A Stochastic PCA Algorithm with an Exponential Convergence Rate

Ohad Shamir

Weizmann Institute of Science

NIPS Optimization Workshop
December 2014

Principal Component Analysis

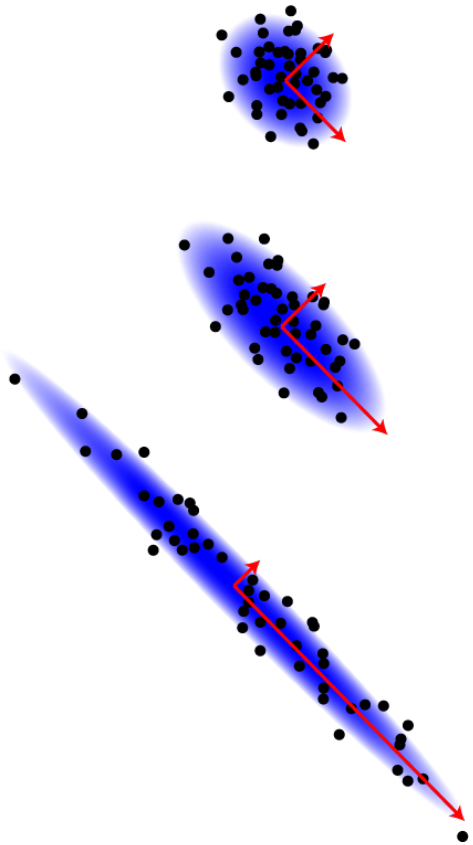
PCA

- Input: $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$
- Goal: Find k directions with most variance

$$\max_{U \in \mathbb{R}^{d \times k} : U^\top U = I} \frac{1}{n} \sum_{i=1}^{n} \left\| U^\top \mathbf{x} \right\|^2$$

For $k = 1$: Find leading eigenvector of covariance matrix

$$\max_{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| = 1} \mathbf{w}^\top \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{w}$$



Existing Approaches

$$\max_{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|=1} \mathbf{w}^\top \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{w}$$

Regime: n, d "large", non-sparse matrix

Existing Approaches

$$\max_{\mathbf{w}\in\mathbb{R}^d:\|\mathbf{w}\|=1} \mathbf{w}^\top \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{w}$$

Regime: n, d "large", non-sparse matrix

Approach 1: Eigendecomposition

- Compute leading eigenvector of $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ exactly (e.g. via QR decomposition)
- Runtime: $\mathcal{O}(\mathbf{d^3})$

Existing Approaches

Approach 2: Power Iterations

- Initialize $\mathbf{w}_1$ randomly on unit sphere
- For $t = 1, 2, \ldots$
  - $\mathbf{w}'_{t+1} := \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{w}_t = \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}_t, \mathbf{x}_i\rangle \mathbf{x}_i$
  - $\mathbf{w}_{t+1} := \mathbf{w}'_{t+1} / \left\|\mathbf{w}'_{t+1}\right\|$

- $\mathcal{O}\left(\frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations for ϵ-optimality
  - λ: Eigengap
- $\mathcal{O}(\mathbf{nd})$ runtime per iteration
- Overall runtime $\mathcal{O}\left(\frac{\mathbf{nd}}{\lambda}\log\left(\frac{\mathbf{d}}{\epsilon}\right)\right)$

Existing Approaches

Approach 2: Power Iterations

- Initialize $\mathbf{w}_1$ randomly on unit sphere
- For $t = 1, 2, \ldots$
  - $\mathbf{w}'_{t+1} := \left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{w}_t = \frac{1}{n}\sum_{i=1}^n \langle \mathbf{w}_t, \mathbf{x}_i\rangle \mathbf{x}_i$
  - $\mathbf{w}_{t+1} := \mathbf{w}'_{t+1} / \left\|\mathbf{w}'_{t+1}\right\|$

- $\mathcal{O}\left(\frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations for ϵ-optimality
  - λ: Eigengap
- $\mathcal{O}(\mathbf{nd})$ runtime per iteration
- Overall runtime $\mathcal{O}\left(\frac{\mathbf{nd}}{\lambda}\log\left(\frac{\mathbf{d}}{\epsilon}\right)\right)$

Approach 2.5: Lanczos Iterations

- More complex algorithm, but roughly similar iteration runtime and only $\mathcal{O}\left(\frac{1}{\sqrt{\lambda}}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations [Kuczyński and Woźniakowski 1989]
- Overall runtime $\mathcal{O}\left(\frac{\mathbf{nd}}{\sqrt{\lambda}}\log\left(\frac{\mathbf{d}}{\epsilon}\right)\right)$

Existing Approaches

Approach 3: Stochastic/Incremental Algorithms

Example (Oja's algorithm)

- Initialize $\mathbf{w}_1$ randomly on unit sphere
- For $t = 1, 2, \ldots$
  - Pick $i_t \in \{1, \ldots, n\}$ (randomly or otherwise)
  - $\mathbf{w}'_{t+1} := \mathbf{w}_t + \eta_t \mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top \mathbf{w}_t$
  - $\mathbf{w}_{t+1} := \mathbf{w}'_{t+1} / \left\|\mathbf{w}'_{t+1}\right\|$

Also Krasulina 1969; Arora, Cotter, Livescu, Srebro 2012; Mitliagkas, Caramanis, Jain 2013; De Sa, Olukotun, Ré 2014...

Existing Approaches

Approach 3: Stochastic/Incremental Algorithms

Example (Oja's algorithm)

- Initialize $\mathbf{w}_1$ randomly on unit sphere
- For $t = 1, 2, \ldots$
  - Pick $i_t \in \{1, \ldots, n\}$ (randomly or otherwise)
  - $\mathbf{w}'_{t+1} := \mathbf{w}_t + \eta_t \mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top \mathbf{w}_t$
  - $\mathbf{w}_{t+1} := \mathbf{w}'_{t+1} / \left\|\mathbf{w}'_{t+1}\right\|$

Also Krasulina 1969; Arora, Cotter, Livescu, Srebro 2012; Mitliagkas, Caramanis, Jain 2013; De Sa, Olukotun, Ré 2014...

- $\mathcal{O}(d)$ runtime per iteration
- Iteration bounds:
  - Balsubramani, Dasgupta, Freund 2013: $\tilde{\mathcal{O}}\left(\frac{d}{\lambda^2}\left(\frac{1}{\epsilon} + d\right)\right)$
  - De Sa, Olukotun, Ré 2014: For a different SGD method, $\tilde{\mathcal{O}}\left(\frac{d}{\lambda^2 \epsilon}\right)$
- Runtime: $\tilde{\mathcal{O}}\left(\frac{\mathbf{d}^2}{\lambda^2 \epsilon}\right)$

Existing Approaches

Up to constants/log-factors:

Algorithm	Time per iter.	# iter.	Runtime
Exact			d^3
Power/Lanczos	nd	$\frac{1}{\lambda^p}$	$\frac{nd}{\lambda^p}$
Incremental	d	$\frac{d}{\lambda^2\epsilon}$	$\frac{d^2}{\lambda^2\epsilon}$

Main Question

Can we get the best of both worlds? $\mathcal{O}(d)$ time per iteration **and** fast convergence (logarithmic dependence on ϵ?)

Convex Optimization to the Rescue?

Our problem is equivalent to:

$$\min_{\mathbf{w}:\|\mathbf{w}\|=1} \frac{1}{n}\sum_{i=1}^{n}\left(-\langle \mathbf{w}, \mathbf{x}_i\rangle^2\right)$$

Much recent progress in solving strongly convex + smooth problems with finite-sum structure

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

Stochastic algorithms with $\mathcal{O}(d)$ runtime per iteration and exponential convergence

[Le Roux, Schmidt, Bach 2012; Shalev-Shwartz and Zhang 2012; Johnson and Zhang 2013; Zhang, Mahdavi, Jin 2013; Konečný and Richtárik 2013; Xiao and Zhang 2014; Zhang and Xiao, 2014...]

Convex Optimization to the Rescue?

$$\min_{\mathbf{w}:\|\mathbf{w}\|=1} \frac{1}{n}\sum_{i=1}^{n}\left(-\langle \mathbf{w}, \mathbf{x}_i\rangle^2\right)$$

Unfortunately:

- Function not strongly convex, or even convex (in fact, concave everywhere)
- Has > 1 global optima, plateaus...

$\Rightarrow$ Existing results don't work as-is

But: Maybe we can borrow some ideas...

Algorithm

$$\min_{\mathbf{w}:\|\mathbf{w}\|=1} \frac{1}{n}\sum_{i=1}^{n}\left(-\langle \mathbf{w}, \mathbf{x}_i\rangle^2\right)$$

Oja Iteration

- Choose $i_t \in \{1,\ldots,n\}$ at random
- $\mathbf{w}'_{t+1} = \mathbf{w}_t + \eta_t \langle \mathbf{w}_t, \mathbf{x}_{i_t}\rangle \mathbf{x}_{i_t}$
- $\mathbf{w}_{t+1} := \mathbf{w}'_{t+1} / \left\|\mathbf{w}'_{t+1}\right\|$

Essentially projected stochastic gradient descent

Algorithm

Letting $A = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$, update step is

$$\begin{aligned}
\mathbf{w}'_{t+1} &= \mathbf{w}_t + \eta_t \mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top \mathbf{w}_t \\
&= \mathbf{w}_t + \underbrace{\eta_t A \mathbf{w}_t}_{\text{power/gradient step}} + \underbrace{\eta_t \left(\mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top - A \right) \mathbf{w}_t}_{\text{zero-mean noise}}
\end{aligned}$$

Algorithm

Letting $A = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$, update step is

$$\begin{aligned}
\mathbf{w}'_{t+1} &= \mathbf{w}_t + \eta_t \mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top \mathbf{w}_t \\
&= \mathbf{w}_t + \underbrace{\eta_t A \mathbf{w}_t}_{\text{power/gradient step}} + \underbrace{\eta_t \left(\mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top - A \right) \mathbf{w}_t}_{\text{zero-mean noise}}
\end{aligned}$$

Main idea: Replace by

$$\mathbf{w}'_{t+1} = \mathbf{w}_t + \underbrace{\eta A \mathbf{w}_t}_{\text{power/gradient step}} + \underbrace{\eta \left(\mathbf{x}_{i_t} \mathbf{x}_{i_t}^\top - A \right) (\mathbf{w}_t - \tilde{\mathbf{u}})}_{\text{zero-mean noise}}$$

where $\tilde{\mathbf{u}}$ "close" to $\mathbf{w}_t$
(similar to SVRG of Johnson and Zhang (2013))

Algorithm

VR-PCA

- **Parameters:** Step size η, epoch length m
- **Input:** Data set $\{\mathbf{x}_i\}_{i=1}^n$, Initial unit vector $\tilde{\mathbf{w}}_0$
- For $s = 1, 2, \ldots$
  - $\tilde{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top \tilde{\mathbf{w}}_{s-1}$
  - $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$
  - For $t = 1, 2, \ldots, m$
    - Pick $i_t \in \{1, \ldots, n\}$ uniformly at random
    - $\mathbf{w}'_t = \mathbf{w}_{t-1} + \eta\left(\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top(\mathbf{w}_{t-1} - \tilde{\mathbf{w}}_{s-1}) + \tilde{\mathbf{u}}\right)$
    - $\mathbf{w}_t = \frac{1}{\|\mathbf{w}'_t\|}\mathbf{w}'_t$
  - $\tilde{\mathbf{w}}_s = \mathbf{w}_m$

Algorithm

VR-PCA

- **Parameters:** Step size η, epoch length m
- **Input:** Data set $\{\mathbf{x}_i\}_{i=1}^n$, Initial unit vector $\tilde{\mathbf{w}}_0$
- For $s = 1, 2, \ldots$
  - $\tilde{\mathbf{u}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top \tilde{\mathbf{w}}_{s-1}$
  - $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$
  - For $t = 1, 2, \ldots, m$
    - Pick $i_t \in \{1, \ldots, n\}$ uniformly at random
    - $\mathbf{w}'_t = \mathbf{w}_{t-1} + \eta\left(\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top(\mathbf{w}_{t-1} - \tilde{\mathbf{w}}_{s-1}) + \tilde{\mathbf{u}}\right)$
    - $\mathbf{w}_t = \frac{1}{\|\mathbf{w}'_t\|}\mathbf{w}'_t$
  - $\tilde{\mathbf{w}}_s = \mathbf{w}_m$

To get $k > 1$ directions: Either repeat, or perform orthogonal-like iterations:

- Replace all vectors by $k \times d$ matrices
- Replace normalization step by orthogonalization step

Analysis

Theorem

Suppose $\max_i \|\mathbf{x}_i\|^2 \leq r$, *and* A *has leading eigenvector* $\mathbf{v}_1$. *Assuming* $\langle \tilde{\mathbf{w}}_0, \mathbf{v}_1 \rangle \geq \frac{1}{\sqrt{2}}$, *then for any* $\delta, \epsilon \in (0,1)$, *if*

$$\eta \leq \frac{c_1 \delta^2}{r^2}\lambda \quad , \quad m \geq \frac{c_2 \log(2/\delta)}{\eta\lambda} \quad , \quad m\eta^2 r^2 + r\sqrt{m\eta^2 \log(2/\delta)} \leq c_3,$$

(where c_1, c_2, c_3 *are constants) and we run* $T = \left\lceil \frac{\log(1/\epsilon)}{\log(2/\delta)} \right\rceil$ *epochs, then* $\Pr\left(\langle \tilde{\mathbf{w}}_T, \mathbf{v}_1 \rangle^2 \geq 1 - \epsilon \right) \geq 1 - 2\log(1/\epsilon)\delta$

Corollary

Picking η, m appropriately, ϵ-convergence w.h.p.
in $\mathcal{O}\left(d\left(n + \frac{1}{\lambda^2}\right)\log\left(\frac{1}{\epsilon}\right)\right)$ runtime

- Exponential convergence with $\mathcal{O}(d)$-time iterations
- Proportional to # examples plus eigengap
- Proportional to single data pass if $\lambda \geq 1/\sqrt{n}$

Proof Idea

Track decay of $F(\mathbf{w}_t) = 1 - \langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2$

Key Lemma

Assuming $\eta = \alpha\lambda$ and $F(\mathbf{w}_t) \leq 3/4$,

$$\mathbb{E}\left[F(\mathbf{w}_{t+1})|\mathbf{w}_t\right] \;\leq\; \left(1 - \Theta(\alpha\lambda^2)\right) F(\mathbf{w}_t) + \mathcal{O}\left(\alpha^2\lambda^2 F(\tilde{\mathbf{w}}_{s-1})\right).$$

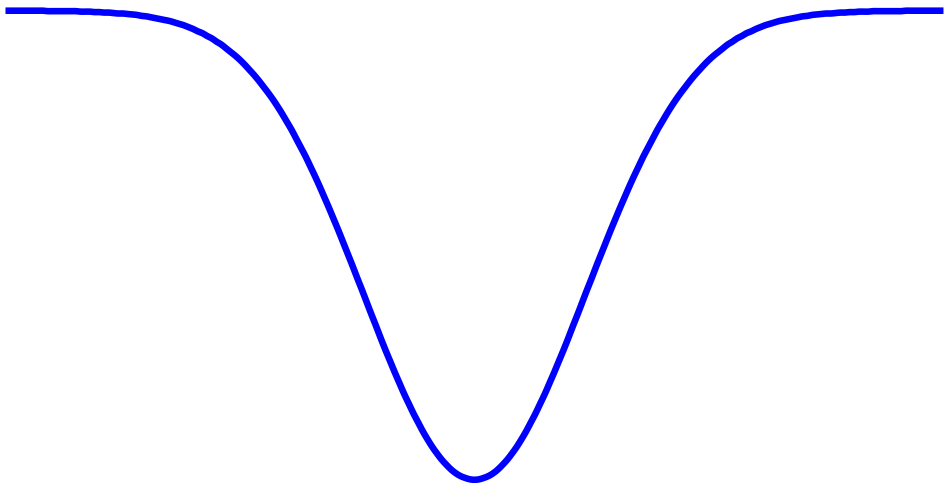
Proof Idea

Track decay of $F(\mathbf{w}_t) = 1 - \langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2$

Key Lemma

Assuming $\eta = \alpha\lambda$ and $F(\mathbf{w}_t) \leq 3/4$,

$$\mathbb{E}\left[F(\mathbf{w}_{t+1})|\mathbf{w}_t\right] \;\leq\; \left(1 - \Theta(\alpha\lambda^2)\right) F(\mathbf{w}_t) + \mathcal{O}\left(\alpha^2\lambda^2 F(\tilde{\mathbf{w}}_{s-1})\right).$$



Proof Idea

Assume $\eta = \alpha\lambda$ ($\alpha \ll 1$)

Proof Idea

Assume $\eta = \alpha\lambda$ ($\alpha \ll 1$)

- Using martingale arguments: W.h.p., never reach "flat" region in $m \leq \mathcal{O}\left(\frac{1}{\alpha^2\lambda^2}\right)$ iterations

Proof Idea

Assume $\eta = \alpha\lambda$ ($\alpha \ll 1$)

- Using martingale arguments: W.h.p., never reach "flat" region in $m \leq \mathcal{O}\left(\frac{1}{\alpha^2\lambda^2}\right)$ iterations
- $\Rightarrow$ For all $t \leq m$
  $\mathbb{E}\left[F(\mathbf{w}_{t+1})|\mathbf{w}_t\right] \leq \left(1 - \Theta(\alpha\lambda^2)\right) F(\mathbf{w}_t) + \mathcal{O}\left(\alpha^2\lambda^2 F(\tilde{\mathbf{w}}_{s-1})\right).$

Proof Idea

Assume $\eta = \alpha\lambda$ ($\alpha \ll 1$)

- Using martingale arguments: W.h.p., never reach "flat" region in $m \leq \mathcal{O}\left(\frac{1}{\alpha^2\lambda^2}\right)$ iterations
- $\Rightarrow$ For all $t \leq m$
  $\mathbb{E}\left[F(\mathbf{w}_{t+1})|\mathbf{w}_t\right] \leq \left(1 - \Theta(\alpha\lambda^2)\right) F(\mathbf{w}_t) + \mathcal{O}\left(\alpha^2\lambda^2 F(\tilde{\mathbf{w}}_{s-1})\right).$
- Carefully unwinding recursion, and using $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$,
  $\mathbb{E}\left[F(\mathbf{w}_m)|\mathbf{w}_0\right] \leq \left(\left(1 - \Theta(\alpha\lambda^2)\right)^m + \mathcal{O}(\alpha)\right) F(\tilde{\mathbf{w}}_{s-1})$

Proof Idea

Assume $\eta = \alpha\lambda$ ($\alpha \ll 1$)

- Using martingale arguments: W.h.p., never reach "flat" region in $m \leq \mathcal{O}\left(\frac{1}{\alpha^2\lambda^2}\right)$ iterations
- $\Rightarrow$ For all $t \leq m$
  $\mathbb{E}\left[F(\mathbf{w}_{t+1})|\mathbf{w}_t\right] \leq \left(1 - \Theta(\alpha\lambda^2)\right) F(\mathbf{w}_t) + \mathcal{O}\left(\alpha^2\lambda^2 F(\tilde{\mathbf{w}}_{s-1})\right).$
- Carefully unwinding recursion, and using $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$,
  $\mathbb{E}\left[F(\mathbf{w}_m)|\mathbf{w}_0\right] \leq \left(\left(1 - \Theta(\alpha\lambda^2)\right)^m + \mathcal{O}(\alpha)\right) F(\tilde{\mathbf{w}}_{s-1})$
- $\Rightarrow$ If $m \geq \Omega\left(\frac{1}{\alpha\lambda^2}\right)$, $F(\mathbf{w}_m)$ smaller than $F(\tilde{\mathbf{w}}_{s-1})$ by some constant factor

Proof Idea

Assume $\eta = \alpha\lambda$ ($\alpha \ll 1$)

- Using martingale arguments: W.h.p., never reach "flat" region in $m \leq \mathcal{O}\left(\frac{1}{\alpha^2\lambda^2}\right)$ iterations
- $\Rightarrow$ For all $t \leq m$
  $\mathbb{E}\left[F(\mathbf{w}_{t+1})|\mathbf{w}_t\right] \leq \left(1 - \Theta(\alpha\lambda^2)\right) F(\mathbf{w}_t) + \mathcal{O}\left(\alpha^2\lambda^2 F(\tilde{\mathbf{w}}_{s-1})\right).$
- Carefully unwinding recursion, and using $\mathbf{w}_0 = \tilde{\mathbf{w}}_{s-1}$,
  $\mathbb{E}\left[F(\mathbf{w}_m)|\mathbf{w}_0\right] \leq \left(\left(1 - \Theta(\alpha\lambda^2)\right)^m + \mathcal{O}(\alpha)\right) F(\tilde{\mathbf{w}}_{s-1})$
- $\Rightarrow$ If $m \geq \Omega\left(\frac{1}{\alpha\lambda^2}\right)$, $F(\mathbf{w}_m)$ smaller than $F(\tilde{\mathbf{w}}_{s-1})$ by some constant factor
- Overall, if $\frac{1}{\alpha\lambda^2} \ll m \ll \frac{1}{\alpha^2\lambda^2}$, every epoch shrinks F by constant factor

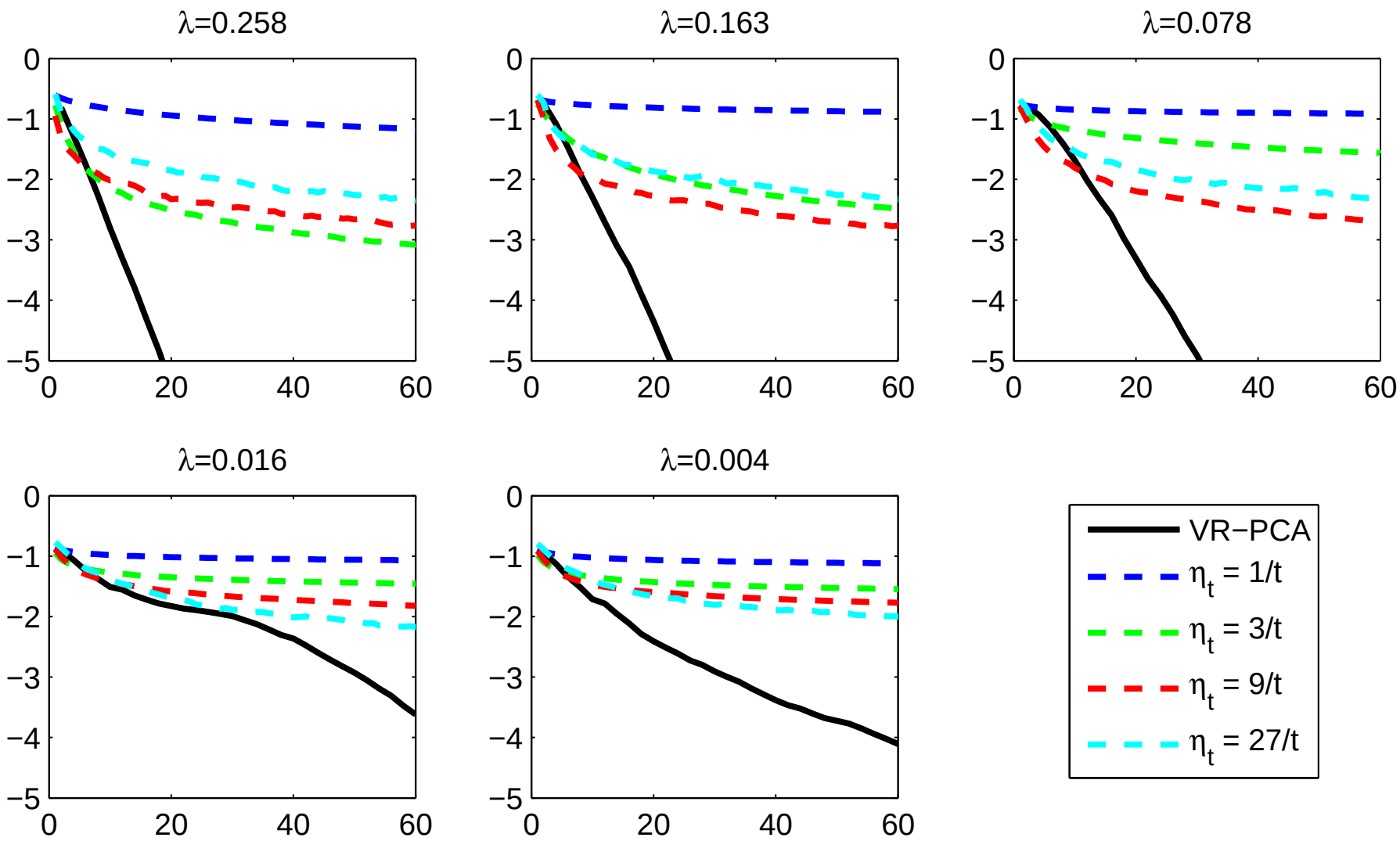
Experiments

Ran VR-PCA with

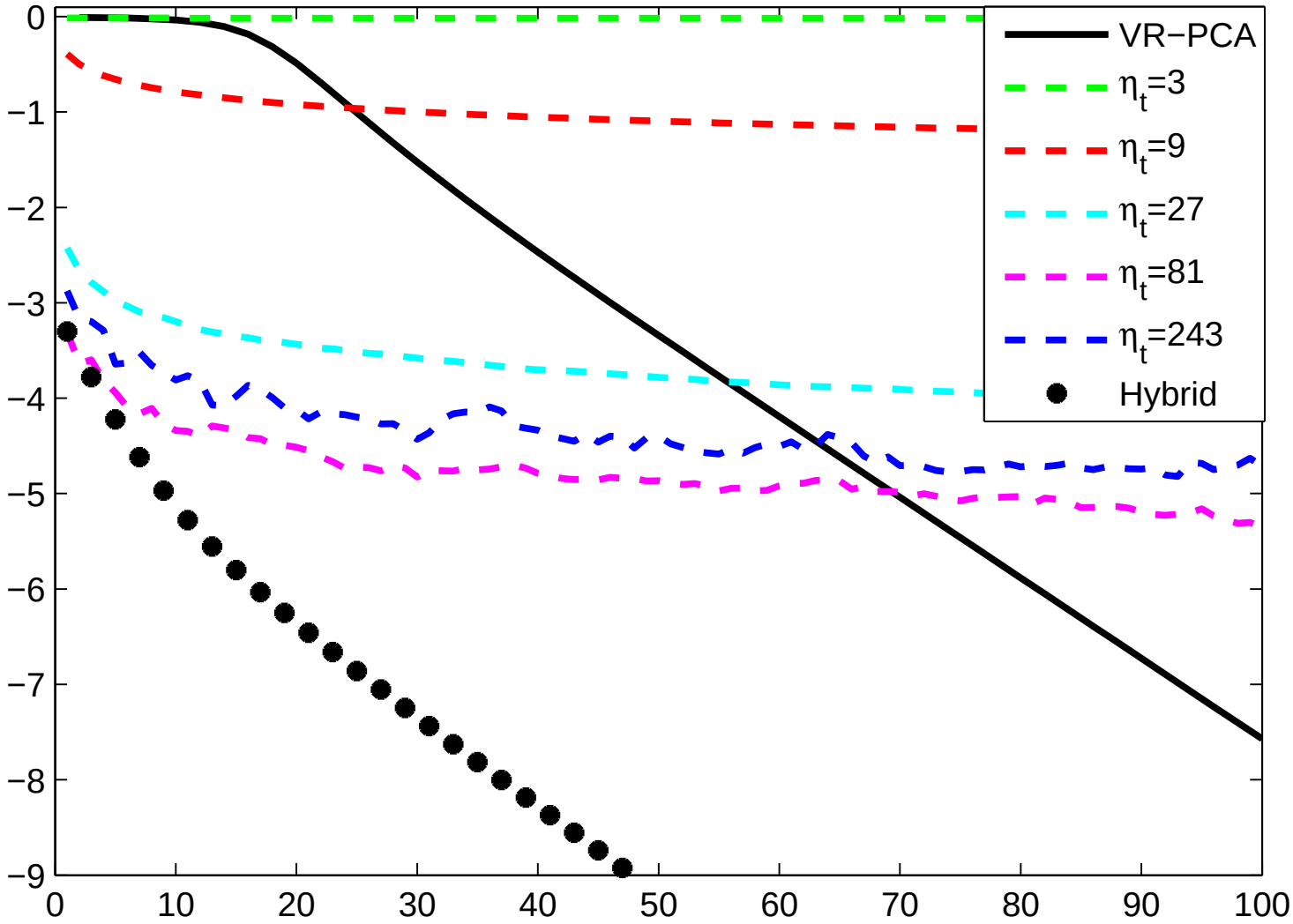
- Random initialization
- $m = n$ (epoch $\approx$ one pass over data)
- $\eta = 0.05/\sqrt{n}$ (based on theory)

Compared to Oja's algorithm with hand-tuned step-size

Experiments



Preliminary Experiments



Work-in-progress and Open Questions

Work-in-progress and Open Questions

- Runtime is $\mathcal{O}(d(n+\frac{1}{\lambda^2})\log\left(\frac{1}{\epsilon}\right))$ can we get $\mathcal{O}(d(n+\frac{1}{\lambda})\log\left(\frac{1}{\epsilon}\right)$ or better, analogous to convex case?
  - Experimentally, required parameters do seem somewhat different
- Generalizing analysis to PCA with several directions
- Theoretically optimal parameters without knowing λ
- Other "fast-stochastic" approaches
- Other non-convex problems

More details: arXiv 1409:2848

THANKS!