Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds

Authored by:

Suvrit Sra
Sashank J. Reddi
Hongyi Zhang

Abstract

We study optimization of finite sums of emph{geodesically} smooth functions on Riemannian manifolds. Although variance reduction techniques for optimizing finite-sums have witnessed tremendous attention in the recent years, existing work is limited to vector space problems. We introduce emph{Riemannian SVRG} (rsvrg), a new variance reduced Riemannian optimization method. We analyze rsvrg for both geodesically emph{convex} and emph{nonconvex} (smooth) functions. Our analysis reveals that rsvrg inherits advantages of the usual SVRG method, but with factors depending on curvature of the manifold that influence its convergence. To our knowledge, rsvrg is the first emph{provably fast} stochastic Riemannian method. Moreover, our paper presents the first non-asymptotic complexity analysis (novel even for the batch setting) for nonconvex Riemannian optimization. Our results have several implications; for instance, they offer a Riemannian perspective on variance reduced PCA, which promises a short, transparent convergence analysis.

1 Paper Body

We study the following rich class of (possibly nonconvex) finite-sum optimization problems: n

min
x2X ?M
f (x) ,
1X fi (x), n i=1
(1)

where (M, g) is a Riemannian manifold with the Riemannian metric g, and X ? M is a geodesically convex set. We assume that each fi : M ! R is geodesically L-smooth (see ?2). Problem (1) generalizes the fundamental machine learning problem of empirical risk minimization, which is usually cast in vector spaces, to a Riemannian setting. It also includes as special cases important problems

such as principal component analysis (PCA), independent component analysis (ICA), dictionary learning, mixture modeling, among others (see e.g., the related work section). The Euclidean version of (1) where M = Rd and g is the Euclidean inner-product has been the subject of intense algorithmic development in machine learning and optimization, starting with the classical work of Robbins and Monro [26] to the recent spate of work on variance reduction [10; 18; 20; 25; 28]. However, when (M, g) is a nonlinear Riemannian manifold, much less is known beyond [7; 38]. When solving problems with manifold constraints, one common approach is to alternate between optimizing in the ambient Euclidean space and ?projecting? onto the manifold. For example, two well-known methods to compute the leading eigenvector of symmetric matrices, power iteration and Oja?s algorithm [23], are in essence projected gradient and projected stochastic gradient algorithms. For certain manifolds (e.g., positive definite matrices), projections can be quite expensive to compute. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

An effective alternative is to use Riemannian optimization1 , which directly operates on the manifold in question. This mode of operation allows Riemannian optimization to view the constrained optimization problem (1) as an unconstrained problem on a manifold, and thus, to be ?projection-free.? More important is its conceptual value: viewing a problem through the Riemannian lens, one can discover insights into problem geometry, which can translate into better optimization algorithms. Although the Riemannian approach is appealing, our knowledge of it is fairly limited. In particular, there is little analysis about its global complexity (a.k.a. non-asymptotic convergence rate), in part due to the difficulty posed by the nonlinear metric. Only very recently Zhang and Sra [38] developed the first global complexity analysis of batch and stochastic gradient methods for geodesically convex functions. However, the batch and stochastic gradient methods in [38] suffer from problems similar to their vector space counterparts. For solving finite sum problems with n components, the full-gradient method requires n derivatives at each step; the stochastic method requires only one derivative but at the expense of slower O(1/?2) convergence to an ?-accurate solution. These issues have motivated much of the recent progress on faster stochastic optimization in vector spaces by using variance reduction [10; 18; 28] techniques. However, all ensuing methods critically rely on properties of vector spaces, whereby, adapting them to work in the context of Riemannian manifolds poses major challenges. Given the richness of Riemannian optimization (it includes vector space optimization as a special case) and its growing number of applications, developing fast stochastic Riemannian optimization is important. It will help us apply Riemannian optimization to large-scale problems, while offering a new set of algorithmic tools for the practitioner?s repertoire. Contributions. We summarize the key contributions of this paper below. ? We introduce Riemannian SVRG (R SVRG), a variance reduced Riemannian stochastic gradient method based on SVRG [18]. We analyze R SVRG for geodesically strongly convex functions through a novel theoretical analysis that accounts for the nonlinear (curved) geometry of the manifold to

yield linear convergence rates. ? Building on recent advances in variance reduction for nonconvex optimization [3; 25], we generalize the convergence analysis of R SVRG to (geodesically) nonconvex functions and also to gradient dominated functions (see ?2 for the definition). Our analysis provides the first stochastic Riemannian method that is provably superior to both batch and stochastic (Riemannian) gradient methods for nonconvex finite-sum problems. ? Using a Riemannian formulation and applying our result for (geodesically) gradient-dominated functions, we provide new insights, and a short transparent analysis explaining fast convergence of variance reduced PCA for computing the leading eigenvector of a symmetric matrix. To our knowledge, this paper provides the first stochastic gradient method with global linear convergence rates for geodesically strongly convex functions, as well as the first non-asymptotic convergence rates for geodesically nonconvex optimization (even in the batch case). Our analysis reveals how manifold geometry, in particular curvature, impacts convergence rates. We illustrate the benefits of R SVRG by showing an application to computing leading eigenvectors of a symmetric matrix and to the task of computing the Riemannian centroid of covariance matrices, a problem that has received great attention in the literature [5; 16; 38]. Related Work. Variance reduction techniques, such as control variates, are widely used in Monte Carlo simulations [27]. In linear spaces, variance reduced methods for solving finite-sum problems have recently witnessed a huge surge of interest [e.g. 4; 10; 14; 18; 20; 28; 36]. They have been shown to accelerate stochastic optimization for strongly convex objectives, convex objectives, nonconvex fi (i 2 [n]), and even when both f and fi (i 2 [n]) are nonconvex [3; 25]. Reddi et al. [25] further proved global linear convergence for gradient dominated nonconvex problems. Our analysis is inspired by [18; 25], but applies to the substantially more general Riemannian optimization setting. References of Riemannian optimization can be found in [1; 33], where analysis is limited to asymptotic convergence (except [33, Theorem 4.2] which proved linear rate convergence for first-order line search method with bounded and positive definite hessian). Stochastic Riemannian optimization has 1 Riemannian optimization is optimization on a known manifold structure. Note the distinction from manifold learning, which attempts to learn a manifold structure from data. We briefly review some Riemannian optimization applications in the related work.

2

been previously considered in [7; 21], though with only asymptotic convergence analysis, and without any rates. Many applications of Riemannian optimization are known, including matrix factorization on fixed-rank manifold [32; 34], dictionary learning [8; 31], optimization under orthogonality constraints [11; 22], covariance estimation [35], learning elliptical distributions [30; 39], and Gaussian mixture models [15]. Notably, some nonconvex Euclidean problems are geodesically convex, for which Riemannian optimization can provide similar guarantees to convex optimization. Zhang and Sra [38] provide the first global complexity analysis for first-order Riemannian algorithms, but their analysis is restricted to geodesically convex problems with full or stochastic gradients. In contrast, we propose R SVRG, a variance reduced Riemannian stochastic

gradient algorithm, and analyze its global complexity for both geodesically convex and nonconvex problems. In parallel with our work, [19] also proposed and analyzed R SVRG specifically for the Grassmann manifold. Their complexity analysis is restricted to local convergence to strict local minima, which essentially corresponds to our analysis of (locally) geodesically strongly convex functions.

2

Preliminaries

Before formally discussing Riemannian optimization, let us recall some foundational concepts of Riemannian geometry. For a thorough review one can refer to any classic text, e.g.,[24]. A Riemannian manifold (M, g) is a real smooth manifold M equipped with a Riemannain metric g. The metric g induces an inner product structure in each tangent space Tx M associated with every x 2 M. We denote the inner product of u, v 2 Tx M as hu, vi , gx (u, v); and the norm p hu,vi of u 2 Tx M is defined as kuk , gx (u, u). The angle between u, v is defined as arccos kukkvk . A geodesic is a constant speed curve : [0, 1] ! M that is locally distance minimizing. An exponential map Expx : Tx M ! M maps v in Tx M to y on M, such that there is a geodesic d with (0) = x, (1) = y and ? (0) , dt (0) = v. If between any two points in X ? M there is a unique geodesic, the exponential map has an inverse Expx 1 : X ! Tx M and the geodesic is the unique shortest path with kExpx 1 (y)k = kExpy 1 (x)k the geodesic distance between x, y 2 X .

Parallel transport yx : Tx M ! Ty M maps a vector v 2 Tx M to yx v 2 Ty M, while preserving norm, and roughly speaking, ?direction,? analogous to translation in Rd . A tangent vector of a geodesic remains tangent if parallel transported along . Parallel transport preserves inner products. v

x

v

Expx (v)

x

y y xv

Figure 1: Illustration of manifold operations. (Left) A vector v in Tx M is mapped to Expx (v); (right) A vector v in Tx M is parallel transported to Ty M as

y x v.

The geometry of a Riemannian manifold is determined by its Riemannian metric tensor through various characterization of curvatures. Let u, v 2 Tx M be linearly independent, so that they span a two dimensional subspace of Tx M. Under the exponential map, this subspace is mapped to a two dimensional submanifold of U ? M. The sectional curvature ?(x, U) is defined as the Gauss curvature of U at x. As we will mainly analyze manifold trigonometry, for worst-case analysis, it is sufficient to consider sectional curvature. Function Classes. We now define some key terms. A set X is called geodesically convex if for any x, y 2 X , there is a geodesic with (0) = x, (1) = y and (t) 2 X for t 2 [0, 1]. Throughout the paper, we assume that the function f in (1) is defined on a geodesically convex set X on a Riemannian manifold M.

We call a function f : X ! R geodesically convex (g-convex) if for any x, y 2 X and any geodesic such that (0) = x, (1) = y and (t) 2 X for t 2 [0, 1], it holds that f ((t)) ? (1

t)f (x) + tf (y). 3

It can be shown that if the inverse exponential map is well-defined, an equivalent definition is that for any x, y 2 X , f (y) f (x) + hgx , Expx 1 (y)i, where gx is a subgradient of f at x (or the gradient if f is differentiable). A function f : X ! R is called geodesically ?-strongly convex (?-strongly g-convex) if for any x, y 2 X and subgradient gx , it holds that f (y)

f (x) + hgx , Expx 1 (y)i + ?2 kExpx 1 (y)k2 .

We call a vector field g : X ! Rd geodesically L-Lipschitz (L-g-Lipschitz) if for any x, y 2 X , kg(x)

x y g(y)k

? LkExpx 1 (y)k,

where xy is the parallel transport from y to x. We call a differentiable function f : X ! R geodesically L-smooth (L-g-smooth) if its gradient is L-g-Lipschitz, in which case we have f (y) ? f (x) + hgx , Expx 1 (y)i +

1 2 L 2 kExpx (y)k .

We say f : X ! R is ? -gradient dominated if x? is a global minimizer of f and for every x 2 X f (x)

f (x?) ? ? krf (x)k2 .

(2)

We recall the following trigonometric distance bound that is essential for our analysis: Lemma 1 ([7; 38]). If a, b, c are the side lengths of a geodesic triangle in a Riemannian manifold with sectional curvature lower bounded by ?min , and A is the angle between sides b and c (defined through inverse exponential map and inner product in tangent space), then p —?min —c 2 p a ? b2 + c2 2bc cos(A). (3) tanh(—?min —c) An Incremental First-order Oracle (IFO) [2] in (1) takes an i 2 [n] and a point x 2 X , and returns a pair (fi (x), rfi (x)) 2 R ? Tx M. We measure non-asymptotic complexity in terms of IFO calls.

3

Riemannian SVRG

In this section we introduce R SVRG formally. We make the following standing assumptions: (a) f attains its optimum at x? 2 X ; (b) X is compact, and the diameter of X is bounded by D, that is, maxx,y2X d(x, y) ? D; (c) the sectional curvature in X is upper bounded by ?max , and within X the exponential map is invertible; and (d) the sectional curvature in X is lower bounded by ?min . We define the following key geometric constant that capture the impact of manifold curvature: p (—?min —D p , if ?min ¡ 0, ?= (4) tanh(—?min —D) 1, if ?min 0, We note that most (if not all) practical manifold optimization problems can satisfy these assumptions. Our proposed R SVRG algorithm is shown in Algorithm 1. Compared with the Euclidean SVRG, it differs in two key aspects: the variance reduction step uses parallel transport to combine gradients from different tangent spaces; and the exponential map is used (instead of the update xs+1 ?vts+1). t 3.1

Convergence analysis for strongly g-convex functions

In this section, we analyze global complexity of R SVRG for solving (1), where each fi (i 2 [n]) is g-smooth and f is strongly g-convex. In this case, we show that R SVRG has linear convergence rate. This is in contrast with the O(1/t) rate of Riemannian stochastic gradient algorithm for strongly g-convex functions [38]. Theorem 1. Assume in (1) each fi is L-g-smooth, and f is ?-strongly g-convex, then if we run Algorithm 1 with Option I and parameters that satisfy 3??L2 (1 + 4?? 2 2??)m (? 5??L2) + ¡1 2 ? 2??L ? 2??L2 then with S outer loops, the Riemannian SVRG algorithm produces an iterate xa that satisfies ?=

Ed2 (xa , x?) ? ?S d2 (x0 , x?). 4

Algorithm 1: R SVRG (x0 , m, ?, S) Parameters: update frequency m, learning rate ?, number of epochs S initialize x ? 0 = x0 ; for s = 0, 1, . . . , S 1 do xs+1 =x ?s ; 0 P s+1 g = n1 n xs); i=1 rfi (? for t = 0, 1, . . . , m 1 do Randomly pick it 2 {1, . . . , n}; xs+1

t vts+1 = rfit (xs+1) rfit (? xs) t x ?s s+1 s+1 xt+1 = Expxs+1 ?vt ;

end Set x ?s+1 = xs+1 m ;

g s+1 ;

t

end Option I: output xa = x ?S ; 1 S 1 Option II: output xa chosen uniformly randomly from {{xs+1 }m t t=0 }s=0 .

The proof of Theorem 1 is in the appendix, and takes a different route compared with the original SVRG proof [18]. Specifically, due to the nonlinear Riemannian metric, we are not able to bound the squared norm of the variance reduced gradient by f (x) f (x?). Instead, we bound this quantity by the squared distances to the minimizer, and show linear convergence of the iterates. A bound on E[f (x) f (x?)] is then implied by L-g-smoothness, albeit with a stronger dependence on the condition number. Theorem 1 leads to the following more digestible corollary on the global complexity of the algorithm: Corollary 1. With ?assumptions as in Theorem 1 and properly chosen parameters, after ? 2 1 O (n + ?L ?2) log(?) IFO calls, the output xa satisfies E[f (xa)

f (x?)] ? ?.

We give a proof with specific parameter choices in the appendix. Observe the dependence on ? in our result: for ?min ¡ 0, we have ? ¿ 1, which implies that negative space curvature adversarially affects convergence rate; while for ?min 0, we have ? = 1, which implies that for nonnegatively curved manifolds, the impact of curvature is not explicit. In the rest of our analysis we will see a similar effect of sectional curvature; this phenomenon seems innate to manifold optimization (also see [38]). In the analysis we do not assume each fi to be g-convex, which resulted in a worse dependence on the condition number. We note that a similar result was obtained in linear space [12]. However, we will see in the next section that by generalizing the analysis for gradient dominated functions in [25], we are able to greatly improve this dependence. 3.2

Convergence analysis for geodesically nonconvex functions

In this section, we analyze global complexity of R SVRG for solving (1), where each fi is only required to be L-g-smooth, and neither fi nor f need be g-convex. We measure convergence to a stationary point using krf (x)k2 fol-

lowing [13]. Note, however, that here rf (x) 2 Tx M and krf (x)k is defined via the inner product in Tx M. We first note that Riemannian-SGD on nonconvex L-g-smooth problems attains O(1/?2) convergence as SGD [13] holds; we relegate the details to the appendix. Recently, two groups independently proved that variance reduction also benefits stochastic gradient methods for nonconvex smooth finite-sum optimization problems, with different analysis [3; 25]. Our analysis for nonconvex R SVRG is inspired by [25]. Our main result for this section is Theorem 2. Theorem 2. Assume in (1) each fi is L-g-smooth, the sectional curvature in X is lower bounded by ?min , and we run Algorithm 1 with Option II. Then there exist universal constants ?0 2 (0, 1), ? ¿ 0 such that if we set ? = ?0 /(Ln?1 ? ?2) (0 ¡ ?1 ? 1 and 0 ? ?2 ? 2), m = bn3?1 /2 /(3?0 ? 1 2?2)c and T = mS, we have E[krf (xa)k2] ?

Ln?1 ? ?2 [f (x0) f (x?)] , T?

where x? is an optimal solution to (1).

5

Algorithm 2: GD-SVRG(x0 , m, ?, S, K) Parameters: update frequency m, learning rate ?, number of epochs S, K, x0 for k = 0, . . . , K 1 do xk+1 = R SVRG(xk , m, ?, S) with Option II; end Output: xK

The key challenge in proving Theorem 2 in the Riemannian setting is to incorporate the impact of using a nonlinear metric. Similar to the g-convex case, the nonlienar metric impacts the convergence, notably through the constant ? that depends on a lower-bound on sectional curvature. Reddi et al. [25] suggested setting ?1 = 2/3, in which case we obtain the following corollary. Corollary 2. With assumptions and parameters in Theorem 2, choosing ?1 = 2/3, the IFO complexity for achieving an ?-accurate solution is: ? O n + (n2/3 ? 1 ?2 /?) , if ?2 ? 1/2, IFO calls = O n? 2?2 1 + (n2/3 ? ?2 /?) , if ?2 ¿ 1/2. Setting ?2 = 1/2 in Corollary 2 immediately leads to Corollary 3: Corollary 3. With assumptions in Theorem 2 and ?1 = 2/3, ?2 = 1/2, the IFO complexity for achieving an ?-accurate solution is O n + (n2/3 ? 1/2 /?) . The same reasoning allows us to also capture the class of gradient dominated functions (2), for which Reddi et al. [25] proved that SVRG converges linearly to a global optimum. We have the following corresponding theorem for R SVRG: Theorem 3. Suppose that in addition to the assumptions in Theorem 2, f is ? -gradient dominated. Then there exist universal constants ?0 2 (0, 1), ? ¿ 0 such that if we run Algorithm 2 with 1/2 0 ? = ?0 /(Ln2/3 ? 1/2), m = bn/(3?0)c, S = d(6 + 18? ?0 /(?n1/3)e, we have n 3)L? ? E[krf (xK)k2] ? 2

E[f (xK)

f (x?)] ? 2

K K

krf (x0)k2 , [f (x0)

f (x?)].

We summarize the implication of Theorem 3 as follows (note the dependence on curvature): Corollary 4. With Algorithm 2 and the parameters in Theorem 3, the IFO complexity to compute an ?-accurate solution for a gradient dominated function f is O((n + L? ? 1/2 n2/3) log(1/?)). A typical example of gradient dominated function is a strongly g-convex function (see appendix). Specifically,

we have the following corollary, which prove linear convergence rate of R SVRG with the same assumptions as in Theorem 1, improving the dependence on the condition number. Corollary 5. With Algorithm 2 and the parameters in Theorem 3, the IFO complexity to compute an ?-accurate solution for a ?-strongly g-convex function f is O((n + ? 1 L? 1/2 n2/3) log(1/?)).

4 4.1

Applications Computing the leading eigenvector

In this section, we apply our analysis of R SVRG for gradient dominated functions (Theorem 3) to fast eigenvector computation, a fundamental problem that is still being actively researched in the big-data setting [12; 17; 29]. For the problem of computing the leading eigenvector, i.e., ?X n ? min x¿ zi zi¿ x , x¿ Ax = f (x), (5) x¿ x=1

i=1

existing analyses for state-of-the-art algorithms typically result in O(1/ 2) dependence on the eigengap of A, as opposed to the conjectured O(1/) dependence [29], as well as the O(1/) dependence of power iteration. Here we give new support for the O(1/) conjecture. Note that Problem (5) seen as one in Rd is nonconvex, with negative semidefinite Hessian everywhere, and has nonlinear constraints. However, we show that on the hypersphere Sd 1 Problem (5) is unconstrained, and has gradient dominated objective. In particular we have the following result: 6

0 Theorem 4. Suppose A has eigenvalues 1 ¿ 2 ? ? ? = 1 d and 2 , and x is drawn 0 uniformly randomly on the hypersphere. Then with probability 1 p, x falls in a Riemannian ball of a global optimum of the objective function, within which the objective function is O(pd2)-gradient dominated.

We provide the proof of Theorem 4 in appendix. Theorem 4 gives new insights for why the conjecture might be true ? once it is shown that with a constant stepsize and with high probability (both independent of) the iterates remain in such a Riemannian ball, applying Corollary 4 one can immediately prove the O(1/) dependence conjecture. We leave this analysis as future work. Next we show that variance reduced PCA (VR-PCA) [29] is closely related to R SVRG. We implement Riemannian SVRG for PCA, and use the code for VR-PCA in [29]. Analytic forms for exponential map and parallel transport on hypersphere can be found in [1, Example 5.4.1; Example 8.1.1]. We conduct well-controlled experiments comparing the performance of two algorithms. Specifically, to investigate the dependence of convergence on , for each = 10 3 /k where k = 1, . . . , 25, we generate a d ? n matrix Z = (z1 , . . . , zn) where d = 103 , n = 104 using the method Z = U DV ¿ where U, V are orthonormal matrices and D is a diagonal matrix, as described in [29]. Note that A has the same eigenvalues as D2 . All the data matrices share the same U, V and only differ in (thus also in D). We also fix the same random initialization x0 and random seed. We run both algorithms on each matrix for 50 epochs. For every five epochs, we estimate the number of epochs required to double its accuracy 2 . This number can serve as an indicator of the global complexity of the algorithm. We plot this number for different epochs against 1/ , shown in Figure 2. Note that the performance of RSVRG and VR-PCA with the same stepsize is very

similar, which implies a close connection x+v of the two. Indeed, the update kx+vk used in [29] and others is a well-known approximation to the exponential map Expx (v) with small stepsize (a.k.a. retraction). Also note the complexity of both algorithms seems to have an asymptotically linear dependence on 1/ .

#epochs required
accuracy
10 -4 10 -6 10
-8
0
2
4
RSVRG
100
RSVRG VR-PCA
6
1-5 11-15 21-25 31-35 41-45
50
0
0
1
2
1//
#IFO calls #10 5
VR-PCA
100
3 #10
4
#epochs required
/ = 1e-3
10 -2
1-5 11-15 21-25 31-35 41-45
50
0
0
1
2
1//
3 #10
4

Figure 2: Computing the leading eigenvector. Left: RSVRG and VR-PCA are indistinguishable in terms of IFO complexity. Middle and right: Complexity appears to depend on 1/ . x-axis shows the inverse of eigengap , y-axis shows the estimated number of epochs required to double the accuracy. Lines represent different epoch index. All variables are controlled except for .

4.2

Computing the Riemannian centroid

In this subsection we validate that R SVRG converges linearly for averaging PSD matrices under the Riemannian metric. The problem for finding the Riemannian centroid of a set of o PSD matrices n Pn n ? n 1/2 1/2 2 {Ai }i=1 is X = arg minX?0 f (X; {Ai }i=1) , i=1 k log(X Ai X)kF where X is also a PSD matrix. This is a geodesically strongly convex problem, yet nonconvex in Euclidean space. It has been studied both in matrix computation and in various applications [5; 16]. We use the same experiment setting as described in [38] 3 , and compare R SVRG against Riemannian full gradient (RGD) and stochastic gradient (RSGD) algorithms (Figure 3). Other methods for this problem include the relaxed Richardson iteration algorithm [6], the approximated joint diagonalization algorithm [9], and Riemannian Newton and quasi-Newton type methods, notably the limited-memory Riemannian 2

?

f (x) Accuracy is measured by f (x) , i.e. the relative error between the objective value and the optimum. —f (x?)— We measure how much the error has been reduced after each five epochs, which is a multiplicative factor c ¡ 1 on the error at the start of each five epochs. Then we use log(2)/ log(1/c) ? 5 as the estimate, assuming c stays constant. 3 We generate 100 ? 100 random PSD matrices using the Matrix Mean Toolbox [6] with normalization so that the norm of each matrix equals 1.

7

BFGS [37]. However, none of these methods were shown to greatly outperform RGD, especially in data science applications where n is large and extremely small optimization error is not required.

10

0

N=100,Q=1e2

RGD RSGD RSVRG

10

-5

10

5

10

0

RGD RSGD RSVRG

10 0

1000

2000

#IFO calls

N=100,Q=1e8

-5

10

5

10

0

RGD RSGD RSVRG

10 0
1000
2000
N=1000,Q=1e2
-5
10
5
10
0
5000
#IFO calls
10000
N=1000,Q=1e8
RGD RSGD RSVRG
10 0
#IFO calls
accuracy
5
accuracy
10
accuracy
accuracy

Note that the objective is sum of squared Riemannian distances in a non-positively curved space, thus is (2n)-strongly g-convex and (2n?)-g-smooth. According to the proof of Corollary 1 (see appendix) the optimal stepsize for R SVRG is O(1/(? 3 n)). For all the experiments, we initialize all 1 the algorithms using the arithmetic mean of the matrices. We set ? = 100n , and choose m = n in Algorithm 1 for R SVRG, and use suggested parameters from [38] for other algorithms. The results suggest R SVRG has clear advantage in the large scale setting.

-5
0
5000
10000
#IFO calls

Figure 3: Riemannian mean of PSD matrices. N : number of matrices, Q: conditional number of each

matrix. x-axis shows the actual number of IFO calls, y-axis show f (X) f (X ?) in log scale. Lines show the performance of different algorithms in colors. Note that R SVRG achieves linear convergence and is especially advantageous for large dataset.

5

Discussion

We introduce Riemannian SVRG, the first variance reduced stochastic gradient algorithm for Riemannian optimization. In addition, we analyze its global

complexity for optimizing geodesically strongly convex, convex, and nonconvex functions, explicitly showing their dependence on sectional curvature. Our experiments validate our analysis that Riemannian SVRG is much faster than full gradient and stochastic gradient methods for solving finite-sum optimization problems on Riemannian manifolds. Our analysis of computing the leading eigenvector as a Riemannian optimization problem is also worth noting: a nonconvex problem with nonpositive Hessian and nonlinear constraints in the ambient space turns out to be gradient dominated on the manifold. We believe this shows the promise of theoretical study of Riemannian optimization, and geometric optimization in general, and we hope it encourages other researchers in the community to join this endeavor. Our work also has limitations ? most practical Riemannian optimization algorithms use retraction and vector transport to efficiently approximate the exponential map and parallel transport, which we do not analyze in this work. A systematic study of retraction and vector transport is an important topic for future research. For other applications of Riemannian optimization such as low-rank matrix completion [34], covariance matrix estimation [35] and subspace tracking [11], we believe it would also be promising to apply fast incremental gradient algorithms in the large scale setting. Acknowledgment: SS acknowledges support of NSF grant: IIS-1409802. HZ acknowledges support from the Leventhal Fellowship.

2 References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. Princeton University Press, 2009. [2] A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 78?86, 2015. [3] Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. arXiv:1603.05643, 2016. [4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In Advances in Neural Information Processing Systems, pages 773?781, 2013. [5] R. Bhatia. Positive Definite Matrices. Princeton University Press, 2007. [6] 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. [7] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. Automatic Control, IEEE Transactions on, 58(9):2217?2229, 2013.

8

[8] A. Cherian and S. Sra. Riemannian dictionary learning and sparse coding for positive definite matrices. arXiv:1507.02772, 2015. [9] M. Congedo, B. Afsari, A. Barachant, and M. Moakher. Approximate joint diagonalization and geometric mean of symmetric positive definite matrices. PloS one, 10(4):e0121423, 2015. [10] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In NIPS, pages 1646?1654, 2014. [11] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. SIAM jour-

nal on Matrix Analysis and Applications, 20(2):303?353, 1998. [12] D. Garber and E. Hazan. Fast and simple pca via convex optimization. arXiv preprint arXiv:1509.05647, 2015. [13] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341?2368, 2013. [14] P. Gong and J. Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. arXiv preprint arXiv:1406.1102, 2014. [15] R. Hosseini and S. Sra. Matrix manifold optimization for Gaussian mixtures. In NIPS, 2015. [16] B. Jeuris, R. Vandebril, and B. Vandereycken. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. Electronic Transactions on Numerical Analysis, 39:379?402, 2012. [17] C. Jin, S. M. Kakade, C. Musco, P. Netrapalli, and A. Sidford. Robust shift-and-invert preconditioning: Faster and more sample efficient algorithms for eigenvector computation. arXiv:1510.08896, 2015. [18] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in Neural Information Processing Systems, pages 315?323, 2013. [19] H. Kasai, H. Sato, and B. Mishra. Riemannian stochastic variance reduced gradient on grassmann manifold. arXiv preprint arXiv:1605.07367, 2016. [20] J. Kone?cn‘y and P. Richt?rik. Semi-stochastic gradient descent methods. arXiv:1312.1666, 2013. [21] X. Liu, A. Srivastava, and K. Gallivan. Optimal linear representations of images for object recognition. IEEE TPAMI, 26(5):662?666, 2004. [22] M. Moakher. Means and averaging in the group of rotations. SIAM journal on matrix analysis and applications, 24(1):1?16, 2002. [23] E. Oja. Principal components, minor components, and linear neural networks. Neural Networks, 5(6): 927?935, 1992. [24] P. Petersen. Riemannian geometry, volume 171. Springer Science & Business Media, 2006. [25] S. J. Reddi, A. Hefny, S. Sra, B. P?cz?s, and A. Smola. Stochastic variance reduction for nonconvex optimization. arXiv:1603.06160, 2016. [26] H. Robbins and S. Monro. A stochastic approximation method. Annals of Mathematical Statistics, 22: 400?407, 1951. [27] R. Y. Rubinstein and D. P. Kroese. Simulation and the Monte Carlo method, volume 707. John Wiley & Sons, 2011. [28] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. arXiv:1309.2388, 2013. [29] O. Shamir. A Stochastic PCA and SVD Algorithm with an Exponential Convergence Rate. In International Conference on Machine Learning (ICML-15), pages 144?152, 2015. [30] S. Sra and R. Hosseini. Geometric optimisation on positive definite matrices for elliptically contoured distributions. In Advances in Neural Information Processing Systems, pages 2562?2570, 2013. [31] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. arXiv:1511.04777, 2015. [32] M. Tan, I. W. Tsang, L. Wang, B. Vandereycken, and S. J. Pan. Riemannian pursuit for big matrix recovery. In International Conference on Machine Learning (ICML-14), pages 1539?1547, 2014. [33] C. Udriste. Convex functions and optimization methods on Riemannian manifolds, volume 297. Springer Science & Business Media, 1994. [34] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. SIAM Journal on Optimization, 23(2):1214?1236, 2013. [35] A. Wiesel. Geodesic convexity and covariance estimation. IEEE Transactions

on Signal Processing, 60 (12):6182?6189, 2012. [36] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization, 24(4):2057?2075, 2014. [37] X. Yuan, W. Huang, P.-A. Absil, and K. Gallivan. A riemannian limited-memory bfgs algorithm for computing the matrix geometric mean. Procedia Computer Science, 80:2147?2157, 2016. [38] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. arXiv:1602.06053, 2016. [39] T. Zhang, A. Wiesel, and M. S. Greco. Multivariate generalized Gaussian distribution: Convexity and graphical models. Signal Processing, IEEE Transactions on, 61(16):4141?4148, 2013.

9