F-008

Riemannian stochastic variance reduced gradient on Grassmann manifold

Hiroyuki KASAI[†] **Hiroyuki SATO**[‡] **Bamdev MISHRA**[§]

1. Introduction

A general loss minimization problem is defined as $\min_w f(w)$, where $f(w) := \frac{1}{N}\sum_{n=1}^{N} f_n(w)$, w is the model variable, N is the number of samples, and $f_n(w)$ is the loss incurred on n-th sample. The *full gradient decent* (GD) algorithm requires evaluations of N derivatives, i.e., $\sum_{n=1}^{N} \nabla f_n(w)$, per iteration, which is computationally heavy when N is very large. A popular alternative is to use only one derivative $\nabla f_n(w)$ for n-th sample, which is the basis of the *stochastic gradient descent* (SGD) algorithm. Recently, Stochastic variance reduced gradient (SVRG) [1] and its variants have been proposed to accelerate the convergence of SGD. However, all these cases assume that search space is Euclidean. Meanwhile, this paper particularly examines the problems where the variables have a manifold structure. Optimization on *Riemannian manifolds* $\mathcal{M}$ has shown state-of-the-art performance, where the problem $\min_{w\in\mathcal{M}} f(w)$ is solved as an *unconstrained optimization problem* defined over the Riemannian manifold search space [2]. In this paper, building upon *Riemannian stochastic gradient* algorithm (R-SGD) [3], we propose a novel (and to the best of our knowledge, the first) extension of SVRG in the Euclidean space to the Riemannian manifold search space (R-SVRG). This extension is not trivial and requires particular consideration in dealing with averaging, addition and subtraction of multiple gradients at different points on the manifold $\mathcal{M}$. To this end, this paper specifically focuses on the *Grassmann manifold* $\mathrm{Gr}(r, d)$, which is the set of r-dimensional linear subspaces in $\mathbb{R}^d$. For this particular purpose, the model variable w is instead denoted as $\mathbf{U} \in \mathrm{Gr}(r, d)$ through the paper. Nonetheless, our proposed algorithm and the analysis presented can be generalized to other compact Riemannian manifolds.

2. Problems on Grassmann manifold

We focus on three popular problems on the Grassmann manifold, which are the PCA, low-rank matrix completion, and the Karcher mean computation problems. This section especially addresses the PCA problem. The details of the other problems are in [4].

Given an orthonormal matrix projector $\mathbf{U} \in \mathrm{St}(r, d)$, the PCA problem is to minimize the sum of squared residual errors between projected data points and the original data as $\min_{\mathbf{U}\in\mathrm{St}(r,d)} \frac{1}{N}\sum_{n=1}^{N} \|\boldsymbol{x}_n - \mathbf{U}\mathbf{U}^T\boldsymbol{x}_n\|_2^2$, where $\boldsymbol{x}_n$ is a data vector of size $d \times 1$. This is equivalent to maximizing $\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n^T\mathbf{U}\mathbf{U}^T\boldsymbol{x}_n$. Here, the critical points in the space $\mathrm{St}(r, d)$ are not isolated because the cost function remains unchanged under the group action $\mathbf{U} \mapsto \mathbf{U}\mathbf{O}$ for all orthogonal matrices $\mathbf{O}$ of size $r \times r$. Subsequently, the PCA problem is an optimization problem on the Grassmann manifold $\mathrm{Gr}(r, d)$.

[†]The University of Electro-Communications, JAPAN
[‡]Tokyo University of Science, JAPAN
[§]Amazon Development Centre India, India

3. R-SVRG on Grassmann manifold

An element of the Grassmann manifold $\mathrm{Gr}(r, d)$ is an r-dimensional subspace of $\mathbb{R}^d$, is represented by a $d\times r$ matrix $\mathbf{U}$ with orthonormal columns, i.e., $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, and is identified with a set of $d \times r$ orthogonal matrices $[\mathbf{U}] := \{\mathbf{U}\mathbf{O}_r : \mathbf{O} \in \mathcal{O}(r)\}$. In other words, $\mathrm{Gr}(r, d) := \mathrm{St}(r, d)/\mathcal{O}(r)$, where $\mathrm{St}(r, d)$ is the Stiefel manifold that is the set of matrices of size $d \times r$ with orthonormal columns. The Grassmann manifold has the structure of a Riemannian quotient manifold [2].

In computing a sequence of points on $\mathrm{Gr}(r, d)$ in the proposed R-SVRG, we use the *exponential map* to construct a geodesic curve, which is given [2] by

$$\mathbf{U}(t) \;=\; [\mathbf{U}(0)\mathbf{V} \;\; \mathbf{W}] \begin{bmatrix} \cos t\Sigma \\ \sin t\Sigma \end{bmatrix} \mathbf{V}^T, \qquad (1)$$

where $\xi = \mathbf{W}\Sigma\mathbf{V}^T$ is the rank-r singular value decomposition of a tangent vector ξ at $\mathbf{U}(0)$. The $\cos(\cdot)$ and $\sin(\cdot)$ operations are only on the diagonal entries. The *logarithm map*, which is the inverse of the exponential map, of $\mathbf{U}(t)$ at $\mathbf{U}(0)$ on $\mathrm{Gr}(r, d)$ is given [2] by

$$\xi \;=\; \log_{\mathbf{U}(0)}(\mathbf{U}(t)) \;=\; \mathbf{W}\arctan(\Sigma)\mathbf{V}^T, \quad (2)$$

where $\mathbf{W}\Sigma\mathbf{V}^T$ is the rank-r singular value decomposition of $(\mathbf{U}(t) - \mathbf{U}(0)\mathbf{U}(0)^T\mathbf{U}(t))(\mathbf{U}(0)^T\mathbf{U}(t))^{-1}$. The *parallel translation* of $\zeta \in T_{\mathbf{U}(0)}\mathrm{Gr}(r, d)$ on the Grassmann manifold along the geodesic with ξ, which is used in computing the modified Riemannian stochastic gradient, is given in closed form [2] by

$$\zeta(t) \;=\; \Big([\mathbf{U}(0)\mathbf{V} \;\; \mathbf{W}] \begin{bmatrix} -\sin t\Sigma \\ \cos t\Sigma \end{bmatrix} \mathbf{W}^T + (\mathbf{I} - \mathbf{W}\mathbf{W}^T)\Big)\,\zeta.$$

We now describe our proposed R-SVRG algorithm, where the modified Riemannian stochastic gradient ξ_t^s is calculated by parallel-translating $\mathrm{grad} f(\tilde{\mathbf{U}}^{s-1})$ and $\mathrm{grad} f_{i_t^s}(\tilde{\mathbf{U}}^{s-1})$ along $\log_{\tilde{\mathbf{U}}^{s-1}}(\mathbf{U}_{t-1}^s)$ as

$$\xi_t^s \;=\; \mathrm{grad} f_{i_t^s}(\mathbf{U}_{t-1}^s) - P_\gamma^{\mathbf{U}_{t-1}^s \leftarrow \tilde{\mathbf{U}}^{s-1}}\Big(\mathrm{grad} f_{i_t^s}(\tilde{\mathbf{U}}^{s-1}) - \mathrm{grad} f(\tilde{\mathbf{U}}^{s-1})\Big). \qquad (3)$$

Finally, the overall algorithm with a fixed step-size is summarized in **Algorithm 1**. Meanwhile, SVRG needs full gradient calculation every epoch at the beginning. This poses a bigger overhead than the ordinal SGD algorithm at the beginning, and eventually, causes *cold-start* property on them. To avoid this, the papers in the Euclidean space proposes to use standard SGD updating only for first epoch. This paper also adopts this simple modification of R-SVRG, denoted as R-SVRG+.

Algorithm 1 Algorithm for R-SVRG [4].

Require: Update frequency $m_s > 0$ and step-size $\eta > 0$.

Initialize $\tilde{\mathbf{U}}^0$.
for $s = 1, 2, \ldots$ **do**
  Calculate the Riemannian full gradient $\mathrm{grad} f(\tilde{\mathbf{U}}^{s-1})$.
  Store $\mathbf{U}_0^s = \tilde{\mathbf{U}}^{s-1}$.
  for $t = 1, 2, \ldots, m_s$ **do**
    Choose $i_t^s \in \{1, \ldots, N\}$ uniformly at random.
    Calculate the tangent vector ζ from $\tilde{\mathbf{U}}^{s-1}$ to $\mathbf{U}_{t-1}^s$ by logarithm mapping in (2).
    Calculate the modified Riemannian stochastic gradient ξ_t^s by (3).
    Update $\mathbf{U}_t^s$ from $\mathbf{U}_{t-1}^s$ as $\mathbf{U}_t^s = \mathrm{Exp}_{\mathbf{U}_{t-1}^s}(-\eta \xi_t^s)$ with the exponential mapping (1).
  end for
  option I: $\tilde{\mathbf{U}}^s = g_{m_s}(\mathbf{U}_1^s, \ldots, \mathbf{U}_{m_s}^s)$ (or $\tilde{\mathbf{U}}^s = \mathbf{U}_t^s$ for randomly chosen $t \in \{1, \ldots, m_s\}$).
  option II: $\tilde{\mathbf{U}}^s = \mathbf{U}_{m_s}^s$.
end for

4. Main result: convergence analysis

Our main results of convergence analysis of **Algorithm 1** are as follows [4]. The former is on global convergence and the latter is on local convergence.

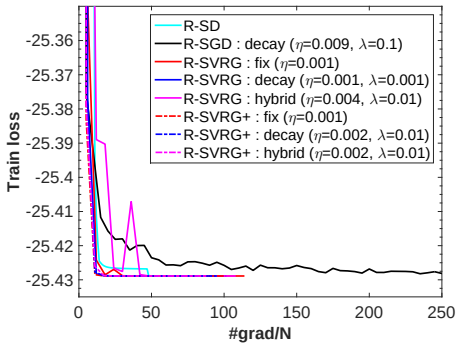
Theorem 1. *Consider* **Algorithm 1** *on a connected Riemannian manifold $\mathcal{M}$ with injectivity radius uniformly bounded from below by $I > 0$. Assume that the sequence of step-sizes $(\eta_t^s)_{m_s \geq t \geq 1, s \geq 1}$ satisfies the condition that $\sum (\eta_t^s)^2 < \infty$ and $\sum \eta_t^s = +\infty$. Suppose there exists a compact set K such that $\mathbf{U}_t^s \in K$ for all $t \geq 0$. We also suppose that the gradient is bounded on K, i.e., there exists $A > 0$ such that for all $\mathbf{U}_t^s \in K$ and $i_t^s \in Z$ we have $\|\mathrm{grad} f(\mathbf{U}_t^s)\| \leq A/3$. Then $f(\mathbf{U}_t^s)$ converges a.s. and $\mathrm{grad} f(\mathbf{U}_t^s) \to 0$ a.s.*

Theorem 2. *Let $\mathcal{M}$ be the Grassmann manifold and $\mathbf{U}^* \in \mathcal{M}$ be a non-degenerate local minimizer of f (i.e., $\mathrm{grad} f(\mathbf{U}^*) = 0$ and the Hessian $\mathrm{Hess} f(\mathbf{U}^*)$ of f at $\mathbf{U}^*$ is positive definite). Assume that there exists a convex neighborhood $\mathcal{U}$ of $\mathbf{U}^* \in \mathcal{M}$ and a positive real number σ such that the smallest eigenvalue of the Hessian of f at each $\mathbf{U} \in \mathcal{U}$ is not less than σ. When each $\mathrm{grad} f_n$ is β-Lipschitz continuously differentiable and $\eta > 0$ is sufficiently small such that $0 < \eta(\sigma - 14\eta\beta^2) < 1$, it then follows that for any sequence $\{\tilde{\mathbf{U}}^s\}$ generated by the algorithm converging to $\mathbf{U}^*$, there exists $K > 0$ such that for all $s > K$,*

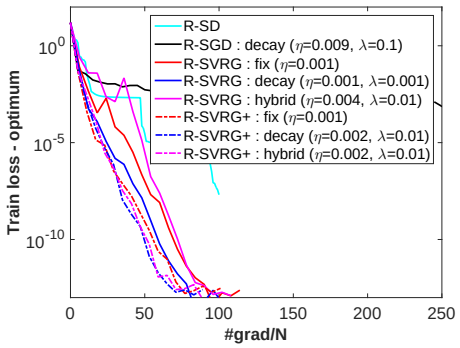
$$\mathbb{E}[(\mathrm{dist}(\tilde{\mathbf{U}}^s, \mathbf{U}^*))^2] \leq \frac{4(1 + 8m\eta^2\beta^2)}{\eta m(\sigma - 14\eta\beta^2)} \mathbb{E}[(\mathrm{dist}(\tilde{\mathbf{U}}^{s-1}, \mathbf{U}^*))^2].$$

5. Numerical comparisons

This section compares the performance of R-SVRG with R-SGD and the Riemannian steepest descent algorithm, i.e., R-SD [2]. We consider both *fixed* step-size as well as *decay* step-size sequences. We also consider a *hybrid* step-size sequence that follows the decay step-size. Figures 1(a) and (b) show the results of the training loss and *optimality gap*, respectively, where $N = 10000$, $d = 20$, and $r = 5$. The optimality gap evaluates the performance against the minimum loss, which is obtained by the Matlab function `pca`. Figure (a) shows the enlarged results of the training loss, where all algorithms of R-SVRG(+) yield better convergence properties. Among the step-size sequences of R-SVRG(+), the hybrid sequence shows the best performance among all. Between R-SVRG and R-SVRG+, the latter shows superior performance for all step-size sequences. For the optimality gap plots in Figure (b), the results follow similar trends as those of training loss plots.



(a) Train loss (enlarged).



(b) Optimality gap.

Figure 1: Performance evaluations on PCA problem.

References

[1] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

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

[4] H. Kasai, H. Sato, and B. Mishra. Riemannian stochastic variance reduced gradient on grassmann manifold. *arXiv preprint: arXiv:1605.07367*, 2016.