Riemannian stochastic variance reduced gradient on Grassmann manifold

Hiroyuki Kasai*, Hiroyuki Sato§ and Bamdev Mishra†

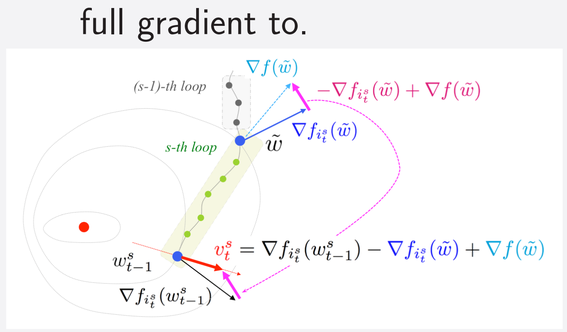
*The University of Electro-Communications, Japan §Tokyo university of Science, Japan
†Amazon Development Centre India, India

Summary (our contributions)

- Address stochastic gradient descent (SGD) algorithm for empirical risk minimization problem as
$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} f_i(w).$$
  - Paritularlly, structured problems on manifolds, i.e., $w \in \mathcal{M}$ [1].
- Propose Riemannian SVRG (R-SVRG).
  - Extend SVRG in the Euclidean into Riemannian manifolds.
- Global and Local convergence analyses for (i) exponential mapping and parallel translation, and (ii) retraction and vector transport.
- Show effectiveness of R-SVRG from numerical comparisons.

Stochastic variance reduced gradient (SVRG) [2]

- Stochastic gradient descent (SGD)
  - High scalability to large-scale data because per-iteration complexity does not depend on n.
  - But, slow convergence property (sub-linear) with decaying stepseize
- Stochastic variance reduced gradient (SVRG)
  - Basic idea
    - Reduce the variance of stochastic gradients.
    - No need to store all gradients.
    - Linear convergence rate
  - Basic strategy
    - Periodically, calculate and store a full gradient.
    - Every iteration adjusts a stochastic gradient by the latest full gradient to.



SVRG pseudo-code

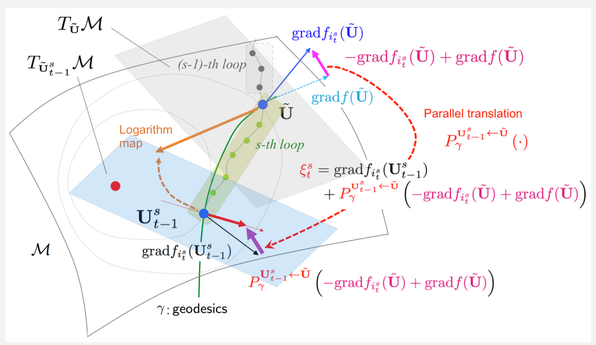
1: Initial iterate $w_0^0 \in \mathcal{M}$.
2: **for** $s = 1, 2, \ldots$ (outer loop) **do**
3: Store $\tilde{w} = w_t^{s-1}$.
4: Store $\nabla f(\tilde{w})$.
5: **for** $t = 1, \ldots, m_s$ (inner) **do**
6: Calculate modified stochastic gradient
$$v_t^s = \underbrace{\nabla f_{i_t^s}(w_{t-1}^s)}_{\text{single gradient at } w_{t-1}^s} - \overbrace{\nabla f_{i_t^s}(\tilde{w})}^{\text{single gradient}} + \underbrace{\nabla f(\tilde{w})}_{\text{full gradient}}.$$
7: Update $w_t^s = w_{t-1}^s - \alpha v_t^s$.
8: **end for**
9: set $\tilde{w} = w_t^s$ for randomly chosen t.
10: **end for**

Riemannian stochastic variance reduced gradient (R-SVRG) [3]

- Address various structured problems
  - E.g., PCA, matrix compleiton, subspace tracking, spectral clustering.
- Focus on the Grassmann manifold $\mathbf{Gr}(r, d)$.
  - The set of r-dimensional linear subspaces in $\mathbb{R}^d$.
  - Can be generalized to other compact Riemannian manifolds.
- Notations

	SVRG	R-SVRG
Model parameter	$w_{t-1}^s \in \mathbb{R}^n$	$\mathbf{U}_{t-1}^s \in \mathbf{Gr}(r, d)$
Edge point of outer loop	$\tilde{w} \in \mathbb{R}^n$	$\tilde{\mathbf{U}} \in \mathbf{Gr}(r, d)$
Stochastic gradient	$\nabla f_{i_t^s}(w_{t-1}^s) \in \mathbb{R}^n$	$\mathbf{grad} f_{i_t^s}(\mathbf{U}_{t-1}^s) \in T_{\mathbf{U}_{t-1}^s}\mathbf{Gr}(r, d)$
Modified stochastic gradient	$v_t^s \in \mathbb{R}^n$	$\xi_t^s \in T_{\mathbf{U}_{t-1}^s}\mathbf{Gr}(r, d)$

- Agorithm basis
  - Build on Riemannian stochastic gradient [4].
  - Straightforward modification of stochastic gradient
    - Extend SVRG case: $v_t^s = \nabla f_{i_t^s}(w_{t-1}^s) - \nabla f_{i_t}(\tilde{w}) + \nabla f(\tilde{w})$
    $$\xi_t^s = \mathbf{grad} f_{i_t^s}(\mathbf{U}_{t-1}^s) - \mathbf{grad} f_{i_t^s}(\tilde{\mathbf{U}}) + \mathbf{grad} f(\tilde{\mathbf{U}})$$
    - Meaningless because manifolds are not vector space.
  - Proposed modification
    - Transport vectors at $\tilde{\mathbf{U}}$ into the current tangent space at $\mathbf{U}_{t-1}^s$ by parallel translation or vector transport, then add them.
    $$\xi_t^s = \mathbf{grad} f_{i_t^s}(\mathbf{U}_{t-1}^s) + \underbrace{\mathcal{T}}_{\text{parallel-translation / vector transport operator}} \left(-\mathbf{grad} f_{i_t^s}(\tilde{\mathbf{U}}) + \mathbf{grad} f(\tilde{\mathbf{U}})\right)$$



Convergence analyses

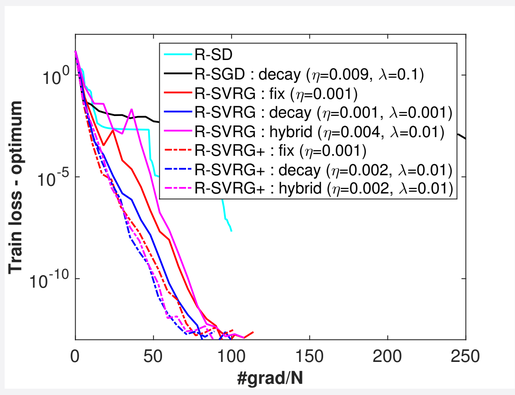
- Global convergence analysis with decaying step-sizes.
  - Guarantee that the iteration globally converges to a critical point starting from any initialization point.
- Local convergence rate analysis under fixed step-size.
  - Consider the rate in neighborhood of a local minimum.
  - Assume that Lipschitz smoothness and lower bound of Hessian hold only in this neighborhood.
  - Exponential mapping and paralell translation
  $$\mathbb{E}[(\mathrm{dist}(\tilde{\mathbf{U}}^s, \mathbf{U}^*))^2] \leq \frac{4(1 + 8m\alpha^2\beta^2)}{\alpha m(\sigma - 14\eta\beta^2)} \mathbb{E}[(\mathrm{dist}(\tilde{\mathbf{U}}^{s-1}, \mathbf{U}^*))^2].$$
  - Retraction and vector transport
  $$\mathbb{E}[(\mathrm{dist}(\tilde{\mathbf{U}}^s, \mathbf{U}^*))^2] \leq \frac{4(1 + 16m\alpha^2(\beta^2 + C^2a^2))}{\alpha m(\sigma - 28\alpha(\beta^2 + C^2a^2))} \mathbb{E}[(\mathrm{dist}(\tilde{\mathbf{U}}^{s-1}, \mathbf{U}^*))^2],$$
  where a and C are some constants satisfying $\|\mathcal{T}_\eta(\xi) - P_\eta(\xi)\| \leq a\|\xi\|\|\eta\|$ and $\|\mathbf{grad} f_i(\mathbf{U})\| \leq C$ for $\mathbf{U}$ sufficiently close to the local minimum and for $\xi, \eta \in T_{\mathbf{U}}\mathbf{Gr}(r, d)$.

Proof sketch in exponential mapping and paralell translation case

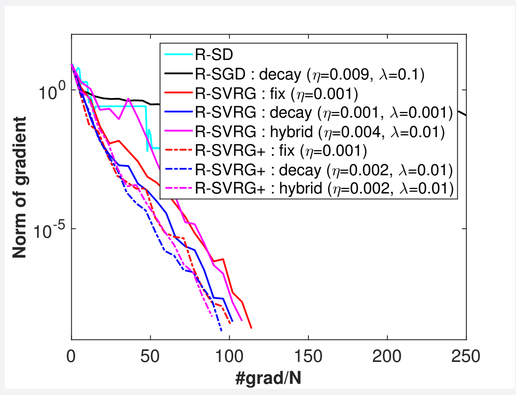
- Obtain below by assuming the smallest eigenvalue σ of Hessian of f as
$$f(z) \geq f(w) + \langle \mathrm{Exp}_w^{-1}(z), \mathbf{grad} f(w) \rangle_w + \frac{\sigma}{2} \|\mathrm{Exp}_w^{-1}(z)\|_w^2, \ w, z \in \mathcal{U}. \quad (1)$$
- Obtain the variance of ξ_t^s from β-Lipschitz continuity as
$$\mathbb{E}_{i_t^s}[\|\xi_t^s\|^2] \leq \beta^2(14(\mathrm{dist}(w_{t-1}^s, w^*))^2 + 8\mathrm{dist}(\tilde{w}^{s-1}, w^*))^2) \quad (2)$$
- Obtain the expectation of the decrease of the distance to the solution in the inner iteration from the lemma for a geodesic triangle in an Alexandrov space as
$$\mathbb{E}_{i_t^s}\left[(\mathrm{dist}(\mathbf{U}_t^s, \mathbf{U}^*))^2 - (\mathrm{dist}(\mathbf{U}_{t-1}^s, \mathbf{U}^*))^2\right]$$
$$\leq \mathbb{E}_{i_t^s}[(\mathrm{dist}(\mathbf{U}_{t-1}^s, \mathbf{U}_t^s))^2 + 2\eta\langle \mathbf{grad} f(\mathbf{U}_{t-1}^s), \mathrm{Exp}_{\mathbf{U}_{t-1}^s}^{-1}(\mathbf{U}^*)\rangle_{\mathbf{U}_{t-1}^s}]. \quad (3)$$
- Putting (1)&(2) into (3) with summing over the inner loop finally yields the decrease of the distance to the solution in the outer iteration.

Numerical comparisons

- Evaluate the case of exponential mapping and parallel translation.
  - The tools in the Grassmann manifold have closed-form expressions [1].
- PCA problem (Synthtic dataset: $n = 10000$, $d = 20$, and $r = 5$)

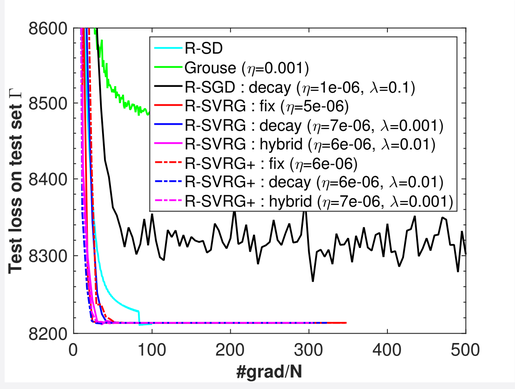


(a) Optimality gap.

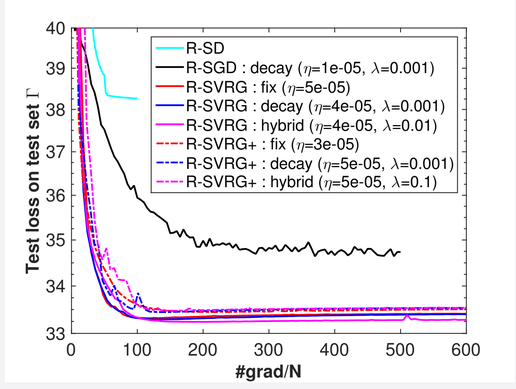


(b) Norm of gradient.

- Matrix completion problem (Jester and MovieLens10M)



(a) Jester dataset.



(b) MovieLens10M dataset.

Reference

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ, 2008.
[2] R. Johnson and T. Zhang. "Accelerating stochastic gradient descent using predictive variance reduction," In NIPS, pages 315-323, 2013.
[3] H. Kasai, H. Sato, and B. Mishra, "Riemannian stochastic variance reduced gradient on Grassmann manifold," arXiv preprint: arXiv:1605.07367, 2016.
[4] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. IEEE Trans. on Automatic Control, 58(9):2217-2229, 2013.

kasai@is.uec.ac.jp, bamdevm@amazon.com