# Riemannian Stochastic Recursive Gradient Algorithm

Hiroyuki Kasai (The University of Electro-Communications, Japan), Hiroyuki Sato (Kyoto University, Japan), Bamdev Mishra (Microsoft, India)

## Problem of interest

- Consider

$$\min_{w \in \mathcal{M}} \left\{ f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w) \right\}.$$

  - $n$ is # of samples.
  - $w$ is on Riemannian manifold $\mathcal{M}$ [1].
- Applications include, for example,
  - matrix/tensor completion, subspace tracking.

## Contributions

- Propose Riemannian stochastic recursive gradient algorithm (R-SRG).
- Our analysis deals with both (strongly) retraction-convex and non-convex functions.
- Our analysis considers computationally efficient retraction and vector transport instead of the more restrictive exponential mapping and parallel translation.
- Obtained rates for both functions.
- The obtained total complexity is the first result with respect to retraction and vector transport.

## Proposed R-SRG algorithm

- Generalization of [4] into a Riemannian setting.
- Update $w_{t+1} = R_{w_t}(-\alpha_t v_t)$, where $v_t$ is recursively updated as

$$v_t = \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}_{w_{t-1}}^{w_t}(\underbrace{\mathrm{grad} f_{i_t}(w_{t-1}) - v_{t-1}}_{\text{at previous } w_{t-1}}).$$

  - $R_w$: retraction.
  - $\mathcal{T}_{w_{t-1}}^{w_t}$: vector transport from $w_{t-1}$ to $w_t$.
  - c.f. R-SVRG algorithm [2,3]

$$\xi_t = \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}_{\tilde{w}}^{w_t}(\underbrace{\mathrm{grad} f_{i_t}(\tilde{w}) - \mathrm{grad} f(\tilde{w})}_{\text{at outer loop edge } \tilde{w}}).$$

- $\mathbb{E}[v_t|\mathcal{F}_t] \neq \mathrm{grad} f(w_t)$, but $\mathbb{E}[v_t] = \mathbb{E}[\mathrm{grad} f(w_t)]$.
- R-SRG+: Adaptive length of inner-loop.
  - Exploit the linearly convergent property of $\|v_t\|_{w_t}$ in retraction-convex functions.
  - Stop inner loop when the norm of $v_t$ decreases below the threshold of that of $v_0$.

## Advantages of R-SRG

- Transport vectors from the previous iterate.
  - R-SVRG transports between two distant iterates [2,3].
  - More notable than the Euclidean case [4].
- Computationally efficient due to no calculation of inverse of retraction.
- Provide an accelerated variant R-SRG+ in retraction-convex functions.
- Applicable to a wider range of manifolds (e.g., the Stiefel and fixed-rank manifolds).

### R-SRG algorithm

**Require:** Update frequency $m$ and sequence $\{\alpha_t\}$ with $\alpha_t > 0$.
Initialize $\tilde{w}^0$.
**for** $s = 1, 2, \ldots$ **do**
  Store $w_0 = \tilde{w}^{s-1}$.
  Calculate Riemannian full gradient $\mathrm{grad} f(w_0)$.
  Store $v_0 = \mathrm{grad} f(w_0)$.
  Update $w_1 = R_{w_0}(-\alpha_0 v_0)$.
  **for** $t = 1, 2, \ldots, m-1$ **do**
    Choose $i_t \in [n]$ uniformly at random.
    Calculate $v_t = \mathrm{grad} f_{i_t}(w_t) - \mathcal{T}_{w_{t-1}}^{w_t}(\mathrm{grad} f_{i_t}(w_{t-1}) - v_{t-1})$.
    Update $w_{t+1} = R_{w_t}(-\alpha_t v_t)$.
  **end for**
  Set $\tilde{w}^s = w_{t'}$ for randomly chosen $t' \in [n]$.
**end for**

## Key lemmas

- **Lemma 3.6:** (Difference between parallel translation and vector transport.) There exists a constant $\theta > 0$ such that

$$\|\mathcal{T}_\eta \xi - P_\eta \xi\|_z \leq \theta \|\xi\|_w \|\eta\|_w,$$

  where $\xi, \eta \in T_w\mathcal{M}$ and $R_w(\eta) = z$.
- **Lemma 3.7:** (Retraction $L_l$-Lipschitz.) There exists a constant $L_l > 0$ such that

$$\|P(\gamma)_z^w \mathrm{grad} f(z) - \mathrm{grad} f(w)\|_w \leq L_l \|\eta\|_w,$$

  where $L_l = C_h(1 + C_\eta \theta)$, with $C_\eta$ being the upper bound of the norm of $\eta$ for $\eta \in T_w\mathcal{M}$.
- **Lemma 3.8:** (Inner product.) There exists a constant $\nu > 0$ such that

$$\langle \xi, \mathrm{Exp}_w^{-1}(z) \rangle_w \leq \langle \xi, R_w^{-1}(z) \rangle_w + \nu \|R_w^{-1}(z)\|_w^2,$$

  where $\xi \in T_w\mathcal{M}$ and $\|\xi\|_w \leq 2C_g$, where $C_g$ is a constant in Assumption (1.4).

## Comparison of total complexity to achieve $\mathbb{E}[\|\mathrm{grad} f(w_T)\|^2] \leq \epsilon$

- Vector transport

| Function type | R-SRG (Proposed) |
|---|---|
| Retraction convex | $\mathcal{O}((n+\frac{1}{\epsilon})\log(\frac{1}{\epsilon}) / \log(c(1-\frac{\beta}{L^2})))$ |
| Retraction $\mu$-strongly convex | $\mathcal{O}((n+\kappa)\log(\frac{1}{\epsilon}) / \log(c(1-\frac{\beta}{L^2})))$ |
| Non-convex | $\mathcal{O}(n + \frac{L^2\rho_l^2+\theta^2}{\epsilon^2})$ |
| $\tau$-gradient dominated | $\mathcal{O}((n+\tau^2(L^2\rho_l^2+\theta^2))\log(\frac{1}{\epsilon}))$ |

- Parallel translation

| Function type | R-SRG (Proposed) | R-SVRG [3] |
|---|---|---|
| Geodesically convex | $\mathcal{O}((n+\frac{1}{\epsilon})\log(\frac{1}{\epsilon}))$ | – |
| Geodesically $\mu$-strongly convex | $\mathcal{O}((n+\kappa)\log(\frac{1}{\epsilon}))$ | $\mathcal{O}((n+\zeta\kappa^2)\log(\frac{1}{\epsilon}))$ |
| Non-convex | $\mathcal{O}(n+\frac{L^2}{\epsilon^2})$ | $\mathcal{O}(n+\zeta^{\frac{1}{2}}n^{\frac{2}{3}}/\epsilon)$ |
| $\tau$-gradient dominated | $\mathcal{O}((n+\tau^2L^2)\log(\frac{1}{\epsilon}))$ | $\mathcal{O}((n+L\tau\zeta^{\frac{1}{2}}n^{\frac{2}{3}})\log(\frac{1}{\epsilon}))$ |

$\beta = \mathcal{O}((L_l+\theta+L)\theta+\nu L)$, $\mathcal{O}(\beta/L^2) = \mathcal{O}(\rho_l\theta/L)$, $\kappa = L/\mu$, $\rho_l = L_l/L$

## Obtained convergence rates

- Linear rate for (strongly) retraction-convex functions.
- Linear rate of $\|v_t\|_{w_t}$ in the inner loop for retraction-convex functions.
- Sublinear rate in a single outer loop for general non-convex functions.
- Linear rate in for gradient-dominated functions.

## Parameters' impact on complexity

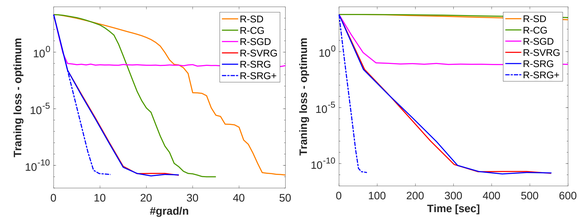
- Total complexity increases drastically when
  - $\rho_l (= L_l/L)$ increases, deviating from 1, when retraction curve deviates from geodesic.
  - $\theta$ deviates from 0 when vectors from vector transport deviate from those of parallel translation.
- Those deviations more strongly influence when non-convex cases than convex case.
- The complexities retain values similar to the case of exponential mapping and parallel translation.

## Comparison with R-SVRG

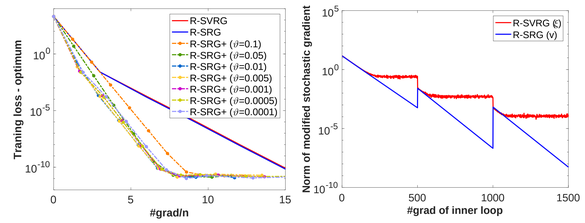
- ✓ Does not depend a curvature parameter $\zeta (\geq 1)$ as R-SVRG.
- ✓ Superior to R-SVRG in the geodesically $\mu$-strongly convex case.
- ✗ Inferior to R-SVRG comparing with a single outer loop for non-convex functions.
- ✓ Superior to R-SVRG with regard to $n$.
- ✓ Superior to R-SVRG for $\tau$-gradient dominated functions.

## Numerical evaluations

- Riemannian centroid problem

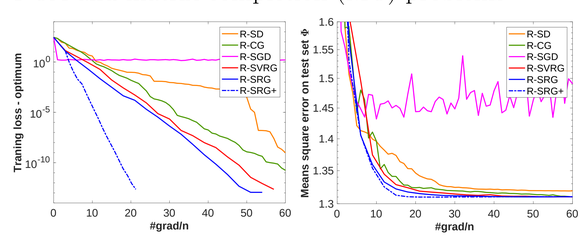



(a) Opti. gap vs. # of gradient evaluations. (b) Opti. gap vs. processing time.




(c) Influence of $\vartheta$ on R-SRG+. (d) Norm of modified stochastic gradient.

- PCA and matrix completion (MC) problems




(a) Optimality gap for PCA problem. (b) Test MSE for MC problem (Jester).

## References

1 Absil, P.-A., Mahony, R., and Sepulchre, R. Optimization Algorithms on Matrix Manifolds. Princeton University Press, 2008.
2 Sato, H., Kasai, H., and Mishra, B. Riemannian stochastic variance reduced gradient. arXiv preprint: arXiv:1702.05594, 2017.
3 Zhang, H., Reddi, S. J., and Sra, S. Fast stochastic optimization on Riemannian manifolds. In NIPS, 2016.
4 Nguyen, L. M., Liu, J., Scheinberg, K., and Takac, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In ICML, 2017.

The University of Electro-Communications UEC
KYOTO UNIVERSITY FOUNDED 1897
Microsoft