Better models in optimization

John Duchi (based on joint work with Feng Ruan and Hilal Asi)
Stanford University

August 2018

Outline

Motivating experiments

Models in optimization

Stochastic optimization
- Stability is better
- Nothing gets worse
- Adaptivity in easy problems

Revisiting experimental results

Phase retrieval and composite optimization (if time)

Outline

Stochastic gradient methods

The problem in this talk:

$$\begin{aligned}&\underset{x}{\text{minimize}}\ F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s)\\&\text{subject to}\ x \in X\end{aligned}$$

Stochastic gradient methods

The problem in this talk:

$$\begin{aligned} &\underset{x}{\text{minimize}}\ F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s) \\ &\text{subject to}\ x \in X \end{aligned}$$

Stochastic gradient method:

$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k; S_k)$$

Stochastic gradient methods

The problem in this talk:

$$\begin{aligned} &\operatorname*{minimize}_x \; F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s) \\ &\text{subject to } x \in X \end{aligned}$$

Stochastic gradient method:

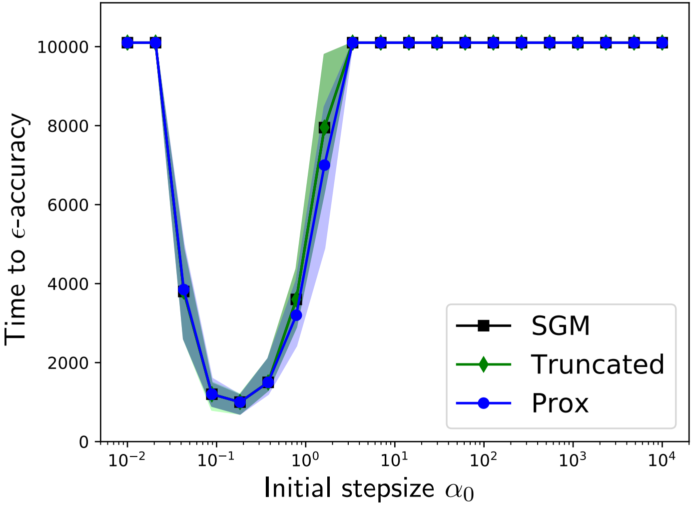
$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k; S_k)$$

Why we use this?

- Easy to analyze?
- Default in software packages and simple to implement?
- It works?

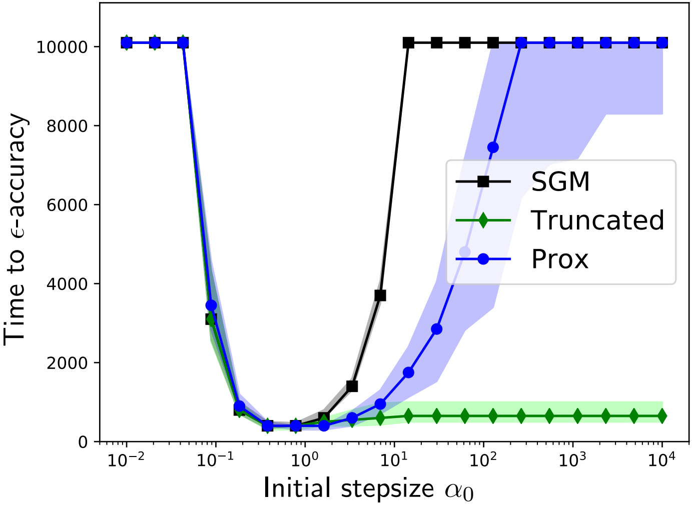
Linear regression

$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$



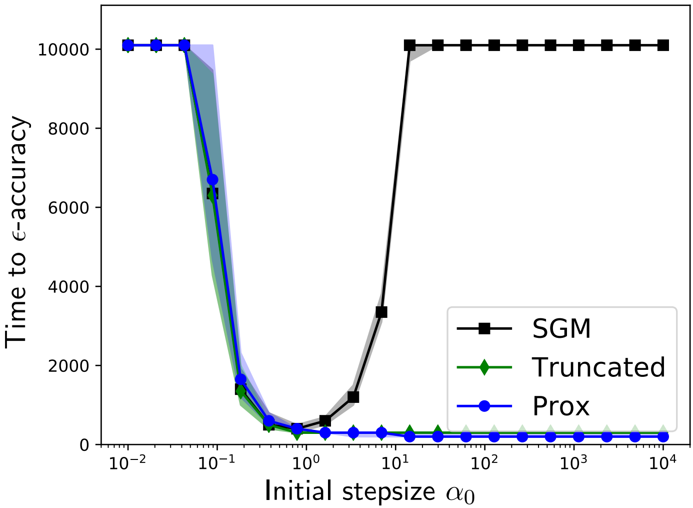
Linear regression

$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$



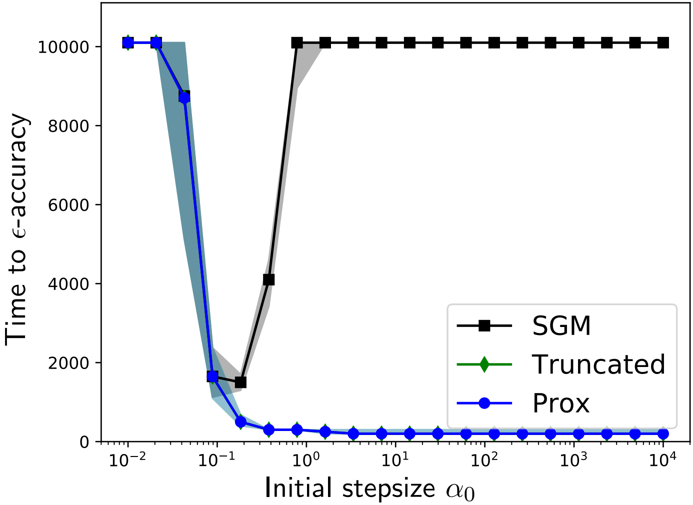
Linear regression

$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$



Absolute loss regression

$$F(x) = \frac{1}{m}\sum_{i=1}^{m} |a_i^T x - b_i|$$



Outline

Motivating experiments

Models in optimization

Stochastic optimization
- Stability is better
- Nothing gets worse
- Adaptivity in easy problems

Revisiting experimental results

Phase retrieval and composite optimization (if time)

Optimization methods

How do we solve optimization problems?

1. Build a "good" but simple local model of f
2. Minimize the model (perhaps regularizing)

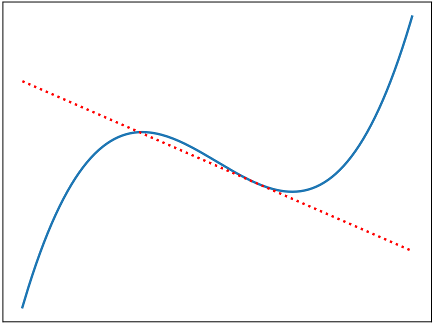
Optimization methods

How do we solve optimization problems?

1. Build a "good" but simple local model of f
2. Minimize the model (perhaps regularizing)

Gradient descent: Taylor (first-order) model

$$f(y) \approx f_x(y) := f(x) + \nabla f(x)^T (y - x)$$



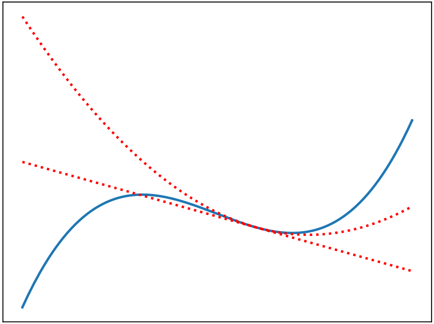
Optimization methods

How do we solve optimization problems?

1. Build a "good" but simple local model of f
2. Minimize the model (perhaps regularizing)

Newton's method: Taylor (second-order) model

$$f(y) \approx f_x(y) := f(x) + \nabla f(x)^T (y - x) + (1/2)(y - x)^T \nabla^2 f(x)(y - x)$$



Generic(ish) optimization methods

Iterate

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|^2 \right\}$$

Generic(ish) optimization methods

Iterate

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

- Proximal point method ($f_x = f$) [Rockafellar 76]
- Gradient descent ($f_x(y) = f(x) + \langle \nabla f(x), y - x \rangle$)
- Newton ($f_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(x - y)^T \nabla^2 f(x)(x - y)$)
- Prox-linear ($f_x(y) = h(c(x) + \nabla c(x)^T (y - x))$)

The aProx family for stochastic optimization

Iterate:

- Sample $S_k \stackrel{\text{iid}}{\sim} P$
- Update by minimizing model

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|^2 \right\}$$

The aProx family for stochastic optimization

Iterate:

- Sample $S_k \overset{\text{iid}}{\sim} P$
- Update by minimizing model

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

Examples:

- Stochastic gradient method
- Stochastic proximal-point (implicit gradient) method, $f_{x_k}(x) = f(x)$ [Rockafellar 76; Kulis & Bartlett 10; Karampatziakis & Langford 11; Bertsekas 11; Toulis & Airoldi 17; Ryu & Boyd 16]
- Stochastic prox-linear methods [D. & Ruan 18; Asi & D. 18]

Models in stochastic optimization

Stochastic gradient method

$$f_x(y; s) = f(x; s) + \langle f'(x; s), y - x \rangle \;\; \text{for some } f'(x; s) \in \partial f(x; s)$$

Models in stochastic optimization

Stochastic gradient method

$$f_x(y; s) = f(x; s) + \langle f'(x; s), y - x \rangle \text{ for some } f'(x; s) \in \partial f(x; s)$$

Conditions on our models (convex case)

i. Convex model:

$$y \mapsto f_x(y; s) \quad \text{is convex}$$

ii. Lower bound:

$$f_x(y; s) \le f(y; s)$$

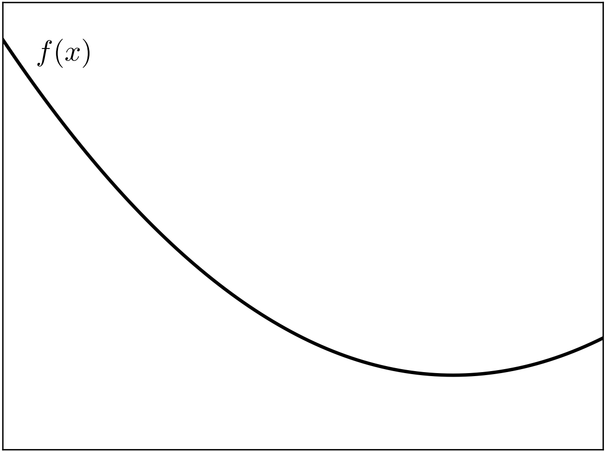
iii. Local correctness:

$$f_x(x; s) = f(x; s) \quad \text{and} \quad \partial f_x(x; s) \subset \partial f(x; s)$$

[D. & Ruan 17; Davis & Drusvyatskiy 18]

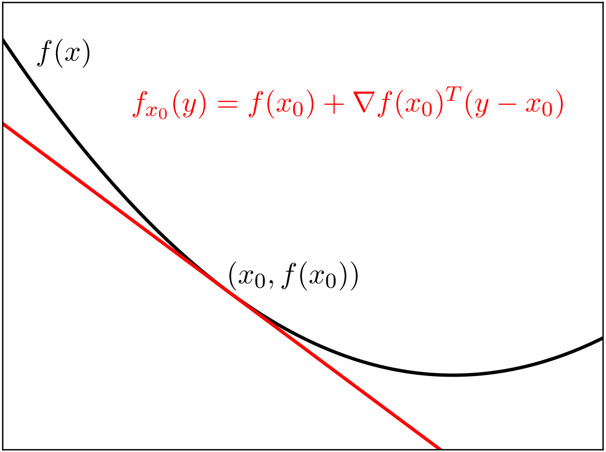
Modeling conditions

Model $f_x(y)$ of f near x



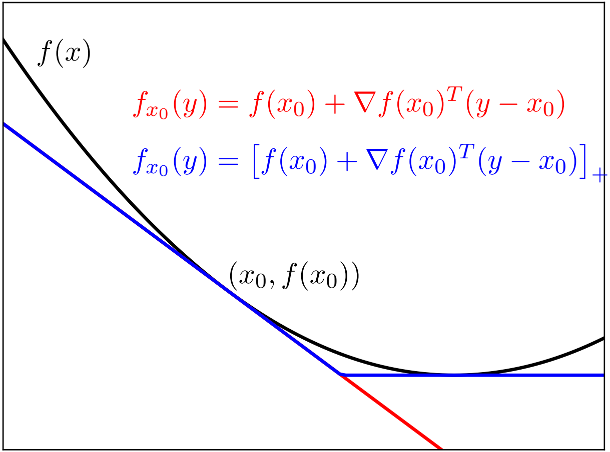
Modeling conditions

Model $f_x(y)$ of f near x

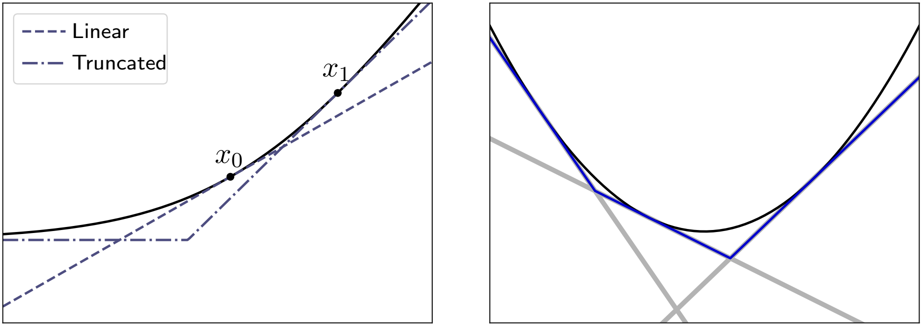


Modeling conditions

Model $f_x(y)$ of f near x



Models in stochastic optimization



i. (Sub)gradient: $f_x(y) = f(x) + \langle f'(x), y - x\rangle$

ii. Truncated: $f_x(y) = (f(x) + \langle f'(x), y - x\rangle) \vee \inf_x f(x)$

iii. Bundle/multi-line: $f_x(y) = \max\{f(x_i) + \langle f'(x_i), x - x_i\rangle\}$

The aProx family

Iterate:

- Sample $S_k \stackrel{\text{iid}}{\sim} P$
- Update by minimizing model

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|^2 \right\}$$

Outline

Motivating experiments

Models in optimization

Stochastic optimization

- Stability is better
- Nothing gets worse
- Adaptivity in easy problems

Revisiting experimental results

Phase retrieval and composite optimization (if time)

The aProx family

Iterate:

- Sample $S_k \stackrel{\text{iid}}{\sim} P$
- Update by minimizing model

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|^2 \right\}$$

Non-divergence

Example

Let

$$b_i = a_i^T x^\star$$

for $i = 1, 2, \ldots, m$.

- Iterate stochastic gradient method on $\frac{1}{2m}\sum_{i=1}^{m}(a_i^T x - b_i)^2$

Non-divergence

Example

Let

$$b_i = a_i^T x^\star$$

for $i = 1, 2, \ldots, m$.

- Iterate stochastic gradient method on $\frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$
- for all iterations

$$(x_{k+1} - x^\star) = (I - \alpha_k a_i a_i^T)(x_k - x^\star)$$

Non-divergence

Example

Let

$$b_i = a_i^T x^\star$$

for $i = 1, 2, \ldots, m$.

- Iterate stochastic gradient method on $\frac{1}{2m}\sum_{i=1}^m (a_i^T x - b_i)^2$
- for all iterations

$$(x_{k+1} - x^\star) = (I - \alpha_k a_i a_i^T)(x_k - x^\star)$$

- If $\alpha_1, \alpha_2, \ldots$ too large, may diverge exponentially at first: if $\Sigma = m^{-1}\sum_{i=1}^m a_i a_i^T$,

$$\mathbb{E}[x_{k+1} - x^\star] = \prod_{i=1}^{k}(\alpha_i \Sigma - I)x^\star$$

Non-divergence

Example

Let

$$b_i = a_i^T x^\star$$

for $i = 1, 2, \ldots, m$.

- Iterate stochastic gradient method on $\frac{1}{2m}\sum_{i=1}^m (a_i^T x - b_i)^2$
- for all iterations

$$(x_{k+1} - x^\star) = (I - \alpha_k a_i a_i^T)(x_k - x^\star)$$

- If $\alpha_1, \alpha_2, \ldots$ too large, may diverge exponentially at first: if $\Sigma = m^{-1}\sum_{i=1}^m a_i a_i^T$,

$$\mathbb{E}[x_{k+1} - x^\star] = \underbrace{\prod_{i=1}^k (\alpha_i \Sigma - I)}_{\text{exponential?}} x^\star$$

Stability guarantees

Use full stochastic-proximal method,

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$

Theorem (Asi & D. 18)

Assume $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ *is non-empty and* $\mathbb{E}[\|f'(x^\star; S)\|^2] \le \sigma^2$*. Then*

$$\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \le \operatorname{dist}(x_0, \mathcal{X}^\star)^2 + \sigma^2 \sum_{i=1}^{k} \alpha_i^2$$

Stability guarantees

Use full stochastic-proximal method,

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$

Theorem (Asi & D. 18)

Assume $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ *is non-empty and* $\mathbb{E}[\|f'(x^\star; S)\|^2] \le \sigma^2$*. Then*

$$\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \le \operatorname{dist}(x_0, \mathcal{X}^\star)^2 + \sigma^2 \sum_{i=1}^{k} \alpha_i^2$$

Theorem (Asi & D. 18)

Under the same assumptions,

$$\sup_k \operatorname{dist}(x_k, \mathcal{X}^\star) < \infty \quad \text{and} \quad \operatorname{dist}(x_k, \mathcal{X}^\star) \stackrel{a.s.}{\to} 0.$$

Stability guarantees under growth

Assume that local strong convexity

$$f(y;s) \geq f(x;s) + \langle f'(x;s), y - x \rangle + \frac{1}{2}(x-y)^T \Sigma(s)(x-y)$$

holds with $\mathbb{E}[\Sigma(S)] = \overline{\Sigma} \succ 0$

Theorem (Asi & D. 18)

The stochastic proximal-point method satisfies

$$\mathbb{E}[\|x_{k+1} - x^\star\|_2^2 \mid x_k] \leq (1 - c\alpha_k)\,\|x_k - x^\star\|_2^2 + \sigma^2\alpha_k^2.$$

and

$$\mathbb{E}[\|x_k - x^\star\|_2^2] \lesssim \sigma^2 k \alpha_k^2.$$

Stability guarantees under growth

Assume that local strong convexity

$$f(y;s) \geq f(x;s) + \langle f'(x;s), y - x \rangle + \frac{1}{2}(x-y)^T \Sigma(s)(x-y)$$

holds with $\mathbb{E}[\Sigma(S)] = \overline{\Sigma} \succ 0$

Theorem (Asi & D. 18)

The stochastic proximal-point method satisfies

$$\mathbb{E}[\|x_{k+1} - x^\star\|_2^2 \mid x_k] \leq (1 - c\alpha_k) \|x_k - x^\star\|_2^2 + \sigma^2 \alpha_k^2.$$

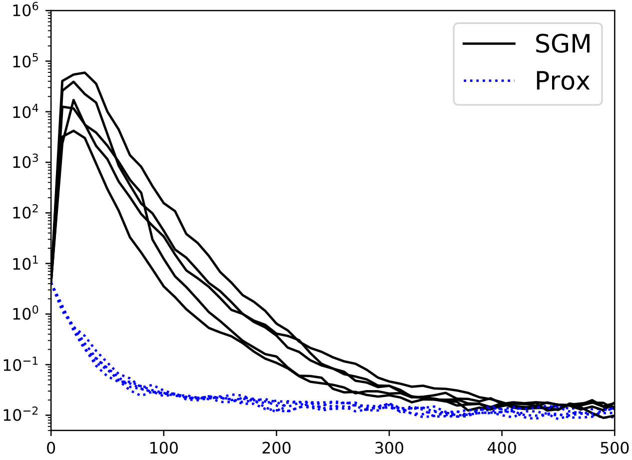
and

$$\mathbb{E}[\|x_k - x^\star\|_2^2] \lesssim \sigma^2 k \alpha_k^2.$$

(Always converging toward optimum)

Example behaviors

On least-squares objective $F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$



A few additional stability guarantees

- Do not need full proximal method, just accurate enough approximations
- Do not need convexity; some forms of weak convexity sufficient for stability

Classical asymptotic analysis

Theorem (Polyak & Juditsky 92)

Let F be convex and strongly convex in a neighborhood of $x^\star$, and assume that $f(x;S)$ are globally smooth. For x_k generated by stochastic gradient method,

$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k}(x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1}\,\mathrm{Cov}(\nabla f(x^\star;S))\nabla^2 F(x^\star)^{-1}\right).$$

New asymptotic analysis

Theorem (Asi & D. 18)

Let F be convex and strongly convex in a neighborhood of $x^\star$, and assume that $f(x; S)$ are smooth near $x^\star$. Then if x_k remain bounded and the models $f_{x_k}(\cdot; S_k)$ satisfy our conditions,

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (x_i - x^\star) \stackrel{d}{\rightsquigarrow} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \nabla^2 F(x^\star)^{-1}\right).$$

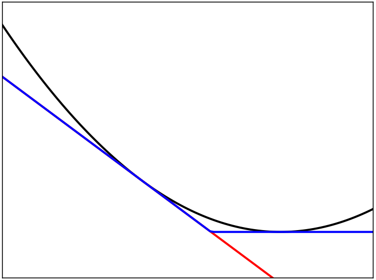
New asymptotic analysis

Theorem (Asi & D. 18)

Let F be convex and strongly convex in a neighborhood of $x^\star$, and assume that $f(x; S)$ are smooth near $x^\star$. Then if x_k remain bounded and the models $f_{x_k}(\cdot; S_k)$ satisfy our conditions,

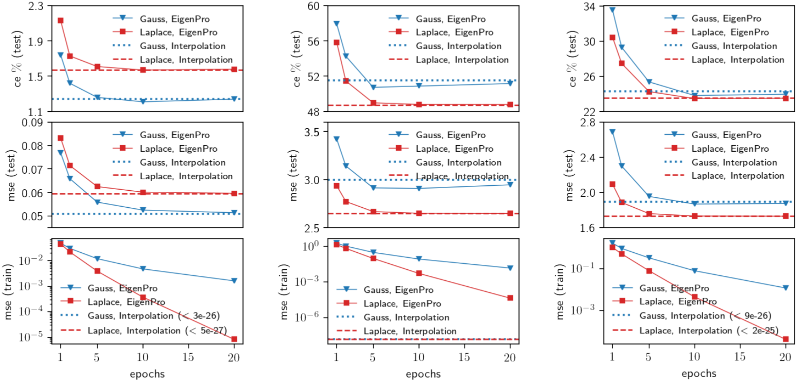
$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}\left(0, \nabla^2 F(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \nabla^2 F(x^\star)^{-1}\right).$$

- Optimal by local minimax theorem [Hájek 72; Le Cam 73; D. & Ruan 18]
- Key insight: subgradients of $f_{x_k}(\cdot; S_k)$ close to $\nabla f(x_k; S_k)$



What is an easy problem?

- Interpolation problems [Belkin, Hsu, Mitra 18; Ma, Bassily, Belkin 18]
- Overparameterized linear systems (Kaczmarz algorithms) [Strohmer & Vershynin 09; Needell, Srebro, Ward 14; Needell & Tropp 14]
- Random projections for linear constraints [Leventhal & Lewis 10]



What is an easy problem?

$$\underset{x}{\text{minimize}}\ \ F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s)$$

What is an easy problem?

$$\underset{x}{\text{minimize}} \; F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s)$$

Definition: Problem is *easy* if there exists $x^\star$ such that $f(x^\star; S) = \inf_x f(x; S)$ with probability 1. [Schmidt & Le Roux 13; Ma, Bassily, Belkin 18; Belkin, Rakhlin, Tsybakov 18]

What is an easy problem?

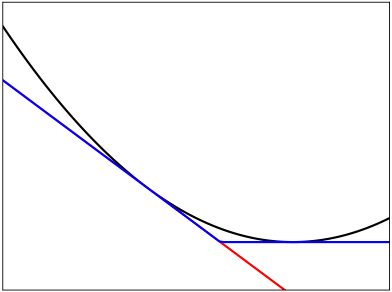
$$\underset{x}{\text{minimize}} \ F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s)$$

Definition: Problem is *easy* if there exists $x^\star$ such that $f(x^\star;S) = \inf_x f(x;S)$ with probability 1. [Schmidt & Le Roux 13; Ma, Bassily, Belkin 18; Belkin, Rakhlin, Tsybakov 18]

One additional condition

iv. The models f_x satisfy

$$f_x(y;s) \geq \inf_{x^\star \in X} f(x^\star;s)$$



Easy strongly convex problems

Theorem (Asi & D. 18)

Let the function F satisfy the growth condition

$$F(x) \geq F(x^\star) + \frac{\lambda}{2} \operatorname{dist}(x, X^\star)^2$$

where $X^\star = \operatorname{argmin}_x F(x)$, and be easy. Then

$$\mathbb{E}[\operatorname{dist}(x_k, X^\star)^2] \leq \max\left\{\exp\left(-c \sum_{i=1}^{k} \alpha_i\right), \exp\left(-ck\right)\right\} \operatorname{dist}(x_1, X^\star)^2.$$

Easy strongly convex problems

Theorem (Asi & D. 18)

Let the function F satisfy the growth condition

$$F(x) \geq F(x^\star) + \frac{\lambda}{2} \operatorname{dist}(x, X^\star)^2$$

where $X^\star = \operatorname{argmin}_x F(x)$, and be easy. Then

$$\mathbb{E}[\operatorname{dist}(x_k, X^\star)^2] \leq \max\left\{\exp\left(-c\sum_{i=1}^{k}\alpha_i\right), \exp\left(-ck\right)\right\} \operatorname{dist}(x_1, X^\star)^2.$$

- Adaptive no matter the stepsizes
- Most other results (e.g. for SGM [Schmidt & Le Roux 13; Ma, Bassily, Belkin 18]) require careful stepsize choices

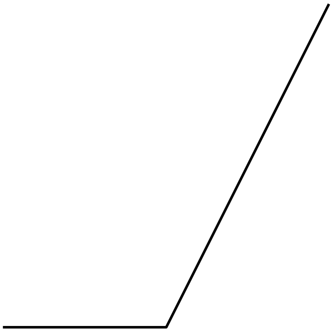
Sharp convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m} \sum_{i=1}^{m} \left[1 - a_i^T x\right]_+$



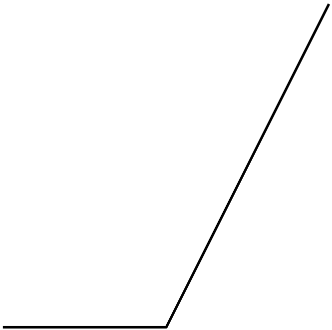
Sharp convex problems

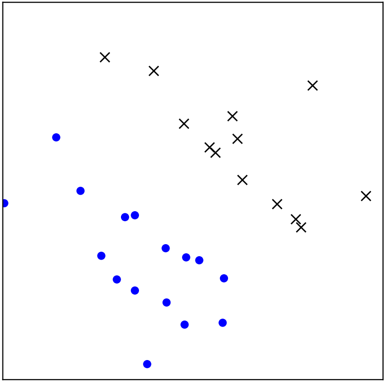
Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m} \sum_{i=1}^{m} \left[1 - a_i^T x\right]_+$





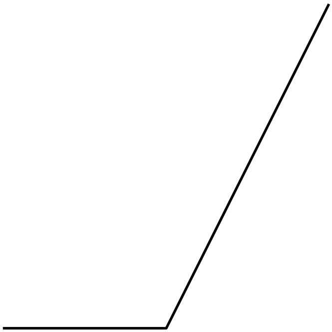
Sharp convex problems

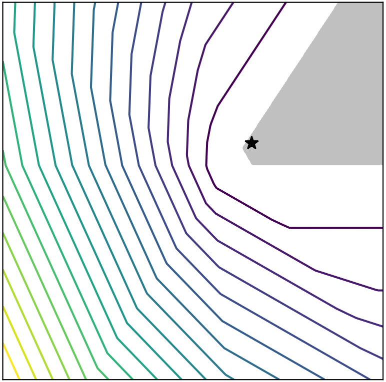
Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m}\sum_{i=1}^{m}\left[1 - a_i^T x\right]_+$





Sharp convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m}\sum_{i=1}^{m} \left[1 - a_i^T x\right]_+$
- Projection onto intersections: $F(x) = \frac{1}{m}\sum_{i=1}^{m} \operatorname{dist}(x, C_i)$

Sharp convex problems

Definition: An objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$. [Ferris 88; Burke & Ferris 95]

- Piecewise linear objectives
- Hinge loss $F(x) = \frac{1}{m} \sum_{i=1}^m \left[1 - a_i^T x\right]_+$
- Projection onto intersections: $F(x) = \frac{1}{m} \sum_{i=1}^m \operatorname{dist}(x, C_i)$

Theorem (Asi & D. 18)

Let F have sharp growth and be easy. Then

$$\mathbb{E}[\operatorname{dist}(x_{k+1}, X^\star)^2] \leq \max\left\{\exp(-ck), \exp\left(-c\sum_{i=1}^k \alpha_i\right)\right\} \operatorname{dist}(x_1, X^\star)^2.$$

Outline

Motivating experiments

Models in optimization

Stochastic optimization
- Stability is better
- Nothing gets worse
- Adaptivity in easy problems

Revisiting experimental results

Phase retrieval and composite optimization (if time)

Methods

Iterate

$$x_{k+1} = \operatorname*{argmin}_{x} \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \left\| x - x_k \right\|_2^2 \right\}$$

Methods

Iterate

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f_{x_k}(x; S_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}$$

- Stochastic gradient

$$f_{x_k}(x; S_k) = f(x_k; S_k) + \langle f'(x_k; S_k), x - x_k \rangle$$

- Truncated gradient ($f \geq 0$):

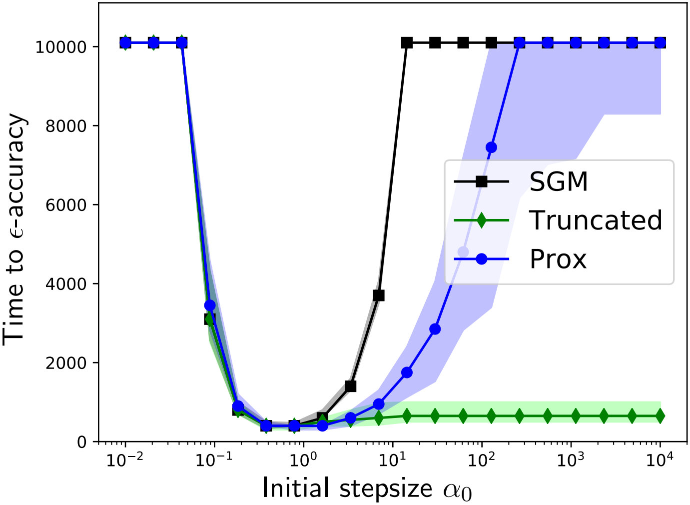
$$f_{x_k}(x; S_k) = \left[f(x_k; S_k) + \langle f'(x_k; S_k), x - x_k \rangle \right]_+$$

- (Stochastic) proximal point

$$f_{x_k}(x; S_k) = f(x; S_k)$$

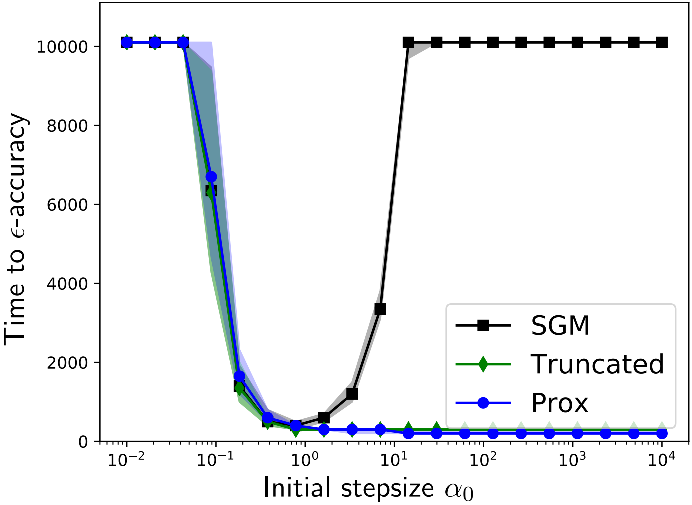
Linear regression with low noise

$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$

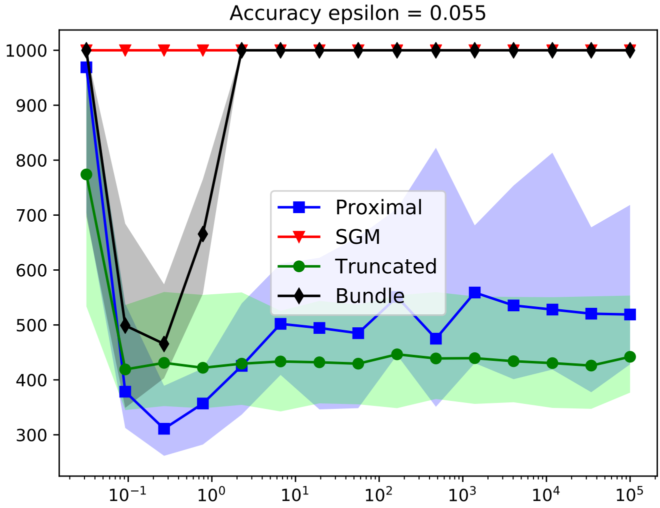


Linear regression with no noise

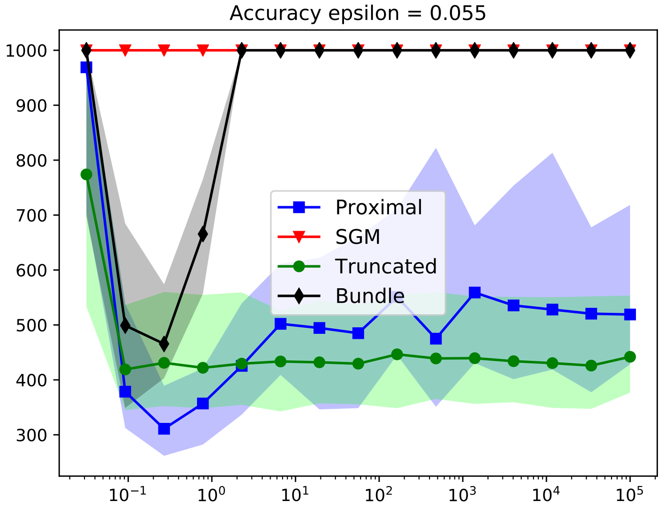
$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$



Linear regression with “poor” conditioning



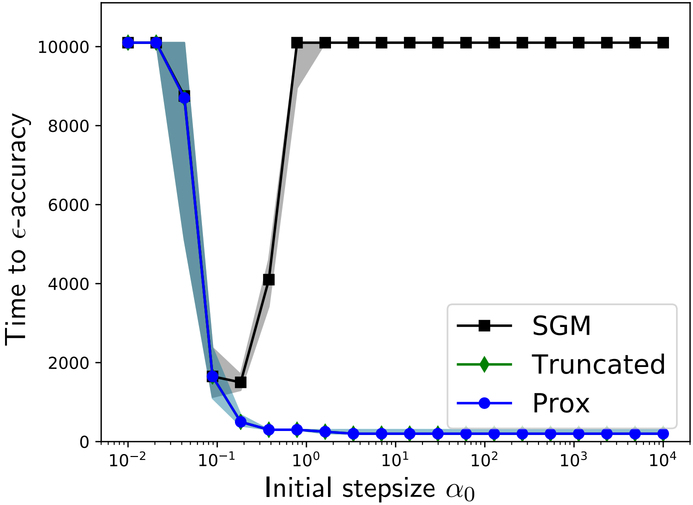
Linear regression with “poor” conditioning



Poor conditioning? $\kappa(A) = 15$

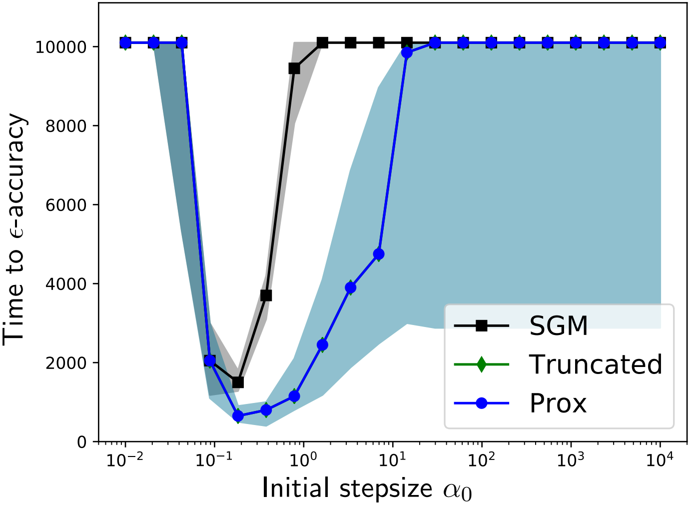
Absolute loss regression with no noise

$$F(x) = \frac{1}{m} \sum_{i=1}^{m} |a_i^T x - b_i|$$



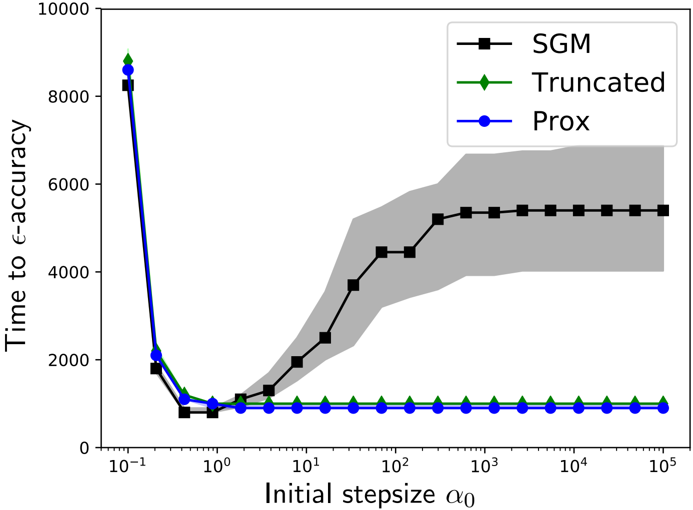
Absolute loss regression with noise

$$F(x) = \frac{1}{m}\sum_{i=1}^{m} |a_i^T x - b_i|$$



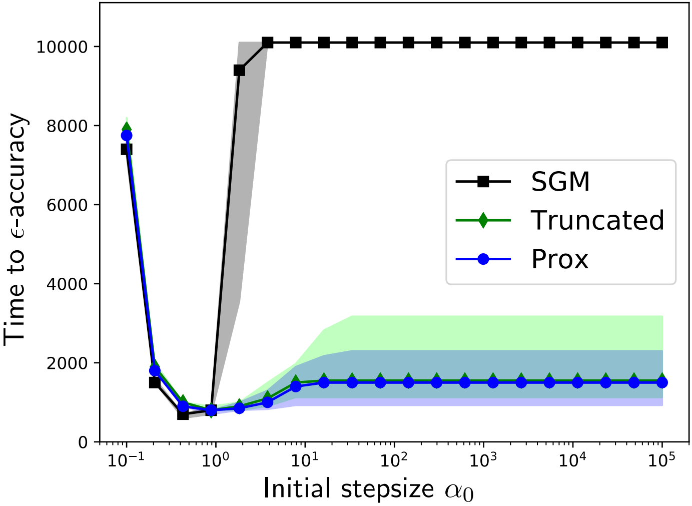
Multiclass hinge loss: no noise

$$f(x;(a,l)) = \max_{i \neq l} \left[1 + \langle a, x_i - x_l \rangle\right]_+$$



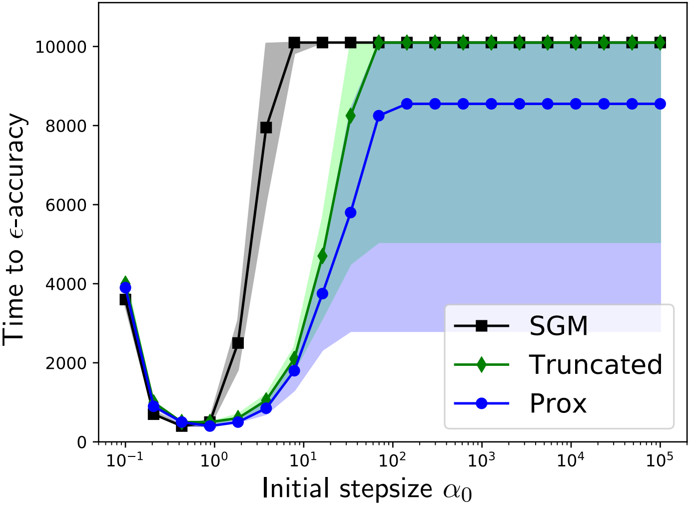
Multiclass hinge loss: small label flipping

$$f(x;(a,l)) = \max_{i \neq l} \left[1 + \langle a, x_i - x_l \rangle\right]_+$$

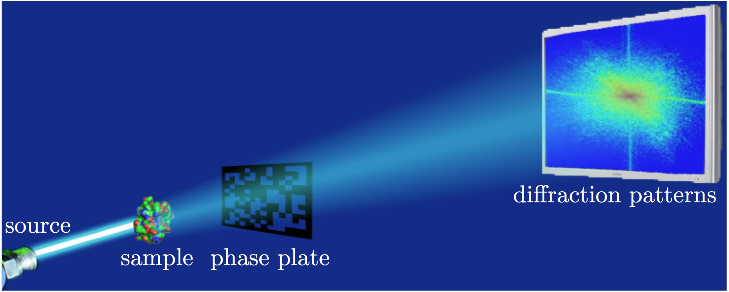


Multiclass hinge loss: substantial label flipping

$$f(x; (a, l)) = \max_{i \neq l} [1 + \langle a, x_i - x_l \rangle]_+$$

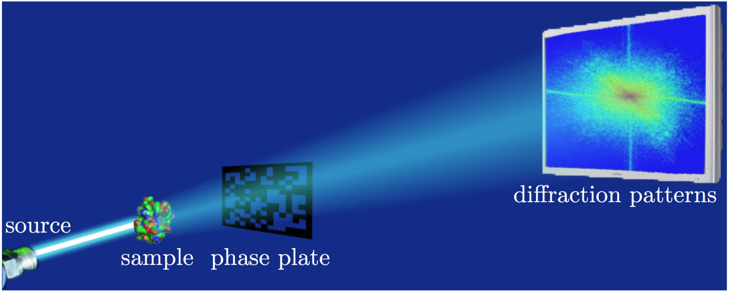


(Robust) Phase retrieval



[Candès, Li, Soltanolkotabi 15]

(Robust) Phase retrieval



[Candès, Li, Soltanolkotabi 15]

Observations (usually)

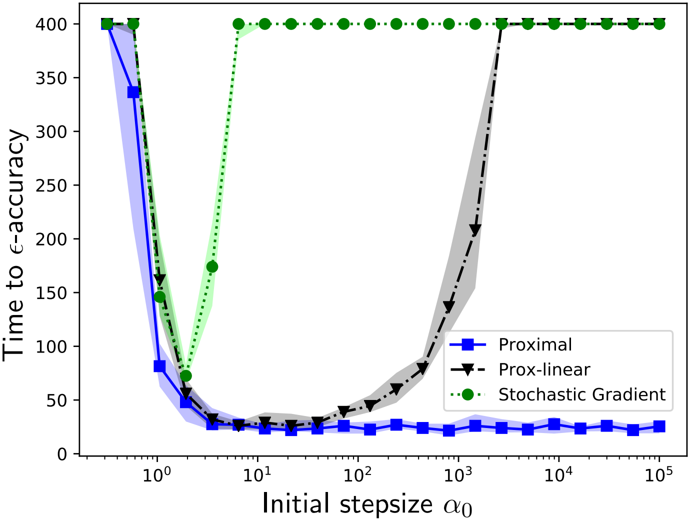
$$b_i = \langle a_i, x^\star \rangle^2$$

yield objective

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i|$$

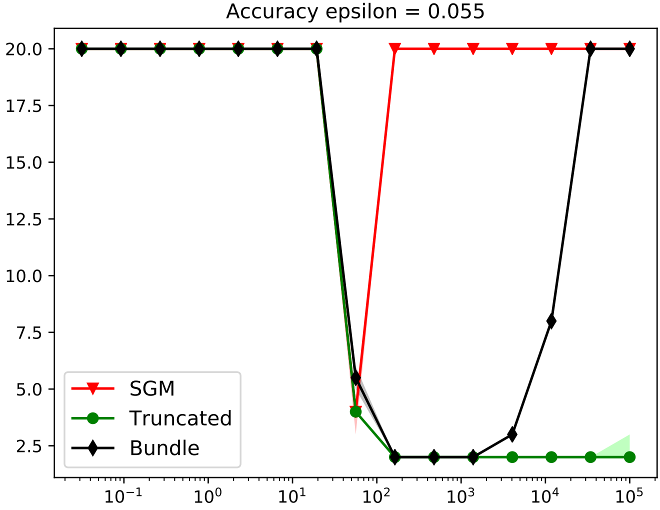
Phase retrieval without noise

$$F(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i|$$

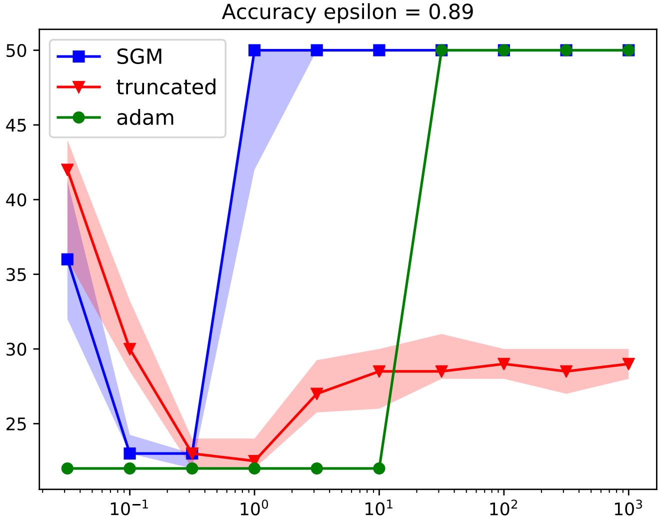


Matrix completion without noise

$$F(x, y) = \sum_{i,j \in \Omega} |\langle x_i, y_j \rangle - M_{ij}|$$

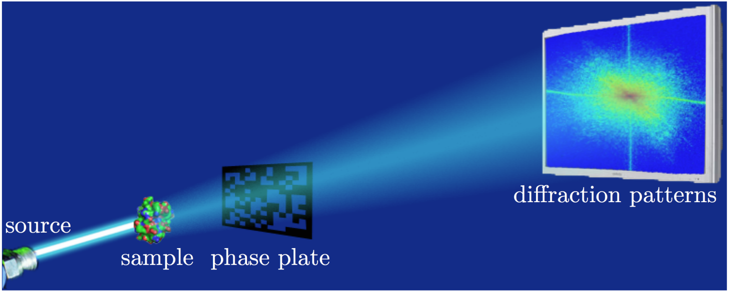


Obligatory CIFAR Experiment



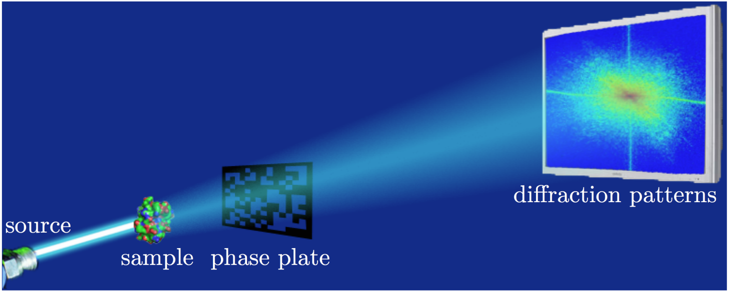
Outline

(Robust) Phase retrieval



[Candès, Li, Soltanolkotabi 15]

(Robust) Phase retrieval



[Candès, Li, Soltanolkotabi 15]

Observations (usually)

$$b_i = \langle a_i, x^\star \rangle^2$$

yield objective

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i|$$

Robust phase retrieval problems

Data model: true signal $x^\star \in \mathbb{R}^n$, noise $\xi_i = 0$ most of the time

$$b_i = \langle a_i, x^\star \rangle^2 + \xi_i$$

Robust phase retrieval problems

Data model: true signal $x^\star \in \mathbb{R}^n$, noise $\xi_i = 0$ most of the time

$$b_i = \langle a_i, x^\star \rangle^2 + \xi_i$$

Goal: solve

$$\underset{x}{\text{minimize}} \;\; f(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i|$$

Robust phase retrieval problems

Data model: true signal $x^\star \in \mathbb{R}^n$, noise $\xi_i = 0$ most of the time

$$b_i = \langle a_i, x^\star \rangle^2 + \xi_i$$

Goal: solve

$$\underset{x}{\text{minimize}} \;\; f(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i|$$

Composite problem: $f(x) = \frac{1}{m} \|\phi(Ax) - b\|_1 = h(c(x))$ where $\phi(\cdot)$ is elementwise square,

$$h(z) = \frac{1}{m} \|z\|_1 , \quad c(x) = \phi(Ax) - b$$

Composite optimization problems (other model-able structures)

The problem:

$$\underset{x}{\text{minimize}} \;\; f(x) := h(c(x))$$

where

$$h : \mathbb{R}^m \to \mathbb{R} \text{ is convex} \quad \text{and} \quad c : \mathbb{R}^n \to \mathbb{R}^m \text{ is smooth}$$

[Fletcher & Watson 80; Fletcher 82; Burke 85; Wright 87; Lewis & Wright 15; Drusvyatskiy & Lewis 16]

Modeling composite problems

Now we make a *convex* model

$$f(x) = h(c(x))$$

Modeling composite problems

Now we make a *convex* model

$$f(x) = h(\underbrace{c(x)}_{\text{linearize}})$$

Modeling composite problems

Now we make a *convex* model

$$f(y) \approx h(c(x) + \nabla c(x)^T (y - x))$$

Modeling composite problems

Now we make a *convex* model

$$f(y) \approx h(\underbrace{c(x) + \nabla c(x)^T (y - x)}_{= c(y) + O(\|x - y\|^2)})$$

Modeling composite problems

Now we make a *convex* model

$$f_x(y) := h\left(c(x) + \nabla c(x)^T (y - x)\right)$$

Modeling composite problems

Now we make a *convex* model

$$f_x(y) := h\left(c(x) + \nabla c(x)^T(y - x)\right)$$

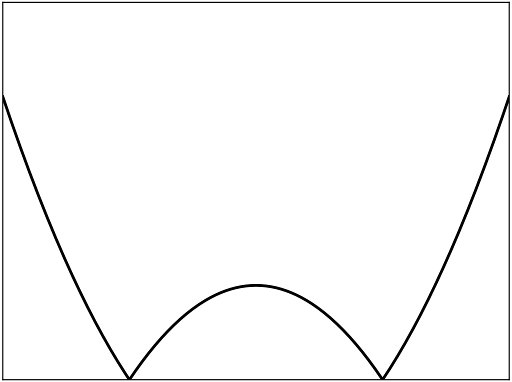
[Burke 85; Drusvyatskiy, Ioffe, Lewis 16]

Modeling composite problems

Now we make a *convex* model

$$f_x(y) := h\left(c(x) + \nabla c(x)^T (y - x)\right)$$

Example: $f(x) = |x^2 - 1|$, $h(z) = |z|$ and $c(x) = x^2 - 1$

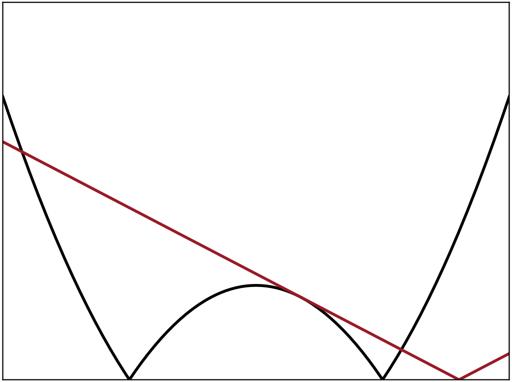


Modeling composite problems

Now we make a *convex* model

$$f_x(y) := h\left(c(x) + \nabla c(x)^T (y - x)\right)$$

Example: $f(x) = |x^2 - 1|$, $h(z) = |z|$ and $c(x) = x^2 - 1$

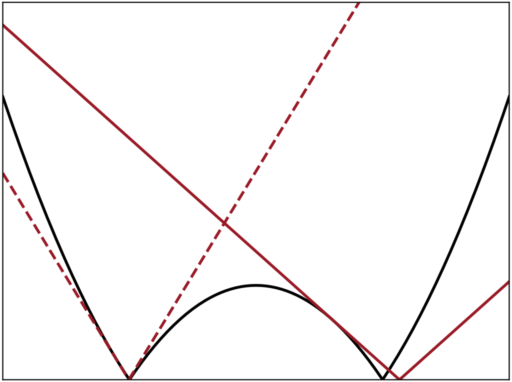


Modeling composite problems

Now we make a *convex* model

$$f_x(y) := h\left(c(x) + \nabla c(x)^T (y - x)\right)$$

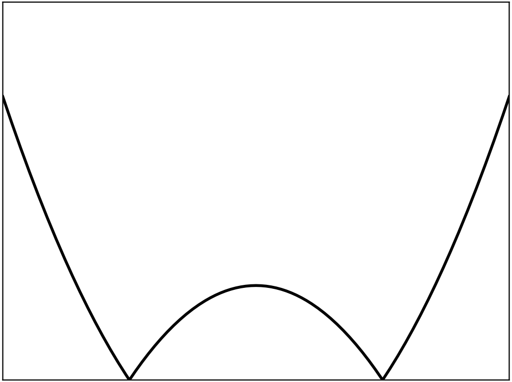
Example: $f(x) = |x^2 - 1|$, $h(z) = |z|$ and $c(x) = x^2 - 1$



Recent analysis: weakly convex case

Definition: A function F is ρ-weakly convex if for all x_0,

$$F(x) + \frac{\rho}{2}\|x - x_0\|^2 \quad \text{is convex}$$



Recent analysis: weakly convex case

Definition: A function F is ρ-weakly convex if for all x_0,

$$F(x) + \frac{\rho}{2} \|x - x_0\|^2 \quad \text{is convex}$$

Examples:

- F has $\nabla^2 F(x) \succeq -\lambda I$, then F is λ-weakly convex
- $f(x) = h(c(x))$ for h convex, M-Lipschitz and c smooth with ∇c L-Lipschitz is $L \cdot M$-weakly convex

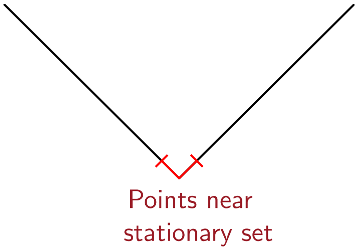
Recent analysis: weakly convex case

Definition: A function F is ρ-weakly convex if for all x_0,

$$F(x) + \frac{\rho}{2} \|x - x_0\|^2 \quad \text{is convex}$$

Typical convergence guarantee:
iterates x_k *close* to stationary points

$$X_\epsilon^\star := \{x \mid \mathrm{dist}(0, \partial f(x)) \leq \epsilon\}$$



Recent analysis: weakly convex case

Definition: A function F is ρ-weakly convex if for all x_0,

$$F(x) + \frac{\rho}{2} \|x - x_0\|^2 \quad \text{is convex}$$

Theorem (Davis & Drusvyatskiy 18, paraphrased)

Let random functions f be Lipschitz and ρ-weakly convex. Let x_k be generated by model-based method satisfying conditions,

$$X_\epsilon^\star = \{x \mid \mathrm{dist}(0, \partial F(x)) \leq \epsilon\},$$

and choose index $i^\star = i$ with probability $\alpha_i / \sum_{j=1}^k \alpha_j$. Then roughly

$$\mathbb{E}[\mathrm{dist}(x_{i^\star}, X_\epsilon^\star)^2] \lesssim \frac{1 + \sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}$$

Generalized asymptotic analysis: weakly convex case

Theorem (Asi & D., 2018)

Let F be ρ-weakly convex, and assume that

$$\mathbb{E}[\|f'(x;S)\|^2] \le C_1\|F'(x)\|^2 + C_2.$$

Let $X^\star_\epsilon = \{x \mid \mathrm{dist}(0, \partial F(x)) \le \epsilon\}$. Choose index $i^\star = i$ with probability $\alpha_i / \sum_{j=1}^k \alpha_j$. If the iterates x_k remain bounded, then with probability 1,

$$\mathbb{E}[\mathrm{dist}(x_{i^\star}, X^\star_\epsilon)^2 \mid x_1, x_2, \ldots] \lesssim \frac{1 + \sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}.$$

Generalized asymptotic analysis: weakly convex case

Theorem (Asi & D., 2018)

Let F be ρ-weakly convex, and assume that

$$\mathbb{E}[\|f'(x;S)\|^2] \le C_1\|F'(x)\|^2 + C_2.$$

Let $X_\epsilon^\star = \{x \mid \mathrm{dist}(0, \partial F(x)) \le \epsilon\}$. Choose index $i^\star = i$ with probability $\alpha_i / \sum_{j=1}^k \alpha_j$. If the iterates x_k remain bounded, then with probability 1,

$$\mathbb{E}[\mathrm{dist}(x_{i^\star}, X_\epsilon^\star)^2 \mid x_1, x_2, \ldots] \lesssim \frac{1 + \sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i}.$$

Iterates remain bounded with stochastic proximal-point-like algorithms

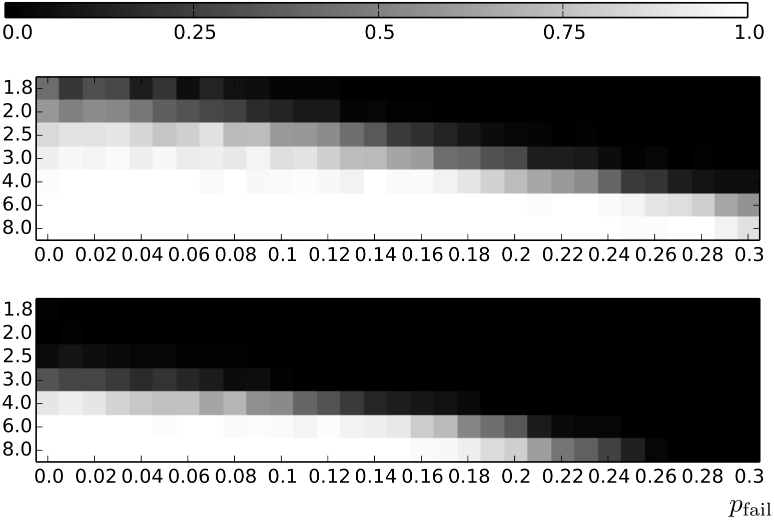
Experiment: corrupted measurements

- Data generation: dimension $n = 200$,

$$a_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_n) \quad \text{and} \quad b_i = \begin{cases} 0 & \text{w.p. } p_{\text{fail}} \\ \langle a_i, x^\star \rangle^2 & \text{otherwise} \end{cases}$$

  (most confuses our initialization method)
- Compare to Zhang, Chi, Liang's Median-Truncated Wirtinger Flow (designed specially for standard Gaussian measurements)
- Look at success probability against m/n (note that $m \geq 2n - 1$ is necessary for injectivity)

Experiment: corrupted measurements



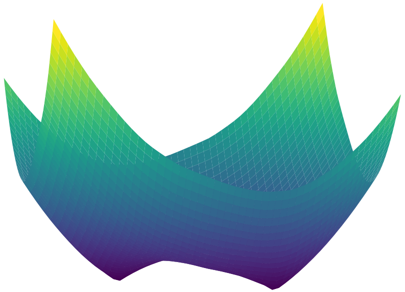
Sharp weakly convex problems

Example: Suppose that

$$b_i = \langle a_i, x^\star \rangle^2, \quad i = 1, \ldots, m.$$

Then

$$F(x) := \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - b_i| \geq F(x^\star) + \lambda \, \mathrm{dist}(x, \{-x^\star, x^\star\}).$$



Sharp weakly convex problems

Definition: An weakly convex objective F is *sharp* if

$$F(x) \geq F(x^\star) + \lambda \operatorname{dist}(x, X^\star)$$

for $X^\star = \operatorname{argmin} F(x)$ and x near $X^\star$. [Ferris 88; Burke & Ferris 95]

Theorem (Asi & D. 18)

Assume that F is weakly convex, has sharp growth, and is easy. If x_k converges to $X^\star = \operatorname{argmin}_x F(x)$ and models f_{x_k} satisfy all conditions, then

$$\limsup_k \frac{\operatorname{dist}(x_k, X^\star)}{(1-\lambda)^k} < \infty.$$

Conclusions

- Perhaps blind application of stochastic gradient methods is not the right answer
- Care and better modeling can yield improved performance
- Computational efficiency important in model choice

Conclusions

- Perhaps blind application of stochastic gradient methods is not the right answer
- Care and better modeling can yield improved performance
- Computational efficiency important in model choice

Questions

- More satisfying adaptation results?
- Parallelism?