Nonconvex Stochastic Scaled Gradient Descent and Generalized Eigenvector Problems

Chris Junchi Li[◇] Michael I. Jordan[◇]

[◇] University of California, Berkeley

Algorithm Design

- Consider the following general constrained nonconvex optimization problem:

$$\min_{\boldsymbol{v}} F(\boldsymbol{v}), \qquad \text{subject to } \boldsymbol{v} \in \mathcal{C}$$

- SGD performs the following update at the t-th step ($t \geq 1$):

$$\boldsymbol{v}_t = \Pi_{\mathcal{C}}\left[\boldsymbol{v}_{t-1} - \eta \widetilde{\nabla} F(\boldsymbol{v}_{t-1}; \zeta_t)\right]$$

where $\Pi_{\mathcal{C}}[\cdot]$ denotes projection operator onto $\mathcal{C}$, and $\widetilde{\nabla} F(\boldsymbol{v}_{t-1}; \zeta_t)$ denotes unbiased gradient estimator of $\nabla F(\boldsymbol{v}_{t-1})$

- What if there is no access to $\widetilde{\nabla} F(\boldsymbol{v}_{t-1}; \zeta_t)$, but instead stochastic vector $\Gamma(\boldsymbol{v}; \zeta)$ as unbiased estimate of *scaled* gradient:

$$\mathbb{E}_{\zeta}[\Gamma(\boldsymbol{v}; \zeta)] = D(\boldsymbol{v}) \nabla F(\boldsymbol{v})$$

- Generalized eigenvector computation (GEV) (Principal component analysis (PCA), Partial least squares regression, Fisher's linear discriminant analysis (LDA), canonical correlation analysis (CCA), etc.)

Stochastic Scaled-Gradient Descent

- SSGD performs the update:

$$\boldsymbol{v}_t = \Pi_{\mathcal{C}}\left[\boldsymbol{v}_{t-1} - \eta\Gamma(\boldsymbol{v}_{t-1}; \zeta_t)\right] \ \text{ where } \mathbb{E}_{\zeta}[\Gamma(\boldsymbol{v}; \zeta)] = D(\boldsymbol{v}) \nabla F(\boldsymbol{v})$$

- **Example**: Generalized Rayleigh quotient given a unit spherical constraint:

$$\min_{\boldsymbol{v}} -\frac{\boldsymbol{v}^\top \mathbf{A} \boldsymbol{v}}{\boldsymbol{v}^\top \mathbf{B} \boldsymbol{v}} \qquad \text{subject to } \boldsymbol{v} \in \mathcal{S}^{d-1} = \{\boldsymbol{v} \in \mathbb{R}^d : \ \|\boldsymbol{v}\| = 1\}$$

- The first-order derivative with respect to $\boldsymbol{v}$

$$\nabla_{\boldsymbol{v}}\left[-\frac{\boldsymbol{v}^\top \mathbf{A} \boldsymbol{v}}{\boldsymbol{v}^\top \mathbf{B} \boldsymbol{v}}\right] = -\frac{(\boldsymbol{v}^\top \mathbf{B} \boldsymbol{v})\mathbf{A}\boldsymbol{v} - (\boldsymbol{v}^\top \mathbf{A} \boldsymbol{v})\mathbf{B}\boldsymbol{v}}{(1/2)(\boldsymbol{v}^\top \mathbf{B} \boldsymbol{v})^2}$$

- Replacing the denominator, denoted as $D(\boldsymbol{v})$, by the constant 1:

$$\boldsymbol{v}_t = \Pi_{\mathcal{S}^{d-1}}\left[\boldsymbol{v}_{t-1} + \eta\left((\boldsymbol{v}_{t-1}^\top \widetilde{\mathbf{B}}' \boldsymbol{v}_{t-1})\widetilde{\mathbf{A}}\boldsymbol{v}_{t-1} - (\boldsymbol{v}_{t-1}^\top \widetilde{\mathbf{A}} \boldsymbol{v}_{t-1})\widetilde{\mathbf{B}}'\boldsymbol{v}_{t-1}\right)\right]$$

where the bracketed term is unbiased estimate of $-\Gamma(\boldsymbol{v}; \zeta)$

Previous Works

- Oja's online PCA iteration [Oja82] (Special case where $\widetilde{\mathbf{B}}$ is taken as $\mathbf{I}$)
- Procedures for efficient online canonical eigenvectors estimation has been explored [AMMS17, GGS+19, CLY+19].
- [BPF+18] studied the CCA problem and proposed a two-time-scale online iteration ("Gen-Oja"), obtained $1/\sqrt{N}$.

Our Contributions

- We propose the (SSGD) algorithm—which generalizes the classical SGD algorithm and has a wider range of applications.
- We provide a local convergence analysis for convex spherical-constraint objective functions. Starting with a warm initialization, matches a known information-theoretic lower bound[MBM18].
- By applying SSGD to the GEV problem, we give a positive answer to the question raised by [ACLS12] regarding to the existence of an efficient online GEV algorithm. Specifically, in the case of CCA, our SSGD algorithm uses as few as two samples at each update, does not incur intermediate and expensive computational cost while achieving a polynomial convergence rate guarantee

Theoretical Results: Assumptions

Initialization:

$$\|\boldsymbol{v}_0 - \boldsymbol{v}^*\| \leq \min\left\{\frac{D\mu}{2^5 \rho}, \delta\right\} \qquad (1)$$

Assumption (Smoothness Assumption): For any $\boldsymbol{v} \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$, we assume that $D(\boldsymbol{v})$ is L_D-Lipschitz, $F(\boldsymbol{v})$ is L_F-Lipschitz, $\nabla F(\boldsymbol{v})$ is L_K-Lipschitz and $\nabla^2 F(\boldsymbol{v})$ is L_Q-Lipschitz, where L_D, L_F, L_K, L_Q are fixed positive constants.

Assumption (Sub-Weibull Tail): For some fixed $\mathcal{V} \in (0, \infty)$ and for all $\boldsymbol{v} \in \mathcal{C}$, we assume that the stochastic vectors $\Gamma(\boldsymbol{v}; \zeta)$ satisfy

$$\mathbb{E} \exp\left(\left\|\frac{\Gamma(\boldsymbol{v}; \zeta)}{\mathcal{V}}\right\|^{\alpha}\right) \leq 2 \qquad (2)$$

Finite-Sample Convergence Rate

Corollary (Finite-Sample): Assume Assumptions 1 and 2 and the initialization condition (1). For fixed positive constants ϵ and sample size T, set the step size as $\eta(T) = \Theta\left(\frac{\log T}{D\mu T}\right)$, satisfying some scaling condition, there exists an event $\mathcal{H}$ with

$$\mathbb{P}(\mathcal{H}) \geq 1 - \left(14 + 8\left(\frac{3}{\alpha}\right)^{\frac{2}{\alpha}} \log^{-\frac{\alpha+2}{\alpha}} \epsilon^{-1}\right) T\epsilon,$$

such that on the event $\mathcal{H}$ the iterates generated by the SSGD algorithm satisfy

$$\|\boldsymbol{v}_T - \boldsymbol{v}^*\| \lesssim \frac{G_\alpha \mathcal{V}}{D\mu} \log^{\frac{\alpha+2}{2\alpha}} \epsilon^{-1} \sqrt{\frac{\log T}{T}}.$$

- In the case of CCA, the ($\alpha = 1/2$) sub-Weibull parameter $\mathcal{V}$ in that case scales with $\sqrt{d}$ and thus the local rate is the minimax-optimal rate $O(\sqrt{d/T})$ up to a polylogarithmic factor.

Asymptotic Normality via Trajectory Averaging

Assumption (Mean-Squared Smoothness): There exists a positive constant L_S such that for all $\boldsymbol{v}, \boldsymbol{v}' \in \{\boldsymbol{v} : \|\boldsymbol{v}\| \leq 1, \|\boldsymbol{v} - \boldsymbol{v}^*\| \leq \delta\}$ and $t \geq 1$, we have for $\boldsymbol{\zeta}$

$$\mathbb{E}\left\|\Gamma(\boldsymbol{v}; \boldsymbol{\zeta}) - \Gamma(\boldsymbol{v}'; \boldsymbol{\zeta})\right\|^2 \leq L_S^2 \|\boldsymbol{v} - \boldsymbol{v}'\|^2$$

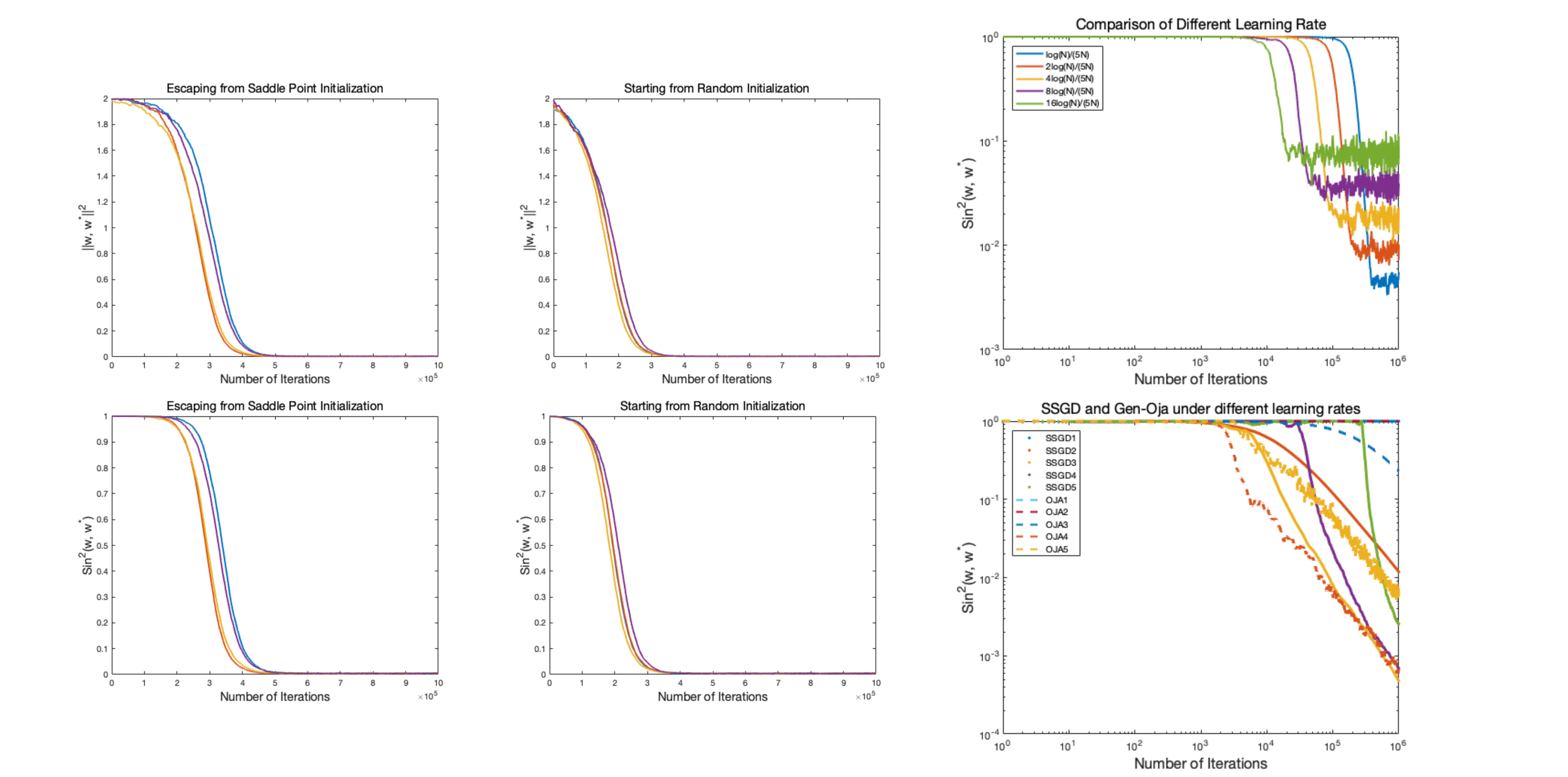
Theorem (Asymptotic Normality): Assume Assumptions 1, 2, 3 and initialization condition (1). If we choose the step size η such that $\eta \to 0$ as the total sample size $T \to \infty$, where

$$T\eta^2 \log^{\frac{2\alpha+4}{\alpha}} T \to 0, \quad T\eta \log^{-\frac{\alpha+2}{\alpha}} T \to \infty \quad \text{a.s.}$$

we obtain Gaussian convergence in distribution:

$$\sqrt{T}\left(\overline{\boldsymbol{v}}_T^{(\eta)} - \boldsymbol{v}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, D^{-2} \cdot \mathcal{M}_*^- \boldsymbol{\Sigma}_* \mathcal{M}_*^-\right)$$

Experiments



- Potential future works: Sharper rate of escape of saddle points for SSGD, study global convergence for generic Riemannian manifolds, etc.

Reference

[ACLS12] Raman Arora, Andrew Cotter, Karen Livescu, and Nathan Srebro. Stochastic optimization for PCA and PLS. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 861–868, 2012.

[AMMS17] Raman Arora, Teodor Vanislavov Marinov, Poorya Mianjy, and Nati Srebro. Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, pages 4775–4784, 2017.

[BPF+18] Kush Bhatia, Aldo Pacchiano, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Gen-Oja: Simple & efficient algorithm for streaming generalized eigenvector computation. In *Advances in Neural Information Processing Systems*, pages 7016–7025, 2018.

[CLY+19] Zhehui Chen, Xingguo Li, Lin Yang, Jarvis Haupt, and Tuo Zhao. On constrained nonconvex stochastic optimization: A case study for generalized eigenvalue decomposition. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 916–925, 2019.

[GGS+19] Chao Gao, Dan Garber, Nathan Srebro, Jialei Wang, and Weiran Wang. Stochastic canonical correlation analysis. *Journal of Machine Learning Research*, 20(167):1–46, 2019.

[MBM18] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[Oja82] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.