Efficient Convex Relaxations for Streaming PCA

Raman Arora, Teodor V. Marinov
Johns Hopkins University, Baltimore, MD 21204

1. Streaming PCA in Stochastic Setting

- Given a sequence of vectors $(\mathbf{x}_t)_{t=1}^{\infty}$ i.i.d. $\mathbf{x}_t \sim \mathcal{D}$.
- Minimize reconstruction error $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}-\mathrm{P}\mathbf{x}\|^2$, where $\mathrm{P} \in \mathcal{P}_k = \{\mathrm{P} : \mathrm{P}^2 = \mathrm{P}, \mathrm{P}^\top = \mathrm{P}, \mathrm{rank}(\mathrm{P}) = k\}$.
- Equivalently solve:

$$\begin{aligned} &\underset{\mathrm{P}\in\mathbb{R}^{d\times d}}{\text{maximize}} && \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\langle \mathrm{P}, \mathbf{x}\mathbf{x}^\top\rangle \\ &\text{subject to} && \mathrm{P}^2 = \mathrm{P}, \mathrm{P}^\top = \mathrm{P}, \mathrm{rank}(\mathrm{P}) = k \end{aligned} \tag{1}$$

- Usually solved by Oja's algorithm:

$$\mathrm{U}_{t+1} \leftarrow (\mathrm{I} + \eta_t \mathbf{x}_t\mathbf{x}_t^\top)\mathrm{U}_t, \mathrm{U}_{t+1} \leftarrow \mathrm{Orth}(\mathrm{U}_{t+1})$$
$$\mathrm{P}_{t+1} \leftarrow \mathrm{U}_{t+1}\mathrm{U}_{t+1}^\top$$

2. A Convex Relaxation

- Convexify $\mathcal{P}_k \to \mathcal{C} = \{\mathrm{P} : \mathrm{Tr}(\mathrm{P}) = k, 0 \preceq \mathrm{P} \preceq \mathrm{I}, \mathrm{P} = \mathrm{P}^\top\}$.
- Can now solve the following convex optimization problem:

$$\begin{aligned} &\underset{\mathrm{P}\in\mathbb{R}^{d\times d}}{\text{maximize}} && \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\langle \mathrm{P}, \mathbf{x}\mathbf{x}^\top\rangle \\ &\text{subject to} && \mathrm{P} \in \mathcal{C} \end{aligned} \tag{2}$$

- Can also add regularization to achieve distribution dependent guarantees:

$$\begin{aligned} &\underset{\mathrm{P}\in\mathbb{R}^{d\times d}}{\text{maximize}} && \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\langle \mathrm{P}, \mathbf{x}\mathbf{x}^\top\rangle - \frac{\lambda}{2}\|\mathrm{P}\|_F^2 \\ &\text{subject to} && \mathrm{P} \in \mathcal{C} \end{aligned} \tag{3}$$

3. Prior Work Guarantees

- Let P^* be the optimal solution to Problem 1, $\mathrm{C} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, and $\Delta(\mathrm{C}) = \lambda_k(\mathrm{C}) - \lambda_{k+1}(\mathrm{C})$.
- Projected Stochastic Gradient Descent on Problem 2 (MSG) is:
$$\mathrm{P}_{t+1} = \Pi_{\mathcal{C}}(\mathrm{P}_t + \eta_t \mathbf{x}_t\mathbf{x}_t^\top)$$
- Projected Stochastic Gradient Descent on Problem 3 (RMSG) is:
$$\mathrm{P}_{t+1} = \Pi_{\mathcal{C}}((1-\lambda\eta_t)\mathrm{P}_t + \eta_t \mathbf{x}_t\mathbf{x}_t^\top)$$
- MSG and RMSG guarantee that $\mathbb{E}\langle \mathrm{P}^* - \mathrm{P}_t, \mathrm{C}\rangle$ is in $\tilde{O}(1/\sqrt{t})$ and $\tilde{O}(1/(\Delta(\mathrm{C})t))$ [2].
- RMSG is statistically optimal, however, MSG and RMSG can take up to $\Omega(d^3)$ computations per iteration!

4. Meta-algorithm and Intuition

Input: Stream of data $\{\mathbf{x}_{t_i}\}$, parameters $\Delta(\mathrm{C})$, probability of failure δ, number of components k
Output: P_T

1: Initialize P_1 from a warm start
2: **for** $t = 1, \ldots, T-1$ **do**
3: $\eta_t = \Theta\left(\frac{1}{\Delta(\mathrm{C})t}\right)$
4: Form unbiased mini-batched estimator of gradient: $\mathrm{C}_t = \frac{1}{n}\sum_{l=1}^{n} \mathbf{x}_{t_l}\mathbf{x}_{t_l}^\top$
5: $\mathrm{P}_{t+1/2} \leftarrow (1 - \frac{\Delta(\mathrm{C})}{2}\eta_t)\mathrm{P}_t + \eta_t\mathrm{C}_t$
6: $\mathrm{P}_{t+1} = \Pi(\mathrm{P}_{t+1/2})$ (gradient descent upgrade) or equivalently
7: $\mathrm{U}_{t+1} = \text{Top-k}\left([\sqrt{1-\Delta(\mathrm{C})\eta_t/2}\,\mathrm{U}_t, \sqrt{\eta_t}\mathbf{X}_t]\right)$
8: $\mathrm{P}_{t+1} = \mathrm{U}_{t+1}\mathrm{U}_{t+1}^\top$ (efficient update)
9: **end for**

Lemma 1 *Let P_t be rank k and suppose $\|C - C_t\| \leq \epsilon$. Then a sufficient condition for P_{t+1} to be rank k is*

$$\langle P_t, C\rangle \geq \langle P^*, C\rangle - \frac{\Delta(C)}{4} + \epsilon(k+1). \tag{4}$$

Intuition behind algorithm: Lemma 1 gives a sufficient conditions for P_{t+1} to be a projection matrix, given that P_t is a projection matrix. The sufficient condition translates to:

- P_t is close enough to P^* in objective and hence the warm start of the algorithm
- C_t is close enough to C with high probability which results in the mini-batched stochastic gradients

5. Main results

High Probability Guarantees

The following holds for Algorithm **MB-MSG**: with probability at least $1-\delta$, for all $t \leq T$

$$\langle \mathrm{P}^* - \mathrm{P}_t, \mathrm{C}\rangle \leq O\left(\frac{k^4 \log(1/\delta)(\log(T))^2}{\sqrt{t + \frac{1}{\gamma}}}\right),$$

where $\gamma = O\left(\frac{\Delta(\mathrm{C})^2}{(k\log(1/\delta))^2}\right)$. Further, it holds that P_t is a rank-k projection matrix.

The following holds for Algorithm **MB-RMSG**: with probability at least $1-\delta$, for all $t \leq T$

$$\langle \mathrm{P}^* - \mathrm{P}_t, \mathrm{C}\rangle \leq \frac{32\log(3e/\delta)}{\Delta(\mathrm{C})^2\left(t + \frac{1}{\gamma} - 1\right)},$$

where $\gamma = \frac{\Delta(\mathrm{C})^3}{128\log(1/\delta)}$. Further, it holds that P_t is a rank-k projection matrix.

Guarantees in Expectation

Let $\mathcal{A}$ be the event that for all $t \in [T]$ it holds that $\|\mathrm{C}_t - \mathrm{C}\| \leq \frac{\Delta(C)}{8(k+1)}$ and P_t is a rank-k projection matrix. Then Algorithm **MB-RMSG** guarantees that $\mathcal{A}$ occurs with probability at least $1-\delta$ and that

$$\mathbb{E}\left[\langle \mathrm{P}^* - \mathrm{P}_T, \mathrm{C}\rangle | \mathcal{A}\right] \leq \tilde{O}\left(\frac{\Delta(\mathrm{C})}{T} + \min(\Delta(\mathrm{C}) \times d, 1)\frac{1}{kT}\right).$$

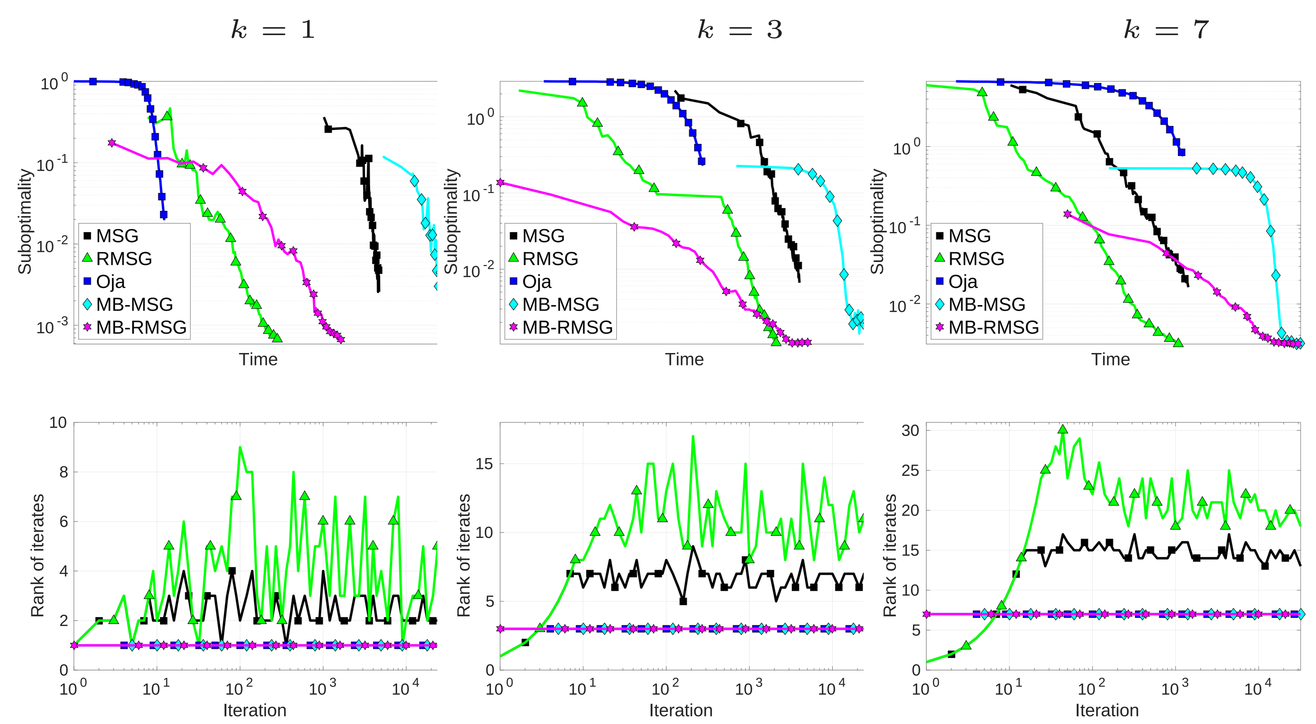
Comparison to Oja's Algorithm

(*Informal [1]*) The following holds for Algorithm **Oja's Algorithm**: with probability at least $1-\delta$, for all $t \leq T$

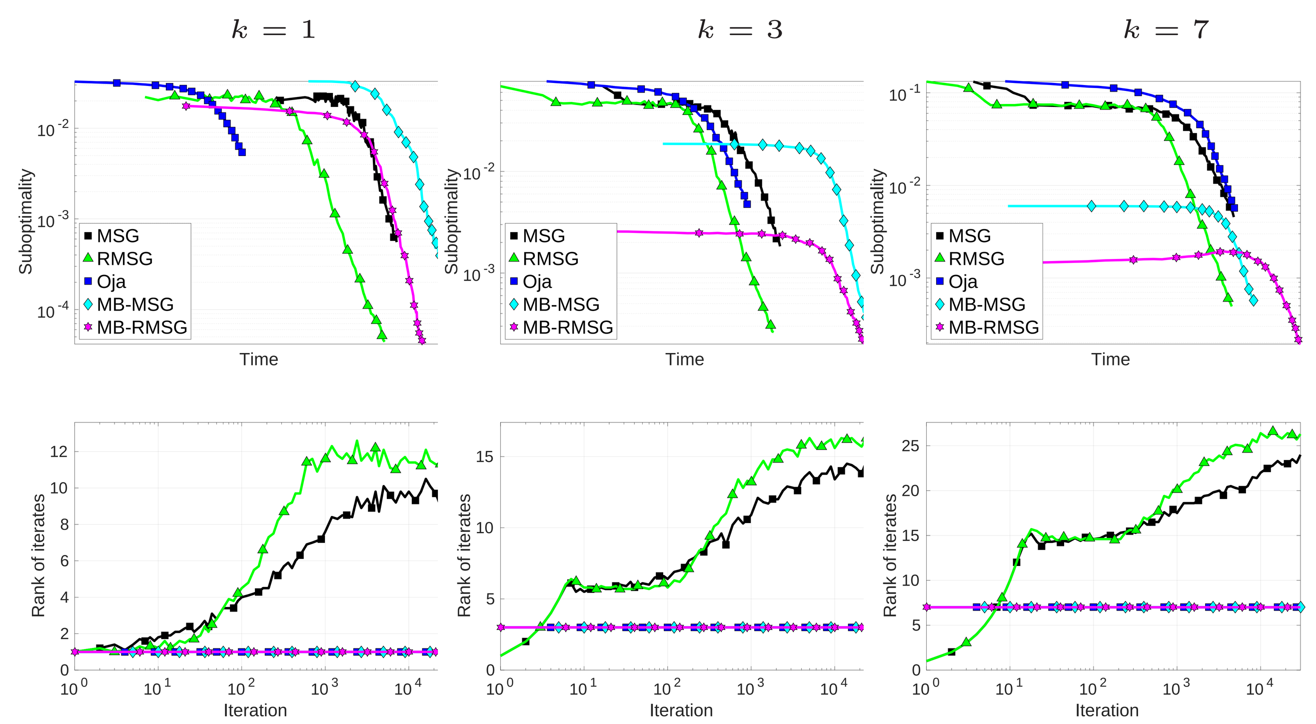
$$\langle \mathrm{P}^* - \mathrm{P}_t, \mathrm{C}\rangle \leq \tilde{O}\left(\frac{1}{t\Delta(\mathrm{C})^2}\right)$$

- Total computational complexity for ϵ-suboptimality for **MB-RMSG** is $\tilde{O}\left(dk^2/(\epsilon\Delta(\mathrm{C})^2) \times \min(d\Delta(\mathrm{C})), 1)\right)$
- Total computational complexity for ϵ-suboptimality for **Oja's Algorithm** is $\tilde{O}\left(dk/(\epsilon\Delta(\mathrm{C})^2)\right)$

6. Experimental Results



Comparisons of Oja, MSG, RMSG, MB-MSG and MB-RMSG on **synthetic data on the left** and **MNIST, on the right**, in terms of runtime (top) and rank of iterates (bottom).



7. REFERENCES

[1] Allen-Zhu, Z. and Li. Y. First efficient convergence for streaming k-pca: a global, gap-free, and near-optimal rate. *FOCS*, 2017.
[2] Mianjy, P. and Arora, R. Stochastic PCA with ℓ_2 and ℓ_1 Regularization. *ICML*, 2018.