian manifolds[2]. Note that the assumption, $\mathbb{E}[\nabla f_n(x_\star) \otimes \nabla f_n(x_\star)|\mathcal{F}_{n-1}] = \Sigma$ a.s. could be slightly relaxed (as detailed in Appendix C.2), but allows us to state our main result more cleanly.

5. Proof Sketch

We provide an overview of the arguments that comprise the proof of Theorem 1 (full details are deferred to Appendix C). We highlight three key steps. First, since we assume the iterates x_n produced from SGD converge to within $\sim O(\sqrt{\gamma_n})$ of $x_\star$, we can perform a Taylor expansion of the recursion in Eq. (1), to relate the points x_n on the manifold $\mathcal{M}$ to vectors Δ_n in the tangent space $T_{x_\star}\mathcal{M}$. This generates a (perturbed) linear recursion governing the evolution of the vectors $\Delta_n \in T_{x_\star}\mathcal{M}$. Recall that as $x_\star$ is unknown, Δ_n is not accessible, but is primarily a tool for our analysis. Second, we can show a fast $O(\frac{1}{n})$ convergence rate for the averaged vectors $\bar{\Delta}_n \in T_{x_\star}\mathcal{M}$, using techniques from the Euclidean setting. Finally, we once again use a local expansion of Eq. (2) to connect the averaged tangent vectors $\bar{\Delta}_n$ to the streaming, Riemannian average $\tilde{\Delta}_n$—transferring the fast rate for the inaccessible vector $\bar{\Delta}_n$ to the computable point $\tilde{x}_n$. Throughout our analysis we extensively use Assumption 2, which restricts the iterates x_n to the subset $\mathcal{X}$.

---

2. Assuming bounded gradients does not contradict Assumption 3, since we are constrained to the compact set $\mathcal{X}$.

5.1. From $\mathcal{M}$ to $T_{x_\star}\mathcal{M}$

We begin by linearizing the progress of the SGD iterates x_n in the tangent space of $x_\star$ by considering the evolution of $\Delta_n = R_{x_\star}^{-1}(x_n)$.

- First, as the Δ_n are all defined in the vector space $T_{x_\star}\mathcal{M}$, Taylor's theorem applied to $R_{x_\star}^{-1} \circ R_{x_n} : T_{x_n}\mathcal{M} \to T_{x_\star}\mathcal{M}$ along with Eq. (1) allows us to conclude that

$$\Delta_{n+1} = \Delta_n - \gamma_{n+1}[\Lambda_{x_\star}^{x_n}]^{-1}(\nabla f_{n+1}(x_n)) + O(\gamma_{n+1}^2).$$

  See Lemma 4 for more details.

- Second, we use the manifold version of Taylor's theorem and appropriate Lipschitz conditions on the gradient to further expand the gradient term $\Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n)$ as

$$\Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n) = \nabla^2 f(x_\star)\Delta_n + \nabla f_{n+1}(x_\star) + \xi_{n+1} + O(\|\Delta_n\|^2),$$

  where the noise term is controlled as $\mathbb{E}[\,\xi_{n+1}|\mathcal{F}_n] = 0$, and $\mathbb{E}[\|\xi_{n+1}\|^2|\mathcal{F}_n] = O(\|\Delta_n\|^2)$. See Lemma 5 for more details.

- Finally, we argue that the operator $[\Lambda_{x_\star}^{x_n}]^{-1}\Gamma_{x_\star}^{x_n} : T_{x_\star}\mathcal{M} \to T_{x_\star}\mathcal{M}$ is a local isometry up to second-order terms: $[\Lambda_{x_\star}^{x_n}]^{-1}\Gamma_{x_\star}^{x_n} = I + O(\|\Delta_n\|^2)$, which crucially rests on the fact R is a second-order retraction. See Lemma 6 for more details.

- Assembling the aforementioned lemmas allows us to derive a (perturbed) linear recursion, governing the tangent vectors $\{\Delta_n\}_{n\geq 0}$ as

$$\Delta_{n+1} = \Delta_n - \gamma_{n+1}\nabla^2 f(x_\star)\Delta_n - \gamma_{n+1}\nabla f_{n+1}(x_\star) - \gamma_{n+1}\xi_{n+1} + O(\|\Delta_n\|^2\gamma_n + \gamma_n^2). \quad (3)$$

  See Theorem 7 for more details.

5.2. Averaging in $T_{x_\star}\mathcal{M}$

Our next step is to prove both asymptotic and non-asymptotic convergence rates for a general, perturbed linear recursion (resembling Eq. (3)) of the form,

$$\Delta_n = \Delta_{n-1} - \gamma_n\nabla^2 f(x_\star)\Delta_{n-1} + \gamma_n(\varepsilon_n + \xi_n + e_n), \quad (4)$$

under appropriate assumptions on the error $\{e_n\}_{n\geq 0}$ and noise $\{\varepsilon_n\}_{n\geq 0}$, $\{\xi_n\}_{n\geq 0}$ sequences detailed in Appendix C.2. Under these assumptions we can derive an asymptotic rate for the average, $\bar{\Delta}_n = \frac{1}{n}\sum_{i=1}^n \Delta_i$, under a first-moment condition on e_n:

$$\sqrt{n}\bar{\Delta}_n \xrightarrow{D} \mathcal{N}(0, \nabla^2 f(x_\star)^{-1}\Sigma\nabla^2 f(x_\star)^{-1}),$$

and, under a slightly stronger second-moment condition on e_n we have:

$$\mathbb{E}[\|\bar{\Delta}_n\|^2] \leq \frac{1}{n}\operatorname{tr}\big[\nabla^2 f(x_\star)^{-1}\Sigma\nabla^2 f(x_\star)^{-1}\big] + O(n^{-2\alpha}) + O(n^{\alpha-2}),$$

where Σ denotes the asymptotic covariance of the noise ε_n. The proof techniques are similar to those of Polyak and Juditsky (1992) and Bach and Moulines (2011) so we do not detail them here. See Theorems 8 and 9 for more details. Note that $\bar{\Delta}_n$ is *not* computable, but does have an interesting interpretation as an upper bound on the Riemannian center-of-mass, $K_n = \arg\min_{x\in\mathcal{M}}\sum_{i=1}^n \big\|R_x^{-1}(x_i)\big\|^2$, of a set of iterates $\{x_n\}_{n\geq 0}$ in $\mathcal{M}$ (see Section C.2.3 and Afsari, 2011, for more details).

5.3. From $T_{x_\star}\mathcal{M}$ back to $\mathcal{M}$

Using the previous arguments, we can conclude that the averaged vector $\bar{\Delta}_n$ obeys both asymptotic and non-asymptotic Polyak-Ruppert-type results. However, $\tilde{\Delta}_n$ is *not* computable. Rather, $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$ corresponds to the computable, Riemannian streaming average $\tilde{x}_n$ defined in Eq. (2). In order to conclude our result, we argue that $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$ and $\bar{\Delta}_n$ are close up to $O(\gamma_n)$ terms. The argument proceeds in two steps:

- Using the fact that $x \to \left\| R_{x_\star}^{-1}(x) \right\|^2$ is retraction convex we can conclude that $\mathbb{E}[\|\Delta_n\|^2] = O(\gamma_n)$ implies that $\mathbb{E}[\|\tilde{\Delta}_n\|^2] = O(\gamma_n)$ as well. See Lemma 11 for more details.

- Then, we can locally expand Eq. (2) to find that,

$$\tilde{\Delta}_{n+1} = \tilde{\Delta}_n + \frac{1}{n+1}(\Delta_{n+1} - \tilde{\Delta}_n) + \tilde{e}_n,$$

  where $\mathbb{E}[\|\tilde{e}_n\|] = O(\frac{\gamma_n}{n+1})$. Rearranging and summing this recursion shows that $\tilde{\Delta}_n = \bar{\Delta}_n + e_n$ for $\mathbb{E}[\|e_n\|] = O(\gamma_n)$, showing these terms are close. See Lemma 12 for details.

6. Applications

We now introduce two applications of our Riemannian iterate-averaging framework.

6.1. Application to Geodesically-Strongly-Convex Functions

In this section, we assume that f is globally geodesically convex over $\mathcal{X}$ and take $R \equiv \mathrm{Exp}$, which allows the derivation of global convergence rates. This function class encapsulates interesting problems such as the matrix Karcher mean problem (Bini and Iannazzo, 2013) which is non-convex in Euclidean space but geodesically strongly convex with an appropriate choice of metric on $\mathcal{M}$.

Zhang and Sra (2016) show for geodesically-convex f, that averaged SGD with step size $\gamma_n \propto \frac{1}{\sqrt{n}}$ achieves the slow $O\left(\frac{1}{\sqrt{n}}\right)$ convergence rate. If in addition, f is geodesically strongly convex on $\mathcal{X}$, they obtain the fast $O(\frac{1}{n})$ rate. However, their result is not *algorithmically robust*, requiring a delicate specification of the step size $\gamma_n \propto \frac{1}{\mu n}$, which is often practically impossible due to a lack of knowledge of μ. Assuming smoothness of f, our iterate-averaging framework provides a means of obtaining a robust and global convergence rate. First, we make the following assumption:

Assumption 7 *The function f is μ-geodesically-strongly-convex on $\mathcal{X}$, for $\mu > 0$, and the set $\mathcal{X}$ is geodesically convex.*

Then using our main result in Theorem 1, with $\gamma_n \propto \frac{1}{n^\alpha}$, we have:

Proposition 2 *Let Assumptions 2, 4, 5, 6, and 7 hold for the iterates evolving in Eq. (1) and Eq. (2) and take the retraction R to be the exponential map Exp. Then,*

$$\mathbb{E}[\|\tilde{\Delta}_n\|^2] \leq \frac{1}{n}\,\mathrm{tr}\big[\nabla^2 f(x_\star)^{-1}\Sigma\nabla^2 f(x_\star)^{-1}\big] + O(n^{-2\alpha}) + O(n^{\alpha-2}).$$

We make several remarks.

- In order to show the result, we first derive a slow rate of convergence for SGD, by arguing that $\mathbb{E}[d^2(x_n, x_\star)] \leq \frac{2C\zeta v^2}{\mu n^\alpha} + O(\exp\big(-c\mu n^{1-\alpha}\big))$ and $\mathbb{E}[d^4(x_n, x_\star)] \leq \frac{4C(3+\zeta)\zeta v^4}{\mu n^{2\alpha}} + O(\exp\big(-c\mu n^{1-\alpha}\big))$ where $c, C > 0$ and $\zeta > 0$ is a geometry-dependent constant (see Proposition 15 for more details). The result follows by combining these results and Theorem 1.

- As in Theorem 1 we also obtain convergence in law and the statistically optimal covariance.

9

- Importantly, taking the step size to be $\gamma_n \propto \frac{1}{\sqrt{n}}$ provides a single, robust algorithm achieving both the slow $O\left(\frac{1}{\sqrt{n}}\right)$ rate in the absence of strong convexity (by Zhang and Sra (2016)) and the fast $O(\frac{1}{n})$ rate in the presence of strong convexity. Thus (Riemannian) averaged SGD automatically adapts to the strong-convexity in the problem without any prior knowledge of its existence (i.e., the value of μ).

6.2. Streaming Principal Component Analysis (PCA)

The framework of geometric optimization is far-reaching, containing even (Euclidean) non-convex problems such as PCA. Recall the classical formulation of streaming k-PCA: we are given a stream of i.i.d. symmetric positive-definite random matrices $H_n \in \mathbb{R}^{d \times d}$ such that $\mathbb{E} H_n = H$, with eigenvalues $\{\lambda_i\}_{1 \leq i \leq d}$ sorted in decreasing order, and hope to approximate the subspace of the top k eigenvectors, $\{v_i\}_{1 \leq i \leq k}$. Sharp convergence rates for streaming PCA (with $k = 1$) were first obtained by Jain et al. (2016) and Shamir (2016a) using the randomized power method. Shamir (2016b) and Allen-Zhu and Li (2017) later extended this work to the more general streaming k-PCA setting. These results are powerful—particularly because they provide *global* convergence guarantees.

For streaming k-PCA, a similar dichotomy to the convex setting exists: in the absence of an eigengap ($\lambda_k = \lambda_{k+1}$) one can only attain the slow $O\left(\frac{1}{\sqrt{n}}\right)$ rate, while the fast $O(\frac{1}{n})$ rate is achievable when the eigengap is positive ($\lambda_k > \lambda_{k+1}$). However, as before, a practically burdensome requirement of these fast $O(\frac{1}{n})$, global-convergence guarantees is that the step sizes of their corresponding algorithms depend explicitly on the unknown eigengap[3] of the matrix H.

By viewing the k-PCA problem as minimizing the Rayleigh quotient, $f(X) = -\frac{1}{2} \operatorname{tr}\left[X^\top H X\right]$, over the Grassmann manifold, we show how to apply the Riemannian iterate-averaging framework developed here to derive, for streaming k-PCA, a fast, *robust* algorithm,

$$X_n = R_{X_{n-1}}\left(\gamma_n H_n X_{n-1}\right) \quad \text{and} \quad \tilde{X}_n = R_{\tilde{X}_{n-1}}\left(\frac{1}{n} X_n X_n^\top \tilde{X}_{n-1}\right). \tag{5}$$

Recall that the Grassmann manifold $\mathcal{G}_{d,k}$ is the set of the k-dimensional subspaces of a d-dimensional Euclidean space which we equip with the projection-like, second-order retraction map $R_X(V) = (X + V)[(X + V)^\top(X + V)]^{-1/2}$. Observe that the randomized power method update (Oja and Karhunen, 1985), $X_n = R_{X_{n-1}}\left(\gamma_n H_n X_{n-1}\right)$, in Eq. (5), is almost identical to the Riemannian SGD update, $X_n = R_{X_{n-1}}\left(\gamma_n(I - X_{n-1}X_{n-1}^\top)H_n X_{n-1}\right)$, in Eq. (1). The principal difference between both is that in the randomized power method, the Euclidean gradient is used instead of the Riemannian gradient. Similarly, the average $\tilde{X}_n = R_{\tilde{X}_{n-1}}\left(\frac{1}{n} X_n X_n^\top \tilde{X}_{n-1}\right)$, considered in Eq. (5), closely resembles the (Riemannian) streaming average in Eq. (2) (see Appendix E.2).

In fact we can argue that the randomized power method, Riemannian SGD, and the classic Oja iteration (the linearization of the randomized power method in γ_n) are equivalent up to $O(\gamma_n^2)$ corrections (see Lemma 17). The average $\tilde{X}_n = R_{\tilde{X}_{n-1}}\left(\frac{1}{n} X_n X_n^\top \tilde{X}_{n-1}\right)$ also admits the same linearization as the Riemannian streaming average up to $O(\gamma_n)$ corrections (see Lemma 19).

Using results from Shamir (2016b) and Allen-Zhu and Li (2017) we can then argue that the randomized power method iterates satisfy a slow rate of convergence under suitable conditions on their initialization. Hence, the present framework is applicable and we can use geometric iterate averaging to obtain a local, robust, accelerated convergence rate. In the following, we will use $\{e_j\}_{1 \leq j \leq k}$ to denote the standard basis vectors in $\mathbb{R}^k$.

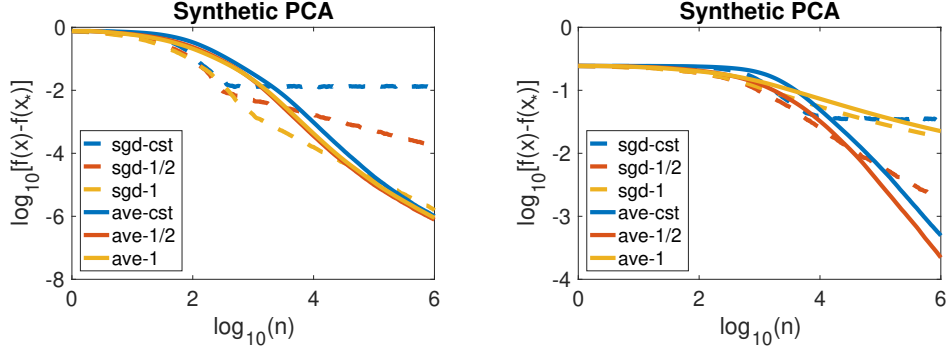3. In this example, the eigengap is analogous to the strong-convexity parameter μ.

Figure 2: Streaming PCA. Left: Well-conditioned problem. Right: Poorly-conditioned problem.

Theorem 3 *Let Assumption 2 hold for the set $\mathcal{X} = \{X : \|X_\star^\top X\|_F^2 \geq k - \eta\}$, for some constant $0 < \eta < \frac{1}{4}$, where $X_\star$ minimizes $f(X)$ over the k-Grassmann manifold. Denote, $\tilde{H}_n = H^{-1/2} H_n H^{-1/2}$, and the 4th-order tensor $C_{ii'jj'} = \mathbb{E}[(v_i^\top \tilde{H}_n v_j)(v_{i'}^\top \tilde{H}_n v_{j'})]$. Further assume that $\|H_n\|_2 \leq 1$ a.s., and that $\lambda_k > \lambda_{k+1}$. Then if X_n and $\tilde{X}_n$ evolve according to Eq. (5), there exists a positive-definite matrix C, such that $\tilde{\Delta}_n = R_{X_\star}^{-1}(\tilde{X}_n)$ satisfies:*

$$\sqrt{n}\tilde{\Delta}_n \xrightarrow{D} \mathcal{N}(0, C) \quad with \quad C = \sum_{j'=1}^{k} \sum_{i'=k+1}^{d} \sum_{j=1}^{k} \sum_{i=k+1}^{d} C_{ii'jj'} \frac{\sqrt{\lambda_i \lambda_j} \cdot \sqrt{\lambda_{i'}\lambda_{j'}}}{(\lambda_j - \lambda_i) \cdot (\lambda_{j'} - \lambda_{i'})} (v_i e_j^\top) \otimes (v_{i'} e_{j'}^\top).$$

We make the following observations:

- If the 4th-order tensor satisfies[4] $C_{ii'jj'} = \kappa \delta_{ii'} \delta_{jj'}$ for constant κ, the aforementioned covariance structure simplifies to $C = \kappa \sum_{j=1}^{k} \sum_{i=k+1}^{d} \frac{\lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2} (v_i e_j^\top) \otimes (v_i e_j^\top)$. This asymptotic variance matches the result of Reiß and Wahl (2016), achieving the same statistical performance as the empirical risk minimizer and matching the lower bound of Cai et al. (2013) obtained for the (Gaussian) spiked covariance model.

- Empirically, even using a constant step size in Eq. (5) appears to yield convergence in a variety of situations; however, we can see a numerical counterexample in Appendix F. We leave it as an open problem to understand the convergence of the iterate-averaged, constant step-size algorithm in the case of Gaussian noise (Bougerol and Lacroix, 1985).

- Assumption 2 could be relaxed using a martingale concentration result showing the iterates X_n are restricted to $\mathcal{X}$ with high probability similar to the work of Shamir (2016a) and Allen-Zhu and Li (2017).

Note that we could also derive an analogous result to Theorem 3 for the (averaged) Riemannian SGD algorithm in Eq. (1) and Eq. (2). However, we prefer to present the algorithm in Eq. (5) since it is simpler and directly averages the (commonly used) randomized power method.

7. Experiments

Here, we illustrate our results on a synthetic, streaming k-PCA problem using the SGD algorithm defined in Eq. (5). We take $k=10$ and $d=50$. The stream $H_n \in \mathbb{R}^d$ is normally-distributed with a covariance matrix H with random eigenvectors, and eigenvalues decaying as $1/i^\alpha + \beta$, for $i \leq k$, and $1/(i-1)^\alpha$, for $i > k$ and $\alpha, \beta \geq 0$. All results are averaged over ten repetitions.

---

4. For example if $H_n = h_n h_n^\top$ for $h_n \sim \mathcal{N}(0, \Sigma)$ – so H_n is a rank-one stream of Gaussian random vectors – this condition is satisfied. See the proof of Theorem 3 for more details.
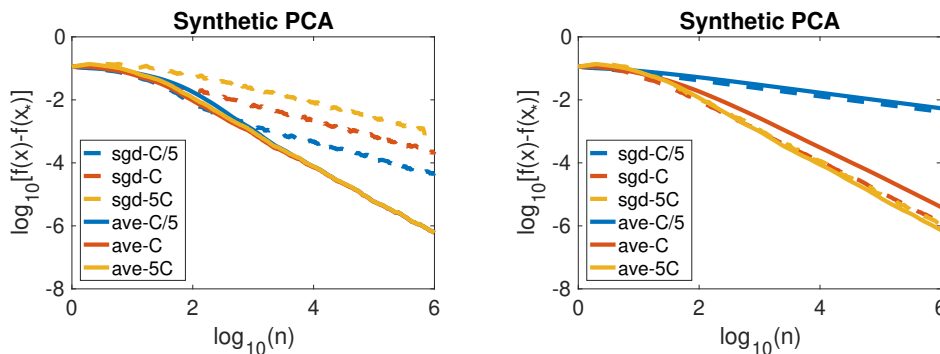
Figure 3: Robustness to constant in step size. Left: step size proportional to $n^{-1/2}$. Right: step size proportional to n^{-1}.

Robustness to Conditioning. In Figure 2 we consider two covariance matrices with different conditioning and we compare the behavior of SGD and averaged SGD for different step sizes (constant (cst), proportional to $1/\sqrt{n}$ (-1/2) and $1/n$ (-1)). When the covariance matrix is well-conditioned, with a large eigengap (left plot), we see that SGD converges at a rate which depends on the step size whereas averaged SGD converges at a $O(1/n)$ rate independently of the step-size choice. For poorly conditioned problems (right plot), the convergence rate deteriorates to $1/\sqrt{n}$ for non-averaged SGD with step size $1/\sqrt{n}$, and averaged SGD with both constant and $1/\sqrt{n}$ step sizes. The $1/n$ step size performs poorly with and without averaging.

Robustness to Incorrect Step-Size. In Figure 3 we consider a well-conditioned problem and compare the behavior of SGD and averaged SGD with step size proportional to $C/\sqrt{n}$ and C/n to investigate the robustness to the choice of the constant C. For both algorithms we take three different constant prefactors $C/5$, C and $5C$. For the step size proportional to $C/\sqrt{n}$ (left plot), both SGD and averaged SGD are robust to the choice of C. For SGD, the iterates eventually converge at a $1/\sqrt{n}$ rate, with a constant offset proportional to C. However, averaged SGD enjoys the fast rate $1/n$ for all choices of C. For the step size proportional to C/n (right plot), if C is too small, the rate of convergence is extremely slow for SGD and averaged SGD.

8. Conclusions

We have constructed and analyzed a geometric framework on Riemannian manifolds that generalizes the classical Polyak-Ruppert iterate-averaging scheme. This framework is able to accelerate a sequence of slowly-converging iterates to an iterate-averaged sequence with a robust $O(\frac{1}{n})$ rate. We have also presented two applications, to the class of geodesically-strongly-convex optimization problems and to streaming k-PCA. Note that our results only apply locally, requiring the iterates to be constrained to lie in a compact set $\mathcal{X}$. Considering a projected variant of our algorithm as in Flammarion and Bach (2017) is a promising direction for further research that may allow us to remove this restriction. Another interesting direction is to provide a global-convergence result for the iterate-averaged PCA algorithm presented here.

References

P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM J. Optim.*, 22(1): 135–158, 2012.

P-A Absil, R. Mahony, and R. Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae*, 80(2):199–220, 2004.

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

P-A Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

B. Afsari. Riemannian L^p center of mass: existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.*, 139(2):655–673, 2011.

Z. Allen-Zhu and Y. Li. First efficient convergence for streaming k-PCA: a global, gap-free, and near-optimal rate. In *Proceedings of the 58th Symposium on Foundations of Computer Science*, FOCS '17, 2017.

A. Aswani, P. Bickel, and C. Tomlin. Regression on manifolds: estimation of the exterior derivative. *Ann. Statist.*, 39(1):48–81, 2011.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

A. Benveniste, P. Priouret, and M. Métivier. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

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

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

L. Bottou. Online algorithms and stochastic approximations. In *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.

P. Bougerol and J. Lacroix. *Products of Random Matrices with Applications to Schrödinger Operators*, volume 8 of *Progress in Probability and Statistics*. Birkhäuser, 1985.

N. Boumal. On intrinsic Cramér-Rao bounds for Riemannian submanifolds and quotient manifolds. *IEEE Trans. Signal Process.*, 61(7):1809–1821, 2013.

N. Boumal and P.-A. Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems 24*, pages 406–414. 2011.

T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: optimal rates and adaptive estimation. *Ann. Statist.*, 41 (6):3074–3110, 2013.

M. P. Do Carmo. *Differential Geometry of Curves and Surfaces*. Courier Dover Publications, 2016.

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

V. Fabian. On asymptotic normality in stochastic approximation. *Ann. Math. Statist*, 39:1327–1332, 1968.

N. Flammarion and F. Bach. Stochastic composite least-squares regression with convergence rate $O(1/n)$. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 831–875. PMLR, 07–10 Jul 2017.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 910–918, 2015.

W. Huang, K. 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.

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

P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford. Streaming PCA: matching matrix Bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *Conference on Learning Theory*, pages 1147–1164, 2016.

H. Kushner and G G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

M. Moakher. Means and averaging in the group of rotations. *SIAM Journal on Matrix Analysis and Applications*, 24(1):1–16, 2002.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.

Y. Nesterov and J.-P. Vial. Confidence level solutions for stochastic programming. *Automatica J. IFAC*, 44(6):1559–1568, 2008.

M. B. Nevelson and R. Z. Hasminski. *Stochastic Approximation and Recursive Estimation*. American Mathematical Society, 1973.

E. Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, Nov 1982.

E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *J. Math. Anal. Appl.*, 106(1):69–84, 1985.

B. T. Polyak. A new method of stochastic approximation type. *Avtomatika i Telemekhanika*, 51(7): 98–107, 1990.

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

M. Reiß and M. Wahl. Non-asymptotic upper bounds for the reconstruction error of PCA. *arXiv preprint arXiv:1609.03779*, 2016.

W. Ring and B. Wirth. Optimization methods on Riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

T. Sakai. *Riemannian Geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, 1996.

H. Sato, H. Kasai, and B. Mishra. Riemannian stochastic variance reduced gradient. *arXiv preprint arXiv:1702.05594*, 2017.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

O. Shamir. Convergence of stochastic gradient descent for PCA. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 257–265. PMLR, 20–22 Jun 2016a.

O. Shamir. Fast stochastic algorithms for SVD and PCA: convergence properties and convexity. In *International Conference on Machine Learning*, pages 248–256, 2016b.

S. T. Smith. Covariance, subspace, and intrinsic Cramér-Rao bounds. *IEEE Trans. Signal Process.*, 53(5):1610–1630, 2005.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method. *IEEE Trans. Inform. Theory*, 63(2):885–914, 2017.

C. Udriste. *Convex Functions and Optimization Methods on Riemannian Manifolds*, volume 297. Springer Science & Business Media, 1994.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

S. Waldmann. Geometric wave equations. *arXiv preprint arXiv:1208.4706*, 2012.

B. Yang. Projection approximation subspace tracking. *Trans. Sig. Proc.*, 43(1):95–107, January 1995.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.

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

Appendix A. Appendices

In Appendix B we provide the proof of Theorem 1. In Appendix C we prove the relevant lemmas mirroring the proof sketch in Section 5. In Appendix D we provide proofs of the results for the application discussed in Section 6.1 about geodesically-strongly-convex optimization. Section E contains background and proofs of results discussed in Section 6.2 regarding streaming k-PCA. Section F contains further experiments on synthetic PCA showing a counterexample to the convergence of averaged, constant step-size SGD mentioned in Section 7 in the main text.

Throughout this section we will denote a sequence of vectors X_n to be $X_n = O(f_n)$, for scalar functions f_n, if there exists a constant $C > 0$, such that $\|X_{n+1}\| \le C f_n$ for all $n \ge 0$ almost surely.

Appendix B. Proofs for Section 2

Here we provide the proof of Theorem 1. The first statement follows by combining Theorems 7, 8, Lemma 12 and Slutsky's theorem. The second statement follows by using Theorems 7, 9, and Lemma 13.

Appendix C. Proofs for Section 5

Here we detail the proofs results necessary to conclude our main result sketched in Section 5.

C.1. Proofs in Section 5.1

We begin with the proofs of the geometric lemmas detailed in Section 5.1, showing how to linearize the progress of the SGD iterates x_n in the tangent space of $x_\star$ by considering the evolution of $\Delta_n = R_{x_\star}^{-1}(x_n)$. Note that since by Assumption 2, for all $n \ge 0$, $x_n \in \mathcal{X}$, the vectors Δ_n all belong to the compact set $R_{x_\star}^{-1}(\mathcal{X})$.

In the course of our argument it will be useful to consider the function $F_{x,y}(\eta_x) = R_y^{-1} \circ R_x(\eta_x) : T_x\mathcal{M} \to T_x\mathcal{M}$ (which crucially is a function defined on a vector space) and further $DR_x(\eta_x) : T_x\mathcal{M} \to T_{R_x(\eta_x)}\mathcal{M}$, the linearization of the retraction map. The first recursion we will study is that of $\Delta_{n+1} = F_{x_n, x_\star}(-\gamma_{n+1}\nabla f_{n+1}(x_n))$:

Lemma 4 *Let Assumption 2 hold. If $\Delta_n = R_{x_\star}^{-1}(x_n)$ for a sequence of iterates evolving as in Eq. (1), then there exists a constant $C_{manifold} > 0$ depending on $\mathcal{X}$ such that,*

$$\Delta_{n+1} = \Delta_n - \gamma_{n+1}[\Lambda_{x_\star}^{x_n}]^{-1}(\nabla f_{n+1}(x_n)) + \gamma_{n+1}g_n,$$

where $\|g_n\| \le \gamma_{n+1}C_{manifold}\|\nabla f_{n+1}(x_n)\|^2$.

Proof Using the chain rule for the differential of a mapping on a manifold and the first-order property of the retraction ($DR_x(0_x) = I_{T_x\mathcal{M}}$) we have that:

$$DF_{x,y}(0_x) = D(R_y^{-1} \circ R_x)(0_x) = DR_y^{-1}(R_x(0_x)) \circ DR_x(0_x)$$
$$= [DR_y(R_y^{-1}(R_x(0_x)))]^{-1} \circ I_{T_x\mathcal{M}} = [DR_y(R_y^{-1}(x))]^{-1} = [\Lambda_y^x]^{-1},$$

where the last line follows by the inverse function theorem on the manifold $\mathcal{M}$. Smoothness of the retraction implies the Hessian of $F_{x,y}$ is uniformly bounded in norm on the compact set

$F_{x,y}^{-1}(R_{x_\star}^{-1}(\mathcal{X}))$. We use C_{manifold} to denote a bound on the operator norm of the Hessian of $F_{x,y}$ in this compact set. In the present situation, we have that $\Delta_{n+1} = F_{x_n,x_\star}(-\gamma_{n+1}\nabla f_{n+1}(x_n))$. Since $F_{x_n,x_\star}$ is a function defined on vector spaces the result follows using a Taylor expansion, $F_{x_n,x_\star}(0) = \Delta_n$, the previous statements regarding the differential of $F_{x_n,x_\star}$, and the uniform bounds on the second-order terms. In particular, the second-order term in the Taylor expansion is upper bounded as $\gamma_{n+1}C_{\text{manifold}}\|\nabla f_{n+1}(x_n)\|^2$ so the bound on the error term g_n follows. ∎

We now further develop this recursion by also considering an asymptotic expansion of the gradient term near the optima.

Lemma 5 *Let Assumptions 4, 5, and 6 hold. If $\Delta_n = R_{x_\star}^{-1}(x_n)$ for a sequence of iterates evolving as in Eq. (1), then there exist sequences $\{\tilde{\xi}_n\}_{n\geq 0}$ and $\{\tilde{e}_n\}_{n\geq 0}$ such that*

$$\Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n) = \nabla^2 f(x_\star)\Delta_n + \nabla f_{n+1}(x_\star) + \tilde{\xi}_{n+1} + \tilde{e}_{n+1},$$

where $\mathbb{E}[\tilde{\xi}_{n+1}|\mathcal{F}_n] = 0$, $\mathbb{E}[\|\tilde{\xi}_{n+1}\|^2|\mathcal{F}_n] \leq 4L\|\Delta_n\|^2$ and $\tilde{e}_{n+1}$ such that $\|\tilde{e}_{n+1}\| \leq \frac{M}{2}\|\Delta_n\|^2$.

Proof We begin with the decomposition:

$$
\begin{aligned}
\nabla^2 f(x_\star)\Delta_n &= \Gamma_{x_n}^{x_\star}\nabla f(x_n) - \nabla f(x_\star) + [\nabla^2 f(x_\star)\Delta_n - \Gamma_{x_n}^{x_\star}\nabla f(x_n) - \nabla f(x_\star)] \\
&= \Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n) - \nabla f_{n+1}(x_\star) + [\nabla^2 f(x_\star)\Delta_n - \Gamma_{x_n}^{x_\star}\nabla f(x_n) - \nabla f(x_\star)] \\
&\quad + [\Gamma_{x_n}^{x_\star}\nabla f(x_n) - \nabla f(x_\star) - \Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n) + \nabla f_{n+1}(x_\star)].
\end{aligned}
$$

Under Assumption 4, using the manifold version of Taylor's theorem (see Absil et al. (2009) Lemma 7.4.8) we have for $\tilde{e}_{n+1} = \nabla^2 f(x_\star)\Delta_n - \Gamma_{x_n}^{x_\star}\nabla f(x_n)$, that

$$\|\tilde{e}_{n+1}\| \leq \frac{M}{2}\|\Delta_n\|^2.$$

Denoting $\tilde{\xi}_{n+1} = [\Gamma_{x_n}^{x_\star}\nabla f(x_n) - \nabla f(x_\star) - \Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n) + \nabla f_{n+1}(x_\star)]$, Assumption 5 directly implies that $\mathbb{E}[\tilde{\xi}_{n+1}|\mathcal{F}_n] = 0$. Finally, using Assumption 6 and the elementary inequality $2\mathbb{E}[A \cdot B|\mathcal{F}_n] \leq \mathbb{E}[A^2|\mathcal{F}_n] + \mathbb{E}[B^2|\mathcal{F}_n]$ for square-integrable random variables A, B shows that,

$$
\begin{aligned}
\mathbb{E}[\|\tilde{\xi}_{n+1}\|^2|\mathcal{F}_n] &\leq 2\|\Gamma_{x_n}^{x_\star}\nabla f(x_n) - \nabla f(x_\star)\|^2 + 2\mathbb{E}\left[\|\Gamma_{x_n}^{x_\star}\nabla f_{n+1}(x_n) - \nabla f_{n+1}(x_\star)\|^2|\mathcal{F}_n\right] \\
&\leq 4L^2\|\Delta_n\|^2.
\end{aligned}
$$

∎

The last important step to conclude a linear recursion in Δ_n is to argue that the operator composition $[\Lambda_{x_\star}^{x_n}]^{-1}\Gamma_{x_\star}^{x_n} : T_{x_\star}\mathcal{M} \to T_{x_\star}\mathcal{M}$, is in fact an isometry (up to 2nd-order terms) since x_n is close to $x_\star$. The following argument crucially uses the fact that $R_{x_\star}$ is a second-order retraction.

Lemma 6 *Let Assumption 2. Let $\Delta_n = R_{x_\star}^{-1}(x_n)$ for a sequence $\{x_n\}_{n\geq 0}$ evolving as in Eq. (1). Then there exists a trilinear operator $K(\cdot, \cdot, \cdot)$ such that*

$$[\Lambda_{x_\star}^{x_n}]^{-1}\Gamma_{x_\star}^{x_n} = I - K(\Delta_n, \Delta_n, \cdot) + O(\|\Delta_n\|^3).$$

18

As noted in the proof, when the exponential map is used as the retraction, the operator K is precisely the Riemann curvature tensor $R_{x_\star}(\Delta_n, \cdot)\Delta_n$ (up to a constant prefactor).

Proof We derive a Taylor expansion for the operator composition $[\Gamma_x^y]^{-1}\Lambda_x^y$ when y is close to x. Consider the function $G(v) = [\Gamma_x^{R_x(v)}]^{-1}\Lambda_x^{R_x(v)} : T_x\mathcal{M} \to L(T_x\mathcal{M})$ where $L(T_x\mathcal{M})$ denotes the set of linear maps on the vector space $T_x\mathcal{M}$. Now, recall that $\Gamma_x^{R_x(tv)}$ is precisely the parallel translation operator along the curve $\gamma(t) = R_y(tv)$. From Proposition 8.1.2 by Absil et al. (2009), we have that

$$\frac{d}{dt}G(tv)|_{t=0} = \frac{d}{dt}[\Gamma_x^{R_x(tv)}]^{-1}\Lambda_x^{R_x(tv)}|_{t=0} = \nabla_{\dot{\gamma}(0)}DR_y,$$

where ∇ denotes the Levi-Civita connection (see also the proof of Absil et al. (2009, Lemma 7.4.7) and Do Carmo (2016, Chapter 2, Exercise 2)). Using the definition of the covariant derivative ∇_v along a vector v, and of the acceleration vector field $\ddot{\gamma}$ (Absil et al., 2009, Section 5.4) we have that

$$\nabla_{\dot{\gamma}(0)}DR_y = \frac{D}{dt}DR_y(\gamma(t))|_{t=0} = \frac{D^2}{dt^2}R_y(tv)|_{t=0} = 0,$$

since R is a second-order retraction. Thus, $\frac{d}{dt}G(tv)|_{t=0} = 0$.

We use K to denote the symmetric trilinear map $d^2G(0)$, where $K(v, v, \cdot) = \frac{1}{2}\frac{d^2}{dt^2}G(tv)|_{t=0}$. Thus, since G is smooth and the iterates are restricted to $\mathcal{X}$ by Assumption 2, a Taylor expansion gives, for $v \in R_{x_\star}^{-1}(\mathcal{X})$, $G(v) = G(0) + K(v, v, \cdot) + O(\|v\|^3)$. For $x = x_\star$ and $v = \Delta_n$, this yields

$$[\Gamma_{x_\star}^{x_n}]^{-1}\Lambda_{x_\star}^{x_n} = I + K(\Delta_n, \Delta_n, \cdot) + O(\|\Delta_n\|^3).$$

Lastly, as $[\Lambda_{x_\star}^{x_n}]^{-1}\Gamma_{x_\star}^{x_n} = \left([\Gamma_{x_\star}^{x_n}]^{-1}\Lambda_{x_\star}^{x_n}\right)^{-1} = \left(I + K(\Delta_n, \Delta_n, \cdot) + O(\|\Delta_n\|^3))\right)^{-1} = I - K(\Delta_n, \Delta_n, \cdot) + O(\|\Delta_n\|^3)$ the conclusion follows. In the special case the exponential map is used as retraction, Waldmann (2012, Theorem A.2.9) directly relates K to the Riemann curvature tensor. They show $K(v, v, \cdot) = -\frac{1}{6}R_{x_\star}(v, \cdot)v$ for $v \in T_{x_\star}\mathcal{M}$. However the result by Waldmann (2012) provides the Taylor expansion up to arbitary order in $\|v\|$. ■

Assembling Lemmas 4, 5 and 6 we obtain the desired linear recursion:

Theorem 7 *Let Assumptions 2, 4, 5, and 6 hold. If $\Delta_n = R_{x_\star}^{-1}(x_n)$ for a sequence of iterates evolving as in Eq. (1), then there exists a martingale-difference sequence $\{\xi_n\}_{n\geq0}$ satisfying $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$, $\mathbb{E}[\|\xi_{n+1}\|^2|\mathcal{F}_n] = O(\|\Delta_n\|^2)$, and an error sequence $\{e_n\}_{n\geq0}$ satisfying $\mathbb{E}[\|e_{n+1}\||\mathcal{F}_n\|] = O(\|\Delta_n\|^2 + \gamma_{n+1})$ and $\mathbb{E}[\|e_{n+1}\|^2|\mathcal{F}_n\|] = O(\|\Delta_n\|^4 + \gamma_{n+1}^2)$ such that*

$$\Delta_{n+1} = \Delta_n - \gamma_{n+1}\nabla^2f(x_\star)\Delta_n - \gamma_{n+1}\nabla f_{n+1}(x_\star) - \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}e_{n+1}.$$

Proof Combining Lemmas 4, 5 and 6,

$$
\begin{aligned}
\Delta_{n+1} &= \Delta_n - \gamma_{n+1}[\Lambda_{x_\star}^{x_n}]^{-1}(\nabla f_{n+1}(x_n)) + \gamma_{n+1}g_n \\
&= \Delta_n - \gamma_{n+1}[\Gamma_{x_n}^{x_\star}\Lambda_{x_\star}^{x_n}]^{-1}\Gamma_{x_n}^{x_\star}(\nabla f_{n+1}(x_n)) + \gamma_{n+1}g_n \\
&= \Delta_n - \gamma_{n+1}[I - K(\Delta_n, \Delta_n, \cdot)] \circ (\nabla^2 f(x_\star)\Delta_n + \nabla f_{n+1}(x_\star) + \tilde{\xi}_{n+1} + \tilde{e}_{n+1}) \\
&\quad + \gamma_{n+1}g_n + O(\gamma_{n+1}\|\Delta_n\|^3) \\
&= \Delta_n - \gamma_{n+1}\nabla^2 f(x_\star)\Delta_n - \gamma_{n+1}\nabla f_{n+1}(x_\star)
\end{aligned}
$$

$$-\gamma_{n+1}\tilde{\xi}_{n+1} + \gamma_{n+1}K(\Delta_n, \Delta_n, \nabla f_{n+1}(x_\star) + \tilde{\xi}_{n+1})$$
$$-\gamma_{n+1}\tilde{e}_{n+1} + \gamma_{n+1}K(\Delta_n, \Delta_n, \nabla^2 f(x_\star)\Delta_n + \tilde{e}_{n+1})$$
$$+\gamma_{n+1}g_n + O(\gamma_{n+1}\|\Delta_n\|^3).$$

Let $\xi_{n+1} = \tilde{\xi}_{n+1} - \gamma_{n+1}K(\Delta_n, \Delta_n, \nabla f_{n+1}(x_\star) + \tilde{\xi}_{n+1})$. By linearity of the map $K(\Delta_n, \Delta_n, \cdot)$, $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$. Moreover by smoothness of the retraction, the tensor K is uniformly bounded in injective norm on the compact set $R_{x_\star}^{-1}(\mathcal{X})$, so $\mathbb{E}[\|\xi_{n+1}\|^2|\mathcal{F}_n] = O(\|\Delta_n\|^2)$.

Let $e_{n+1} = \tilde{e}_{n+1} - K(\Delta_n, \Delta_n, \nabla^2 f(x_\star) + \tilde{e}_{n+1}) - g_n + O(\|\Delta_n\|^3)$. Using Assumptions 2, 6 and the almost sure upper bound on $\tilde{e}_{n+1}$ we have that this term satisfies

$$\mathbb{E}[\|e_{n+1}\|^2|\mathcal{F}_n] = \mathcal{O}\left(\|\Delta_n\|^4 + \gamma_{n+1}^2\right).$$

∎

Note that sharper bounds may be obtained under higher-order assumptions on the moments of the noise. This would provide a sharp constant of the leading asymptotic term of $O(\frac{1}{n})$, when the step-size $\gamma_n = \frac{1}{\sqrt{n}}$ is used.

C.2. Proofs in Section 5.2

Here we provide proofs, in the Euclidean setting, of both asymptotic and non-asymptotic Polyak-Ruppert-type averaging results. We apply these results to the tangent vectors $\Delta \in T_{x_\star}\mathcal{M}$ as described in Section 5.1.

C.2.1. ASYMPTOTIC CONVERGENCE

Throughout this section, we will consider a general linear recursion perturbed by a remainder term e_n of the form:

$$\Delta_n = \Delta_{n-1} - \gamma_n A\Delta_{n-1} + \gamma_n(\varepsilon_n + \xi_n + e_n), \tag{6}$$

for which we will show an asymptotic convergence result under appropriate conditions.

Note that we eventually apply these convergence results to iterates $\Delta_n \in T_{x_\star}\mathcal{M}$, which is a finite-dimensional vector space. In this setting, a probability measure can be defined on a vector space (with inner product) with a covariance operator implicitly depending on the inner product (via the dual map).

We make the following assumptions on the structure of the recursion:

Assumption 8 *A is symmetric positive-definite matrix.*

Assumption 9 *The noise process $\{\varepsilon_n\}$ is a martingale-difference process (with $\mathbb{E}[\varepsilon_n|\mathcal{F}_{n-1}] = 0$ and $\sup_n \mathbb{E}[\varepsilon_n^2] < \infty$), for which there exists $C > 0$ such that $\mathbb{E}[\|\varepsilon_n\|^4|\mathcal{F}_{n-1}] \leq C$ for all $n \geq 0$ and a matrix $\Sigma \succ 0$ such that*

$$\mathbb{E}[\varepsilon_n\varepsilon_n^\top|\mathcal{F}_{n-1}] \xrightarrow{P} \Sigma.$$

Assumption 10 *The noise process $\{\xi_n\}$ is a martingale-difference process (with $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$ and $\sup_n \mathbb{E}[\xi_n^2] < \infty$), and for sufficiently large $n \geq N$, there exists $K > 0$ such that*

$$\mathbb{E}[\|\xi_n\|^2|\mathcal{F}_{n-1}] \leq K\gamma_n \text{ a.s.}$$

with $\gamma_n \to 0$ as $n \to \infty$.

Assumption 11 *For $n \geq 0$*

$$\mathbb{E}[\|e_n\|] = O(\gamma_n).$$

Assumption 12 $\gamma_n \to 0$, $\frac{\gamma_n - \gamma_{n-1}}{\gamma_n} = o(\gamma_n)$ *and* $\sum_{j=1}^{\infty} \frac{\gamma_j}{\sqrt{j}} < \infty$.

The first two conditions in Assumption 12 require that γ_n decrease sufficiently slowly. For example $\gamma_n = \gamma t^{-\alpha}$ with $\frac{1}{2} < \alpha < 1$ satisfy these two conditions but the sequence $\gamma = \gamma t^{-1}$ does not.

We can now derive the asymptotic convergence rate,

Theorem 8 *Let Assumptions 8, 9, 10, 11 and 12 hold for the perturbed linear recursion in Equation (6). Then,*

$$\sqrt{n}\bar{\Delta}_n \overset{D}{\to} \mathcal{N}(0, A^{-1}\Sigma A^{-1}).$$

Proof The argument mirrors the proof of Theorem 2 in Polyak and Juditsky (1992) so we only sketch the primary points. Throughout we will use C to denote an unimportant, global constant that may change line to line.

Consider the purely linear recursion of the form:

$$\Delta_n^1 = \Delta_{n-1}^1 - \gamma_n A \Delta_{n-1}^1 + \gamma_n(\varepsilon_n + \xi_n) \tag{7}$$

$$\bar{\Delta}_n^1 = \frac{1}{n}\sum_{i=0}^{n-1}\Delta_i^1,$$

which satisfies $\Delta_0^1 = \Delta_0$, and approximates the perturbed recursion in Equation (6),

$$\Delta_n = \Delta_{n-1} - \gamma_n A \Delta_{n-1} + \gamma_n(\varepsilon_n + \xi_n) + \gamma_n e_n \tag{8}$$

$$\bar{\Delta}_n = \frac{1}{n}\sum_{i=0}^{n-1}\Delta_i.$$

Now, note that we can show that $\lim_{K\to\infty}\limsup_n \mathbb{E}\left[\|\varepsilon_n\|^2 \mathbb{I}[\|\varepsilon_n\| > K]|\mathcal{F}_{n-1}\right] \overset{p}{\to} 0$, using our (conditional) 4th-moment bound and the (conditional) Cauchy-Schwarz/Markov inequalities, so the relevant assumption in Polyak and Juditsky (1992) is satisfied. Then as the argument in Part 3 of the proof of Theorem 2 in Polyak and Juditsky (1992) shows, under Assumptions 8, 9, 10 the conditions of Proposition (a) of Theorem 1 in Polyak and Juditsky (1992) also hold. This implies the linear process satisfies:

$$\sqrt{n}\bar{\Delta}_n^1 \overset{D}{\to} \mathcal{N}(0, A^{-1}\Sigma A^{-1}).$$

We now argue that the process $\bar{\Delta}_n^1$ and $\bar{\Delta}_n$ are asymptotically equivalent in distribution. First, since the noise process is coupled between Equations 7 and 8, the differenced process obeys a simple (perturbed) linear recursion,

$$\Delta_n - \Delta_n^1 = (I - \gamma_j A)(\Delta_{n-1} - \Delta_{n-1}^1) - \gamma_n e_n.$$

Expanding and averaging this recursion (defining $\delta_n = \bar{\Delta}_n - \bar{\Delta}_n^1$) gives:

$$\Delta_n - \Delta_n^1 = \sum_{j=1}^{n}\Pi_{i=j+1}^{n}(I - \gamma_j A)\gamma_j e_j \implies \delta_n = \frac{1}{n}\sum_{k=1}^{n-1}\sum_{j=1}^{k}[\Pi_{i=j+1}^{k}(I - \gamma_i A)]\gamma_j e_j$$

21

$$\implies \delta_n = \frac{1}{n} \sum_{j=1}^{n-1} \left[\sum_{k=j}^{n-1} \Pi_{i=j+1}^k (I - \gamma_i A) \right] \gamma_j e_j.$$

We can rewrite the recursion for this averaged differenced process as:

$$\sqrt{n}\delta_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} (A^{-1} + w_j^n) e_j,$$

defining,

$$w_j^n = \gamma_j \sum_{i=j}^{n-1} \Pi_{k=j+1}^i (I - \gamma_k A) - A^{-1}.$$

Now if the step-size sequence satisfies the first two conditions of Assumption 12, by Lemma 1 and 2 in Polyak and Juditsky (1992) we have that $\|w_j^n\| \le C$ uniformly. So using Assumption 8 we obtain that:

$$\sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \|(A^{-1} + w_j^t) e_j\| \le C \sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \|e_j\|.$$

An application of the Tonelli-Fubini theorem and Assumption 11 then shows that

$$\mathbb{E}[\sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \|e_j\|] = \sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \mathbb{E}[\|e_j\|] \le C \sum_{j=1}^{\infty} \frac{\gamma_j}{\sqrt{j}} < \infty,$$

by choice of the step-size sequence in Assumption 12. Since $\sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \|e_j\| \ge 0$ and has finite expectation it must be that,

$$\sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \|e_j\| < \infty \implies \sum_{j=1}^{\infty} \frac{1}{\sqrt{j}} \|(A^{-1} + w_j^n) e_j\| < \infty.$$

An application of the Kronecker lemma then shows that

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} \|(A^{-1} + w_j^n) e_j\| \to 0 \implies \sqrt{n}\delta_n \to 0 \text{ a.s.}$$

The conclusion of theorem follows by Slutsky's theorem. ∎

C.2.2. NONASYMPTOTIC CONVERGENCE

Throughout this section, we will consider a general linear recursion perturbed by remainder terms e_n of the form:

$$\Delta_n = \Delta_{n-1} - \gamma_n A \Delta_{n-1} + \gamma_n (\varepsilon_n + \xi_n + e_n). \tag{9}$$

We make the following assumptions on the structure of the recursion:

Assumption 13 *A is symmetric positive-definite matrix, such that $A \succeq \mu I$ for $\mu > 0$.*

Assumption 14 *The noise process $\{\varepsilon_n\}$ is a martingale-difference process (with $\mathbb{E}[\varepsilon_n|\mathcal{F}_{n-1}] = 0$ and $\sup_n \mathbb{E}[\varepsilon_n^2] < \infty$) and a matrix $\Sigma \succ 0$ such that*

$$\mathbb{E}[\varepsilon_n \varepsilon_n^\top | \mathcal{F}_{n-1}] \preccurlyeq \Sigma.$$

Assumption 15 *The noise process $\{\xi_n\}$ is a martingale-difference process (with $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$ and $\sup_n \mathbb{E}[\xi_n^2] < \infty$), and there exists $K > 0$ such that for $n \geq 0$*

$$\mathbb{E}[\|\xi_n\|^2 | \mathcal{F}_{n-1}] \leq K\gamma_n \text{ a.s.}$$

Assumption 16 *There exists M such that for $n \geq 0$, $\mathbb{E}[\|e_n\|^2] \leq M\gamma_n^2$.*

Assumption 17 *The step-sizes take the form $\gamma_n = \frac{C}{n^\alpha}$ for $C > 0$ and $\alpha \in [1/2, 1)$.*

Assumption 18 *There exists $C' > 0$ such that for $n \geq 0$, we have that*

$$\sqrt{\mathbb{E}[\|\Delta_n\|^2]} = O(\sqrt{\gamma_n}) = C'n^{-\alpha/2}$$

Using these Assumptions we can derive the non-asymptotic convergence rate:

Theorem 9 *Let Assumptions 13, 14, 15, 16, 17 and 18 hold for the recursion in Equation 9,*

$$\mathbb{E}[\|\bar{\Delta}_n\|^2] \leq \frac{1}{n}\operatorname{tr}[A^{-1}\Sigma A^{-1}] + O(n^{-2\alpha}) + O(n^{\alpha-2}).$$

Proof The argument mirrors the proof of Theorem 3 in Bach and Moulines (2011) so we only sketch the key points. First, since A is invertible due to Assumption 13, from Equation 9:

$$\Delta_{n-1} = \frac{A^{-1}(\Delta_{n-1} - \Delta_n)}{\gamma_n} + A^{-1}\varepsilon_n + A^{-1}\xi_n + A^{-1}e_n,$$

We now analyze the average of each of the 4 terms separately. Throughout we will use C to denote an unimportant, numerical constant that may change line to line.

- Summing the first term by parts we obtain,

$$\frac{1}{n}\sum_{k=1}^{n}\frac{A^{-1}(\Delta_{k-1} - \Delta_k)}{\gamma_k} = \frac{1}{n}\sum_{k=1}^{n-1}A^{-1}\Delta_k\left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}\right) - \frac{1}{n\gamma_n}A^{-1}\Delta_n + \frac{1}{n\gamma_1}A^{-1}\Delta_0,$$

and using Minkowski's inequality (in L_2) gives,

$$\sqrt{\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^{n}\frac{A^{-1}(\Delta_{k-1} - \Delta_k)}{\gamma_k}\right\|^2} \leq \frac{1}{n\mu}\sum_{k=1}^{n-1}\sqrt{\mathbb{E}\|\Delta_k\|^2}\left|\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}\right| + \frac{\sqrt{\mathbb{E}\|\Delta_n\|^2}}{n\gamma_n\mu} + \frac{\|\Delta_0\|}{n\gamma_1\mu}.$$

Since we choose a sequence of decreasing step-sizes of the form $\gamma = \frac{C}{n^\alpha}$ for $\alpha \in [\frac{1}{2}, 1)$, an application of the Bernoulli inequality shows that $|\gamma_{k+1}^{-1} - \gamma_k^{-1}| = C^{-1}[(k+1)^\alpha - k^\alpha] \leq C^{-1}\alpha k^{\alpha-1}$. By assumption, we have that $\sqrt{\mathbb{E}\|\Delta_n\|^2} \leq Cn^{-\alpha/2}$ so,

$$\sqrt{\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^{n}\frac{A^{-1}(\Delta_k - \Delta_k)}{\gamma_k}\right\|^2} \leq \frac{C\alpha}{n\mu}\sum_{k=1}^{n-1}k^{\alpha/2-1} + \frac{C}{\mu}n^{\alpha/2-1} + \frac{C}{n\mu}\|\Delta_0\|$$

$$\leq \frac{2Cn^{\alpha/2-1}}{\mu} + \frac{Cn^{\alpha/2-1}}{\mu} + \frac{C\|\Delta_0\|}{n\mu}$$

$$\leq \frac{3Cn^{\alpha/2-1}}{\mu} + \frac{C\|\Delta_0\|}{n\mu}.$$

This implies that,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^{n}\frac{A^{-1}(\Delta_{k-1} - \Delta_k)}{\gamma_k}\right\|^2 = O(n^{\alpha-2}).$$

- Using the Assumption 14 and the orthogonality of martingale increments we immediately obtain the leading order term as,

$$\mathbb{E}\|A^{-1}\bar{\varepsilon}_n\|^2 \leq \frac{1}{n}\operatorname{tr}\big[A^{-1}\Sigma A^{-1}\big].$$

- Using Assumption 15 and the orthogonality of martingale increments we obtain,

$$\mathbb{E}\|A^{-1}\bar{\xi}_n\|^2 = \frac{1}{n^2\mu^2}\sum_{k=1}^{n}\mathbb{E}\|\xi_k\|^2 \leq \frac{C}{n^2\mu^2}\sum_{k=0}^{n-1}k^{-\alpha} = O(n^{-(\alpha+1)}).$$

- Using the Minkowski inequality (in L_2), and Assumption 16, we have that

$$\mathbb{E}\|A^{-1}\bar{e}_{n-1}\|^2 \leq \left(\frac{1}{n\mu}\sum_{k=1}^{n}\sqrt{\mathbb{E}\|e_k\|^2}\right)^2 \leq \frac{M^2}{(n\mu)^2}\left(\sum_{k=1}^{n}k^{-\alpha}\right)^2 \leq \frac{M^2}{\mu^2}n^{-2\alpha}.$$

The conclusion follows. ∎

C.2.3. ON THE RIEMANNIAN CENTER OF MASS

Note that $\bar{\Delta}_n$ is *not* computable, but has an interesting interpretation as an upper bound on the Riemannian center of mass (or Karcher mean),

$$K_n = \arg\min_{x\in\mathcal{M}}\frac{1}{n}\sum_{i=1}^{n}\left\|R_x^{-1}(x_i)\right\|^2$$

of a set of iterates $\{x_i\}_{n\geq 0}$ in $\mathcal{M}$. When it exists, computing K_n is itself a nontrivial geometric optimization problem since it does not admit a closed-form solution in general. See Moakher (2002); Bini and Iannazzo (2013); Hosseini and Sra (2015) for more background on the Karcher mean problem. If we consider a "symmetric" retraction R satisfying for $x, y \in \mathcal{X}$ that $\|R_x^{-1}(y)\|^2 = \|R_y^{-1}(x)\|^2$ (which is the case for the exponential map for example), then

Lemma 10 *Let $\{x_i\}_{i=0}^{n}$ be a sequence of iterates contained in $\mathcal{M}$ and let the retraction R be symmetric, then*

$$\|R_{x_\star}^{-1}(K_n)\|^2 \leq 2\|\bar{\Delta}_n\|^2.$$

Proof The first-order optimality condition requires that $\nabla D(K_n) = 0$ where the manifold gradient is given by $\nabla D(x) = \frac{1}{n}\sum_{i=1}^{n} R_x^{-1}(x_i)$. Thus, $\nabla D(x_\star) = \bar{\Delta}_n$. By Assumption 2, the function D is 1-retraction strongly convex. Defining the function $g : t \mapsto D\left(R_{x_\star}(t\frac{R_{x_\star}^{-1}(K_n)}{\|R_{x_\star}^{-1}(K_n)\|})\right)$, we have that at $t_0 = \|R_{x_\star}^{-1}(K_n)\|$,

$$2\|\nabla D(x_\star)\|^2 = 2(g'(t_0) - g'(0))^2 \geq t_0^2 = \|R_{x_\star}^{-1}(K_n)\|^2.$$

∎

Therefore the Riemannian center of mass will enjoy the same convergence rate as $\bar{\Delta}_n$ itself.

C.3. Proofs in Section 5.3

Finally, we would like to asymptotically understand the evolution of the averaged vector $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$, where $\tilde{x}_n$ is the online, streaming iterate average. From Eq. (2) we see that $\tilde{\Delta}_{n+1} = F_{\tilde{x}_n,x_\star}[\frac{1}{n+1}F_{\tilde{x}_n,x_\star}^{-1}(\Delta_{n+1})] = \tilde{F}(\Delta_{n+1})$, defining $\tilde{F}(\cdot) = F_{\tilde{x}_n,x_\star}[\frac{1}{n+1}F_{\tilde{x}_n,x_\star}^{-1}(\cdot)]$.

We first start with a lemma controlling $\|\tilde{\Delta}_n\|$, when x_n locally converges to $x_\star$.

Lemma 11 *Let Assumptions 1 and 2 hold. Consider x_n and $\tilde{x}_n$, which are a sequence of iterates evolving as in Eq. (1) and Eq. (2), and define $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$. Then, $\mathbb{E}[\|\tilde{\Delta}_n\|^2] = O(\gamma_n)$ as well.*

Proof By Assumption 2, the function $x \to \|R_{x_\star}^{-1}(x)\|^2$ is retraction convex in x. Then,

$$\|R_{x_\star}^{-1}(\tilde{x}_n)\|^2 = \|R_{x_\star}^{-1}\left(R_{\tilde{x}_{n-1}}(\frac{1}{n}R_{\tilde{x}_{n-1}}^{-1}(x_n))\right)\|^2 \leq \frac{n-1}{n}\|R_{x_\star}^{-1}(x_{n-1})\|^2 + \frac{1}{n}\|R_{x_\star}^{-1}(\tilde{x}_{n-1})\|^2.$$

A simple inductive argument then shows that $\|R_{x_\star}^{-1}(\tilde{x}_n)\|^2 \leq \frac{1}{n}\sum_{i=0}^{n}\|R_{x_\star}^{-1}(x_i)\|^2$. Using that $\mathbb{E}\|\Delta_n\|^2 = O(\gamma_n)$ (from Assumption 1), and taking expectations shows $\mathbb{E}[\|\tilde{\Delta}_n\|^2] \leq \frac{C}{n}\sum_{i=0}^{n}\gamma_i \leq C\gamma_n$ when we choose a step-size sequence of the form $\gamma_n = \frac{C}{n^\alpha}$. ∎

Finally using an asymptotic expansion we can show that $\tilde{\Delta}_n$ and $\bar{\Delta}_n$ approach each other:

Lemma 12 *Let Assumptions 1 and 2 hold. As before, consider x_n and $\tilde{x}_n$, which are a sequence of iterates evolving as in Eq. (1) and Eq. (2), and define $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$. Then,*

$$\tilde{\Delta}_n = \bar{\Delta} + e_n,$$

where $\mathbb{E}[\|e_n\|] = O(\gamma_n)$.

Proof A similar chain rule computation to Lemma (4) shows that $d\tilde{F}(\Delta) = \frac{1}{n+1}I_{T_{x_\star}\mathcal{M}}$. Now, in addition to $\tilde{\Delta}_{n+1} = F_{\tilde{x}_n,x_\star}[\frac{1}{n+1}F_{\tilde{x}_n,x_\star}^{-1}(\Delta_{n+1})] = \tilde{F}(\Delta_{n+1})$, we also have that $\tilde{\Delta}_n = \tilde{F}(\tilde{\Delta}_n)$ identically. As $\tilde{F}(\cdot)$ is a mapping between vector spaces applying a Taylor expansion to the first expression about $\tilde{\Delta}_n$ gives:

$$\tilde{\Delta}_{n+1} = \tilde{\Delta}_n + \frac{1}{n+1}(\Delta_{n+1} - \tilde{\Delta}_n) + O(D^2\tilde{F}(\Delta)\|\Delta_{n+1} - \tilde{\Delta}_n\|^2). \tag{10}$$

for $\Delta \in R_{x_\star}^{-1}(\mathcal{X})$. Since $\tilde{F}$ is twice-continuously differentiable and $R_{x_\star}^{-1}(\mathcal{X})$ is compact, direct computation of the Hessian using the chain and Leibniz rules shows

$$\tilde{e}_n = O\left((n+1)D^2\tilde{F}(\Delta)\|\Delta_{n+1} - \tilde{\Delta}_n\|^2\right) = O\left((n+1)\left(\frac{1}{(n+1)^2} + \frac{1}{n+1}\right) \cdot \|\Delta_{n+1} - \tilde{\Delta}_n\|^2\right),$$

which implies that

$$\mathbb{E}\|\tilde{e}_n\| = O(\gamma_n),$$

since both $\mathbb{E}[\|\Delta_n\|^2] = O(\gamma_n)$ and $\mathbb{E}[\|\tilde{\Delta}_n\|^2] = O(\gamma_n)$ by Lemma 11. Therefore $(n+1)\tilde{\Delta}_{n+1} = n\tilde{\Delta}_n + \Delta_{n+1} + e_n = \sum_{k=0}^{n+1}\Delta_k + \sum_{k=0}^{n+1}\tilde{e}_k \implies \tilde{\Delta}_{n+1} = \bar{\Delta}_{n+1} + e_{n+1}$ where $e_{n+1} = \frac{\sum_{k=0}^{n+1}\tilde{e}_k}{n+1}$, and $\mathbb{E}[\|e_{n+1}\|] = \mathbb{E}\left[\|\frac{\sum_{k=0}^{n+1}\tilde{e}_k}{n+1}\|\right] \leq \frac{1}{n+1}\sum_{i=0}^{n}\mathbb{E}[\|\tilde{e}_k\|] = O(\gamma_n)$. ∎

This result states that the distance between the streaming average $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$ is close to the computationally intractable $\bar{\Delta}_n$ up to $O(\gamma_n)$ error.

We can prove a slightly stronger statement under a 4th-moment assumption on the iterates that follows identically to the above.

Lemma 13 *Let Assumption 2 hold, and assume the 4th-moment bound $\mathbb{E}[\|\Delta_n\|^4] = O(\gamma_n^2)$. As before, consider x_n and $\tilde{x}_n$, which are a sequence of iterates evolving as in Eq. (1) and Eq. (2), and define $\tilde{\Delta}_n = R_{x_\star}^{-1}(\tilde{x}_n)$. Then,*

$$\mathbb{E}\left[\|\tilde{\Delta}_n - \bar{\Delta}\|^2\right] = O(\gamma_n^2).$$

Proof The proof is almost identical to the proofs of Lemma 11 and 12 so we will be brief. Since the function $x \to x^2$ is convex and nondecreasing over positive support, using Assumption 2, the composition $x \to \|R_{x_\star}^{-1}(x)\|^4$ is also retraction-convex in x. An identical argument to the proof of Lemma 11 then shows that $\mathbb{E}[\|\Delta_n\|^4] = O(\gamma_n^2)$ implies $\mathbb{E}\left[\|\tilde{\Delta}_n\|^4\right] = O(\gamma_n^2)$. Using that $\mathbb{E}\left[\|\tilde{\Delta}_n\|^4\right] = O(\gamma_n^2)$, a nearly identical calculation to Lemma 12 and an application of Minkowski's inequality (in L_2) shows that $\sqrt{\mathbb{E}\left[\|\bar{\Delta}_n - \tilde{\Delta}_n\|^2\right]} = O(\gamma_n)$. The conclusion follows. ∎

Appendix D. Proofs in Section 6

Here we provide further discussion and proofs of results described in Section 6.

D.1. Proofs in Section 6.1

Here we present the proofs of the slow convergence rate (both in 2nd and 4th moments) for SGD applied to geodesically-smooth and strongly-convex functions. As discussed in Section 6.1 we will take the retraction R to be the exponential map throughout this section. Before we begin, we recall the following Lemma from Bach and Moulines (2011),

Lemma 14 *Let $n, m \in \mathbb{N}$ such that $m < n$ and $\alpha \geq 0$. Then,*

$$\frac{1}{2(1-\alpha)}[n^{1-\alpha} - m^{1-\alpha}] \leq \sum_{k=m+1}^{n} n^{-\alpha} \leq \frac{1}{1-\alpha}[n^{1-\alpha} - m^{1-\alpha}].$$

This follows by simply bounding sums via integrals.

With this result we can now show that SGD applied to (local) geodesically-smooth and strongly-convex functions will converge at the "slow" rate with an appropriately decaying size.

Proposition 15 *Let Assumptions 2, 5, 6 and 7 hold for the iterates evolving in Eq. (1). Recalling that $\gamma_n = Cn^{-\alpha}$ where $C > 0$ and $\alpha \in [1/2, 1)$ we have,*

$$\mathbb{E}[d^2(x_n, x_\star)] \leq \frac{2C\zeta\upsilon^2}{\mu n^\alpha} + O(\exp(-c\mu n^{1-\alpha})),$$

and

$$\mathbb{E}[d^4(x_n, x_\star)] \leq \frac{4C(3 + \zeta)\zeta\upsilon^4}{\mu n^{2\alpha}} + O(\exp(-c\mu n^{1-\alpha})).$$

for some $c > 0$, where $\zeta > 0$ is a constant depending on the geometry of $\mathcal{M}$.

Proof Throughout we will use c to denote a global, positive constant that may change from line to line. The quantity $\zeta \equiv \zeta(\kappa, c) = \frac{\sqrt{|\kappa|}c}{\tanh(\sqrt{|\kappa|}c)}$ is a geometric quantity from Zhang and Sra (2016), where κ denotes the sectional curvature of the manifold. Note that bound on the sectional curvature is subsumed by Assumption 2 – since it is a smooth function on $\mathcal{X}$, so is bounded.

Bound on the second moment. We first prove the 2nd-moment bound by following the proof of Theorem 2 by Bach and Moulines (2011), but adapting it to the setting of g-strong convexity. Using Corollary 8 (a generalization of the law of cosines to the manifold setting) by Zhang and Sra (2016) we have

$$d^2(x_{n+1}, x_\star) \leq d^2(x_n, x_\star) + 2\gamma_{n+1}\langle \nabla f_{n+1}(x_n), \mathrm{Exp}_{x_n}^{-1}(x_\star)\rangle + \gamma_{n+1}^2\zeta\|\nabla f_{n+1}(x_n)\|^2, \quad (11)$$

where ζ satisfies $\max_{x \in \mathcal{X}} \zeta(\kappa, d(x, x_\star)) \leq \zeta$. Taking conditional expectations yields

$$\mathbb{E}[d^2(x_{n+1}, x_\star)|\mathcal{F}_n] \leq d^2(x_n, x_\star) + 2\gamma_{n+1}\langle \nabla f(x_n), \mathrm{Exp}_{x_n}^{-1}(x_\star)\rangle + \gamma_{n+1}^2\zeta\mathbb{E}[\|\nabla f_{n+1}(x_n)\|^2|\mathcal{F}_n].$$

Using the definition of g-strong convexity and Assumptions 6,7 we directly get

$$\mathbb{E}[d^2(x_{n+1}, x_\star)|\mathcal{F}_n] \leq (1 - 2\gamma_{n+1}\mu)d^2(x_n, x_\star) + \gamma_{n+1}^2\zeta\upsilon^2.$$

Taking the full expectation, and denoting by $\delta_n = \mathbb{E}[d^2(x_n, x_\star)]$, we obtain the recursion,

$$\delta_n \leq (1 - 2\gamma_n\mu)\delta_{n-1} + \gamma_n^2\zeta\upsilon^2.$$

Unrolling the recursion we have,

$$\delta_n \leq \prod_{i=1}^n (1 - 2\gamma_i\mu)\delta_0 + \zeta\upsilon^2 \sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - 2\mu\gamma_k).$$

Using the elementary inequality $(1 - x) \leq \exp(-x)$ for $x \in \mathbb{R}$, we observe the first term on the right side decreases exponentially fast. To analyze the second term, we split it into two components around $\lfloor n/2 \rfloor$:

$$\sum_{i=1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - 2\mu\gamma_k) = \sum_{i=1}^{\lfloor n/2 \rfloor} \gamma_i^2 \prod_{k=i+1}^n (1 - 2\mu\gamma_k) + \sum_{i=\lfloor n/2 \rfloor+1}^n \gamma_i^2 \prod_{k=i+1}^n (1 - 2\mu\gamma_k). \quad (12)$$

For the first term in Eq. (12), using again $(1 - x) \leq \exp(-x)$ for $x \in \mathbb{R}$

$$
\begin{aligned}
\sum_{i=1}^{\lfloor n/2 \rfloor} \gamma_i^2 \prod_{k=i+1}^{n} (1 - 2\mu\gamma_k) &\leq \prod_{k=\lfloor n/2 \rfloor+1}^{n} (1 - 2\mu\gamma_k) \sum_{i=1}^{\lfloor n/2 \rfloor} \gamma_i^2 \\
&\leq \prod_{k=\lfloor n/2 \rfloor+1}^{n} \exp(-2\mu\gamma_k) \sum_{i=1}^{\lfloor n/2 \rfloor} \gamma_i^2 \\
&\leq \exp\left(-2\mu \sum_{k=\lfloor n/2 \rfloor+1}^{n} \gamma_k \right) \sum_{i=1}^{\lfloor n/2 \rfloor} \gamma_i^2 \\
&\leq C \exp\left(-c\mu n^{1-\alpha} \right) n^{1-2\alpha},
\end{aligned}
$$

using Lemma 14, which will decrease exponentially fast as $n \to \infty$. For the second term,

$$
\begin{aligned}
\sum_{i=\lfloor n/2 \rfloor+1}^{n} \gamma_i^2 \prod_{k=i+1}^{n} (1 - \mu\gamma_k) &\leq \gamma_{\lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor+1}^{n} \gamma_i \prod_{k=i+1}^{n} (1 - \mu\gamma_k) \\
&= \gamma_{\lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor+1}^{n} \frac{1 - (1 - \mu\gamma_i)}{\mu} \prod_{k=i+1}^{n} (1 - \mu\gamma_k) \\
&= \frac{\gamma_{\lfloor n/2 \rfloor}}{\mu} \sum_{i=\lfloor n/2 \rfloor+1}^{n} [\prod_{k=i+1}^{n} (1 - \mu\gamma_k) - \prod_{k=i}^{n}(1 - \mu\gamma_k)] \\
&\leq \frac{\gamma_{\lfloor n/2 \rfloor}}{\mu} [1 - \prod_{k=\lfloor n/2 \rfloor+2}^{n} (1 - \mu\gamma_k)]] \\
&\leq \frac{\gamma_{\lfloor n/2 \rfloor}}{\mu} \leq \frac{2C}{n^\alpha \mu}.
\end{aligned}
$$

The bound on the second moment follows from this last inequality.

Bound on the fourth moment. We now prove the bound on the 4th-moment. We start by expanding the square of Eq. (11),

$$
\begin{aligned}
d^4(x_{n+1}, x_\star) &\leq d^4(x_n, x_\star) + 4\gamma_{n+1}^2 (\langle \nabla f_{n+1}(x_n), \mathrm{Exp}_{x_n}^{-1}(x_\star) \rangle)^2 + \gamma_{n+1}^4 \zeta^2 \|\nabla f_{n+1}(x_n)\|^4 \\
&\quad + 4\gamma_{n+1} \langle \nabla f_{n+1}(x_n), \mathrm{Exp}_{x_n}^{-1}(x_\star) \rangle d^2(x_n, x_\star) + 2\gamma_{n+1}^2 \zeta \|\nabla f_{n+1}(x_n)\|^2 d^2(x_n, x_\star) \\
&\quad\quad + 4\gamma_{n+1}^3 \langle \nabla f_{n+1}(x_n), \mathrm{Exp}_{x_n}^{-1}(x_\star) \rangle \zeta \|\nabla f_{n+1}(x_n)\|^2.
\end{aligned}
$$

Taking conditional expectations and using Cauchy-Schwarz we have,

$$
\begin{aligned}
\mathbb{E}[d^4(x_{n+1}, x_\star) | \mathcal{F}_n] &\leq d^4(x_n, x_\star) + 2(2 + \zeta)\gamma_{n+1}^2 \mathbb{E}[\|\nabla f_{n+1}(x_n)\|^2 | \mathcal{F}_n] d^2(x_n, x_\star) \\
&\quad + \gamma_{n+1}^4 \zeta^2 \mathbb{E}[\|\nabla f_{n+1}(x_n)\|^4 | \mathcal{F}_n] + 4\gamma_{n+1} \langle \nabla f(x_n), \mathrm{Exp}_{x_n}^{-1}(x_\star) \rangle d^2(x_n, x_\star) \\
&\quad + 4\gamma_{n+1}^3 \zeta \mathbb{E}[\|\nabla f_{n+1}(x_n)\|^3 | \mathcal{F}_n] d(x_n, x_\star).
\end{aligned}
$$

Using that f is g-strongly convex (Assumption 7), the 4th-moment bound in Assumption 6, and Jensen's inequality we obtain,

$$\mathbb{E}[d^4(x_{n+1}, x_\star)|\mathcal{F}_n] \leq (1 - 4\gamma_{n+1}\mu)d^4(x_n, x_\star) + 2(2 + \zeta)\gamma_{n+1}^2 d^2(x_n, x_\star)v^2$$
$$+ \gamma_{n+1}^4 \zeta^2 v^4 + 4\gamma_{n+1}^3 \zeta v^3 d(x_n, x_\star).$$

Using the upper bound $4\gamma_{n+1}^3 \zeta v^3 d(x_n, x_\star) \leq 2\gamma_{n+1}^4 \zeta^2 v^4 + 2\gamma_{n+1}^2 v^2 d(x_n, x_\star)^2$, we have,

$$\mathbb{E}[d^4(x_{n+1}, x_\star)|\mathcal{F}_n] \leq (1 - 4\gamma_{n+1}\mu)d^4(x_n, x_\star) + 2(3 + \zeta)v^2 \gamma_{n+1}^2 d^2(x_n, x_\star) + 3\gamma_{n+1}^4 \zeta^2 v^4. \quad (13)$$

Now let us define, $a_n = \mathbb{E}[d^4(x_{n+1}, x_\star)]$, $b_n = \mathbb{E}[d^2(x_{n+1}, x_\star)]$ and $u_n = a_n + \frac{2(3+\zeta)v^2}{\mu}\gamma_{n+1}b_n$. Taking the full expectation of Eq. (13), we can bound u_{n+1} as,

$$\begin{aligned}
u_{n+1} &\leq (1 - \gamma_{n+1}\mu)u_n + 3\gamma_{n+1}^4 \zeta^2 v^4 + \frac{2(3+\zeta)\zeta v^4}{\mu}\gamma_{n+1}^3 + 2(3+\zeta)v^2 \gamma_{n+1}^2 b_n \\
&\quad -(1 - \gamma_{n+1}\mu)\frac{2(3+\zeta)v^2}{\mu}\gamma_{n+1}b_n + (1 - 2\gamma_{n+1}\mu)\frac{2(3+\zeta)v^2}{\mu}\gamma_{n+1}b_n.
\end{aligned}$$

Noting that $2(3+\zeta)v^2 \gamma_{n+1}^2 b_n - (1 - \gamma_{n+1}\mu)\frac{2(3+\zeta)v^2}{\mu}\gamma_{n+1}b_n + (1 - 2\gamma_{n+1}\mu)\frac{2(3+\zeta)v^2}{\mu}\gamma_{n+1}b_n = 0$, we obtain the simple upper-bound on u_{n+1},

$$u_{n+1} \leq (1 - \gamma_{n+1}\mu)u_n + 3\gamma_{n+1}^4 \zeta^2 v^4 + \frac{2(3+\zeta)\zeta v^4}{\mu}\gamma_{n+1}^3.$$

Using $(1 - x) \leq \exp(-x)$ for $x \in \mathbb{R}$, we have,

$$u_{n+1} \leq \exp(-\gamma_{n+1}\mu)u_n + 3\gamma_{n+1}^4 \zeta^2 v^4 + \frac{2(3+\zeta)\zeta v^4}{\mu}\gamma_{n+1}^3.$$

We can unroll this recursion as before,

$$u_n \leq \exp\left(-\mu\sum_{i=1}^n \gamma_i\right)u_0 + \sum_{i=1}^n [3\zeta^2 v^4 \gamma_i^4 + \frac{2(3+\zeta)\zeta v^4}{\mu}\gamma_i^3] \prod_{k=i+1}^n \exp(-\mu\gamma_k).$$

Proceeding exactly as in the proof of the bound on the second moment, we may bound u_n as

$$u_n \leq \frac{2(3+\zeta)\zeta v^4}{\mu}\gamma_{\lfloor\frac{n}{2}\rfloor}^2 + \text{exponentially small remainder terms.}$$

The conclusion follows. ∎

Appendix E. Streaming PCA

Given a sequence of i.i.d. symmetric random matrices $H_n \in \mathbb{R}^{d \times d}$ such that $\mathbb{E}H_n = H$, in the streaming k-PCA problem we hope to approximate the subspace of the top k eigenvectors. Let us denote by $\{\lambda_i\}_{1 \leq i \leq d}$ the eigenvalues of H sorted in decreasing order. Sharp convergence rates and finite sample guarantees for streaming PCA (with $k = 1$) were first obtained by Jain et al. (2016); Shamir (2016a) using the randomized power method (with and without a positive eigengap $\lambda_1 - \lambda_2$). When $\lambda_1 > \lambda_2$, Jain et al. (2016) showed with proper choice of learning rate $\eta_i \sim \tilde{O}\left(\frac{1}{(\lambda_1-\lambda_2)i}\right)$, an ϵ-approximation to the top eigenvector v_1 of H could be found in $\tilde{O}(\frac{\lambda_1}{(\lambda_1-\lambda_2)^2}\frac{1}{\epsilon})$

iterations with constant probability. In the absence of an eigengap, Shamir (2016a) showed a slow rate of convergence $\tilde{O}(\lambda_1/\sqrt{n})$ for the objective function using a step-size choice of $O(1/\sqrt{n})$. Allen-Zhu and Li (2017); Shamir (2016b)[5] later extended these results to the more general streaming k-PCA setting (with $k \geq 1$).

The aforementioned results are quite powerful—because they are *global* convergence results. In particular, they hold for any random initialization and do not require an initialization very close to the optima. In contrast, our framework only provides local results.

However, streaming k-PCA still provides an instructive and practically interesting application of our iterate-averaging framework. An important theme in the following analysis is to leverage to underlying Riemannian structure of the k-PCA problem as a Grassmann manifold.

Throughout this section we will assume that the stream of matrices satisfies the bound $\|H_n\| \leq 1$ a.s.

E.1. Grassmann Manifolds

Preliminaries: We begin by first reviewing the geometry of the Grassmann manifold and proving several useful auxiliary lemmas. We denote the Grassmann manifold $\mathcal{G}_{d,k}$, which is the set of the k-dimensional subspaces of a d-dimensional Euclidean space. Recalling the Stiefel manifold is the submanifold of the orthonormal matrices $\{X \in \mathbb{R}^{d \times k}, X^\top X = I_k\}$), $\mathcal{G}_{d,k}$ can be viewed as the Riemannian quotient manifold of the Stiefel manifold where two matrices are identified as equivalent when their columns span the same subspace. Finally, $\mathcal{G}_{d,k}$ can also be identified with the set of rank k projection matrices $\mathcal{G}_{d,k} = \{X \in \mathbb{R}^{d \times d} \text{ s.t. } X^\top = X, X^2 = X, \text{tr}(X) = k\}$ (see, e.g., Edelman et al., 1998; Absil et al., 2004, for further details). We will use $\mathbf{X}$ to denote an element of $\mathcal{G}_{d,k}$, and X a corresponding member of the equivalence class associated to $\mathbf{X}$, which belongs to the Stiefel manifold. Further, the tangent space at that point $\mathbf{X}$ is given by $T_\mathbf{X}\mathcal{G}_{d,k} = \{Y \in \mathbb{R}^{d \times k}, Y^\top X = 0\}$. In the following we identify $\mathbf{X}$ and X when it is clear from the context.

For our present purposes, we consider the (second-order) retraction map

$$R_X(V) = (X + V)[(X + V)^\top(X + V)]^{-1/2}, \tag{14}$$

which is projection-like mapping onto $\mathcal{G}_{d,k}$. Note we implicitly extend R_X to all matrices in $\mathbb{R}^{d \times d}$, and do not consider it only defined on the tangent space $T_X\mathcal{G}_{d,k}$. For $V \in T_X\mathcal{G}_{d,k}$, we still have $R_X(V) = (X + V)[I_k + V^\top V]^{-1/2}$. If $X^\top Y$ is invertible, then a short computation shows that $R_X^{-1}(Y) = (I - X^\top X X)Y(X^\top Y)^{-1}$ with $\|R_X^{-1}(Y)\|_F^2 = \text{tr}[(X^\top Y X Y^\top)^{-1} - I]$. As we argue next, $\|R_X^{-1}(Y)\|_F^2$ is in fact locally equivalent to the induced, Frobenius norm $d_F(X, Y) = 2^{-1/2}\|XX^\top - YY^\top\|_F^2$.

Measuring Distance on $\mathcal{G}_{d,k}$: It will be useful to have several notions of distance defined on $\mathcal{G}_{d,k}$ between two representative elements X and Y. Let θ_i for $i = 1, \ldots, k$ denote the principal angles between the two subspaces spanned by the columns of X and Y, i.e. $U\cos(\Theta)V^\top$ is the singular value decomposition (SVD) of $X^\top Y$, where Θ is the diagonal matrix of principal angles, and θ the k-vector formed by θ_i. Our first distance of interest will be the arc length (or geodesic distance):

$$d_A(X, Y) = \|\theta\|_2 = \left\|\text{Exp}_X^{-1}(Y)\right\|_2,$$

---

5. Shamir (2016b) does not directly address the streaming setting but his result can be extended to the streaming setting, as remarked by Allen-Zhu and Li (2017).

while the second will be the projected, Frobenius norm:

$$d_F(X,Y) = \|\sin\theta\|_2 = 2^{-1/2}\left\|XX^\top - YY^\top\right\|_F.$$

The distance $d_F(\cdot,\cdot)$ is induced by embedding $\mathcal{G}_{d,k}$ in Euclidean space $\mathbb{R}^{d\times k}$ and inheriting the corresponding Frobenius norm. Lastly, we will also consider the pseudo-distance induced by the retraction $\|R_X^{-1}(Y)\|_F$. Conveniently, for small principal angles we can show these quantities are asymptotically equivalent,

Lemma 16 *Let θ denote the k-vector of principal angles between the subspaces spanned by X and Y. If $\|\theta\|_\infty \leq \frac{\pi}{4}$ then:*

$$\frac{\pi}{2}\|R_X^{-1}(Y)\|_F \geq \frac{\pi}{2}d_F(X,Y) \geq d_A(X,Y) \geq \frac{1}{\sqrt{2}}\|R_X^{-1}(Y)\|_F.$$

Proof In fact $d_A(X,Y) \geq d_F(X,Y)$ uniformly over θ which follows $\sin x \leq x$ for all x. Using the elementary inequality $|\frac{\sin x}{x}| \geq \frac{2}{\pi}$ for $x \in [-\pi/2, \pi/2]$ we immediately obtain that $d_A(X,Y) \leq \frac{\pi}{2}d_F(X,Y)$ if $\|\theta\|_\infty \leq \pi/2$. A direct computation shows:

$$\|R_X^{-1}(Y)\|_F^2 = \mathrm{tr}\Big[(X^\top YY^\top X)^{-1} - I\Big] = \mathrm{tr}\Big[\cos(\Theta)^{-2} - I\Big] = \mathrm{tr}\big[(I - \sin^2(\Theta))^{-1} - I\big].$$

Using that for $x \in (0, \frac{1}{\sqrt{2}})$, $\frac{1}{1-x} \geq 1 + x$ and $\frac{1}{1-x} \leq 1 + 2x$, we obtain for $\|\theta\|_\infty < \pi/4$ that,

$$d_F^2(X,Y) \leq \|R_X^{-1}(Y)\|_F^2 \leq 2d_F^2(X,Y),$$

concluding the argument. ∎

The local equivalence between these quantities will prove useful for relating various algorithms for the streaming k-PCA problem.

Streaming k-PCA in $\mathcal{G}_{d,k}$: Within the geometric framework we can cast the k-PCA problem as minimizing the Rayleigh quotient, $f(X) = -\frac{1}{2}\mathrm{tr}\big[X^\top HX\big]$, over $\mathcal{G}_{d,k}$ as

$$\min_{X\in\mathcal{G}_{d,k}} -\frac{1}{2}\mathrm{tr}\Big[X^\top HX\Big].$$

Edelman et al. (1998) show the (Riemannian) gradient of f[6] is given by

$$\nabla f(X) = -(I - X^\top X)HX.$$

Similarly, the Hessian operator is characterized by the property that if $\Delta \in T_X\mathcal{G}_{d,k}$, then

$$\nabla^2 f(X)[\Delta] = \Delta X^\top HX - (I - XX^\top)H\Delta.$$

It is worth noting that $\nabla f(X) = 0$ for X spanning any subspace or eigenvector of H, but that $\nabla^2 f(X)$ is only positive-definite at the optimum $X_\star$.

---

6. Note that on embedded manifold the Riemannian gradient of a function is given by projection of its Euclidean gradient in the tangent space of the manifold at a point.

E.2. Algorithms for streaming k-PCA

We now have enough background to describe several, well-known iterative algorithms for the streaming k-PCA problem and elucidate their relationship to Riemannian SGD.

Randomized power method (Oja and Karhunen, 1985): corresponds to SGD on the Rayleigh quotient (with step size γ_n) over Euclidean space followed by a projection,

$$X_n = (I_d + \gamma_n H_n)X_{n-1}\big[X_{n-1}^\top(I_d + \gamma_n H_n)^2 X_{n-1}\big]^{-1/2}.$$

Oja iteration (Oja, 1982): corresponds to a first-order expansion in γ_n of the previous randomized power iteration,

$$X_n = X_{n-1} + \gamma_n(I - X_{n-1}X_{n-1}^\top)H_n X_{n-1}.$$

Yang iteration (Yang, 1995): corresponds to a symmetrization of the Oja iteration,

$$X_n = X_{n-1} + \gamma_n(2H_n - X_{n-1}X_{n-1}^\top H_n - H_n X_{n-1}X_{n-1}^\top)X_{n-1}.$$

In the special case that $H_n = h_n h_n^\top$, the Yang iteration can also be related to the unconstrained stochastic optimization of the function $X \mapsto \mathbb{E}\|h_n - XX^\top h_n\|_F^2$.

(Stochastic) Gradient Descent over $\mathcal{G}_{d,k}$ (Bonnabel, 2013): corresponds to directly optimizing the Rayleigh quotient over $\mathcal{G}_{d,k}$ by equipping SGD with either the exponential map Exp or the aforementioned retraction R,

$$X_n = R_{X_{n-1}}\big(\gamma_n(I - X_{n-1}X_{n-1}^\top)H_n X_{n-1}\big).$$

We are now in position to show that for the present problem Riemannian SGD, the randomized power method, and Oja's iteration are equivalent updates up to $O(\gamma_n^2)$ corrections. First note that the randomized power method (with the aforementioned choice of retraction R) can be written as:

$$R_X(\gamma_n H_n X), \tag{15}$$

bearing close resemblance to the Riemannian SGD update,

$$R_X(\gamma_n(I - XX^\top)H_n X). \tag{16}$$

The principal difference between Eq. (16) and Eq. (15) is that in the randomized power method, the Euclidean gradient is used instead of the Riemannian gradient. In the following Lemma we argue that both of the updates in Eq. (16) and Eq. (15) can be directly approximated by the Oja iteration, $X_{n+1} = X_n + \gamma_{n+1}(I - X_n X_n^\top)H_{n+1}X_n + O(\gamma_{n+1}^2)$ up to 2nd-order terms and hence that $\|R_{X_\star}^{-1}(X_{n+1})) - R_{X_\star}^{-1}(R_{X_n}(\gamma_n \nabla f_{n+1}(X_n)))\|_F^2 = O(\gamma_n^2)$. Therefore, a direct modification of Lemma 4 shows that the iterates in Eq. (15) generated from the randomized power method may also be linearized in the tangent space.

Lemma 17 *Alternatively, let $X_n = R_{X_{n-1}}(\gamma_n(I - X_{n-1}X_{n-1}^\top)H_n X_{n-1})$ denote the Riemannian SGD update (equipped with second-order retraction R) or $X_n = R_{X_{n-1}}(H_n X_{n-1})$, the randomized power update. Then both updates satisfy,*

$$X_n = X_n + \gamma_n(I - X_{n-1}X_{n-1}^\top)H_n X_{n-1} + O(\gamma_n^2),$$

and hence are equivalent to the Oja update up to $O(\gamma_n^2)$ terms.

Proof The computation for both the randomized power method and the Riemannian SGD (equipped with the retraction R) are straightforward. For the randomized power method we have that,

$$
\begin{aligned}
X_n &= (I_d + \gamma_n H_n) X_{n-1} \big[X_{n-1}^\top (I_d + \gamma_n H_n)^2 X_{n-1} \big]^{-1/2} \\
&= (I_d + \gamma_n H_n) X_{n-1} \big[I_d + 2\gamma_n X_{n-1}^\top H_n X_{n-1} \big]^{-1/2} + O(\gamma_n^2) \\
&= (I_d + \gamma_n H_n) X_{n-1} \big[I_d - \gamma_n X_{n-1}^\top H_n X_{n-1} \big] + O(\gamma_n^2) \\
&= X_{n-1} + \gamma_n [I_d - X_{n-1} X_{n-1}^\top] H_n X_{n-1} + O(\gamma_n^2).
\end{aligned}
$$

An identical computation shows the same result for the Riemannian SGD update,

$$
\begin{aligned}
X_n &= \big(I_d + \gamma_n [I_d - X_{n-1} X_{n-1}^\top] H_n \big) X_{n-1} \big[X_{n-1}^\top (I_d + \gamma_n [I_d - X_{n-1} X_{n-1}^\top] H_n)^2 X_{n-1} \big]^{-1/2} \\
&= X_{n-1} + \gamma_n [I_d - X_{n-1} X_{n-1}^\top] H_n X_{n-1} + O(\gamma_n^2).
\end{aligned}
$$

∎

Since these two algorithms are identical up to $O(\gamma_n^2)$ corrections we can directly show they will have the same linearization in $T_{X_\star}\mathcal{M}$.

Lemma 18 *Let, $\Delta_n = R_{X_\star}^{-1}(X_n)$, where X_n is obtained from one iteration of the randomized power method, $X_n = R_{X_n}(\gamma_n H_n X_n)$. Then Δ_n obeys,*

$$
\Delta_n = \Delta_{n-1} - \gamma_n \nabla^2 f(x_\star) \Delta_{n-1} + \gamma_n (\varepsilon_n + \xi_n + e_n) + O(\gamma_n^2).
$$

Proof Let $\dot{\Delta}_n = R_{X_\star}^{-1}(Y_n)$, where Y_n is obtained from one iteration of Riemannian SGD (with the aforementioned second-order retraction R), $Y_n = R_{X_n}(\gamma_n \nabla f_n(X_n))$. From Lemma 6 we have that $\dot{\Delta}_n$ may be linearized as

$$
\dot{\Delta}_n = \Delta_{n-1} - \gamma_n \nabla^2 f(x_\star) \Delta_{n-1} + \gamma_n (\varepsilon_n + \xi_n + e_n)
$$

Defining $\square = X_{n-1} + \gamma_n [I_d - X_{n-1} X_{n-1}^\top] H_n X_{n-1}$ for the Oja update, we can then show the randomized power method satisfies,

$$
R_{X_\star}^{-1}(X_n) = Y_n (X_\star^\top Y_n)^{-1} = [\square + O(\gamma_n^2)](X_\star^\top(\square + O(\gamma_n^2)))^{-1} = \square(X_\star^\top \square)^{-1} + O(\gamma_n^2)
$$

This allows us to directly bound the difference between $\dot{\Delta}_n$ and Δ_n using Lemma 17,

$$
\| \dot{\Delta}_n - \Delta_n \|_F = \| R_{X_\star}^{-1}(X_n) - R_{X_\star}^{-1}(Y_n) \|_F = \| Y_n (X_\star^\top Y_n)^{-1} - X_n (X_\star^\top X_n)^{-1} \|_F = O(\gamma_n^2).
$$

Hence the randomized power method iterate X_n obeys the linearization,

$$
\Delta_n = \Delta_{n-1} - \gamma_n \nabla^2 f(x_\star) \Delta_{n-1} + \gamma_n (\varepsilon_n + \xi_n + e_n) + O(\gamma_n^2),
$$

and falls within the scope of our framework. ∎

E.3. Algorithms for Streaming Averaging

Just as there are several algorithms to compute the primal iterate X_n, their are several reasonable ways to compute the streaming average $\tilde{X}_n$. Our general framework directly considers $\tilde{X}_n = R_{\tilde{X}_{n-1}}[\frac{1}{n}R_{\tilde{X}_{n-1}}^{-1}(X_n)]$ which leads to the following update rule:

$$\tilde{X}_n = R_{\tilde{X}_{n-1}}\left[\frac{1}{n}(I - \tilde{X}_{n-1}\tilde{X}_{n-1}^\top)X_n[\tilde{X}_{n-1}X_n]^{-1}\right]. \tag{17}$$

As described in Section C.2.3, the streaming average is an approximation to a corresponding global Riemannian average (which is intractable to compute). Hence, it is reasonable to consider other global Riemannian averages that the streaming average approximates. For instance, the update rule in Eq. (17) is naturally motivated by the global minimization of $X \mapsto \sum_{i=1}^n \|R_X^{-1}(X_i)\|^2$. Considering instead the distance d_F, and attempting to minimizing $X \mapsto \sum_{i=1}^n d_F^2(X, X_i)$, suggests a different averaging scheme. A short computation shows the aforementiond problem can be rewritten as the maximization of the function $2\operatorname{tr}\left[X^\top(\sum_{i=1}^n X_i X_i^\top)X\right]$ and is therefore precisely equivalent to the k-PCA problem. With this in mind, we can directly use the randomized power method to compute the streaming average of the iterates. This leads to the different update rule:

$$\tilde{X}_n = R_{\tilde{X}_{n-1}}\left[\frac{1}{n}X_n X_n^\top \tilde{X}_{n-1}\right], \tag{18}$$

which is exactly one step of randomized power method iteration to compute the first k eigenvectors of the matrix $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ with step size $\gamma_n = \frac{1}{n}$.

Using similar computations to those in Lemma 17 and 18, it can be shown that these iterates are equivalent to those in Eq. (17) since they have a similar linearzation in $T_{X_\star}\mathcal{G}_{d,k}$.

Lemma 19 *Let* $\tilde{X}_n = R_{\tilde{X}_{n-1}}\left[\frac{1}{n}X_n X_n^\top \tilde{X}_{n-1}\right]$ *and* $\tilde{\Delta}_n = R_{X_\star}^{-1}(\tilde{X}_n)$. *Then* $\tilde{X}_n$ *obeys*

$$\tilde{\Delta}_n = \tilde{\Delta}_{n-1} + \frac{1}{n}\left[\tilde{\Delta}_{n-1} + \Delta_n\right] + \frac{1}{n}O\left(\|\Delta_{n-1}\|^2 + \|\tilde{\Delta}_n\|^2 + \frac{1}{n^2}\right).$$

Proof Following a similar approach as in the proof of Lemma 4 we can show that

$$\tilde{\Delta}_n = \tilde{\Delta}_{n-1} + \frac{1}{n+1}[I - X_\star X_\star^\top]X_n X_n^\top \tilde{X}_{n-1} + O(\frac{1}{n^2}).$$

Then a direct expansion shows that for all $\Delta = R_{X_\star}^{-1}(X)$,

$$X = R_{X_\star}(\Delta) = (X_\star + \Delta)[I + \Delta^\top\Delta]^{-1/2} = X_\star + \Delta + O(\|\Delta\|^2),$$

so we obtain (using $\|\Delta_n^\top\tilde{\Delta}_{n-1}\| = O(\|\Delta_n\|^2 + \|\tilde{\Delta}_{n-1}\|^2)$) that

$$[I - X_\star X_\star^\top]X_n X_n^\top \tilde{X}_{n-1} = \Delta_n + \tilde{\Delta}_{n-1} + O(\|\Delta_n\|^2 + \|\tilde{\Delta}_{n-1}\|^2)).$$

∎

This manner of iterate averaging is interesting, not only due to its simplicity, but due to its close connection to the primal, randomized power method. In fact, this averaging method allows us to interpret the aforementioned, streaming, averaged PCA algorithm as a preconditioning method.

Consider the case where we hope to compute the principal k-eigenspace of a poorly conditioned matrix H. The aforementioned streaming, averaged PCA algorithm can be interpreted as the composition of two stages. First, running n steps of the randomized power method with step size $\gamma_n \propto 1/\sqrt{n}$, to produce a well-conditioned matrix $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, and then using the randomized power method with step size $1/n$ to compute the average of the points X_i (which is efficient since the eigengap of $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ is large). The intuition for this first step is formalized in the following remark,

Remark 20 *Assume the sequence of iterates $\{X\}_{i=0}^n$ satisfies $d_F(X_i, X_\star) = O(\sqrt{\gamma})$ for all $i = 1, \ldots, n$ then the eigengap of the averaged matrix $\tilde{X} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ satisfies $\tilde{\lambda}_k - \tilde{\lambda}_{k+1} \geq 1 - kO(\sqrt{\gamma})$ (where we denote by $\{\tilde{\lambda}_i\}_{1\leq i \leq d}$ the eigenvalues of $\tilde{X}$ sorted in decreasing order).*

Proof $X_i X_i^\top = X_\star X_\star^\top + \eta_i$ for η_i satisfying $\|\eta_i\|_F \leq O(\sqrt{\gamma})$ by the definition of the distance $d_F(\cdot, \cdot)$, so it follows that $\tilde{X}_n = X_\star X_\star^\top + \eta$ for η satisfying $\|\eta\|_F \leq O(\sqrt{\gamma})$ using the triangle inequality. Now using the Weyl inequalities (Horn and Johnson, 1990) we have that:

$$\lambda_k(\tilde{X}_n - \eta + \eta) \geq \lambda_k(\tilde{X}_n - \eta) + \lambda_d(\eta),$$

where we denote by $\lambda_k(M)$ the k-th largest eigenvalue of the matrix M. Moreover $\lambda_k(\tilde{X}_n - \eta) = 1$ and $|\lambda_d(\eta)| \leq \|\eta\|_F \leq O(\sqrt{\gamma})$ since the spectral norm is upper bounded by the Frobenius norm, so $\lambda_k(\tilde{X}_n) \geq 1 - O(\sqrt{\gamma})$. Now, recall that each $X_i \in \mathcal{G}_{d,k}$, so $\text{Tr}\left[\tilde{X}_n\right] = \frac{1}{n}\sum_{i=1}^n \text{Tr}\left[X_i X_i^\top\right] = k$ due to the normalization constraint $X_i X_i^\top = I_k$. Thus we must have that $\tilde{\lambda}_{k+1} \leq kO(\sqrt{\gamma})$ which implies $\tilde{\lambda}_k(\tilde{X}_n) - \tilde{\lambda}_{k+1}(\tilde{X}_n) = 1 - kO(\sqrt{\gamma})$. ∎

For these reasons we prefer to use Eq. (18) rather than Eq. (17) in our experiments and our presentation. It is worth noting that since these both iterations are equivalent up to $O(\gamma_n)$ corrections, they will (a) have the same theoretical guarantees in our framework and (b) perform similarly in practice.

E.4. Convergence Results

We are now ready to apply Theorem 3 to the streaming k-PCA problem, with only the tedious task of verifying our various technical assumptions. Since we only seek to derive a local convergence result, we will once again use Assumption 2 which stipulates the iterates X_n are restricted to $\mathcal{X}$ and that the map $X \mapsto \left\|R_{X_\star}^{-1}(X)\right\|_F^2$ is strongly-convex. Here we will take the set $\mathcal{X} = \{Y : d_F(X_\star, Y) \leq \delta\}$, for $\delta > 0$. As previously noted, if the retraction R was taken as the exponential map, the map $X \mapsto \left\|\text{Exp}_{X_\star}^{-1}(X)\right\|_F^2$ would always be retraction strongly convex locally in a ball around y whose radius depends on the curvature, as explained by Afsari (2011). However, we can also verify, that even when R is the aforementioned projection-like retraction it is locally retraction strongly convex,

Remark 21 *Let $k = 1$ and take R as the second-order retraction defined in Eq. (14). Then there exists a constant $\delta > 0$ such that on the set $\mathcal{X} = \{Y : \left|X_\star^\top Y\right| \geq 1 - \delta\}$, $X \mapsto \left\|R_{X_\star}^{-1}(X)\right\|_F^2$ is retraction strongly convex.*

Proof For notational convenience, let ϵ be defined so we consider $\mathcal{Y} = \{Y : \left|X_\star^\top Y\right| \geq 1 - \epsilon\}$. It suffices to show that for $\|X_\star\|_2 = 1$, all Y such that $\|Y\|_2 = 1$ and $\left|X_\star^\top Y\right| \geq 1 - \epsilon$, and all V such that $\|V\|_2 = 1$, the map $g(t) : t \mapsto \left\|R_{X_\star}^{-1}(R_Y(tV))\right\|_F^2$ is convex in t. For $k = 1$, using

the previous formulas for the retraction and its inverse norm, we can explicitly compute the map as $g(t) = \frac{1+t^2}{(X_\star^\top(Y+tV))^2} - 1$. We then have that $g''(t) = \frac{2(3(X_\star V)^2 - 2t(X_\star^\top v)(X_\star^\top y) + (X_\star^\top y)^2)}{(t(X_\star^\top v) + (X_\star^\top y))^4}$. It suffices to show that $3(X_\star^\top V)^2 + (X_\star^\top Y)^2 > 2t(X_\star^\top V)(X_\star^\top Y)$ for all t such that $R_Y(tV) \in \mathcal{Y}$. Since $V \in T_Y\mathcal{M}$ we must have that $V^\top Y = 0$ which implies that $(V^\top X_\star)^2 = (V^\top(Y - X_\star))^2 \leq 2 - 2Y^\top X_\star \leq 2\epsilon$. Hence we have that $2|(X_\star^\top V)(X_\star^\top Y)| \leq 4\epsilon$. On the otherhand, we have that $3(X_\star^\top V)^2 + (X_\star^\top Y)^2 > (X_\star^\top Y)^2 > (1 - \epsilon)^2$. Thus the statement is satisfied for all $|t| \leq \frac{(1-\epsilon)^2}{4\epsilon}$. Using the local equivalence of distances for PCA (Lemma 16), the conclusion holds for some constant $\delta > 0$. ∎

A similar, but lengthier linear algebra computation, also shows the result for $k > 1$.

Theorem 3 requires several other assumptions (namely Assumptions 4 and 6) which are (surprisingly) tedious to verify even in the simple case when f is the quadratic Rayleigh quotient. However, we note Lemma 5, for the streaming k-PCA problem, can also be derived from first principles, circumventing the need to directly check these assumptions.

Remark 22 *If f is the Rayleigh quotient, the conclusion of Lemma 5 follows (without having to directly verify Assumptions 4 and 6), when the stream of matrices satisfies the almost sure bound $\|H_n\| \leq 1$.*

Proof As in the proof of Lemma 19, we use

$$X_n = R_{X_\star}(\Delta_n) = (X_\star + \Delta_n)[I + \Delta_n^\top \Delta_n]^{-1/2} = X_\star + \Delta_n + O(\|\Delta_n\|^2),$$

to simplify $\nabla f_{n+1}(X_n)$. This yields

$$
\begin{aligned}
\nabla f_{n+1}(X_n) &= (I - X_n X_n^\top)H_{n+1}X_n + O(\|\Delta_n\|^2) \\
&= (I - (X_\star + \Delta_n)(X_\star + \Delta_n)^\top)H_{n+1}(X_\star + \Delta_n) \\
&= (I - X_\star X_\star^\top)H_{n+1}X_\star + (I - X_\star X_\star^\top)H_{n+1}\Delta_n - \Delta_n X_\star^\top H_{n+1}X_\star \\
&\quad + X_\star \Delta_n^\top H_{n+1}X_\star + O(\|\Delta_n\|^2).
\end{aligned}
$$

Upon projecting back into $T_{X_*}\mathcal{M}$, the term $X_\star \Delta_n^\top H_n X_\star$ vanishes and we obtain,

$$(I - X_\star X_\star^\top)\nabla f_{n+1}(X_n) = \nabla^2 f(X_\star)\Delta_n + \nabla f_{n+1}(X_*) + \xi_{n+1},$$

with $\xi_{n+1} = (I - X_\star X_\star^\top)H_{n+1}\Delta_n - \Delta_n X_\star^\top H_{n+1}X_\star$ satisfying $\mathbb{E}[\xi_{n+1}|\mathcal{F}_n] = 0$ and $\mathbb{E}[\|\xi_{n+1}\|^2|\mathcal{F}_n] = O(\|\Delta_n\|^2)$ since $\mathbb{E}[\|H - H_{n+1}\|^2|\mathcal{F}_n]$ is bounded (recall that we assume $\|H_n\| \leq 1$ a.s.). ∎

Using results from the work of Allen-Zhu and Li (2017) and Shamir (2016b) we can now argue under appropriate conditions that the randomized power method for the streaming k-PCA problem will converge in expectation to a neighborhood of $X_\star$. Namely,

Lemma 23 (Allen-Zhu and Li (2017) and Shamir (2016b)) *Let X_n denote the iterates of the randomized power method evolving as in Eq. (15). If Assumption 2 holds for $\mathcal{X}$ defined above with $\delta < 1/4$ then,*

$$\mathbb{E}[d_F^2(X_\star, X_n)] = O(\gamma_n^2).$$

Proof This a direct adaptation of Lemma 10 by Shamir (2016b) as explained by Allen-Zhu and Li (2017, Section 3). ∎

Finally, we are able to present the proof of Theorem 3. Since the assumptions of Lemma 23 are satisfied by assumption, and we have asymptotic equivalence of distances from Lemma 16, Assumption 1 is satisfied. Further, since Lemmas 18 and 19 show the linearized process of the randomized power iterates is equivalent to that of Riemannian SGD up to $O(\gamma_n^2)$ corrections, distributional convergence immediately follows from these Lemmas and Theorem 1. Lastly, we can compute the asymptotic variance.

We first compute the inverse of the Hessian of f. Let us consider the following basis of $T_{X_\star}\mathcal{G}_{d,k}$, $\{v_i e_j^\top\}_{k<i\leq d;j\leq k}$, where we denote by v_i the eigenvector of H associated with the eigenvalue λ_i and e_j the jth standard basis vector. Indeed $\{v_i e_j^\top\}_{k<i\leq d;j\leq k}$ is a linear independent set since $\{v_i e_j^\top\}_{i\leq d;j\leq k}$ is a basis of $\mathbb{R}^{d,k}$. Moreover for $k<i\leq d$, $X_\star^\top v_i e_j^\top = 0$ so $v_i e_j^\top \in T_{X_\star}\mathcal{G}_{d,k}$. We conclude this set is a basis from a dimension count. We now compute the projection of the Hessian on this basis

$$
\begin{aligned}
\nabla^2 f(X_\star)[v_i e_j^\top] &= v_i e_j^\top X_\star^\top H X_\star - H v_i e_j^\top \\
&= v_i e_j^\top \operatorname{diag}(\lambda_1,\ldots,\lambda_k) - \lambda_i v_i e_j^\top \\
&= (\lambda_j - \lambda_i) v_i e_j^\top.
\end{aligned}
$$

Therefore

$$
[\nabla^2 f(X_\star)]^{-1}[v_i e_j^\top] = \frac{v_i e_j^\top}{\lambda_j - \lambda_i}.
$$

We now reparametrize $\tilde{H}_n = H^{-1/2} H_n H^{-1/2}$ such that $\mathbb{E}[\tilde{H}_n] = I$. Thus

$$
\begin{aligned}
\nabla f_n(X_\star) &= (I - X_\star X_\star^\top) H_n X_\star \\
&= (I - X_\star X_\star^\top) H^{1/2} \tilde{H}_n H^{1/2} X_\star \\
&= \sum_{i=k+1}^{d} \sqrt{\lambda_i} v_i v_i^\top \tilde{H}_n \sum_{j=1}^{k} \sqrt{\lambda_j} v_j e_j^\top \\
&= \sum_{j=1}^{k} \sum_{i=k+1}^{d} \sqrt{\lambda_i \lambda_j} [v_i^\top \tilde{H}_n v_j] v_i e_j^\top.
\end{aligned}
$$

This yields

$$
[\nabla^2 f(X_\star)]^{-1} \nabla f_n(X_\star) = \sum_{j=1}^{k} \sum_{i=k+1}^{d} \frac{\sqrt{\lambda_i \lambda_j}}{\lambda_j - \lambda_i} [v_i^\top \tilde{H}_n v_j] v_i e_j^\top.
$$

And the asymptotic covariance becomes,

$$
[\nabla^2 f(X_\star)]^{-1} \mathbb{E}[\nabla f_n(X_\star) \nabla f_n(X_\star)^\top][\nabla^2 f(X_\star)]^{-1} =
$$

$$
\sum_{j'=1}^{k} \sum_{i'=k+1}^{d} \sum_{j=1}^{k} \sum_{i=k+1}^{d} \frac{\sqrt{\lambda_i \lambda_j} \cdot \sqrt{\lambda_{i'} \lambda_{j'}}}{(\lambda_j - \lambda_i) \cdot (\lambda_{j'} - \lambda_{i'})} \mathbb{E}\left[\left(v_i^\top \tilde{H}_n v_j\right)\left(v_{i'}^\top \tilde{H}_n v_{j'}\right)\right] (v_i e_j^\top) \otimes (v_{i'} e_{j'}^\top).
$$

It is interesting to note that the tensor structure of $C_{ii'jj'}$ significantly simplifies if we have that $H_n = h_n h_n^\top$ for $h_n \sim \mathcal{N}(0, \Sigma)$ – that H_n is comprised of a rank-one stream of Gaussians. Recall we have $\tilde{H}_n = H^{-1/2} H_n H^{-1/2} = H^{-1/2} x_n x_n^\top H^{-1/2} = x_n'(x_n')^\top$. So for a rank-one stream,

$$C_{ii'jj'} = \mathbb{E}\left[\left(v_i^\top \tilde{H}_n v_j\right)\left(v_{i'}^\top \tilde{H}_n v_{j'}\right)\right] = \mathbb{E}\left[\langle v_i, x_n'\rangle\langle v_{i'}, x_n'\rangle\langle v_j, x_n'\rangle\langle v_{j'}, x_n'\rangle\right].$$

Since $x_n' \sim \mathcal{N}(0, I_d)$ and the law of a jointly multivariate normal random variable is invariant under an orthogonal rotation, the joint distribution of the vector $[\langle v_i, x_n'\rangle, \langle v_{i'}, x_n'\rangle, \langle v_j, x_n'\rangle, \langle v_{j'}, x_n'\rangle] \sim \mathcal{N}(0, I_4)$ for $i \neq i'$ and $j \neq j'$. So the only non-vanishing terms become,

$$C_{ii'jj'} = \delta_{ii'}\delta_{jj'},$$

and the asymptotic covariance reduces to,

$$C = \sum_{j=1}^{k} \sum_{i=k+1}^{d} \frac{\lambda_i \lambda_j}{(\lambda_j - \lambda_i)^2}(v_i e_j^\top) \otimes (v_i e_j^\top).$$

This is precisely the same asymptotic variance given by Reiß and Wahl (2016) and matches the lower bound of Cai et al. (2013) obtained for the (Gaussian) spiked covariance model. However, formally, our result requires the H_n to be a.s. bounded.

Appendix F. Additional Experiments: Numerical Counterexample to Oja Convergence with Constant Step Size

In this section we present an additional experiment which shows empirically that averaged SGD with constant step-size does not always converge for the streaming k-PCA problem. We consider $d = 2$ and a covariance matrice H with random eigenvectors and two eigenvalues $\frac{1 \pm 1/\pi}{4}$. The noise distribution of the stream $\{h_n h_n^\top\}_{n \geq 0}$ uses a more involved construction. Consider $\alpha_n \sim \mathcal{N}(-\pi/2, \pi^2/4)$ and $\beta_n \sim \mathcal{N}(\pi/4, \pi^2/16)$ and $\tau_n \sim \mathcal{B}(1/2)$. Then define θ_n as:

$$\theta_n = \tau_n \alpha_n + (1 - \tau_n)\beta_n,$$

and the stream h_n to be

$$h_n = \left[\frac{\cos(\theta_n)}{\sqrt{(1 - 1/\pi)/2}}, \frac{\sin(\theta_n)}{\sqrt{(1 + 1/\pi)/2}}\right].$$

Figure 4, compares the performance of averaged SGD with constant step size $\gamma = 1$ and the decreasing step size $\gamma_n = \frac{1}{\sqrt{n}}$. We see that with constant step size both SGD and averaged SGD do not converge to the true solution. SGD oscillates around the solution in ball of radius $\sim \gamma$ and averaged SGD does converge but not to the correct solution (although is still contained in a ball of radius $\sim 10^{-4}$ around the correct solution). On the other hand, SGD with decreasing step size behaves just as well as with a Gaussian data stream. SGD converges to the solution at the slow rate $O(1/\sqrt{n})$, while averaged SGD converges at the fast rate $O(1/n)$.

This interesting example shows that constant step size averaged SGD does not converge to the correct solution in all situations. However, it remains a open problem to investigate the convergence properties of constant step size SGD in the Gaussian case.
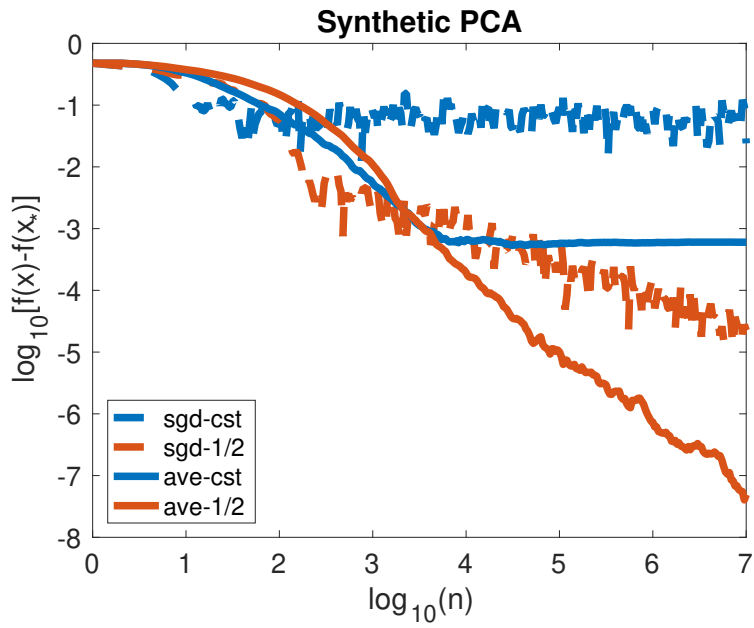
Figure 4: Counterexample for the convergence of averaged SGD with constant step size.