Trends in nonconvex optimization

SUVRIT SRA
Laboratory for Information & Decision Systems (LIDS)
Massachusetts Institute of Technology

Oct 2017, Simons Institute, Berkeley

ml.mit.edu

Ack: Sashank Reddi (Google), Francis Bach (Inria)

Nonconvex problems are …

Nonconvex optimization problem with simple constraints

$$\begin{aligned} \min \quad & \Big(\sum\nolimits_i a_i z_i - s\Big)^2 + \sum\nolimits_i z_i(1 - z_i) \\ \text{s.t.} \quad & 0 \le z_i \le 1, \quad i = 1, \ldots, n. \end{aligned}$$

Question: Is global min of this problem 0 or not?

Nonconvex problems are …

Nonconvex optimization problem with simple constraints

$$\min \quad \left(\sum_i a_i z_i - s\right)^2 + \sum_i z_i(1 - z_i)$$
$$\text{s.t.} \quad 0 \leq z_i \leq 1, \quad i = 1, \ldots, n.$$

Question: Is global min of this problem 0 or not?

Does there exist a subset of $\{a_1, a_2, \ldots, a_n\}$ that sums to s?

Subset-sum problem, well-known NP-Complete prob.

Nonconvex problems are …

Nonconvex optimization problem with simple constraints

$$\min \quad \left(\sum_i a_i z_i - s\right)^2 + \sum_i z_i(1 - z_i)$$
$$\text{s.t.} \quad 0 \leq z_i \leq 1, \quad i = 1, \ldots, n.$$

Question: Is global min of this problem 0 or not?

Does there exist a subset of $\{a_1, a_2, \ldots, a_n\}$ that sums to s?

Subset-sum problem, well-known NP-Complete prob.

$$\min \; x^\top A x, \quad x \geq 0$$

Question: Is x=0 a local minimum or not?

Introduction

What is this talk about?

Some topics in nonconvex ***optimization*** with a bias towards "large-scale" and stuff I know 😀

Introduction

What is this talk about?

Some topics in nonconvex ***optimization*** with a bias towards “large-scale” and stuff I know 😀

What it is not about?

Not encyclopedic coverage of all the trends

Not much about “batch” methods

Not about generalization 😏

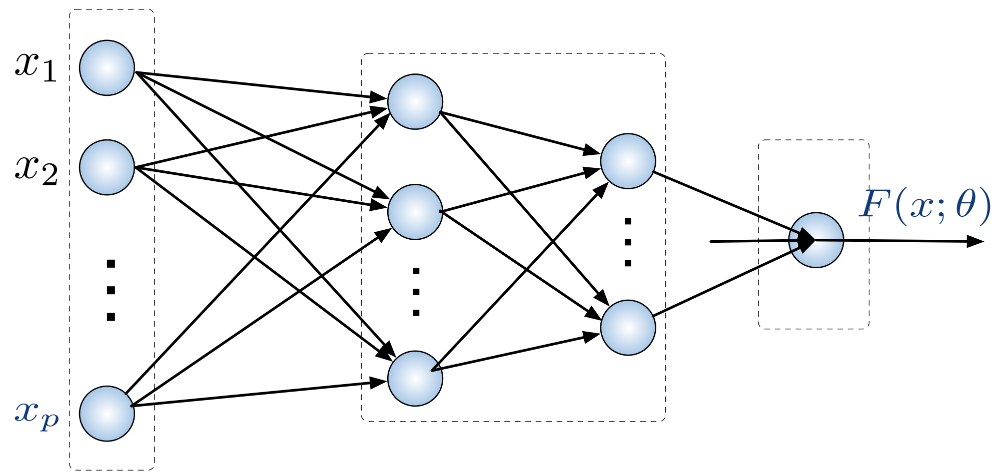
(If I am missing something, please let me know)

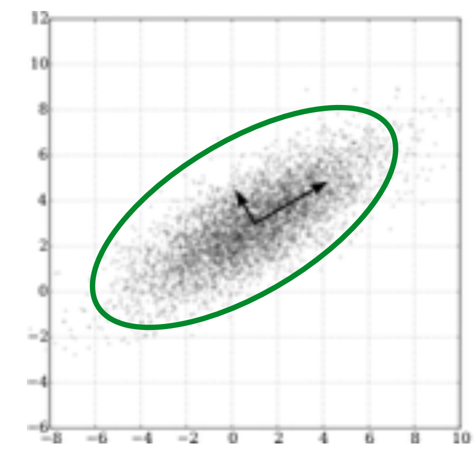
Nonconvex finite-sum problems

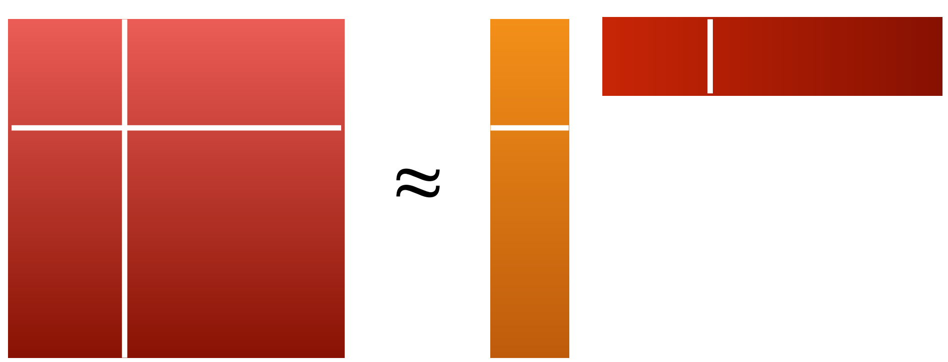
$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, \mathcal{DNN}(x_i, \theta)\big) \quad + \quad \Omega(\theta)$$

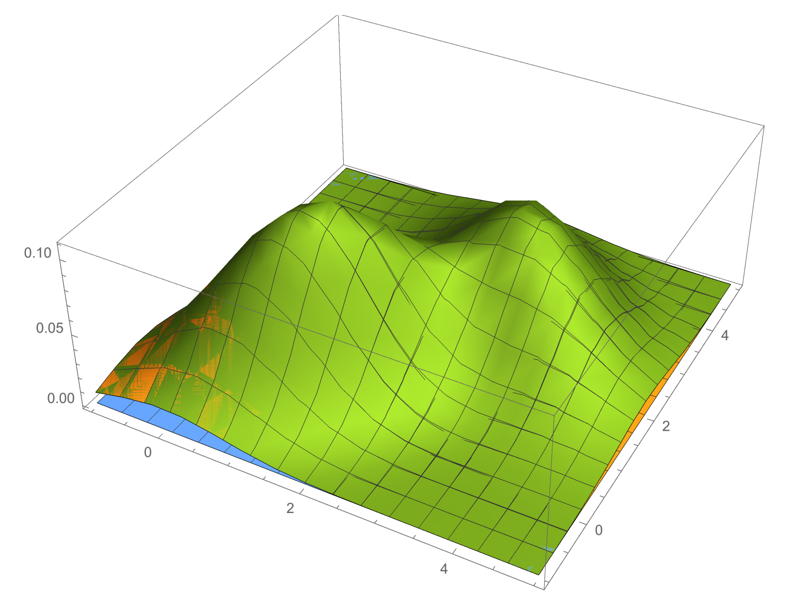
Nonconvex finite-sum problems

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mathcal{DNN}(x_i, \theta)) \quad + \quad \Omega(\theta)$$









Nonconvex ERM / finite-sums

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

Nonconvex ERM / finite-sums

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

Related work

- Original SGD paper *(Robbins, Monro 1951)*
  (asymptotic convergence; no rates)
- SGD with scaled gradients ($\theta_t - \eta_t H_t \nabla f(\theta_t)$) + other tricks: space dilation, *(Shor, 1972)*;Variable metric SGD *(Uryasev 1988)*; AdaGrad *(Duchi, Hazan, Singer, 2012)*; Adam *(Kingma, Ba, 2015)*, and many others…
  (typically asymptotic convergence for nonconvex)
- Large number of other ideas, often for step-size tuning, initialization
  (see e.g., blog post: by S. Ruder on gradient descent algorithms)

Nonconvex ERM / finite-sums

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

Related work

- Original SGD paper *(Robbins, Monro 1951)*
  (asymptotic convergence; no rates)
- SGD with scaled gradients ($\theta_t - \eta_t H_t \nabla f(\theta_t)$) + other tricks: space dilation, *(Shor, 1972)*; Variable metric SGD *(Uryasev 1988)*; AdaGrad *(Duchi, Hazan, Singer, 2012)*; Adam *(Kingma, Ba, 2015)*, and many others…
  (typically asymptotic convergence for nonconvex)
- Large number of other ideas, often for step-size tuning, initialization
  (see e.g., blog post: by S. Ruder on gradient descent algorithms)

Trends: going beyond SGD (theoretically; ultimately in practice too)

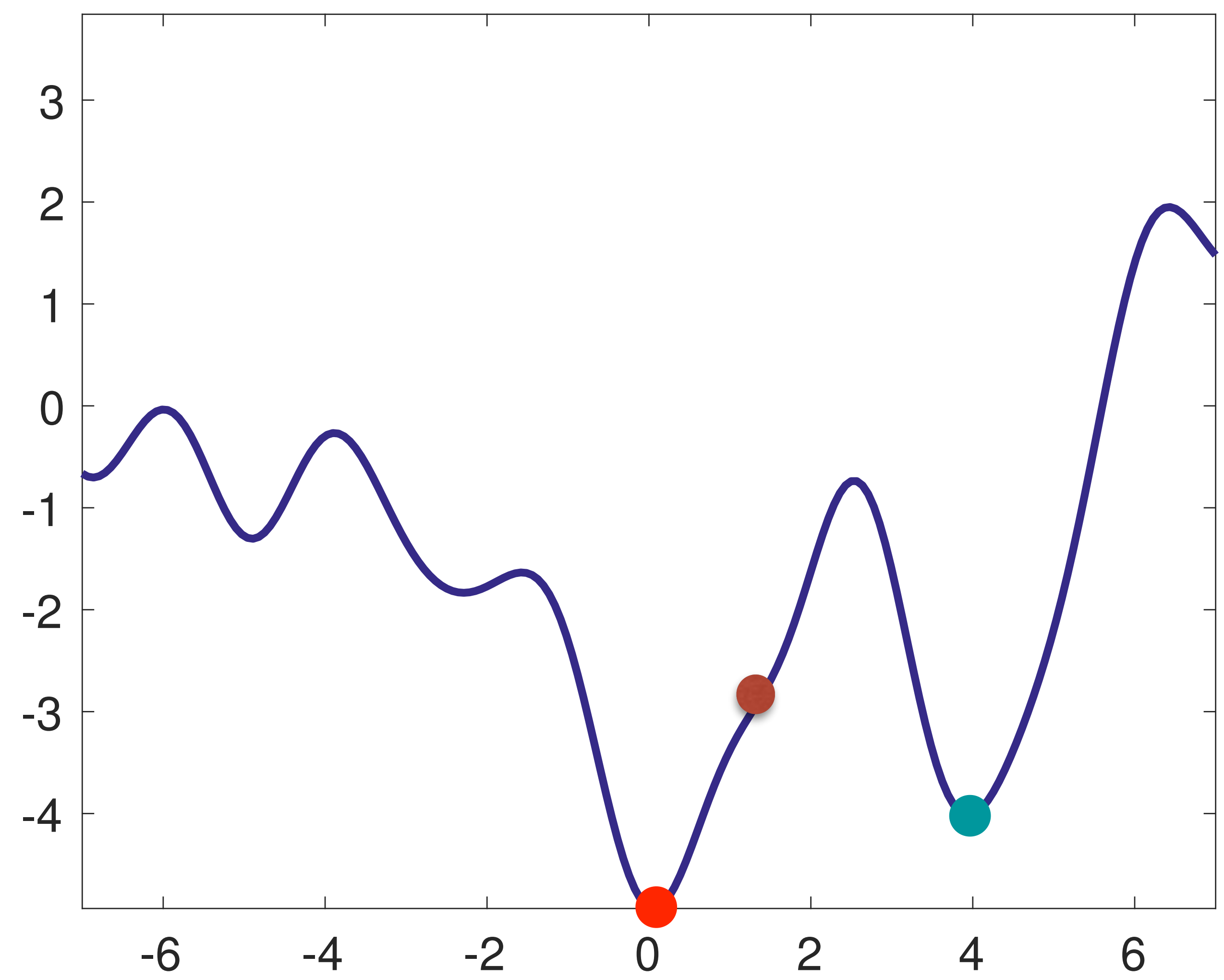
Nonconvex ERM / finite-sums

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

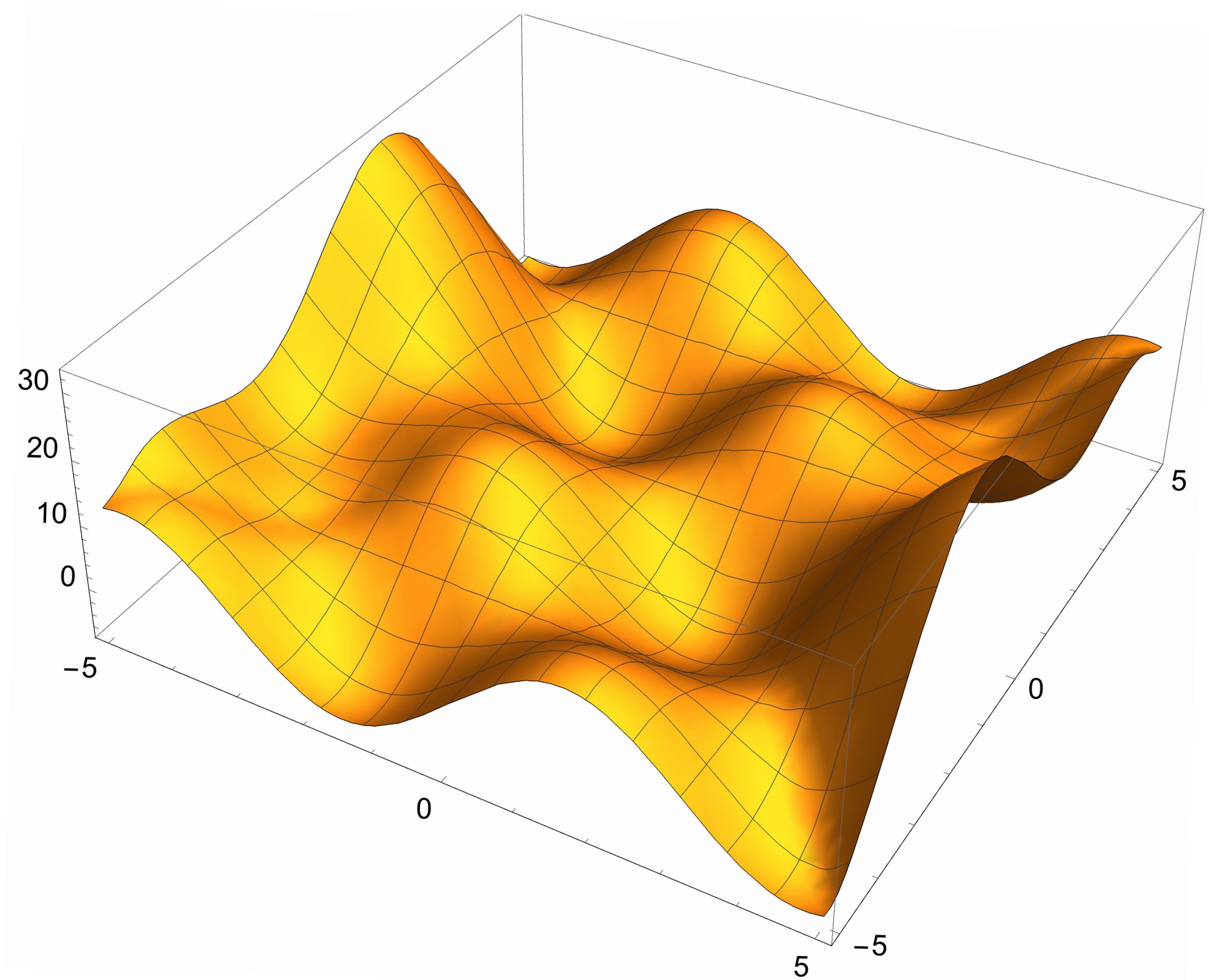
Related work (subset)

- *(Solodov, 1997)* Incremental gradient, smooth nonconvex (asymptotic convergence; no rates proved)
- *(Bertsekas, Tsitsiklis, 2000)* Gradient descent with errors; incremental (see §2.4, *Nonlinear Programming;* no rates proved)
- *(Sra, 2011)* Incremental nonconvex non-smooth (asymptotic convergence only)
- *(Ghadimi, Lan, 2013)* SGD for nonconvex stochastic opt. (first non-asymptotic rates to stationarity)
- *(Ghadimi et al., 2013)* SGD for nonconvex non-smooth stoch. opt. (non-asymptotic rates, but key limitations)

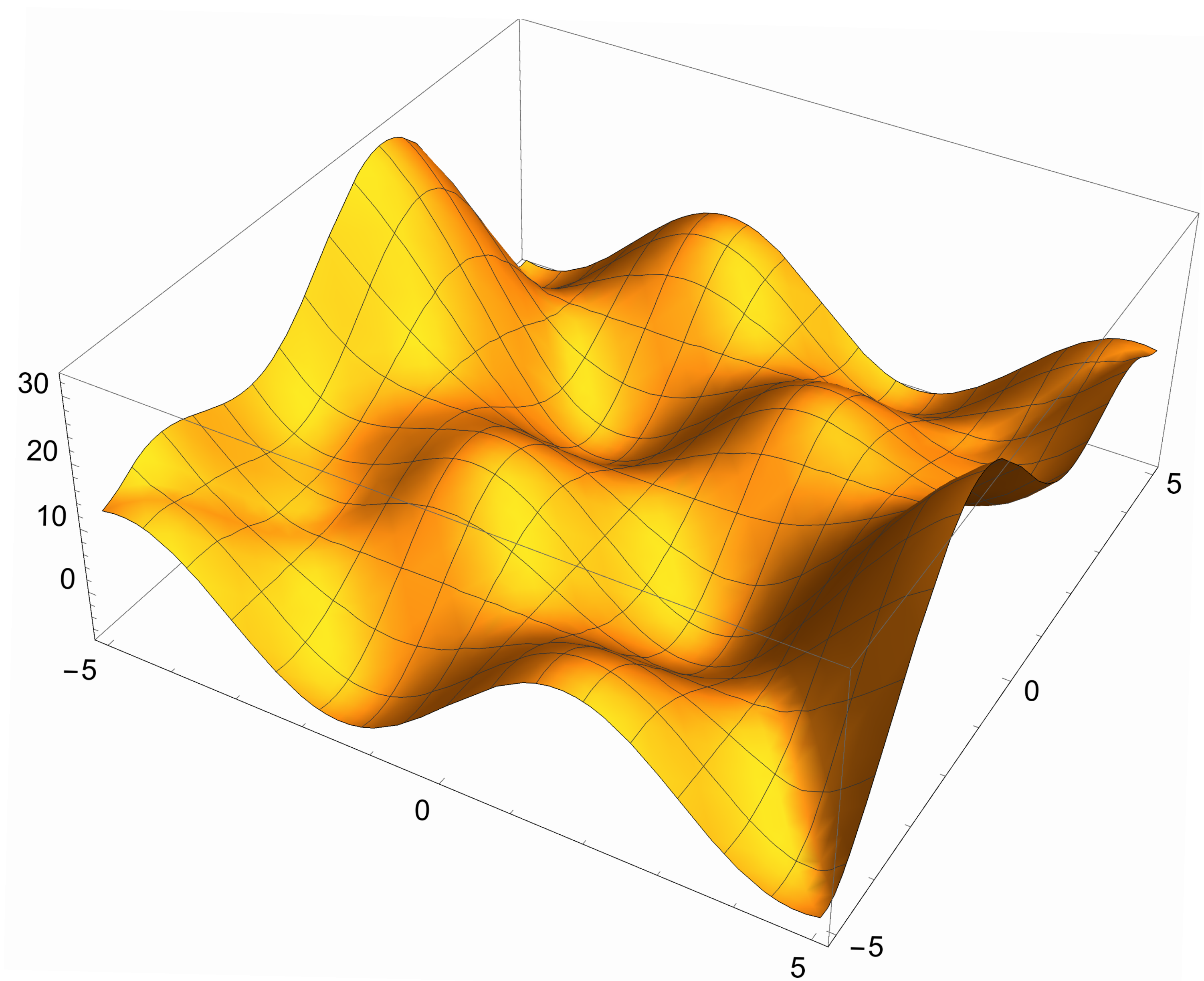
Difficulty of nonconvex optimization



Difficulty of nonconvex optimization



Difficulty of nonconvex optimization

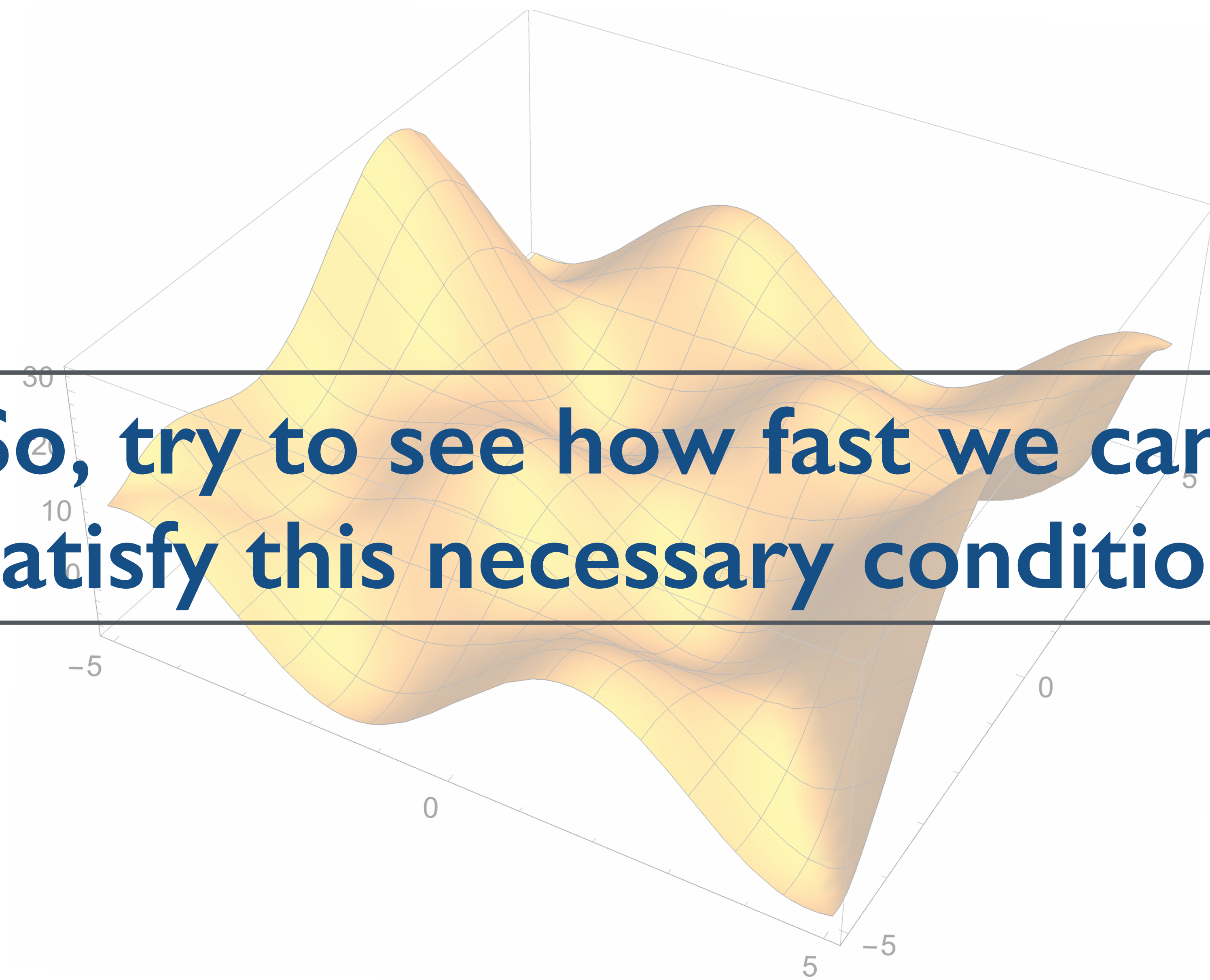


Difficult to optimize, but

$$\nabla g(\theta) = 0$$

necessary condition – local minima, maxima, saddle points satisfy it.

Difficulty of nonconvex optimization



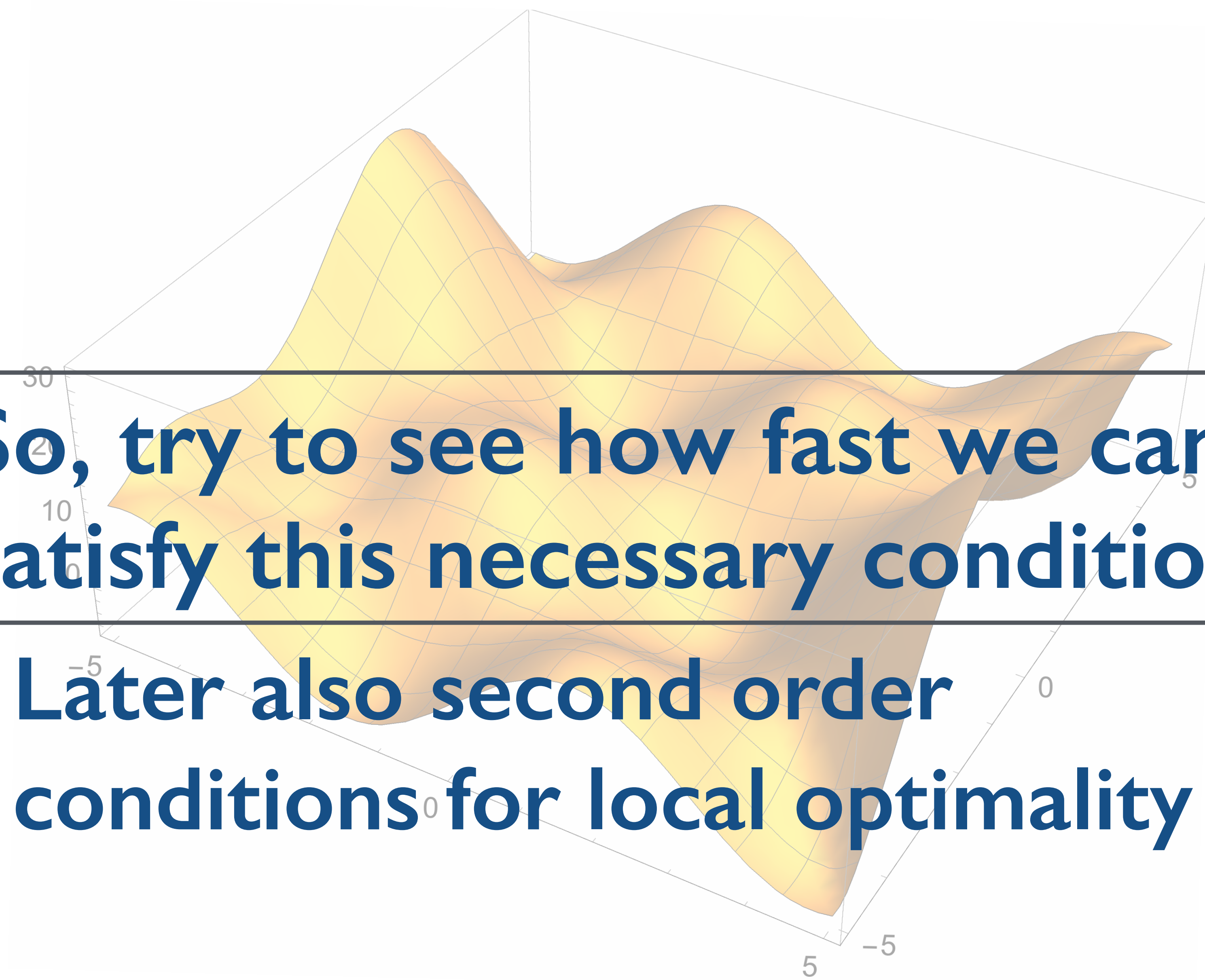
So, try to see how fast we can satisfy this necessary condition

Difficult to optimize, but

$$\nabla g(\theta) = 0$$

necessary condition – local minima, maxima, saddle points satisfy it.

Difficulty of nonconvex optimization



So, try to see how fast we can satisfy this necessary condition

Later also second order conditions for local optimality

Difficult to optimize, but

$$\nabla g(\theta) = 0$$

necessary condition – local minima, maxima, saddle points satisfy it.

Measuring efficiency of nonconvex opt.

Convex: $\mathbb{E}[g(\theta_t) - g^*] \leq \epsilon$ *(optimality gap)*

Nonconvex: $\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$ *(stationarity gap)*

(*Nesterov 2003, Chap 1);*
(Ghadimi, Lan, 2012)

Measuring efficiency of nonconvex opt.

Convex: $\mathbb{E}[g(\theta_t) - g^*] \le \epsilon$ *(optimality gap)*

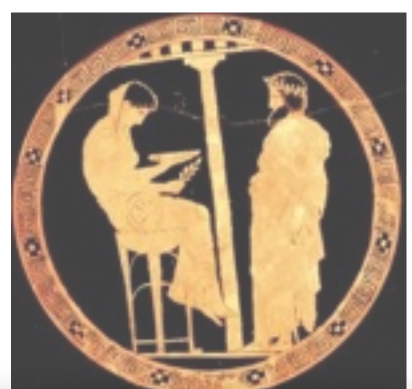
Nonconvex: $\mathbb{E}[\|\nabla g(\theta_t)\|^2] \le \epsilon$ *(stationarity gap)*

(Nesterov 2003, Chap 1); (Ghadimi, Lan, 2012)

Incremental First-order Oracle (IFO)

(Agarwal, Bottou, 2014) (see also: Nemirovski, Yudin, 1983)

(x, i) →



→ $(f_i(x), \nabla f_i(x))$

Measure: #IFO calls to attain ϵ accuracy

IFO Example: SGD vs GD (nonconvex)

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ GD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t)$$

$$\theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

IFO Example: SGD vs GD (nonconvex)

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ GD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t)$$

$$\theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

---

assuming Lipschitz smooth gradients

$$\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$$

IFO Example: SGD vs GD (nonconvex)

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ GD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t) \qquad\qquad \theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

- O(1) IFO calls per iter
- O($1/\epsilon^2$) iterations
- **Total:** O($1/\epsilon^2$) IFO calls
- independent of n

(Ghadimi, Lan, 2013,2014)

---

assuming Lipschitz smooth gradients

$$\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$$

IFO Example: SGD vs GD (nonconvex)

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ GD

SGD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t)$$

- O(1) IFO calls per iter
- O($1/\epsilon^2$) iterations
- **Total:** O($1/\epsilon^2$) IFO calls
- independent of n

(Ghadimi, Lan, 2013,2014)

GD

$$\theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

- O(n) IFO calls per liter
- O($1/\epsilon$) iterations
- **Total:** O(n/ϵ) IFO calls
- depends strongly on n

(Nesterov, 2003; Nesterov 2012)

---

assuming Lipschitz smooth gradients

$$\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$$

IFO Example: SGD vs GD (nonconvex)

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ **GD**

?

SGD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t)$$

- O(1) IFO calls per iter
- O($1/\epsilon^2$) iterations
- **Total:** O($1/\epsilon^2$) IFO calls
- independent of n

(Ghadimi, Lan, 2013,2014)

GD

$$\theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

- O(n) IFO calls per liter
- O($1/\epsilon$) iterations
- **Total:** O(n/ϵ) IFO calls
- depends strongly on n

(Nesterov, 2003; Nesterov 2012)

---

assuming Lipschitz smooth gradients

$$\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$$

Nonconvex finite-sums

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ GD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t)$$

$$\theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

SAG, SVRG, SAGA, et al.

Analysis depends heavily on convexity
(especially for controlling variance)

[Roux, Schmidt, Bach, 2012; Johnson, Zhang 2013; Defazio, Bach, Lacoste-Julien, 2014]
[Gurbuzbalaban, Ozdaglar, Parrilo, 2015 - deterministic]

Nonconvex finite-sums

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

SGD ⟷ GD

$$\theta_{t+1} = \theta_t - \eta \nabla f_{i_t}(\theta_t) \qquad\qquad \theta_{t+1} = x_t - \eta \nabla g(\theta_t)$$

Do these benefits extend to nonconvex finite-sums?

[Roux, Schmidt, Bach, 2012; Johnson, Zhang 2013; Defazio, Bach, Lacoste-Julien, 2014]
[Gurbuzbalaban, Ozdaglar, Parrilo, 2015 - deterministic]

SVRG/SAGA work
(new analysis due to nonconvexity)

Nonconvex SVRG

for s=0 to S-1

$\theta_0^{s+1} \leftarrow \theta_m^s$

$\tilde{\theta}^s \leftarrow \theta_m^s$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s) \Big]$$

end

end

The same algorithm as usual SVRG *(Johnson, Zhang, 2013)*

Nonconvex SVRG

for s=0 to S-1

$\theta_0^{s+1} \leftarrow \theta_m^s$

$\tilde{\theta}^s \leftarrow \theta_m^s$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \dots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s)\Big]$$

end

end

Nonconvex SVRG

for s=0 to S-1

$$\theta_0^{s+1} \leftarrow \theta_m^s$$

$$\tilde{\theta}^s \leftarrow \theta_m^s$$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s)\Big]$$

end

end

Nonconvex SVRG

for s=0 to S-1

$$\theta_0^{s+1} \leftarrow \theta_m^s$$

$$\tilde{\theta}^s \leftarrow \theta_m^s$$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s)\Big]$$

end

end

Nonconvex SVRG

for s=0 to S-1

$$\theta_0^{s+1} \leftarrow \theta_m^s$$

$$\tilde{\theta}^s \leftarrow \theta_m^s$$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{\theta}^s)\Big]$$

end

end

Nonconvex SVRG

for s=0 to S-1

$$\theta_0^{s+1} \leftarrow \theta_m^s$$

$$\tilde{\theta}^s \leftarrow \theta_m^s$$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \underbrace{\nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s)}_{\Delta_t}\Big]$$

end

end

$$\mathbb{E}[\Delta_t] = 0$$

Nonconvex SVRG

for s=0 to S-1

$\theta_0^{s+1} \leftarrow \theta_m^s$

$\tilde{\theta}^s \leftarrow \theta_m^s$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \left[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s)} \right]$$

end

end

Full gradient, computed once every epoch

Nonconvex SVRG

for s=0 to S-1

$\theta_0^{s+1} \leftarrow \theta_m^s$

$\tilde{\theta}^s \leftarrow \theta_m^s$

for t=0 to m-1

Uniformly randomly pick $i(t) \in \{1, \ldots, n\}$

$$\theta_{t+1}^{s+1} = \theta_t^{s+1} - \eta_t \Big[\nabla f_{i(t)}(\theta_t^{s+1}) - \nabla f_{i(t)}(\tilde{\theta}^s) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{\theta}^s)\Big]$$

end

end

Key quantities that determine how the method operates

Full gradient, computed once every epoch

Key ideas for analysis of nc-SVRG

Previous SVRG proofs rely on convexity to control variance

New proof technique – quite general; extends to SAGA, to several other finite-sum nonconvex settings.

[Reddi, Hefny, Sra, Poczos, Smola, 2016]; indp. also [Allen-Zhu, Hazan, 2016]

Key ideas for analysis of nc-SVRG

Previous SVRG proofs rely on convexity to control variance

New proof technique – quite general; extends to SAGA, to several other finite-sum nonconvex settings.

Larger step-size ➥ smaller inner loop
(full-gradient computation dominates epoch)

Smaller step-size ➥ slower convergence
(longer inner loop)

[Reddi, Hefny, Sra, Poczos, Smola, 2016]; indp. also [Allen-Zhu, Hazan, 2016]

Key ideas for analysis of nc-SVRG

Previous SVRG proofs rely on convexity to control variance

New proof technique – quite general; extends to SAGA, to several other finite-sum nonconvex settings.

Larger step-size ➥ smaller inner loop
(full-gradient computation dominates epoch)

Smaller step-size ➥ slower convergence
(longer inner loop)

(Carefully) trading-off #inner-loop iterations m with step-size η leads to lower #IFO calls!

[Reddi, Hefny, Sra, Poczos, Smola, 2016]; indp. also [Allen-Zhu, Hazan, 2016]

Faster nonconvex optimization via VR

Algorithm	Nonconvex (Lipschitz smooth)
SGD	$O\left(\frac{1}{\epsilon^2}\right)$
GD	$O\left(\frac{n}{\epsilon}\right)$
SVRG	$O\left(n + \frac{n^{2/3}}{\epsilon}\right)$
SAGA	$O\left(n + \frac{n^{2/3}}{\epsilon}\right)$
MSVRG	$O\left(\min\left(\frac{1}{\epsilon^2}, \frac{n^{2/3}}{\epsilon}\right)\right)$

$$\mathbb{E}[\|\nabla g(\theta_t)\|^2] \leq \epsilon$$

Remarks

New results for convex case too; additional nonconvex results

[Reddi, Hefny, Sra, Poczos, Smola, ICML 2016]; [Reddi et al. CDC 2016]

Linear rates for nonconvex problems

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

The **Polyak-Łojasiewicz (PL)** class of functions

$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu} \|\nabla g(\theta)\|^2$$

(Polyak, 1963); (Łojasiewicz, 1963)

(More general than many other "restricted" strong convexity uses)

Linear rates for nonconvex problems

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

The **Polyak-Łojasiewicz (PL)** class of functions

$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu}\|\nabla g(\theta)\|^2$$

(Polyak, 1963); (Łojasiewicz, 1963)

Examples:	μ-strongly convex $\Rightarrow$ PL holds Stochastic PCA**, some large-scale eigenvector problems

(More general than many other "restricted" strong convexity uses)

Linear rates for nonconvex problems

$$\min_{\theta \in \mathbb{R}^d} \quad g(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

The **Polyak-Łojasiewicz (PL)** class of functions

$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu} \|\nabla g(\theta)\|^2$$

(Polyak, 1963); (Łojasiewicz, 1963)

Examples:	μ-strongly convex $\Rightarrow$ PL holds Stochastic PCA**, some large-scale eigenvector problems

(More general than many other "restricted" strong convexity uses)

(Karimi, Nutini, Schmidt, 2016) proximal extensions; references

(Attouch, Bolte, 2009) more general Kurdya-Łojasiewicz class

(Bertsekas, 2016) textbook, more "growth conditions"

Linear rates for nonconvex problems

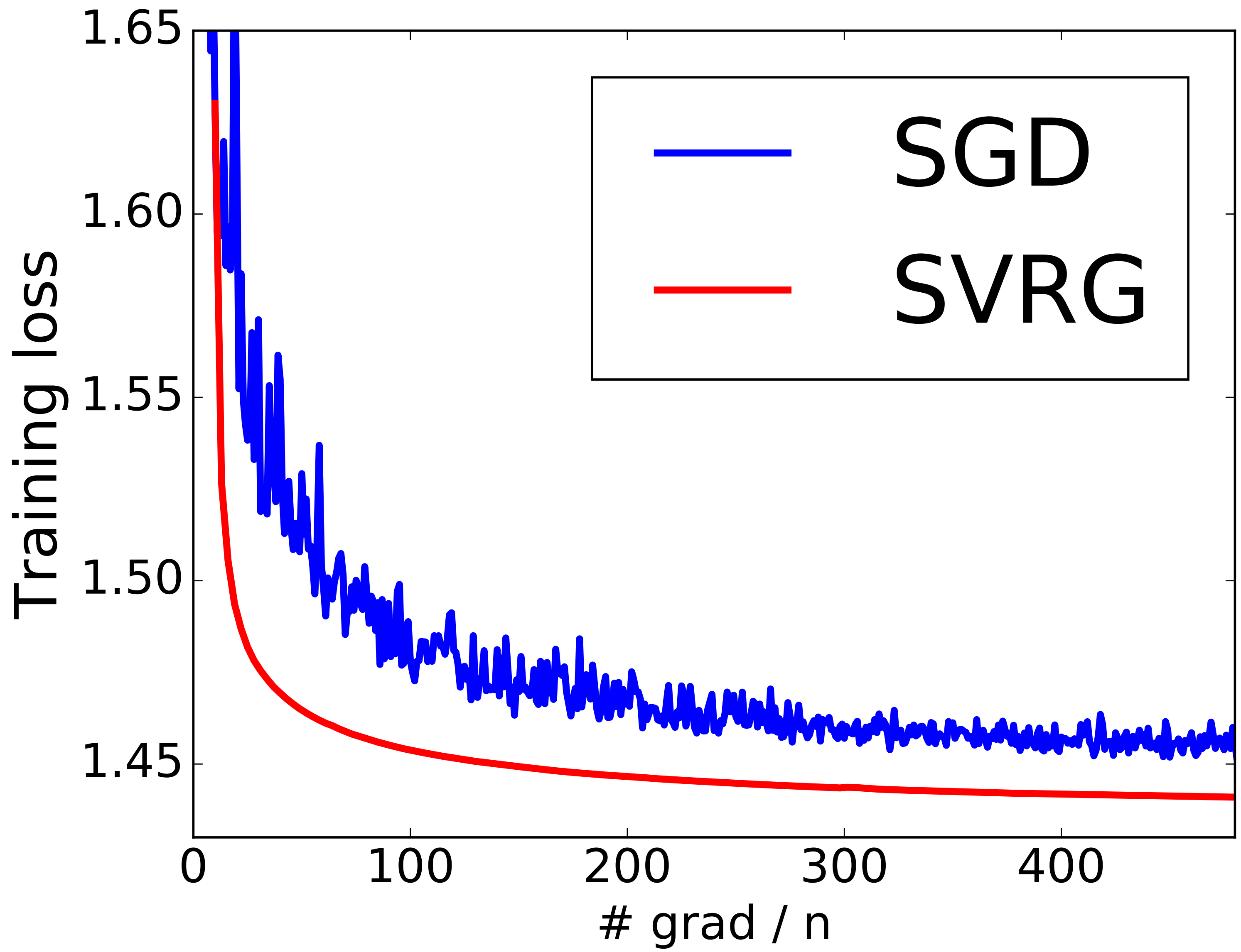
$$g(\theta) - g(\theta^*) \leq \frac{1}{2\mu}\|\nabla g(\theta)\|^2 \quad \Big| \quad \mathbb{E}[g(\theta_t) - g^*] \leq \epsilon$$

Algorithm	Nonconvex	Nonconvex-PL
SGD	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon^2}\right)$
GD	$O\left(\frac{n}{\epsilon}\right)$	$O\left(\frac{n}{2\mu}\log\frac{1}{\epsilon}\right)$
SVRG	$O\left(n + \frac{n^{2/3}}{\epsilon}\right)$	$O\left(\left(n + \frac{n^{2/3}}{2\mu}\right)\log\frac{1}{\epsilon}\right)$
SAGA	$O\left(n + \frac{n^{2/3}}{\epsilon}\right)$	$O\left(\left(n + \frac{n^{2/3}}{2\mu}\right)\log\frac{1}{\epsilon}\right)$
MSVRG	$O\left(\min\left(\frac{1}{\epsilon^2}\right), \frac{n^{2/3}}{\epsilon}\right)$	—

Variant of nc-SVRG attains this fast convergence!

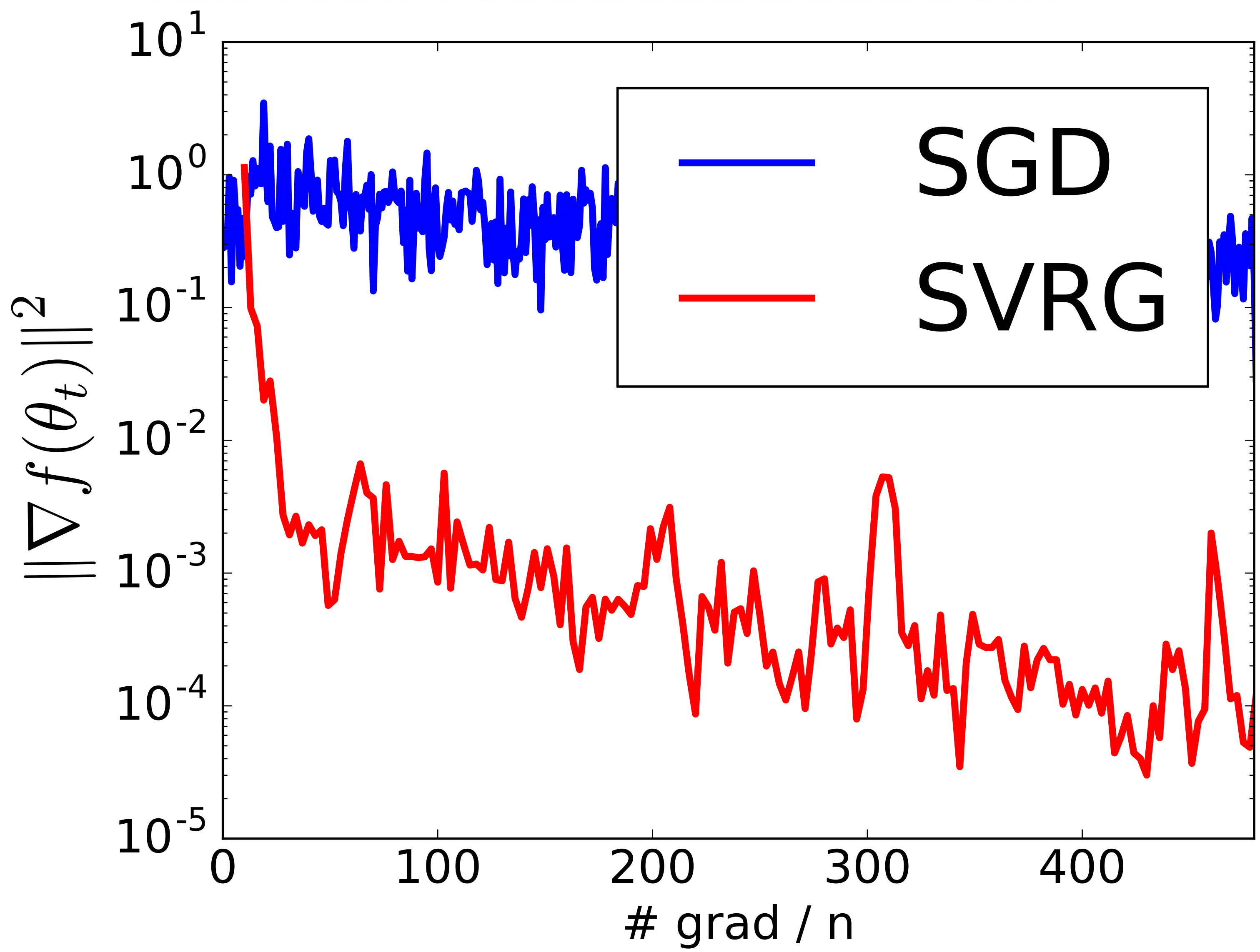
(Reddi, Hefny, Sra, Poczos, Smola, 2016; Reddi et al., 2016)

Empirical results



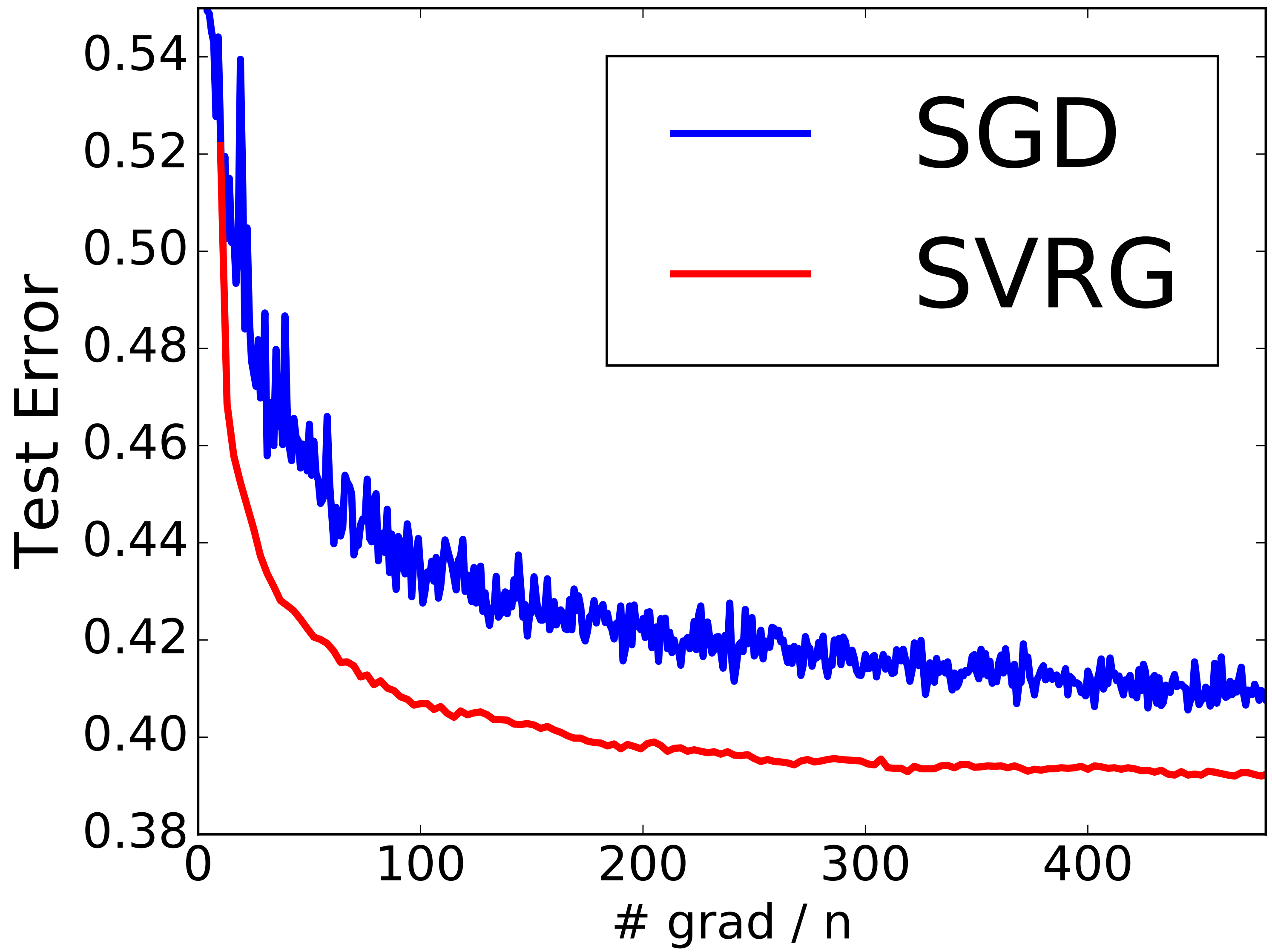
CIFAR10 dataset; 2-layer NN

Empirical results



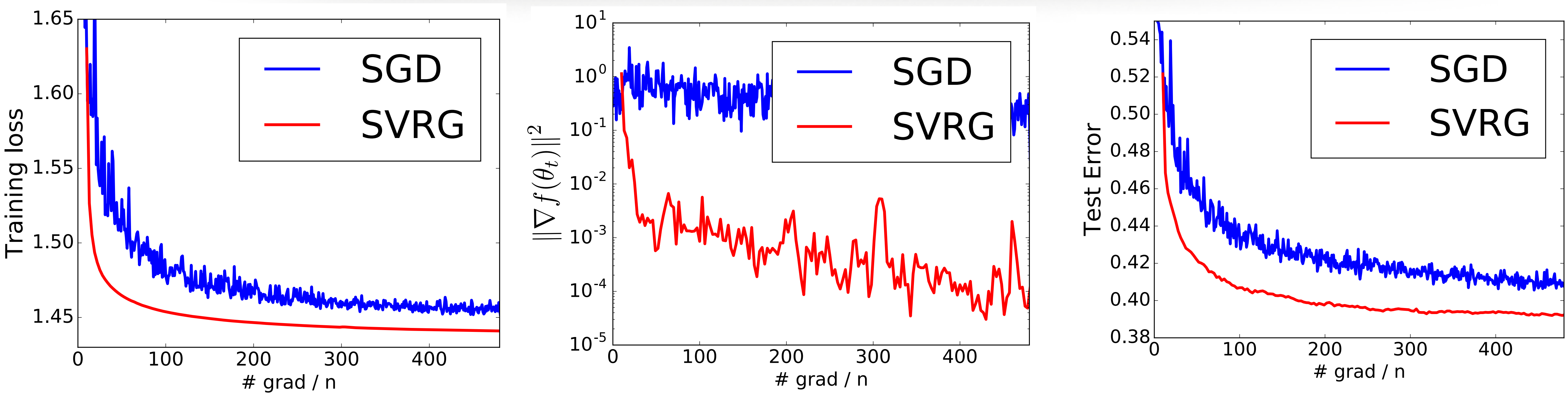
CIFAR10 dataset; 2-layer NN

Empirical results



CIFAR10 dataset; 2-layer NN

Empirical results



CIFAR10 dataset; 2-layer NN

What about deep networks?

Non-smooth surprises!

$$\min_{\theta \in \mathbb{R}^d} \quad \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) + \Omega(\theta)$$

Regularizer, e.g., $\|\cdot\|_1$ for enforcing sparsity of weights (in a neural net, or more generally); or an indicator function of a constraint set, etc.

Nonconvex composite objective problems

$$\min_{\theta \in \mathbb{R}^d} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\text{nonconvex}} + \underbrace{\Omega(\theta)}_{\text{convex}}$$

Nonconvex composite objective problems

$$\min_{\theta\in\mathbb{R}^d} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\textbf{nonconvex}} + \underbrace{\Omega(\theta)}_{\textbf{convex}}$$

Prox-SGD

$$\theta_{t+1} = \mathrm{prox}_{\lambda_t\Omega}\left(\theta_t - \eta_t \nabla f_{i_t}(\theta_t)\right)$$

$$\mathrm{prox}_{\lambda\Omega}(v) := \mathrm{argmin}_u \ \frac{1}{2}\|u-v\|^2 + \lambda\Omega(u)$$

prox: soft-thresholding for $\|\cdot\|_1$; projection for indicator function

Nonconvex composite objective problems

$$\min_{\theta \in \mathbb{R}^d} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\textbf{nonconvex}} + \underbrace{\Omega(\theta)}_{\textbf{convex}}$$

Prox-SGD $\quad \theta_{t+1} = \mathrm{prox}_{\lambda_t \Omega}\left(\theta_t - \eta_t \nabla f_{i_t}(\theta_t)\right)$

Prox-SGD convergence not known!*

$\mathrm{prox}_{\lambda\Omega}(v) := \mathrm{argmin}_u \ \frac{1}{2}\|u - v\|^2 + \lambda\Omega(u)$

prox: soft-thresholding for $\|\cdot\|_1$; projection for indicator function

* Except in special cases (where even rates may be available)

Nonconvex composite objective problems

$$\min_{\theta \in \mathbb{R}^d} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\text{nonconvex}} + \underbrace{\Omega(\theta)}_{\text{convex}}$$

Prox-SGD $\quad \theta_{t+1} = \operatorname{prox}_{\lambda_t \Omega}\left(\theta_t - \eta_t \nabla f_{i_t}(\theta_t)\right)$

Prox-SGD convergence not known!*

prox: soft-thresholding for $\|\cdot\|_1$; projection for indicator function

– Partial results: *(Ghadimi, Lan, Zhang, 2014)*
(using growing minibatches, shrinking step sizes)
– Double loop; projection+subgrad *(Davis, Grimmer, 2017)*

* Except in special cases (where even rates may be available)

Nonconvex composite objective problems

$$\min_{\theta \in \mathbb{R}^d} \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\text{nonconvex}} + \underbrace{\Omega(\theta)}_{\text{convex}}$$

Once again variance reduction to the rescue?

Nonconvex composite objective problems

$$\min_{\theta \in \mathbb{R}^d} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\textbf{nonconvex}} + \underbrace{\Omega(\theta)}_{\textbf{convex}}$$

Once again variance reduction to the rescue?

Prox-SVRG/SAGA converge*
and that too
faster than both SGD and GD!

Nonconvex composite objective problems

$$\min_{\theta \in \mathbb{R}^d} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\theta)}_{\text{nonconvex}} + \underbrace{\Omega(\theta)}_{\text{convex}}$$

Once again variance reduction to the rescue?

Prox-SVRG/SAGA converge*

and that too

faster than both SGD and GD!

The same $O\left(n + \frac{n^{2/3}}{\epsilon}\right)$ once again!

* some care needed

(Reddi, Sra, Poczos, Smola, 2016)

Empirical results: NN-PCA

$$\min_{\|w\|\le 1,\ w\ge 0} -\frac{1}{2}w^\top\left(\sum_{i=1}^n x_i x_i^\top\right)w$$

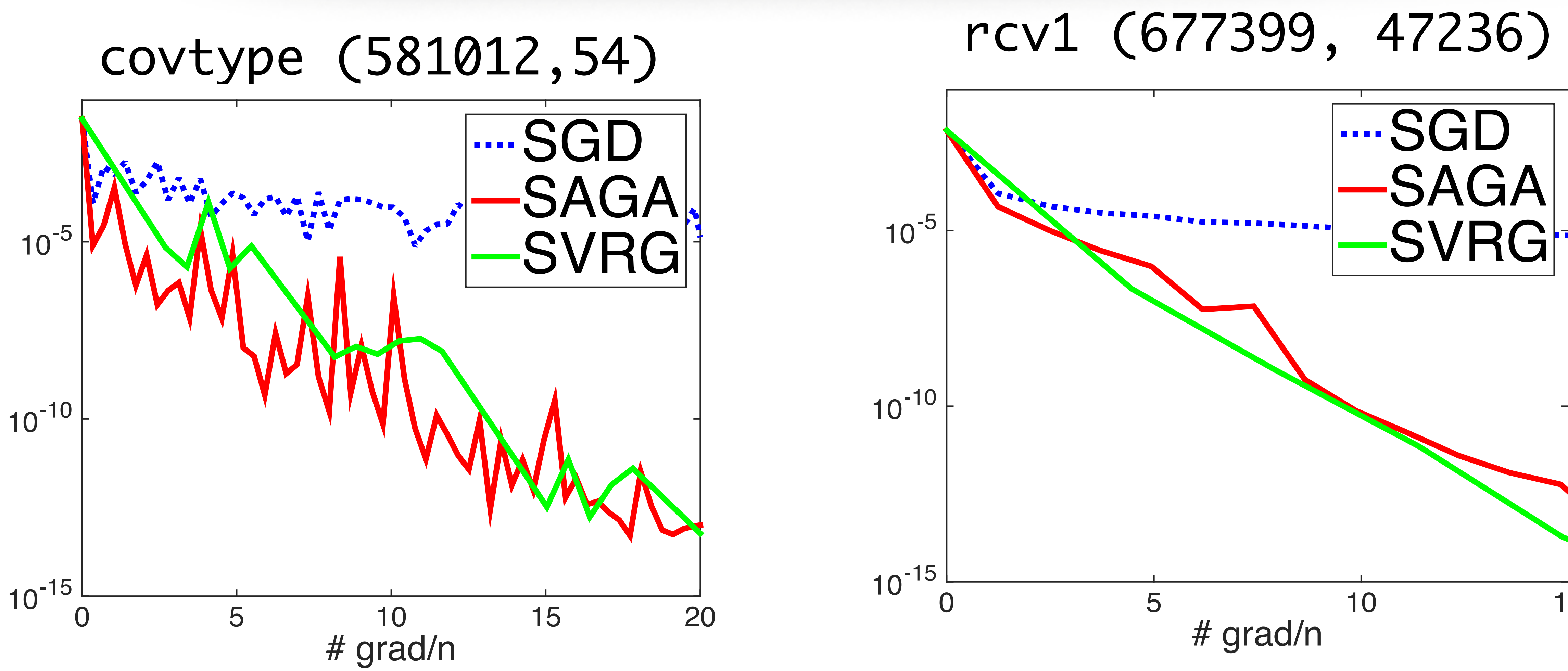
Eigenvecs via SGD: *(Oja, Karhunen 1985)*; via SVRG *(Shamir, 2015,2016)*; *(Garber, Hazan, Jin, Kakade, Musco, Netrapalli, Sidford, 2016);* and many more!

Empirical results: NN-PCA

$$\min_{\|w\|\leq 1,\ w\geq 0} -\frac{1}{2} w^\top \left(\sum_{i=1}^{n} x_i x_i^\top \right) w$$

Eigenvecs via SGD: *(Oja, Karhunen 1985)*; via SVRG *(Shamir, 2015,2016)*; *(Garber, Hazan, Jin, Kakade, Musco, Netrapalli, Sidford, 2016);* and many more!

Empirical results: NN-PCA



y-axis denotes distance $f(\theta) - f(\hat{\theta})$ to an approximate optimum

$$\min_{\|w\|\leq 1,\ w \geq 0} -\frac{1}{2} w^\top \left(\sum_{i=1}^{n} x_i x_i^\top \right) w$$

Eigenvecs via SGD: *(Oja, Karhunen 1985)*; via SVRG *(Shamir, 2015,2016)*; *(Garber, Hazan, Jin, Kakade, Musco, Netrapalli, Sidford, 2016);* and many more!

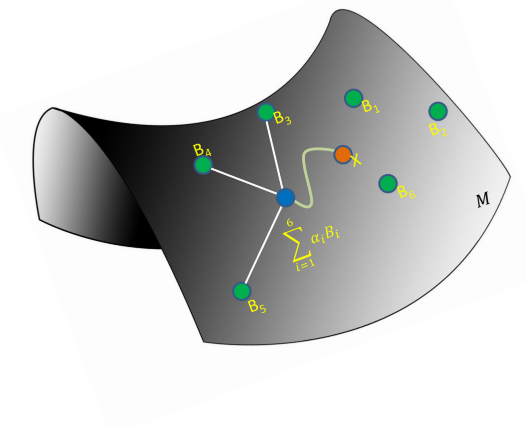
Finite-sum problems with nonconvex g(θ) and params θ lying on a known manifold

$$\min_{\theta \in \mathcal{M}} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

Example: eigenvector problems (the $\|\theta\|=1$ constraint)
problems with orthogonality constraints
low-rank matrices
positive definite matrices / covariances

Nonconvex optimization on manifolds

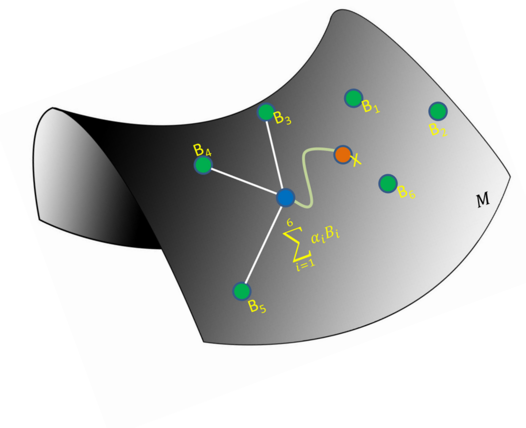
(Zhang, Reddi, Sra, 2016)



$$\min_{\theta \in \mathcal{M}} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

Nonconvex optimization on manifolds

(Zhang, Reddi, Sra, 2016)



$$\min_{\theta \in \mathcal{M}} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$

Related work

- *(Udriste, 1994)* — batch methods; textbook
- *(Edelman, Smith, Arias, 1999)* — classic paper; orthogonality constraints
- *(Absil, Mahony, Sepulchre, 2009)* — textbook; convergence analysis
- *(Boumal, 2014)* — phd thesis, algos, theory, examples
- *(Mishra, 2014)* — phd thesis, algos, theory, examples
- **manopt** — excellent matlab toolbox
- *(Bonnabel, 2013)* — Riemannnian SGD, asymptotic convg.
- and many more!

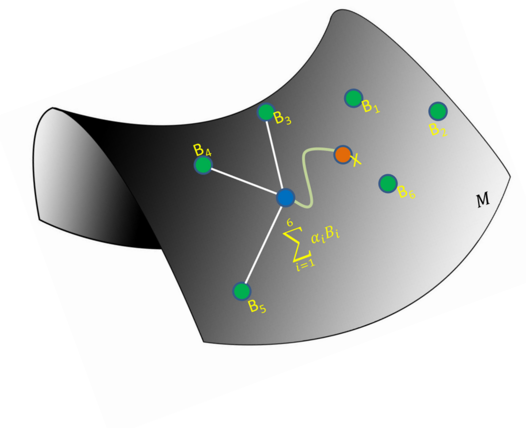
Exploiting manifold structure yields speedups

Nonconvex optimization on manifolds

First non-asymptotic results for general manifolds

(Zhang, Reddi, Sra, 2016)

$$\min_{\theta \in \mathcal{M}} \quad g(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$$



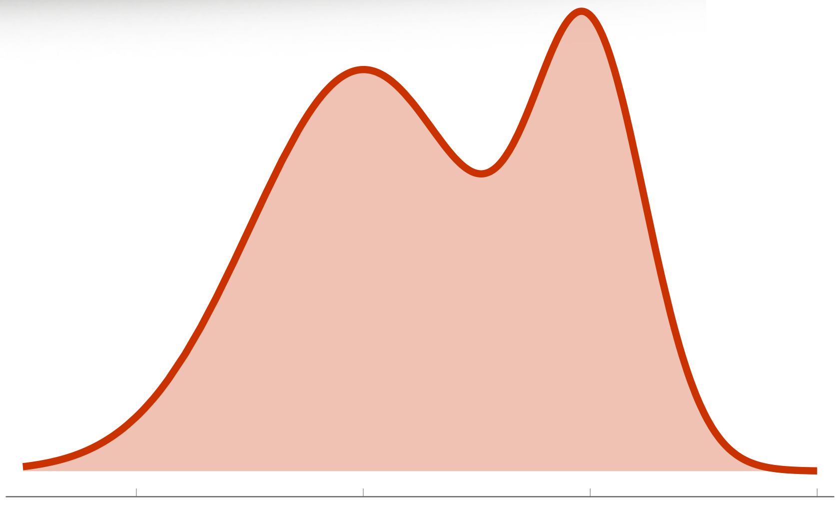
Related work

- *(Udriste, 1994)* — batch methods; textbook
- *(Edelman, Smith, Arias, 1999)* — classic paper; orthogonality constraints
- *(Absil, Mahony, Sepulchre, 2009)* — textbook; convergence analysis
- *(Boumal, 2014)* — phd thesis, algos, theory, examples
- *(Mishra, 2014)* — phd thesis, algos, theory, examples
- manopt — excellent matlab toolbox
- *(Bonnabel, 2013)* — Riemannnian SGD, asymptotic convg.
- and many more!

Exploiting manifold structure yields speedups

Example: Gaussian Mixture Model

$$p_{\mathrm{mix}}(x) := \sum_{k=1}^{K} \pi_k p_{\mathcal{N}}(x; \Sigma_k, \mu_k)$$

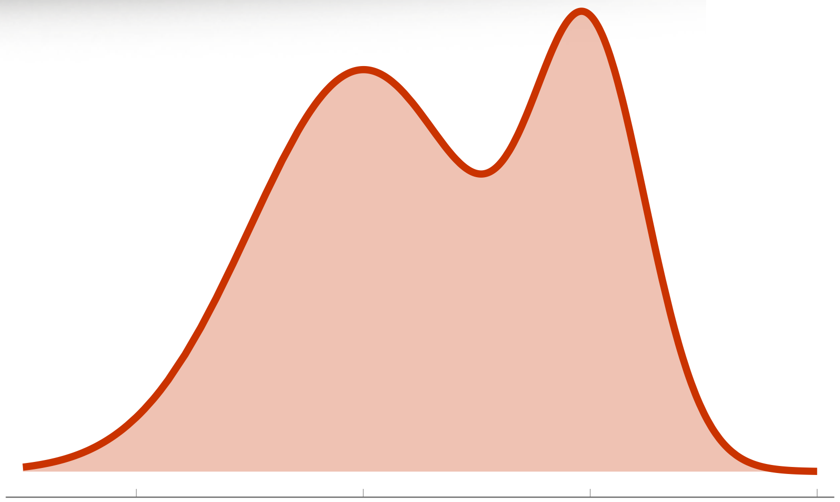


Likelihood $\max \prod_i p_{\mathrm{mix}}(x_i)$

Numerical challenge: positive definite constraint on Σ_k

Example: Gaussian Mixture Model

$$p_{\mathrm{mix}}(x) := \sum_{k=1}^{K} \pi_k p_{\mathcal{N}}(x; \Sigma_k, \mu_k)$$



Likelihood $\max \prod_i p_{\mathrm{mix}}(x_i)$

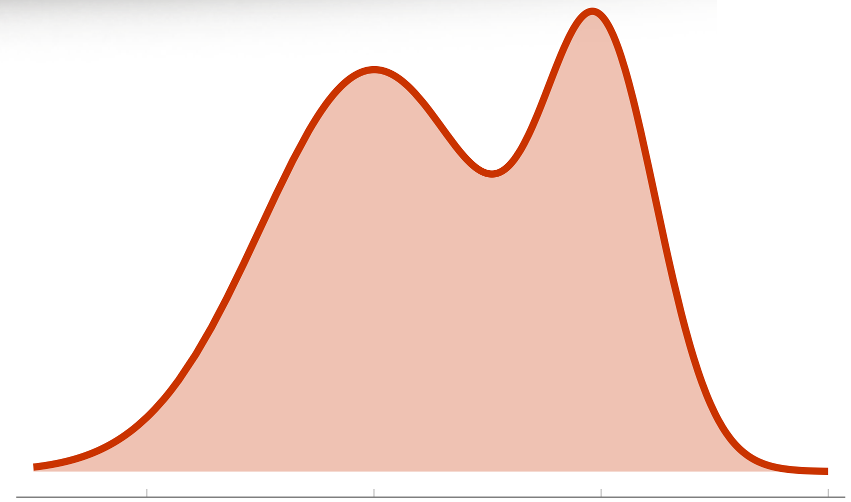
Numerical challenge: positive definite constraint on Σ_k

EM

Algo

Example: Gaussian Mixture Model

$$p_{\text{mix}}(x) := \sum_{k=1}^{K} \pi_k p_{\mathcal{N}}(x; \Sigma_k, \mu_k)$$

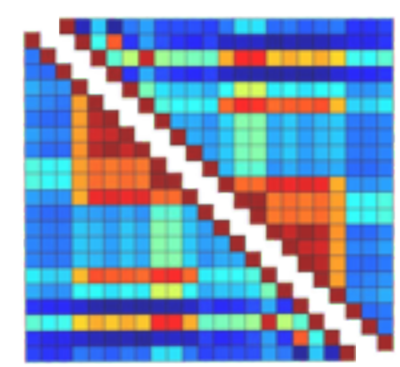


Likelihood $\max \prod_i p_{\text{mix}}(x_i)$

Numerical challenge: positive definite constraint on Σ_k

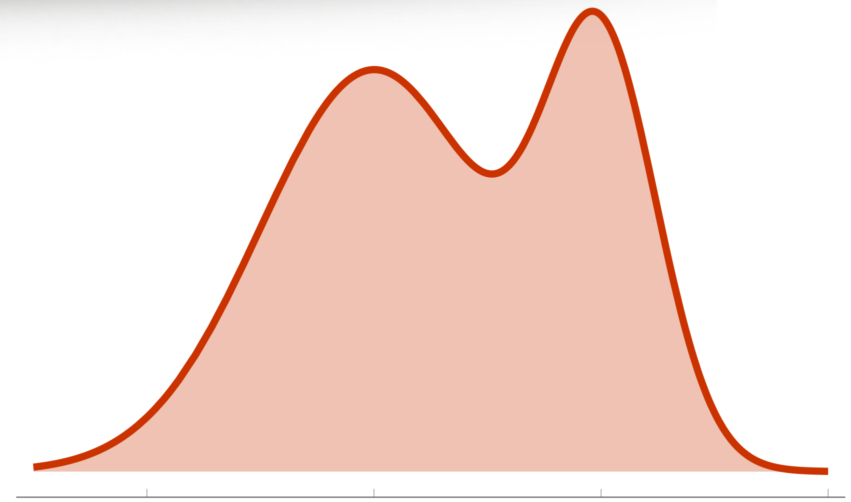
EM Algo

Cholesky LL^T



Example: Gaussian Mixture Model

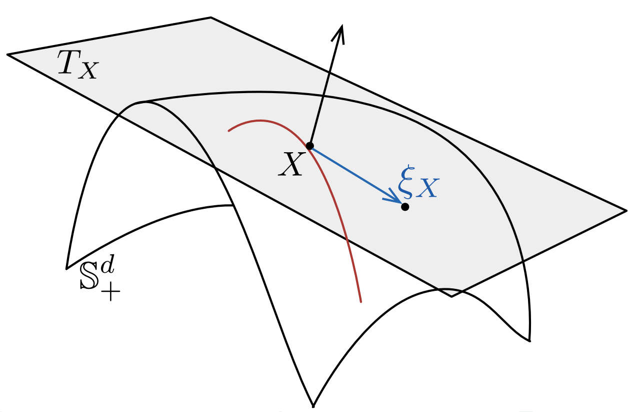
$$p_{\text{mix}}(x) := \sum_{k=1}^{K} \pi_k p_{\mathcal{N}}(x; \Sigma_k, \mu_k)$$



Likelihood $\max \prod_i p_{\text{mix}}(x_i)$

Numerical challenge: positive definite constraint on Σ_k

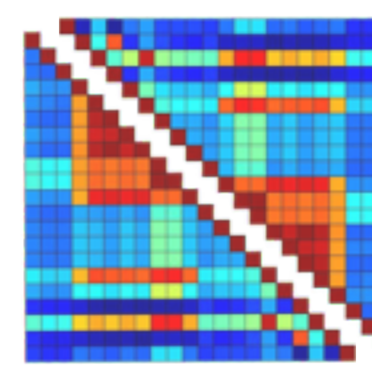
Riemannian
(new)



[Hosseini, Sra, 2015]

EM Algo

Cholesky
LL^T



Careful use of manifold geometry helps!

K
2
5
10

images dataset
d=35,
n=200,000

Riemannian-LBFGS (careful use of geometry)

github.com/utvisionlab/mixest

Careful use of manifold geometry helps!

K	EM
2	17s // 29.28
5	202s // 32.07
10	2159s // 33.05

images dataset
d=35,
n=200,000

Riemannian-LBFGS (careful use of geometry)

github.com/utvisionlab/mixest

Careful use of manifold geometry helps!

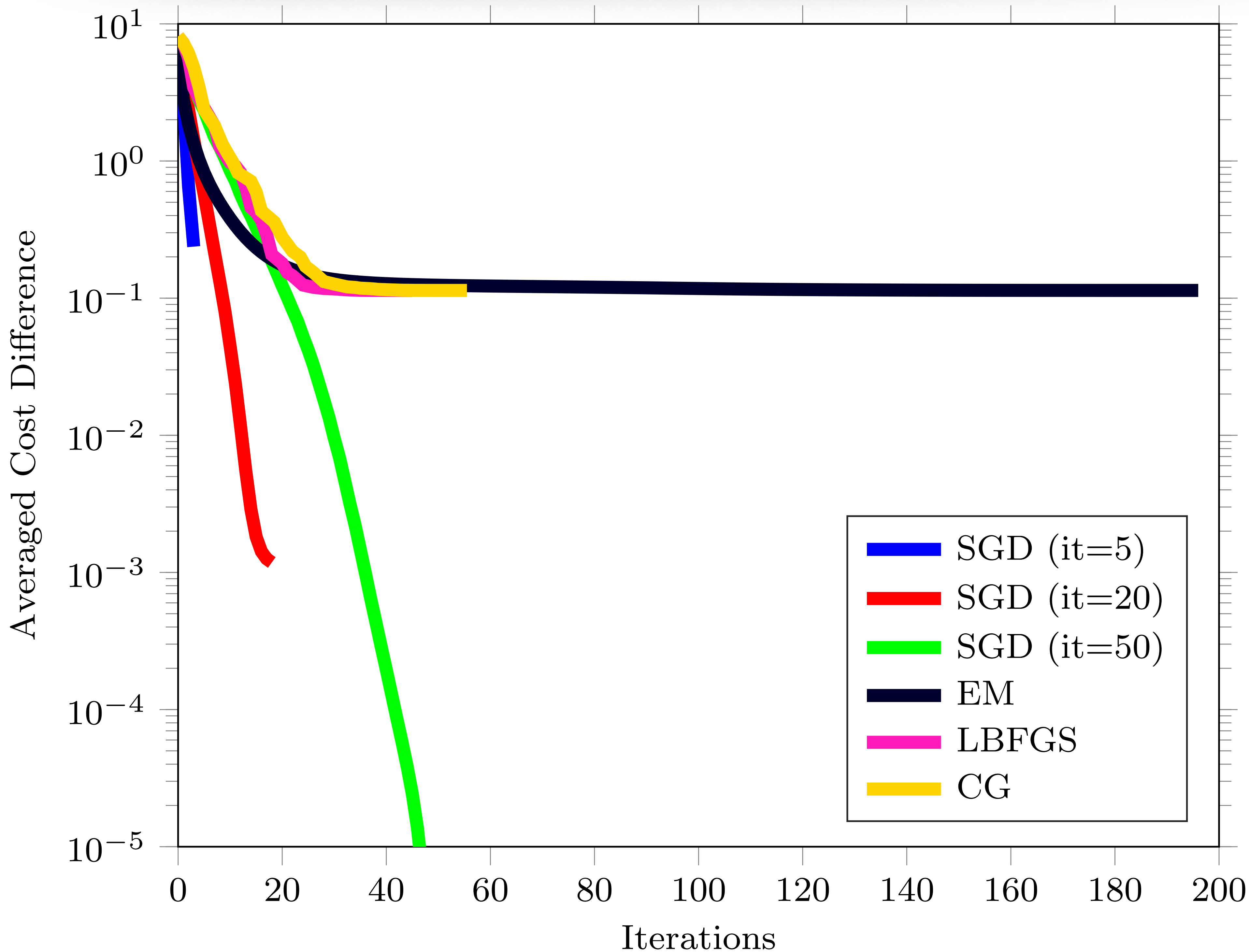
K	EM	R-LBFGS
2	17s // 29.28	**14s** // 29.28
5	202s // 32.07	**117s** // 32.07
10	2159s // 33.05	**658s** // 33.06

images dataset d=35, n=200,000

Riemannian-LBFGS (careful use of geometry)

github.com/utvisionlab/mixest

Riemannian SGD (multi-pass)



[Hosseini, Sra, 2017]

(d=90, n=515345, k=7)

Summary so far

- nc-SVRG/SAGA use fewer #IFO calls than SGD & GD
- Work well in practice
- Easier (than SGD) to use and tune:

  can use constant step-sizes
- Proximal extension holds a few surprises
- SGD and SVRG extend to Riemannian manifolds too

However: careful when using for deep networks

(a topic for another day!)

Beyond stationarity
Escaping saddle points

$$\mathbb{E}[\|\nabla f(x_k)\|^2] \leq \epsilon$$

$$\nabla^2 f(x) \succeq -\epsilon I$$

(epsilon-accurate second order critical)

Escaping saddle points

SGD takes $O(1/\epsilon^2)$ for approximate stationarity does not ensure second order criticality

Noisy SGD + strict-saddles (i.e., Hessian structure)

[Ge, Huang, Jin, Yuan, 2015] bad depend. on dimension

Escaping saddle points

SGD takes $O(1/\epsilon^2)$ for approximate stationarity does not ensure second order criticality

Noisy SGD + strict-saddles (i.e., Hessian structure)

[Ge, Huang, Jin, Yuan, 2015] bad depend. on dimension

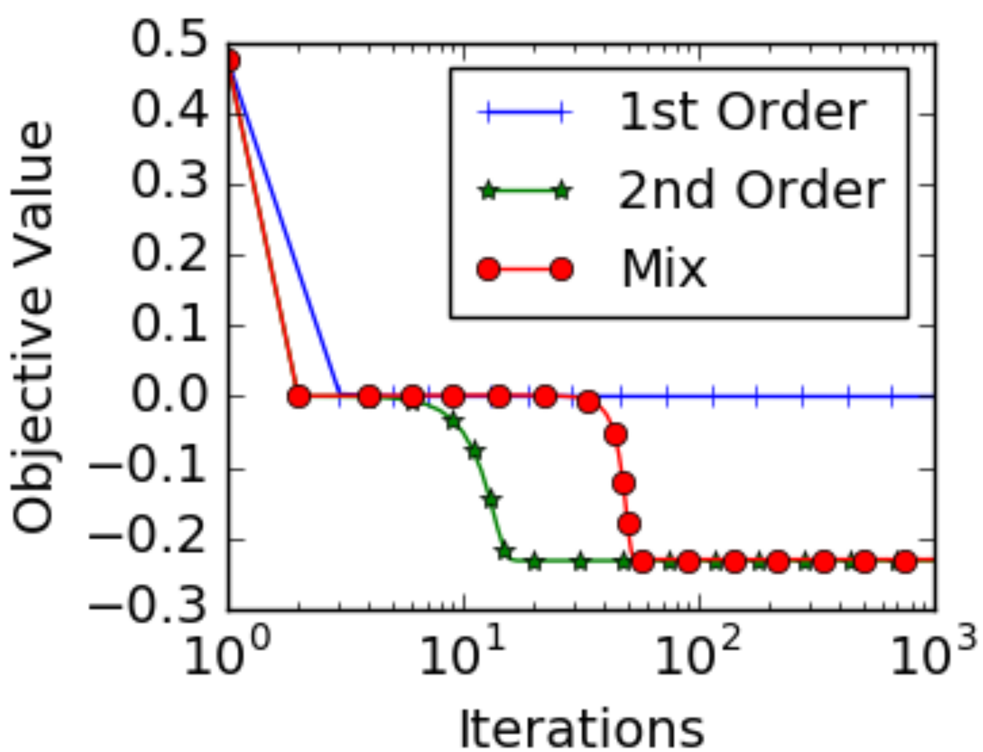
[Carmon, Duchi, Hinder, Sidford, 2016] alternate between 1^{st} and 2^{nd} order methods to escape saddles

Escaping saddle points

SGD takes $O(1/\epsilon^2)$ for approximate stationarity does not ensure second order criticality

Noisy SGD + strict-saddles (i.e., Hessian structure)

[Ge, Huang, Jin, Yuan, 2015] bad depend. on dimension



[Carmon, Duchi, Hinder, Sidford, 2016]

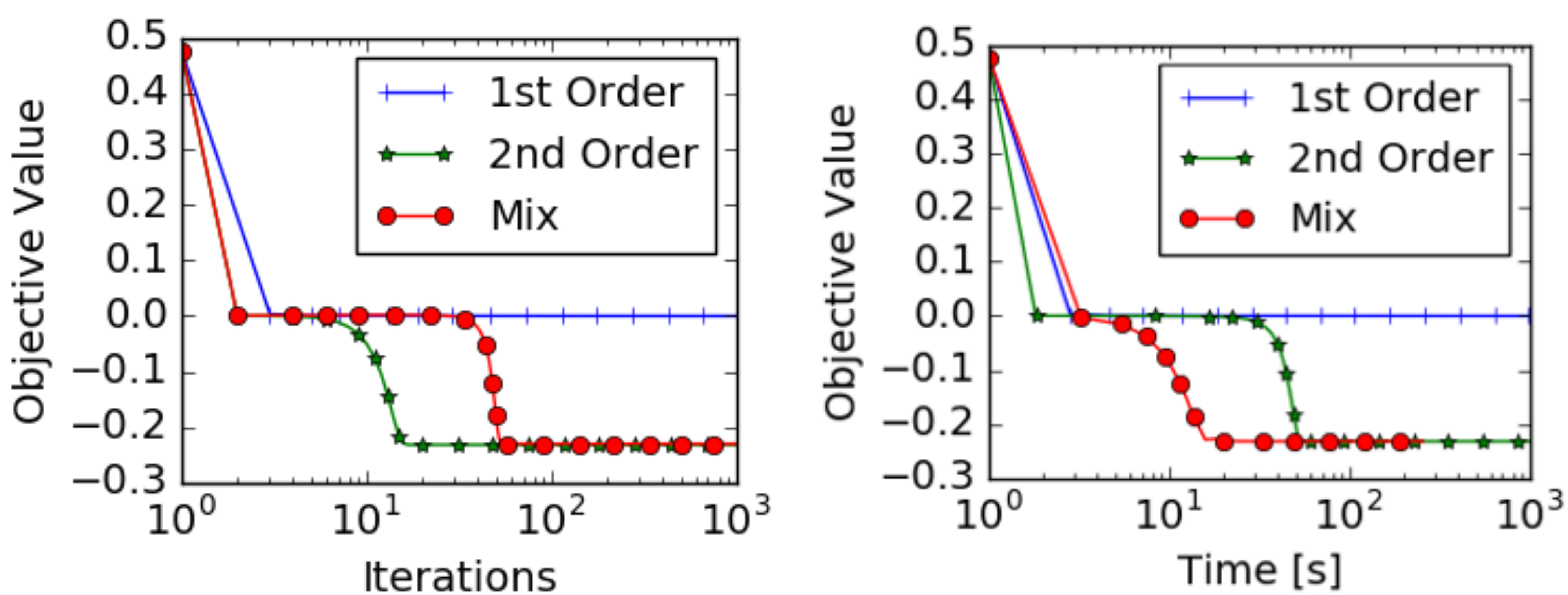
alternate between 1^{st} and 2^{nd} order methods to escape saddles

Escaping saddle points

SGD takes $O(1/\epsilon^2)$ for approximate stationarity does not ensure second order criticality

Noisy SGD + strict-saddles (i.e., Hessian structure)

[Ge, Huang, Jin, Yuan, 2015] bad depend. on dimension



[Carmon, Duchi, Hinder, Sidford, 2016]

alternate between 1^{st} and 2^{nd} order methods to escape saddles

Escaping saddle points: quick summary

Use Cubic Regularization *[Nesterov, Polyak, 2006]*

Try to make it fast *[Agarwal, Allen-Zhu, Bullins, Hazan, Ma, 2016]*

Use noisy gradient methods

[Ge, Huang, Jin, Yuan, 2015] *[Levy, 2016]*

(more careful noise control helps: [Jin, Ge, Netrapalli, Kakade, Jordan, 2017])

Escaping saddle points: quick summary

Use Cubic Regularization *[Nesterov, Polyak, 2006]*

Try to make it fast *[Agarwal, Allen-Zhu, Bullins, Hazan, Ma, 2016]*

Use noisy gradient methods

[Ge, Huang, Jin, Yuan, 2015] *[Levy, 2016]*

(more careful noise control helps: [Jin, Ge, Netrapalli, Kakade, Jordan, 2017])

Carefully mix first-order with second-order methods / info

Batch: *[Carmon, Duchi, Hinder, Sidford, 2016]*

Finite-sums: *[Reddi, Zaheer, Sra, Poczos, Bach, Salakhutdinov, Smola, 2017]*
[Allen-Zhu, 2017]

Escaping saddle points: quick summary

Use Cubic Regularization *[Nesterov, Polyak, 2006]*

Try to make it fast *[Agarwal, Allen-Zhu, Bullins, Hazan, Ma, 2016]*

Use noisy gradient methods

[Ge, Huang, Jin, Yuan, 2015] *[Levy, 2016]*

(more careful noise control helps: [Jin, Ge, Netrapalli, Kakade, Jordan, 2017])

Carefully mix first-order with second-order methods / info

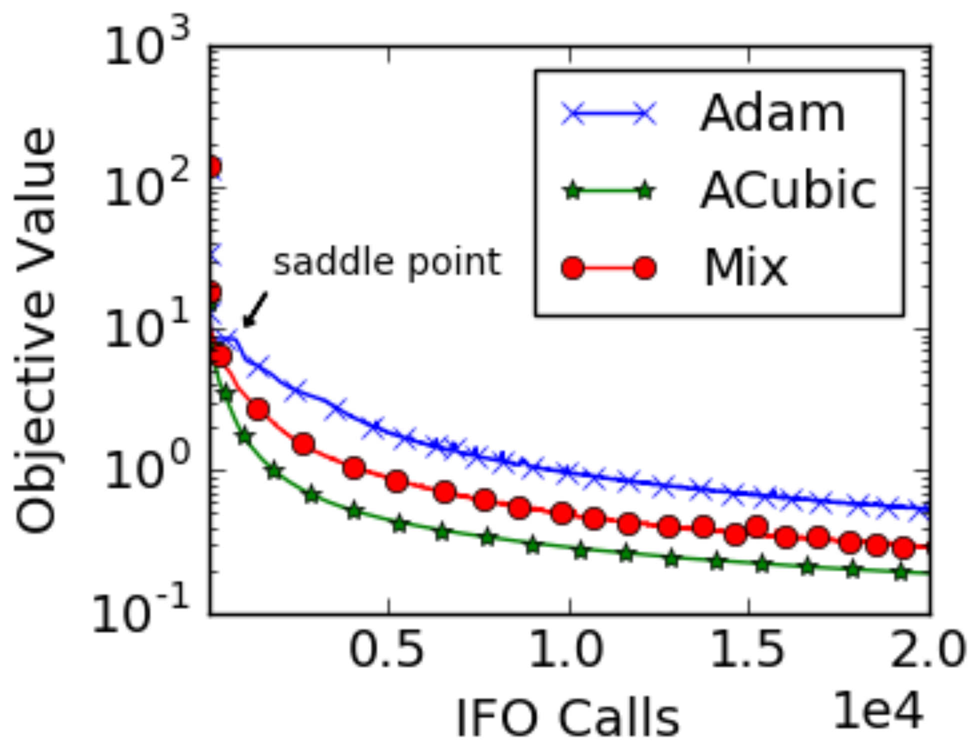
Batch: *[Carmon, Duchi, Hinder, Sidford, 2016]*

Finite-sums: *[Reddi, Zaheer, Sra, Poczos, Bach, Salakhutdinov, Smola, 2017]* *[Allen-Zhu, 2017]*

Third-order smoothness + Hessians

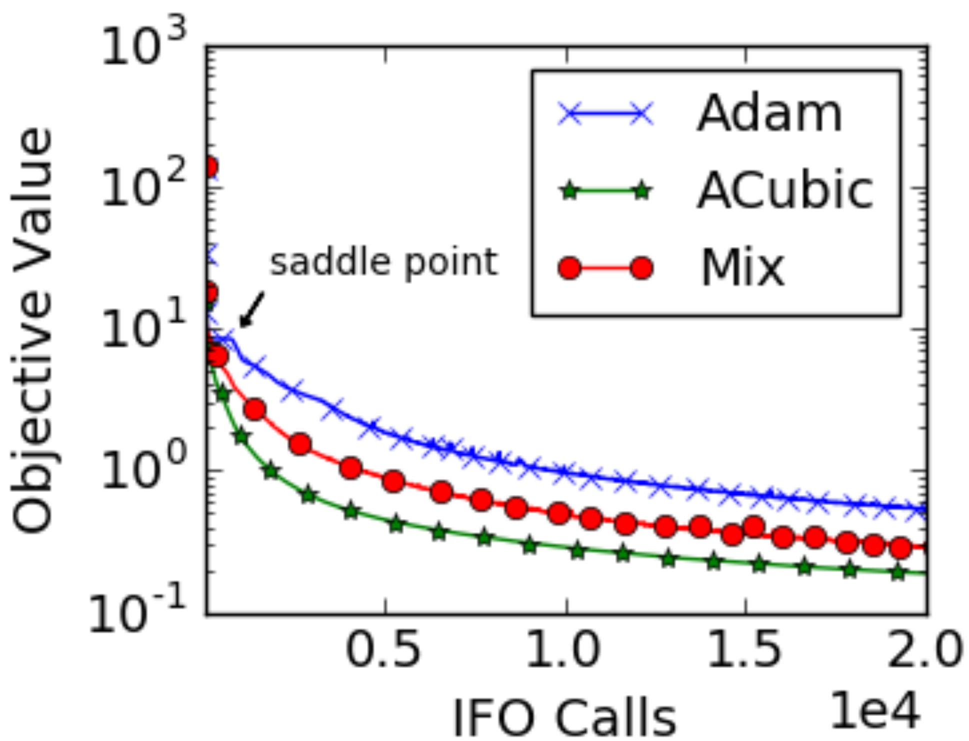
[Carmon, Hinder, Duchi, Sidford, 2017]

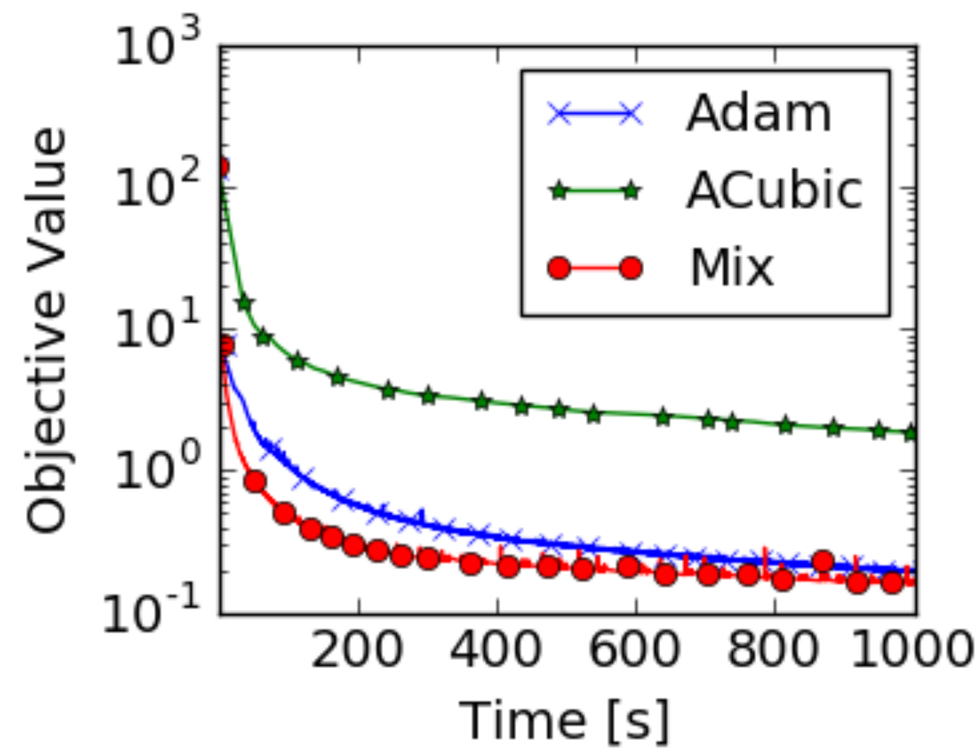
Experiment: deep autoencoders



[Reddi, Zaheer, Sra, Poczos, Bach, Salakhutdinov, Smola, 2017] simple algorithm and analysis

Experiment: deep autoencoders

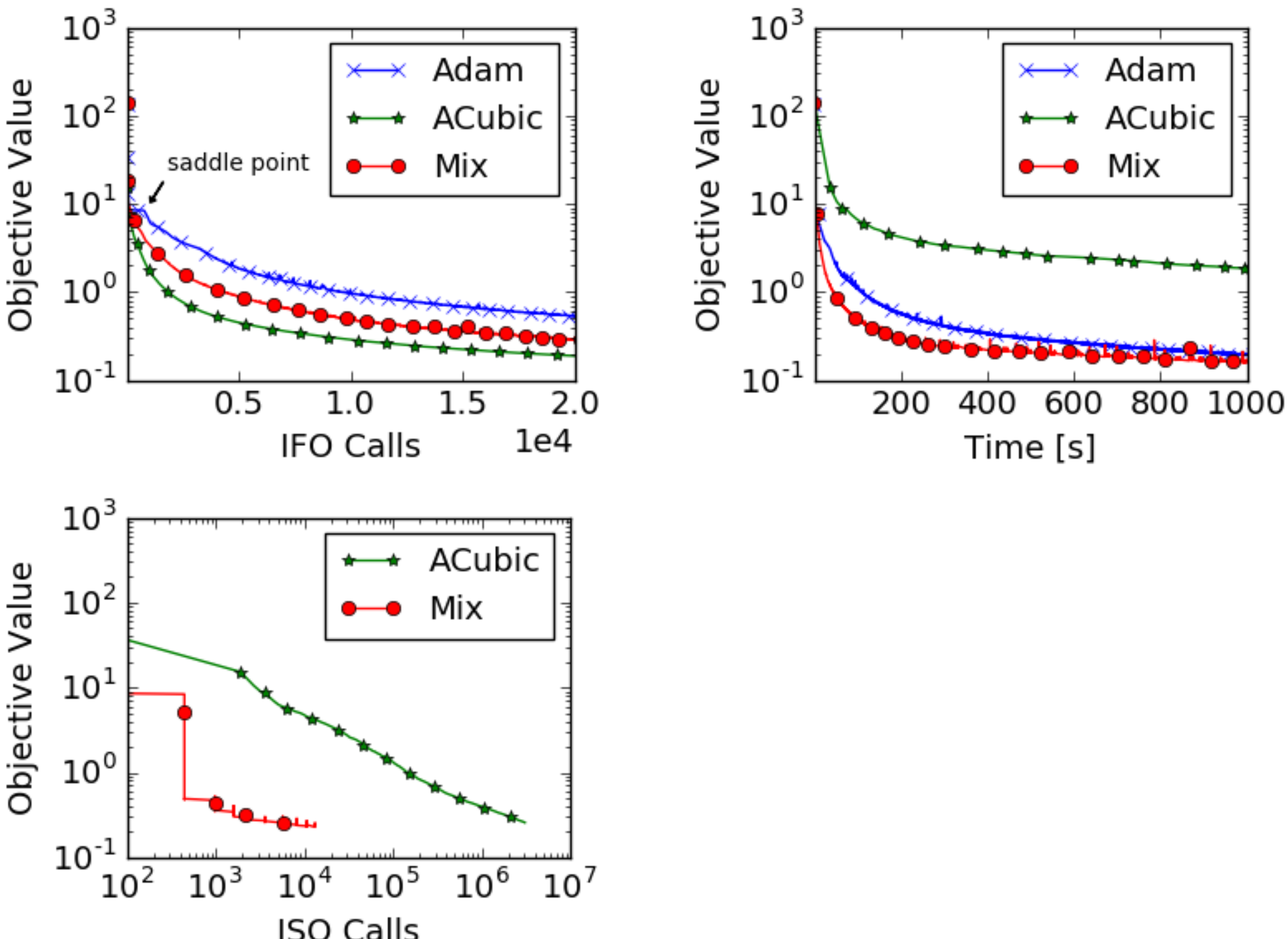




[Reddi, Zaheer, Sra, Poczos, Bach, Salakhutdinov, Smola, 2017]

simple algorithm and analysis

Experiment: deep autoencoders



[Reddi, Zaheer, Sra, Poczos, Bach, Salakhutdinov, Smola, 2017]

simple algorithm and analysis

Much more work, could not cover!

Much more work, could not cover!

- Stochastic quasi-convex optim. *(Hazan, Levy, Shalev-Shwartz, 2015)*
- Nonconvex Frank-Wolfe + SVRG: *(Reddi, Sra, Poczos, Smola,2016)*
- Newton-type + sketching *(Xu, Khosrani, Mahoney, 2016,17)*
- stochastic quasi-Newton methods *(Wang, Ma, Goldfarb, Liu, 2017)*
- nonconvex robust global optimization *(Staib, Jegelka, 2017)*
- accelerated nonconvex methods *(Paquette, Lin, Drusvyatskiy, Mairal, Harchaoui, 2017; Allen-Zhu 2017)*
- global optim. on manifolds *(Zhang, Sra,'16; Zhang, Reddi, Sra, '16)*
- convex relaxations of nonconvex, sums-of-squares, etc..
- momentum + nonconvex + stochastic *(Yang, Lin, Li, 2016)*
- many more, this is just a smattering….

Perspectives

Perspectives

* Impact of non-convexity on generalization

Perspectives

* Impact of non-convexity on generalization
* Non-separable problems, min-max (GAN) problems

Perspectives

* Impact of non-convexity on generalization
* Non-separable problems, min-max (GAN) problems
* Convergence theory, local and global

Perspectives

* Impact of non-convexity on generalization
* Non-separable problems, min-max (GAN) problems
* Convergence theory, local and global
* Lower-bounds for nonconvex finite-sums

Perspectives

* Impact of non-convexity on generalization
* Non-separable problems, min-max (GAN) problems
* Convergence theory, local and global
* Lower-bounds for nonconvex finite-sums
* New applications, models, e.g., manifold optimization

Perspectives

* Impact of non-convexity on generalization
* Non-separable problems, min-max (GAN) problems
* Convergence theory, local and global
* Lower-bounds for nonconvex finite-sums
* New applications, models, e.g., manifold optimization
* Collecting other more "tractable" nonconvex models

Perspectives

* Impact of non-convexity on generalization
* Non-separable problems, min-max (GAN) problems
* Convergence theory, local and global
* Lower-bounds for nonconvex finite-sums
* New applications, models, e.g., manifold optimization
* Collecting other more "tractable" nonconvex models
* Nonconvexity, optimal transport and beyond