# The importance of better models in stochastic optimization

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

Les Houches 2019

# Outline

Motivating experiments

Models in optimization

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

Revisiting experimental results

# Outline

# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.

How much ENERGY spent in this paper?

# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.

How much ENERGY spent in this paper?

|||

How many Toyota Camrys from SF to LA?

# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.

1?

How much ENERGY spent in this paper?

|||

How many Toyota Camrys from SF to LA?

# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.

10?

How much ENERGY spent in this paper?

|||

How many Toyota Camrys from SF to LA?

# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.

100?

How much ENERGY spent in this paper?

|||

How many Toyota Camrys from SF to LA?

# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.

1000?

How much ENERGY spent in this paper?

|||

How many Toyota Camrys from SF to LA?

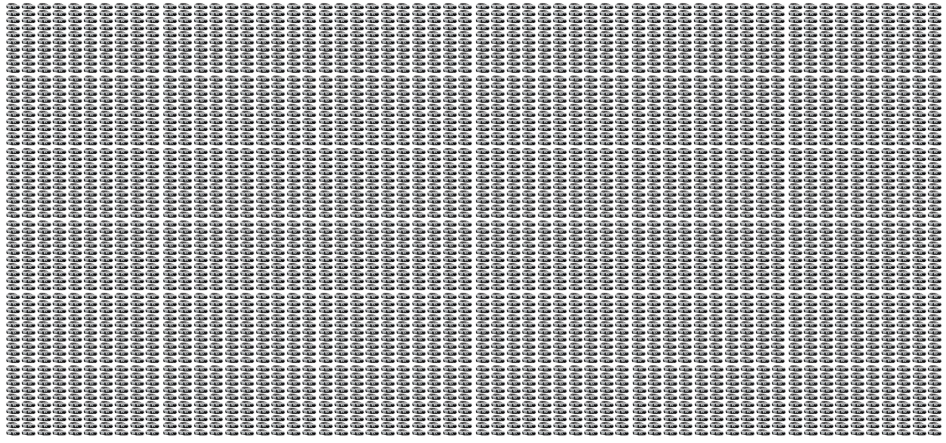
# Why robustness is important

## CAPACITY AND TRAINABILITY IN RECURRENT NEURAL NETWORKS

**Jasmine Collins,* Jascha Sohl-Dickstein & David Sussillo**
Google Brain
Google Inc.
Mountain View, CA 94043, USA
{jlcollins, jaschasd, sussillo}@google.com

ABSTRACT

Two potential bottlenecks on the expressiveness of recurrent neural networks (RNNs) are their ability to store information about the task in their parameters, and to store information about the input history in their units. We show experimentally that all common RNN architectures achieve nearly the same per-task and per-unit capacity bounds with careful training, for a variety of tasks and stacking depths.



**4200**

How much ENERGY spent in this paper?

|||

How many Toyota Camrys from SF to LA?

# 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:**

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

**Weakly convex functions:** for each $s$, some $\rho(s)$ such that

$$f(x;s) + \frac{\rho(s)}{2}\|x\|_2^2$$

convex in $x$

# Stochastic gradient methods

**The problem in this talk:**

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

**Weakly convex functions:** for each $s$, some $\rho(s)$ such that

$$f(x;s) + \frac{\rho(s)}{2}\|x\|_2^2$$

convex in $x$

- add a big enough quadratic, it becomes convex

# 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:**

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

$$\text{subject to } x \in X$$

**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:**

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

**Stochastic gradient method:**

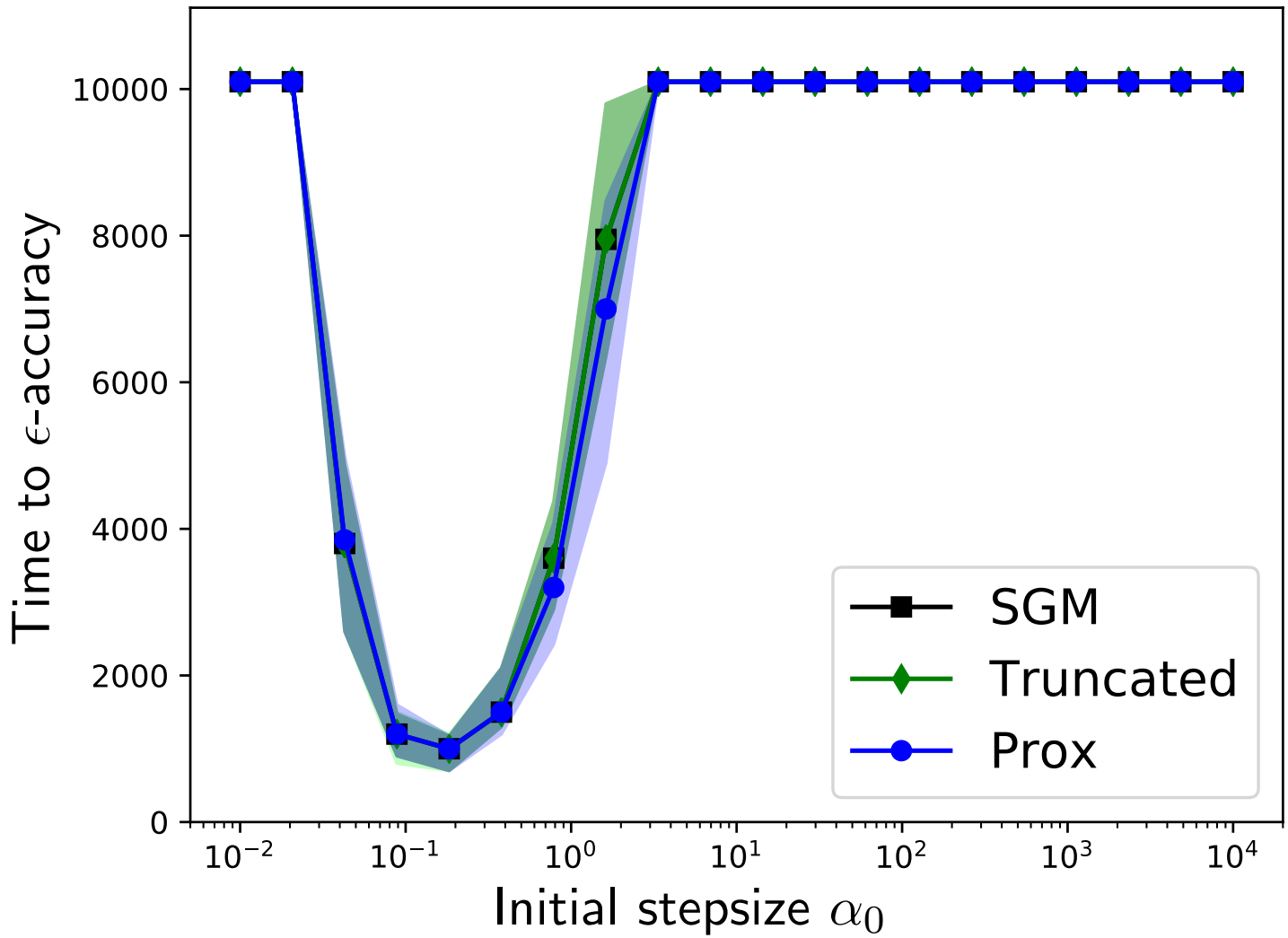
$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k; S_k)$$

**Why we use this?**

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

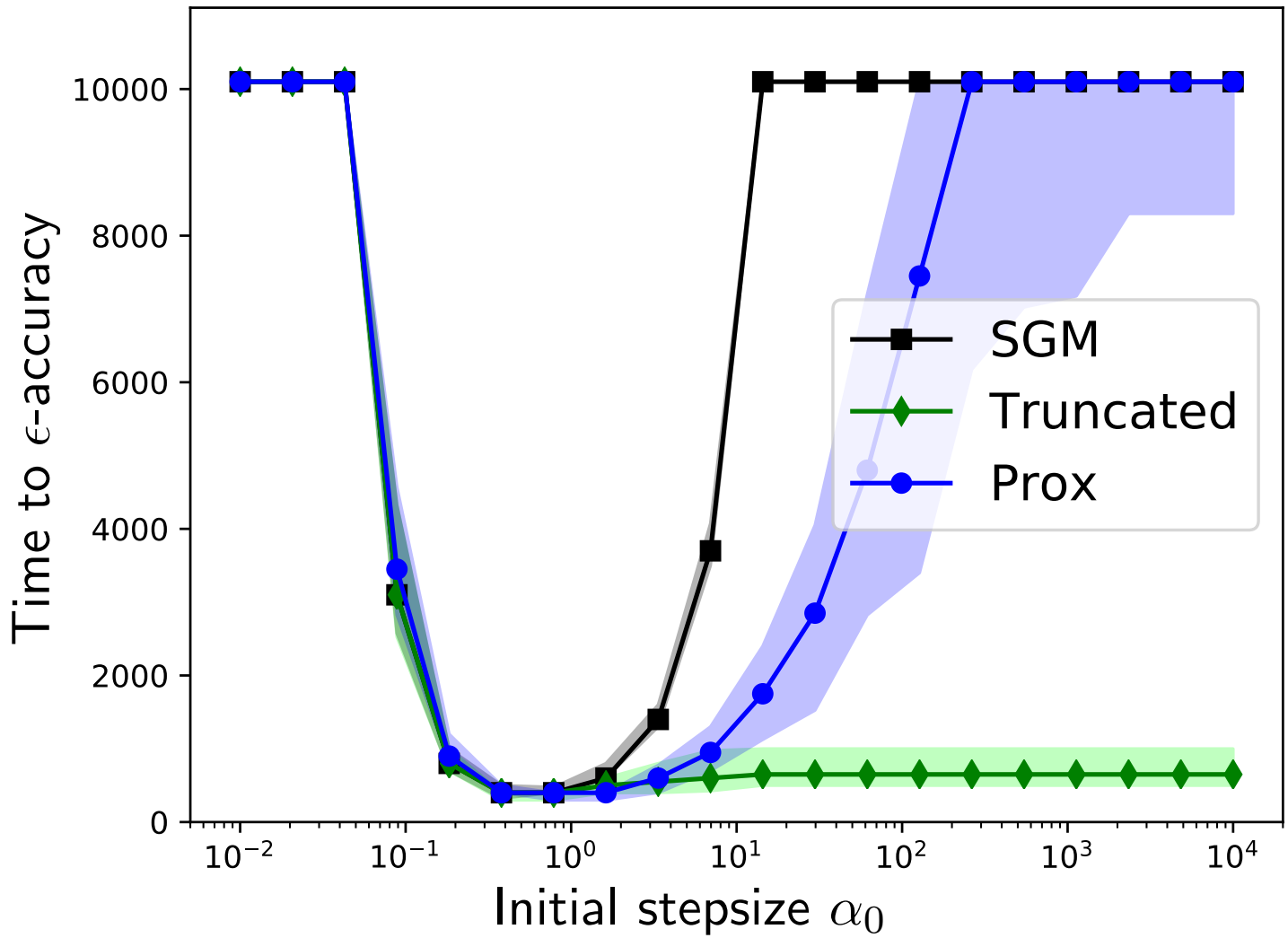
# Linear regression

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



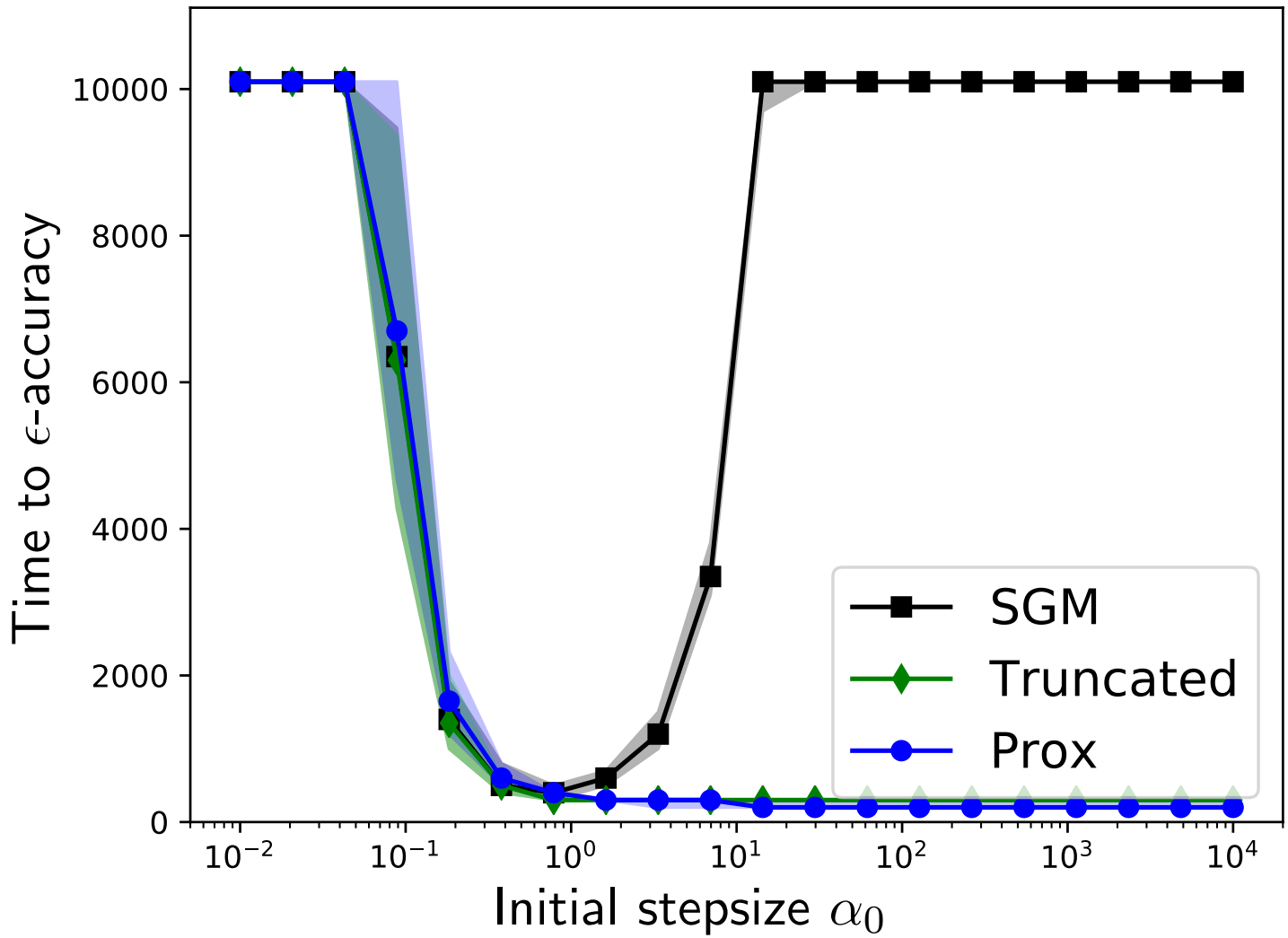
# Linear regression

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



# Linear regression

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



# Outline

# Optimization methods

**How do we solve optimization problems?**

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

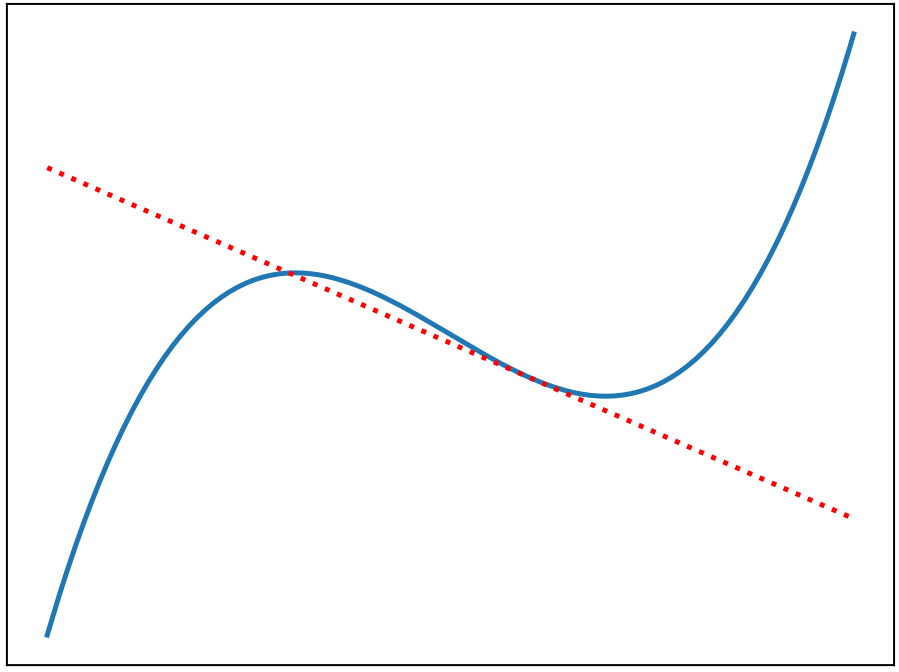
# Optimization methods

**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

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



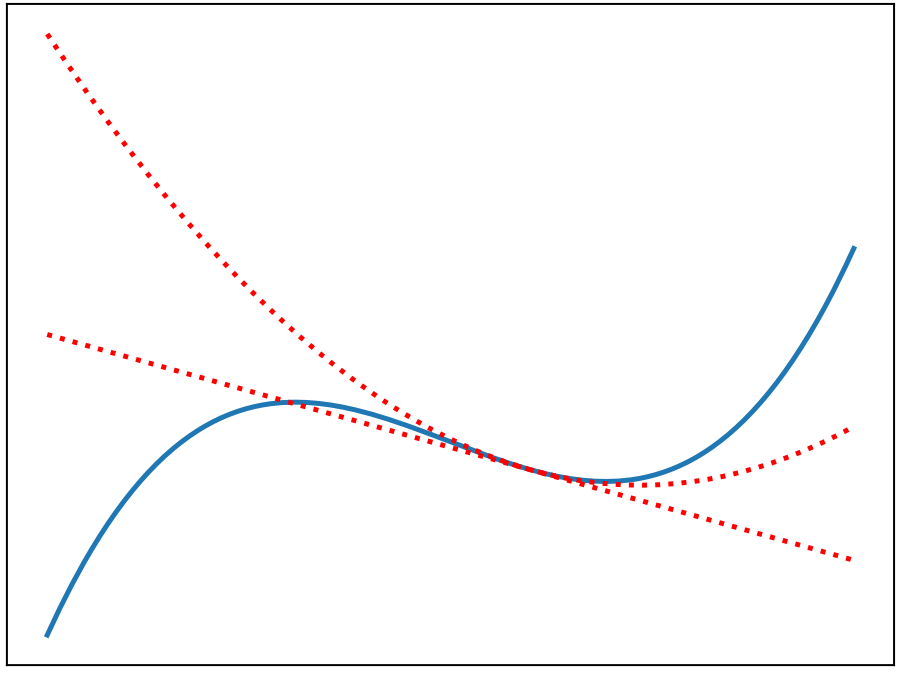
# Optimization methods

**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

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



# 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

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

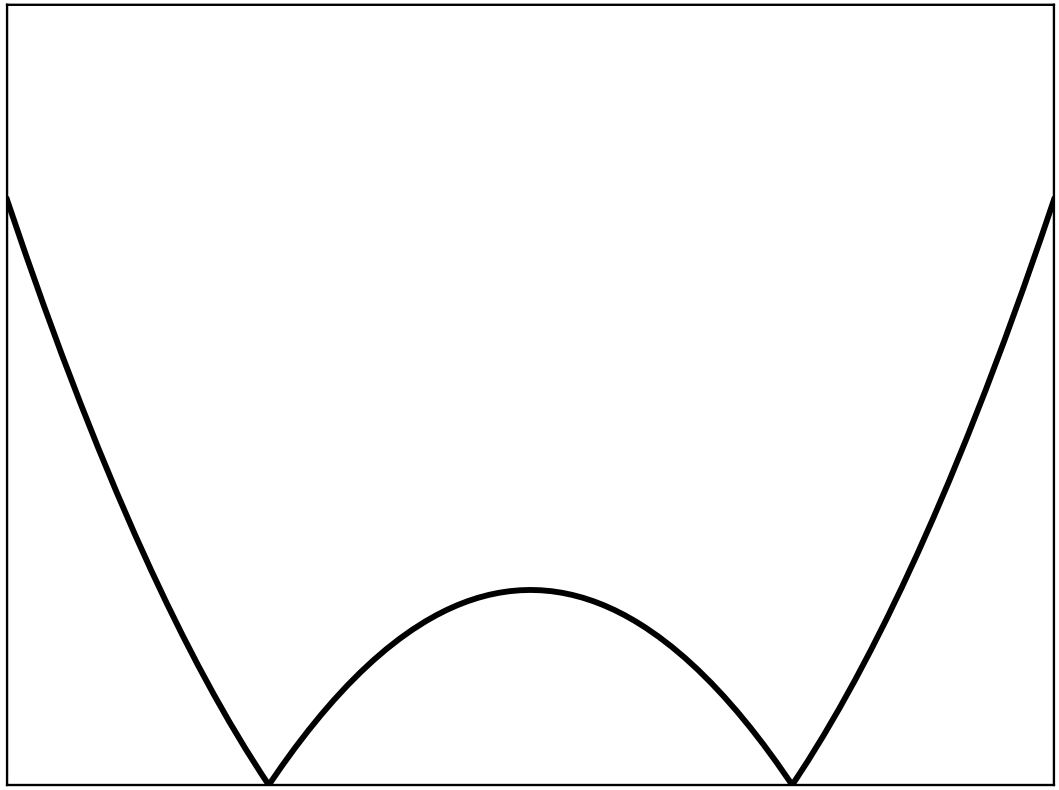
[Burke 85; Drusvyatskiy, Ioffe, Lewis 16]

# 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 $c(x) = x^2 - 1$

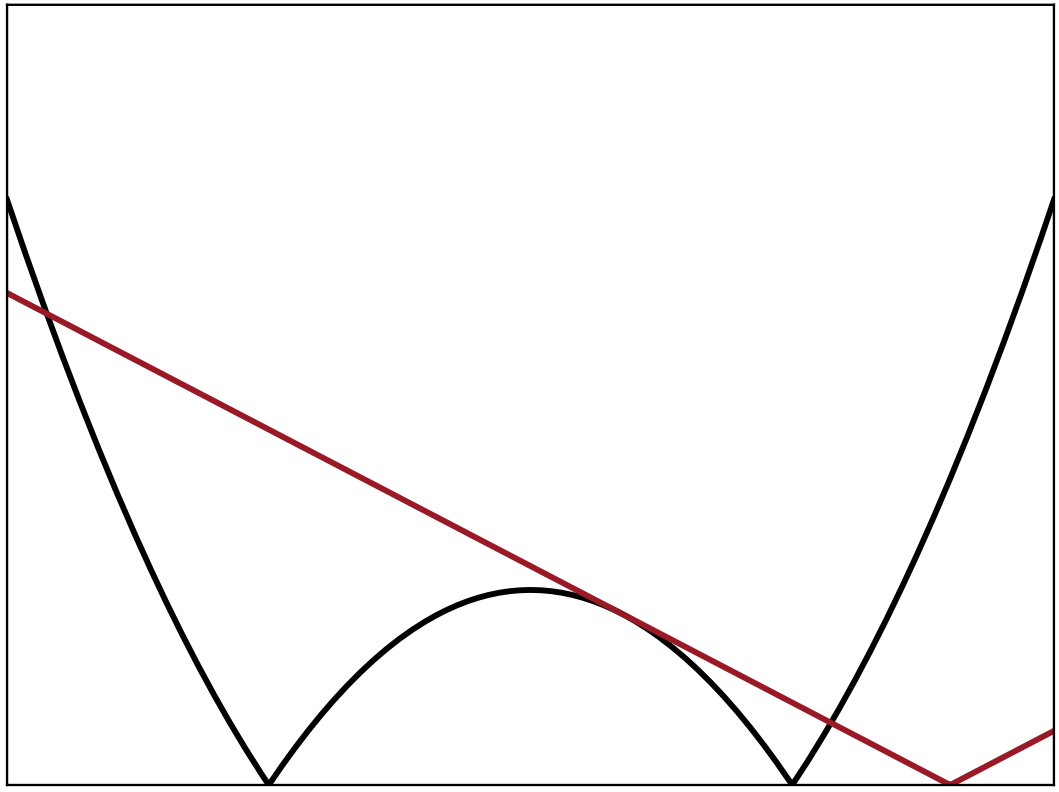


# Modeling composite problems

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 $c(x) = x^2 - 1$

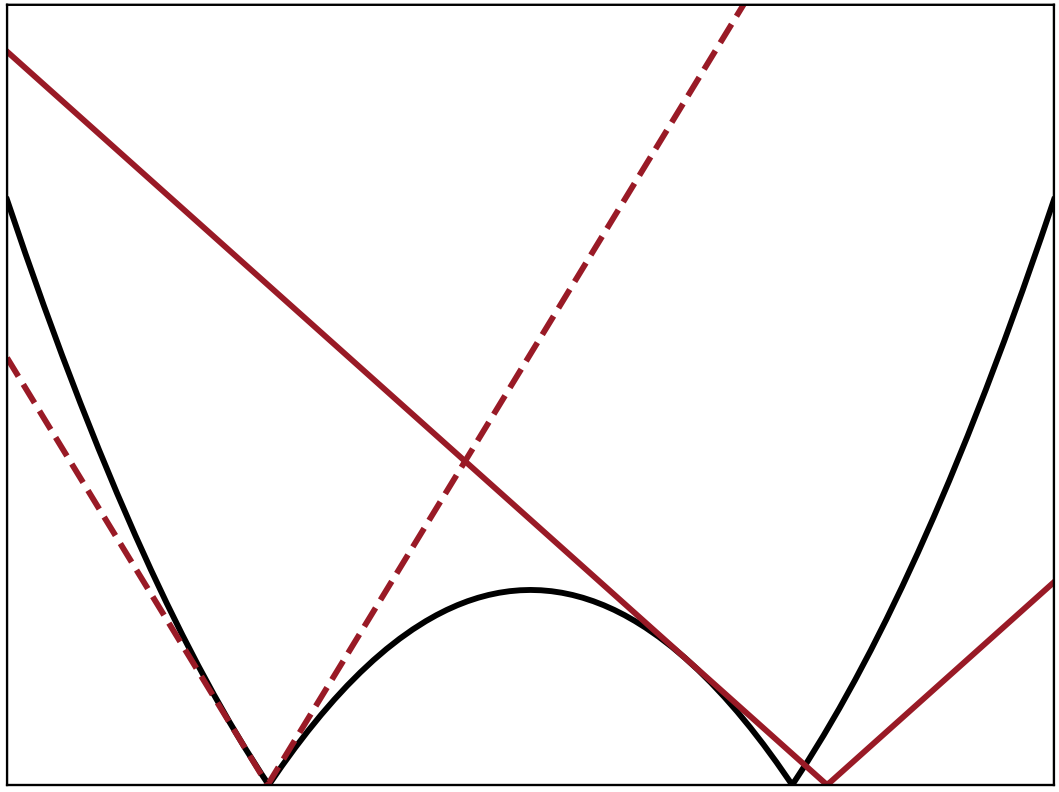


# Modeling composite problems

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 $c(x) = x^2 - 1$



# The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

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

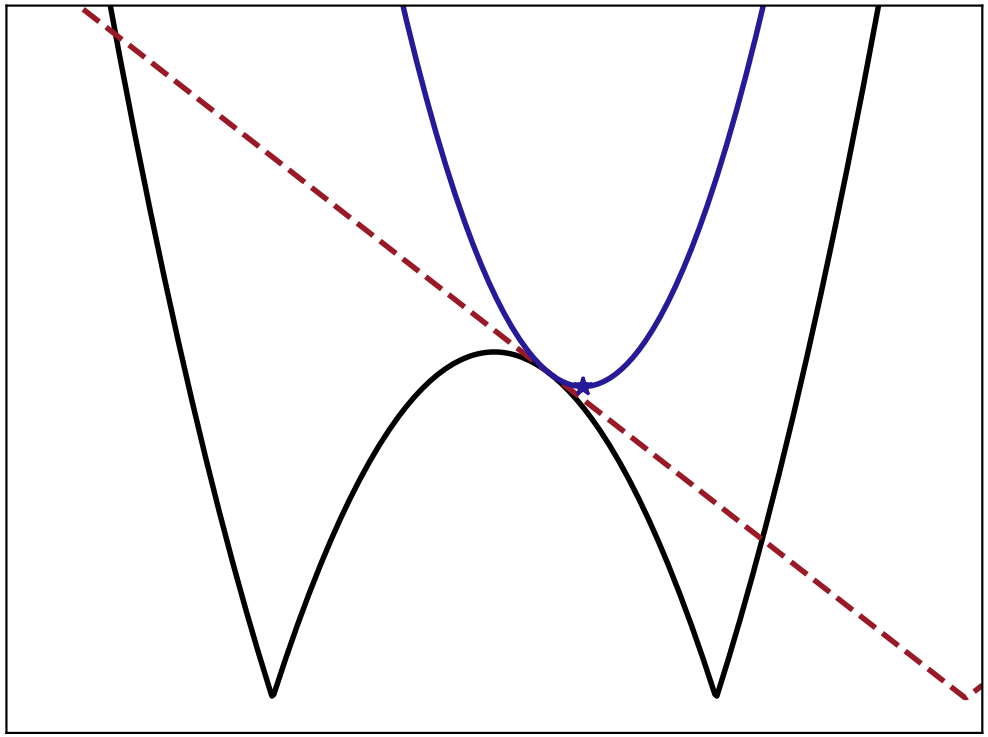
$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

# The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

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

$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

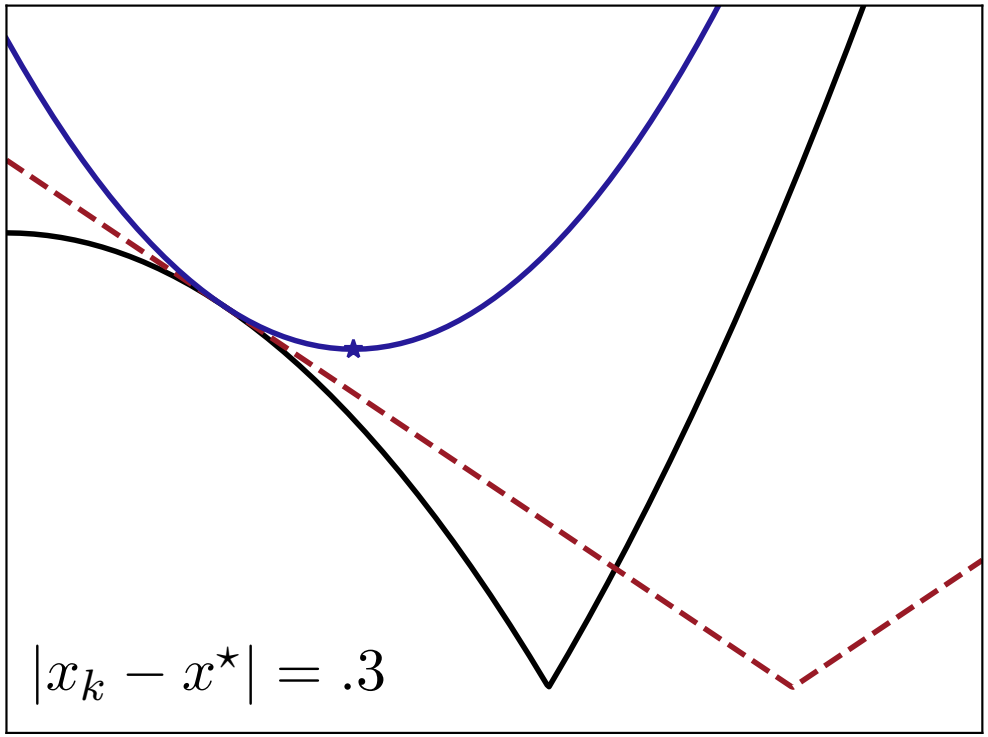


# The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

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

$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

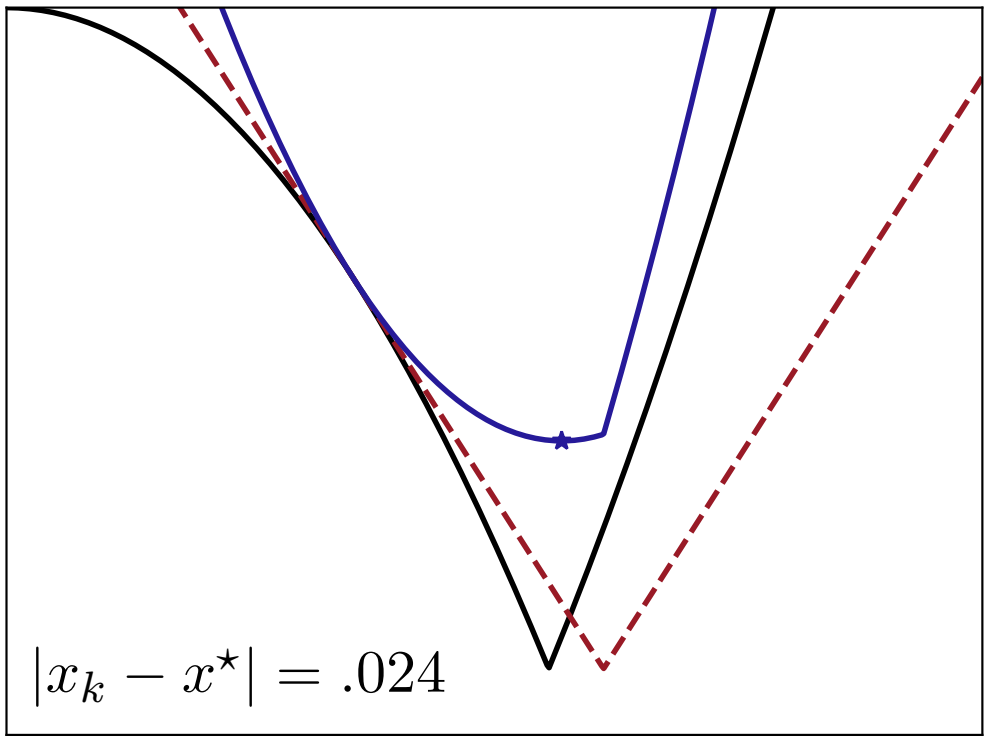


# The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

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

$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

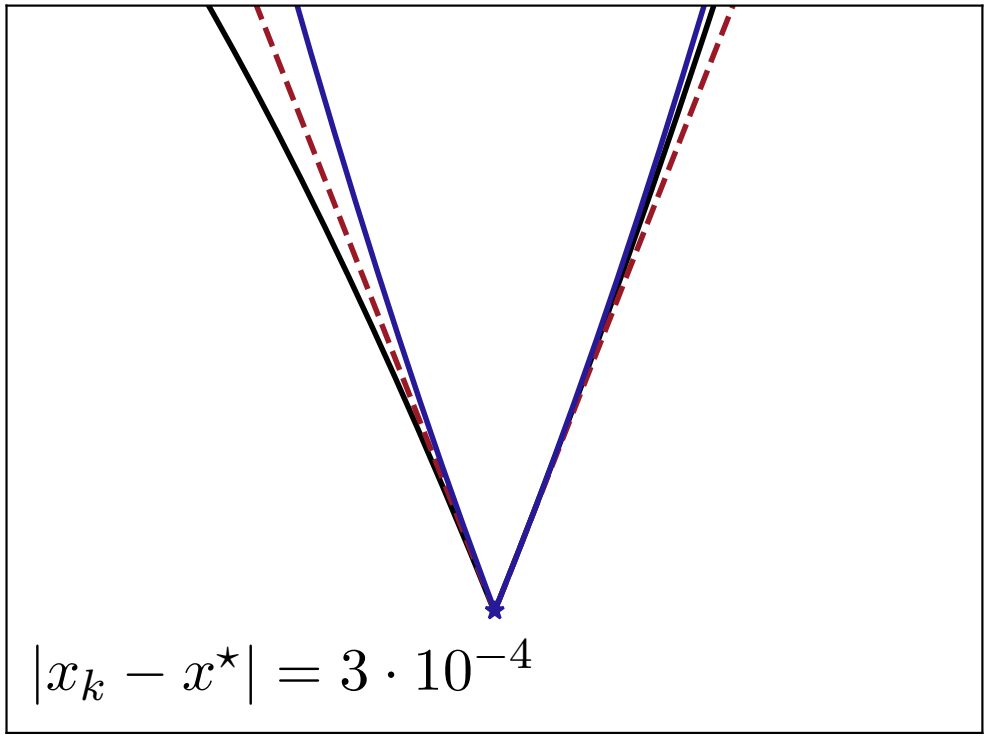


# The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

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

$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

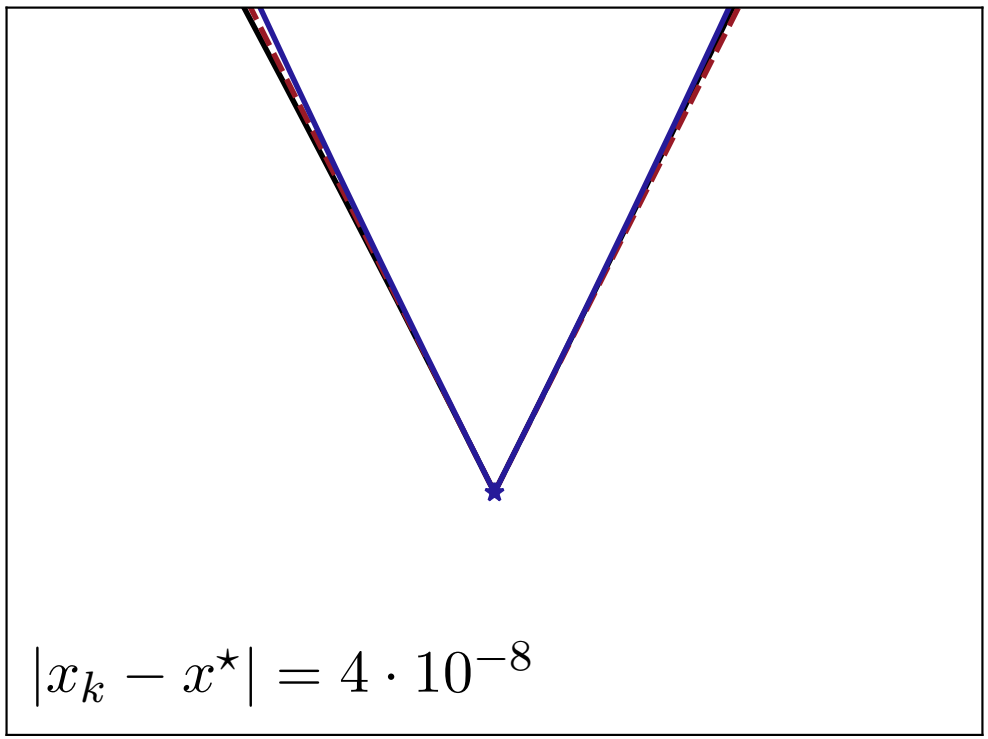


# The prox-linear method [Burke, Drusvyatskiy et al.]

Iteratively (1) form regularized convex model and (2) minimize it

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

$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$



# Generic(ish) optimization methods

Iterate

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

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

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

# 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; Davis & Drusvyatskiy 18; Asi & D. 19]

# Models in stochastic optimization

Conditions on our models (convex case)

i. Convex model:

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

ii. Lower bound:

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

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]

# Models in stochastic optimization

Conditions on our models ($\rho$-weakly convex case)

i. Convex model:

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

ii. Lower bound:

$$f_x(y; s) \leq f(y; s) + \frac{\rho(s)}{2} \|x - y\|_2^2$$

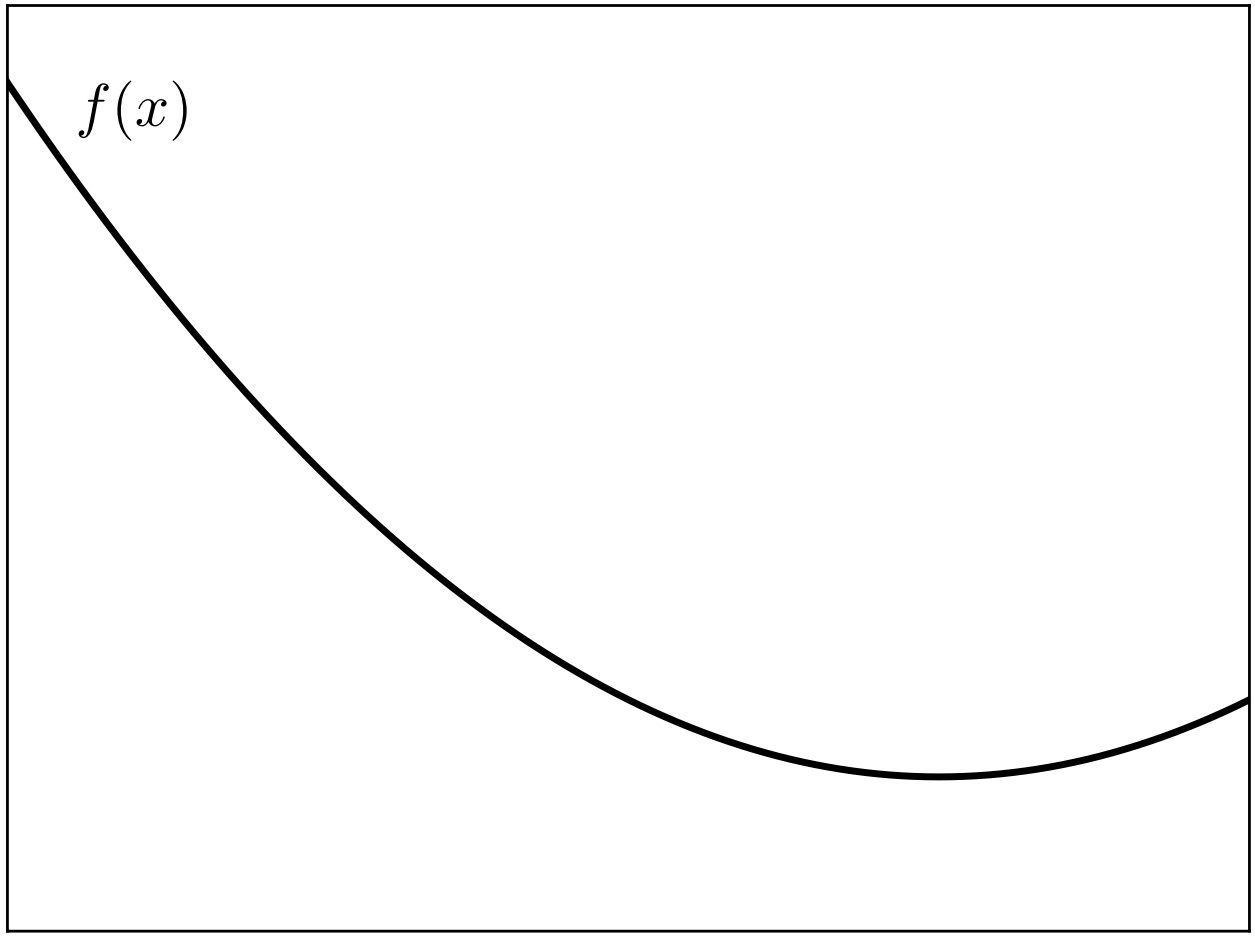
iii. Local correctness:

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

[D. & Ruan 17; Davis & Drusvyatskiy 18; Asi & D. 19]

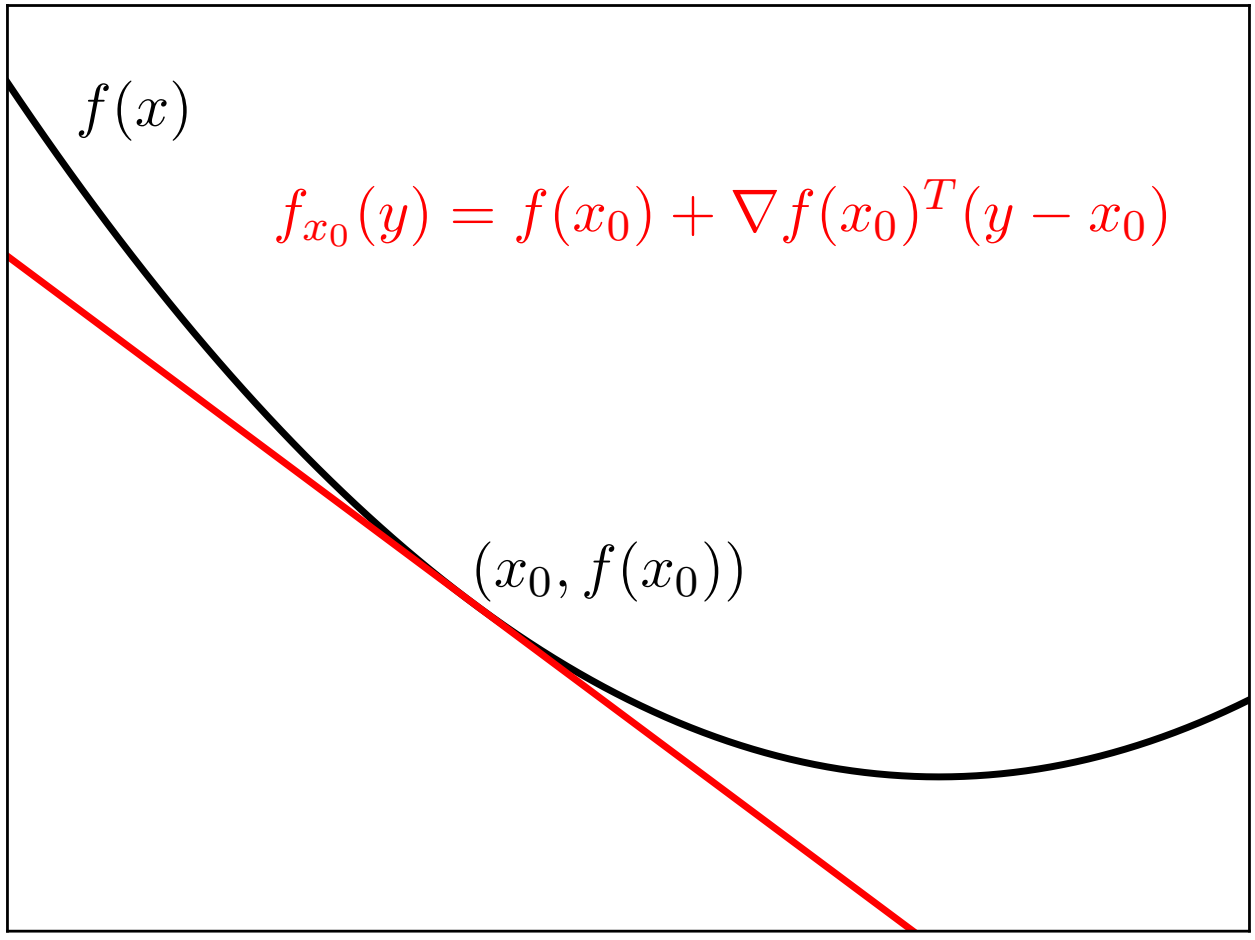
# Modeling conditions

Model $f_x(y)$ of $f$ near $x$



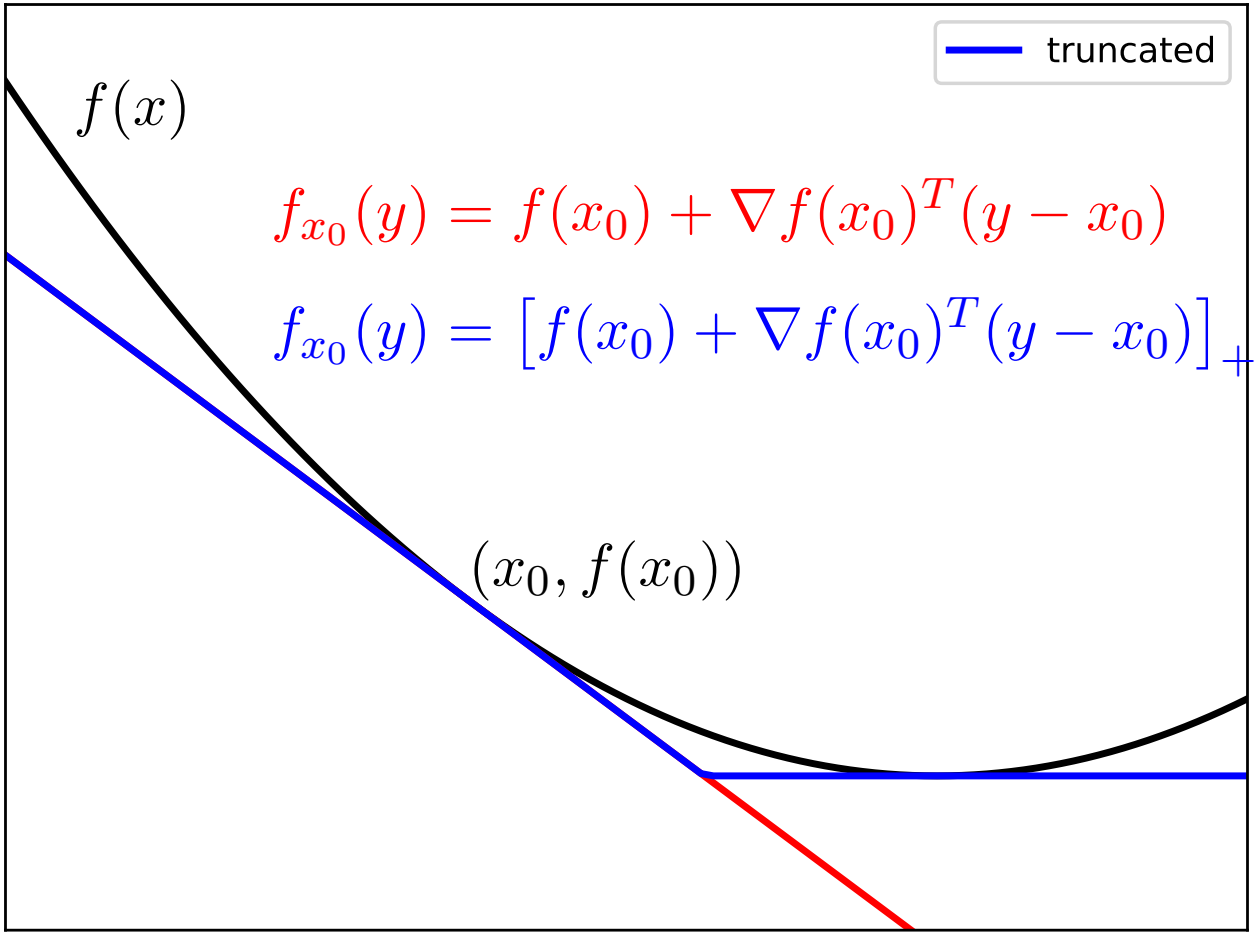
# Modeling conditions

Model $f_x(y)$ of $f$ near $x$

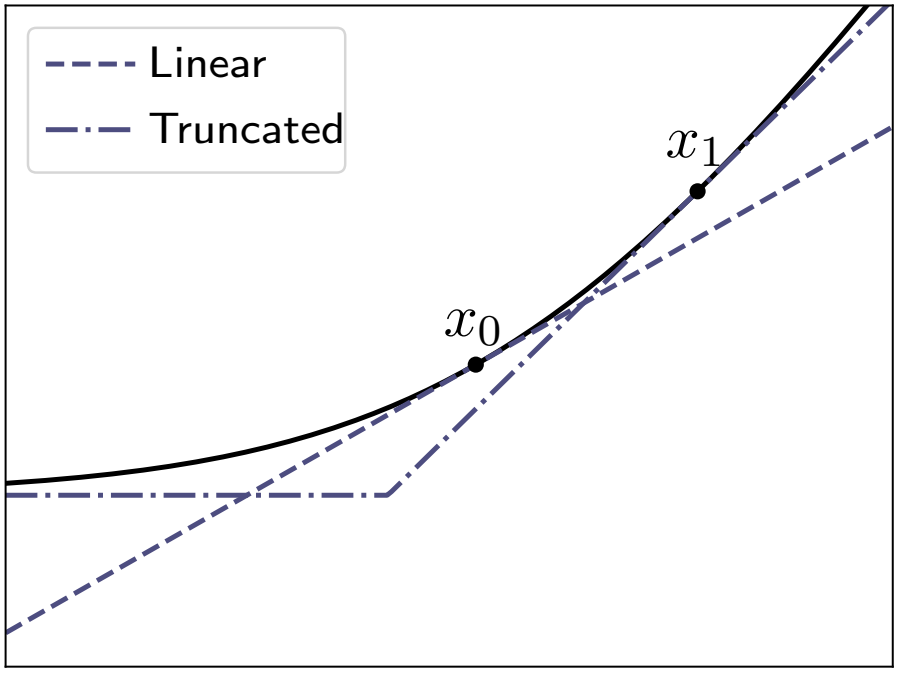


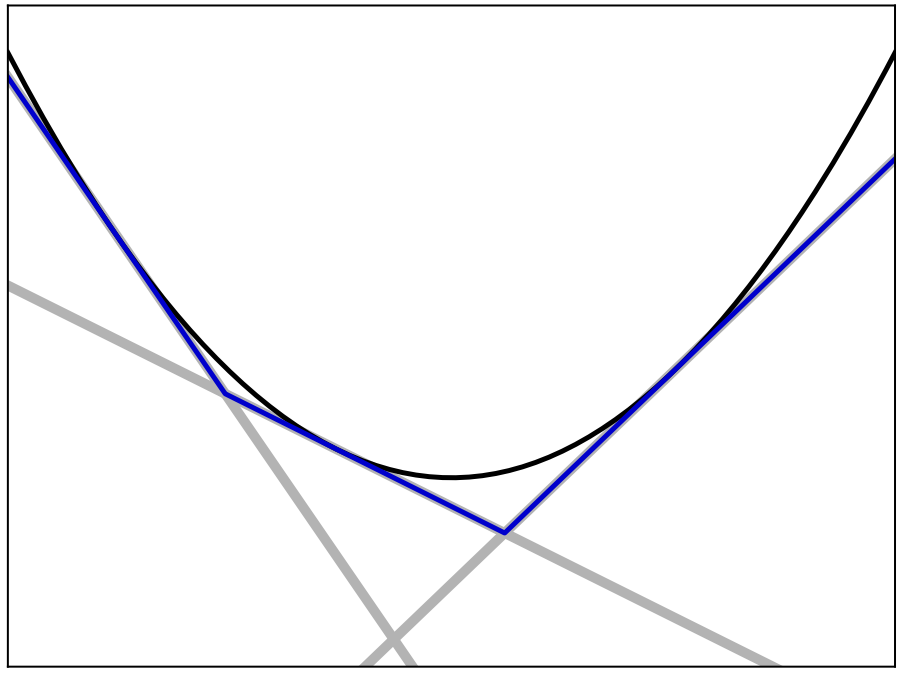
# Modeling conditions

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: $f_x(y) = \max\{f(x_i) + \langle f'(x_i), x - x_i \rangle\}$

iv. Prox-linear: $f_x(y) = h(c(x) + \nabla c(x)^T (y - x))$

# The aProx family

Iterate:

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

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

# Outline

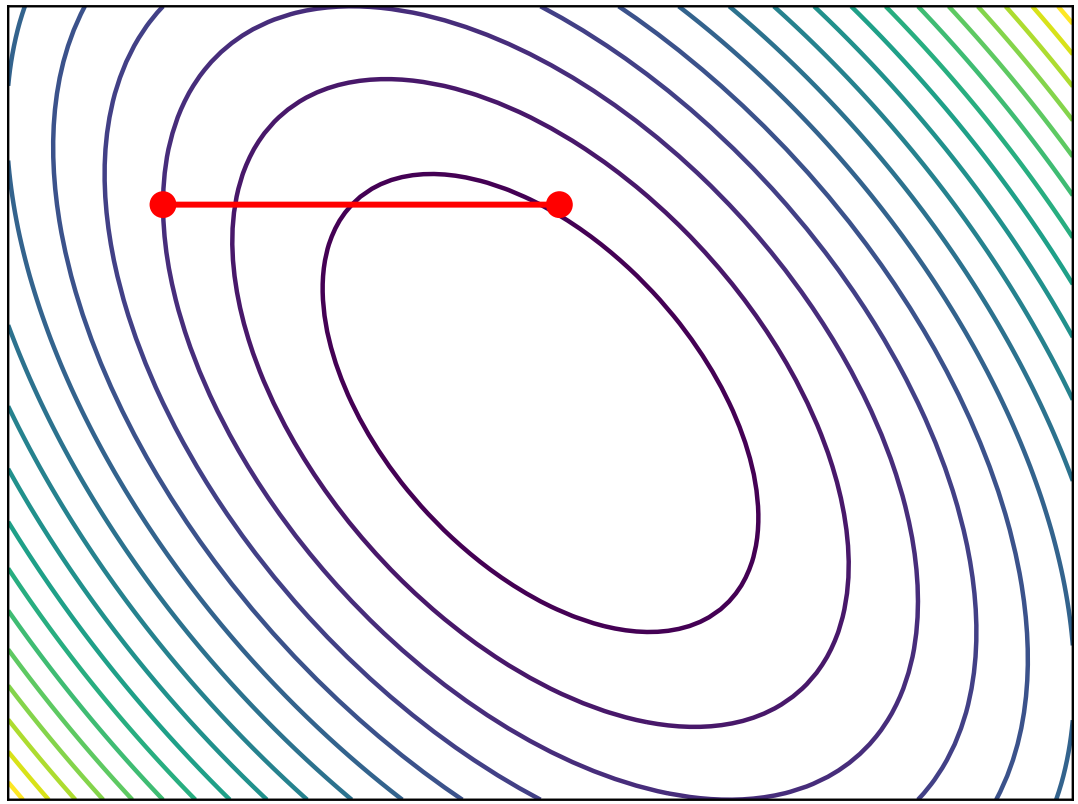
# The aProx family

Iterate:

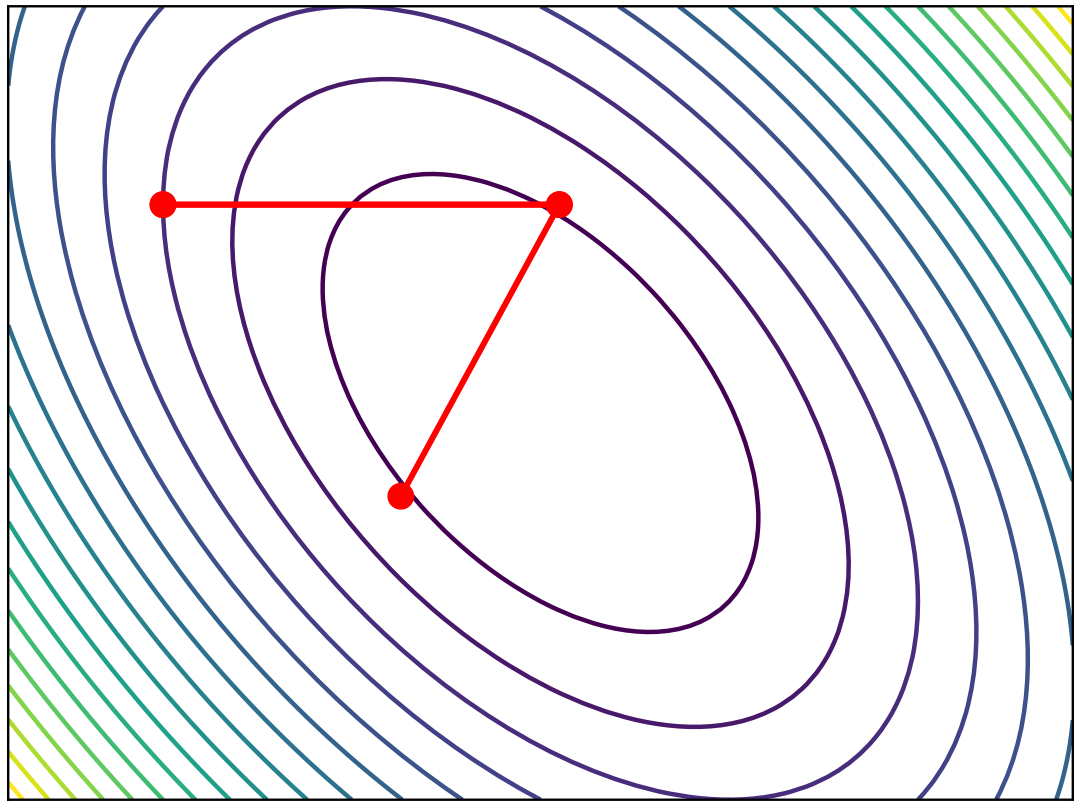
- Sample $S_k \overset{\text{iid}}{\sim} P$
- Update by minimizing model

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

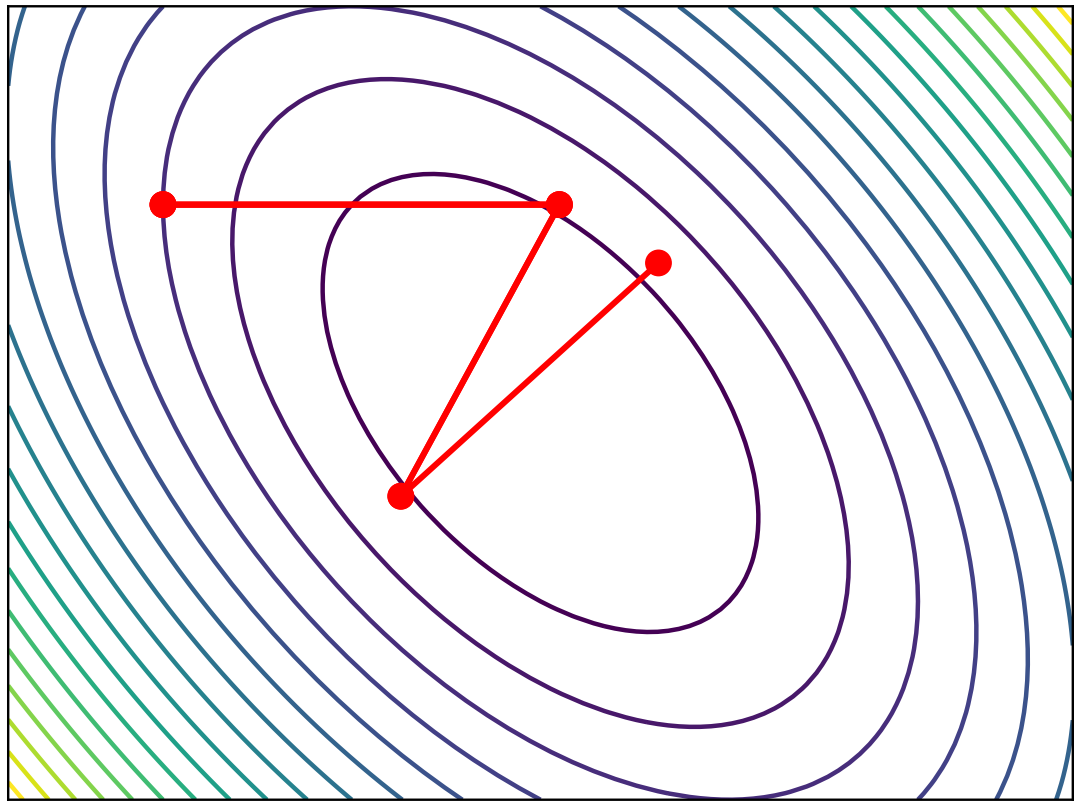
# Divergence of a gradient method



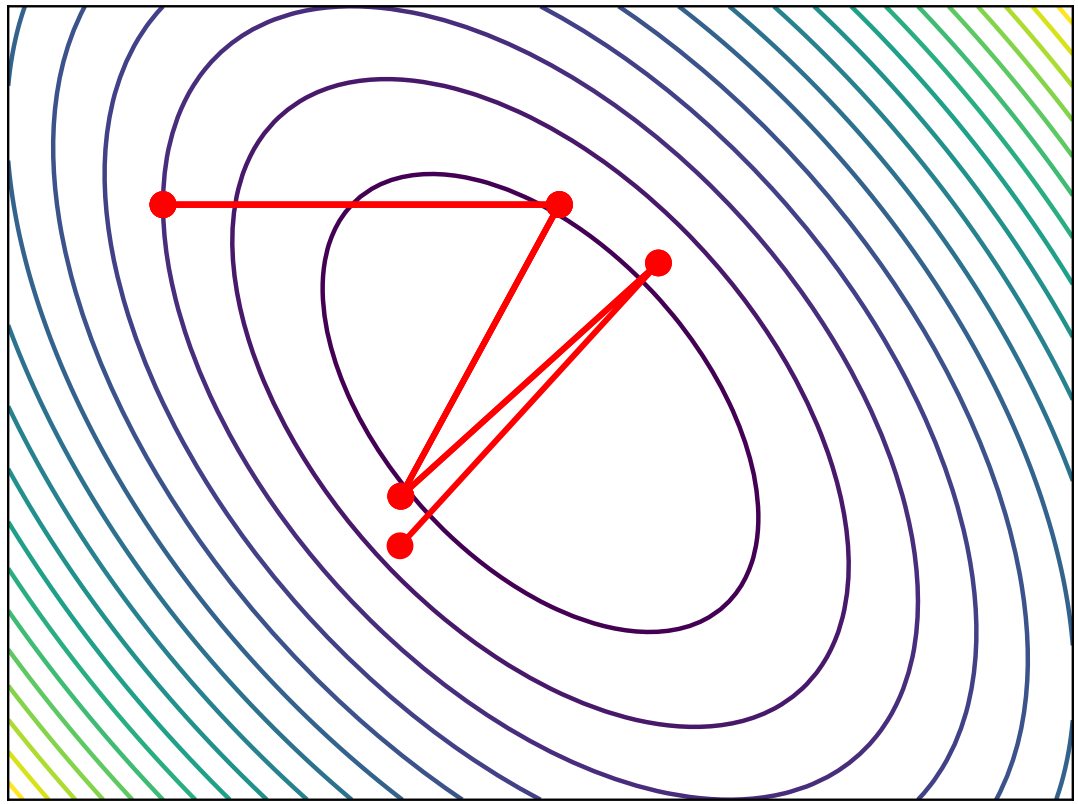
# Divergence of a gradient method



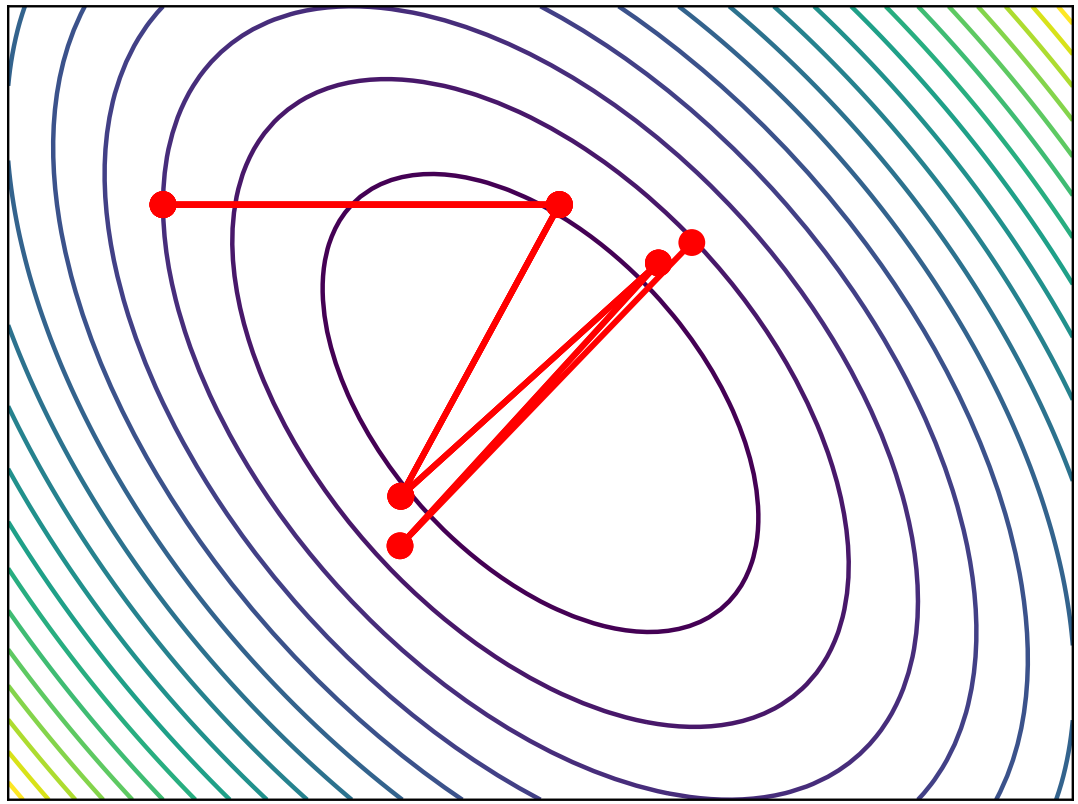
# Divergence of a gradient method



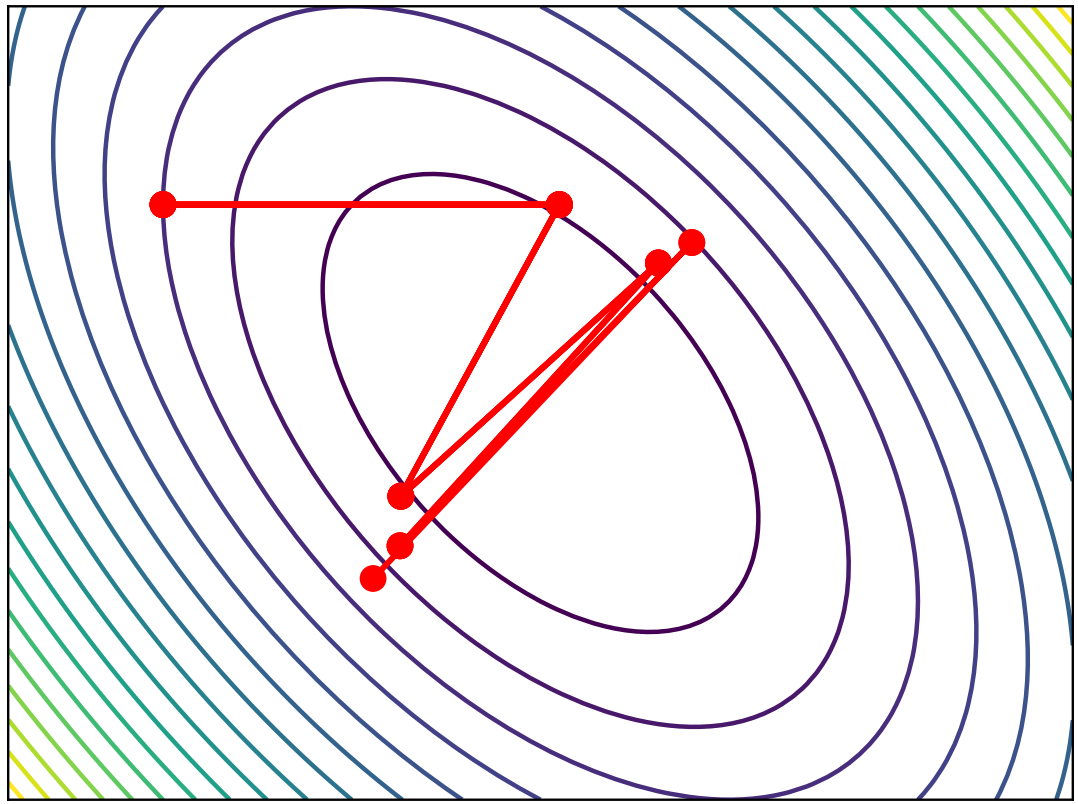
# Divergence of a gradient method



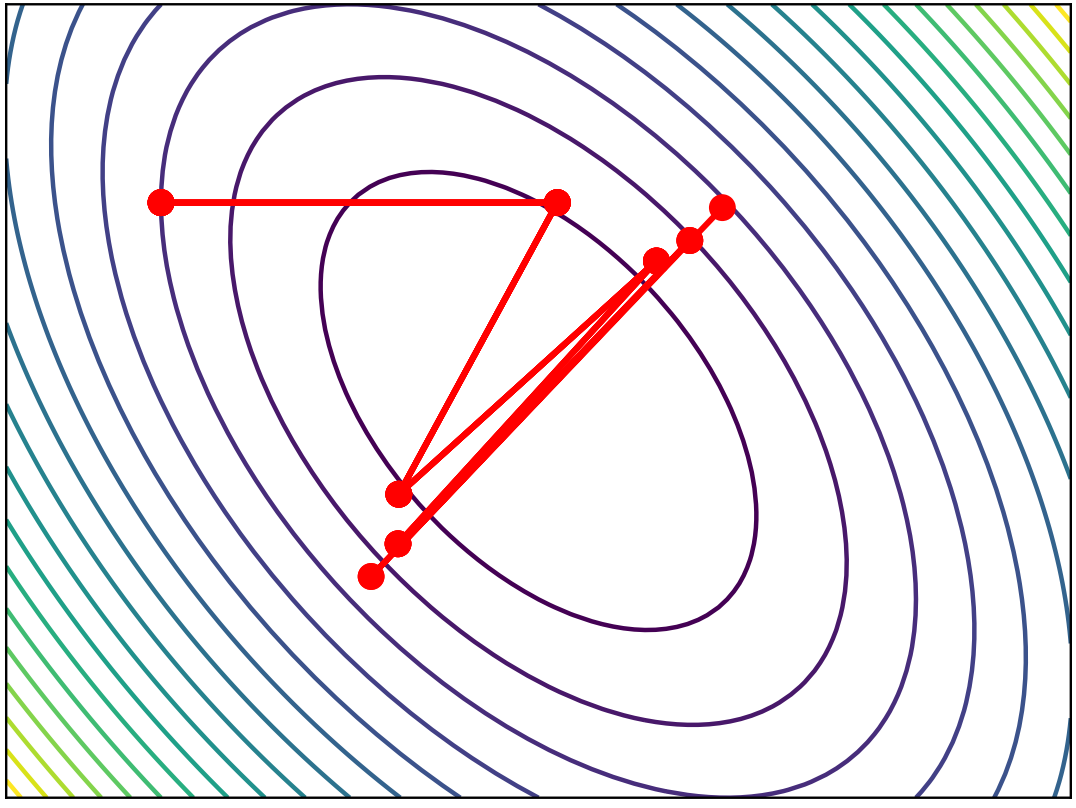
# Divergence of a gradient method



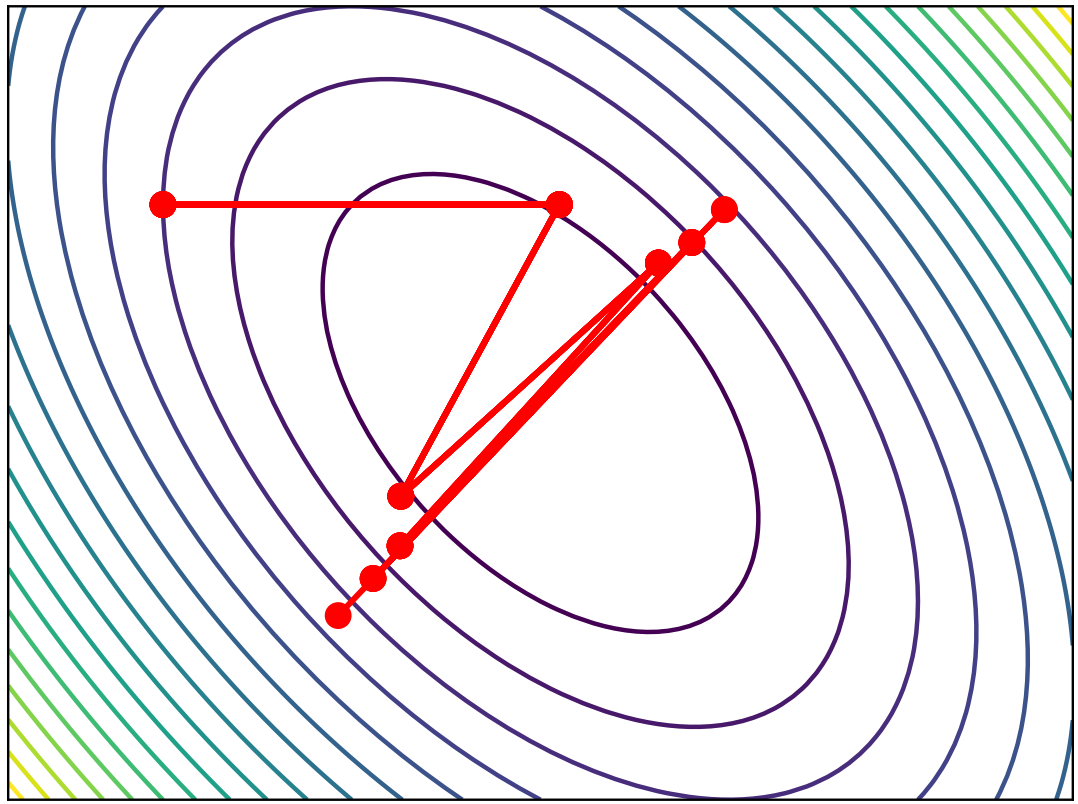
# Divergence of a gradient method



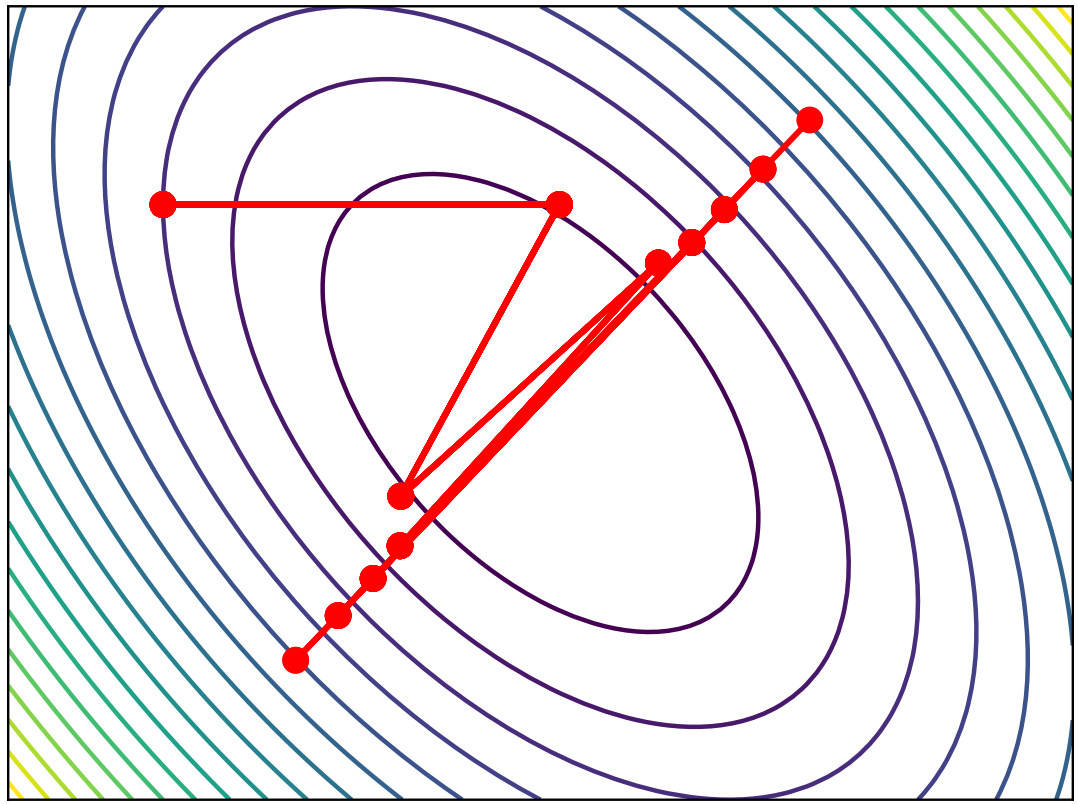
# Divergence of a gradient method



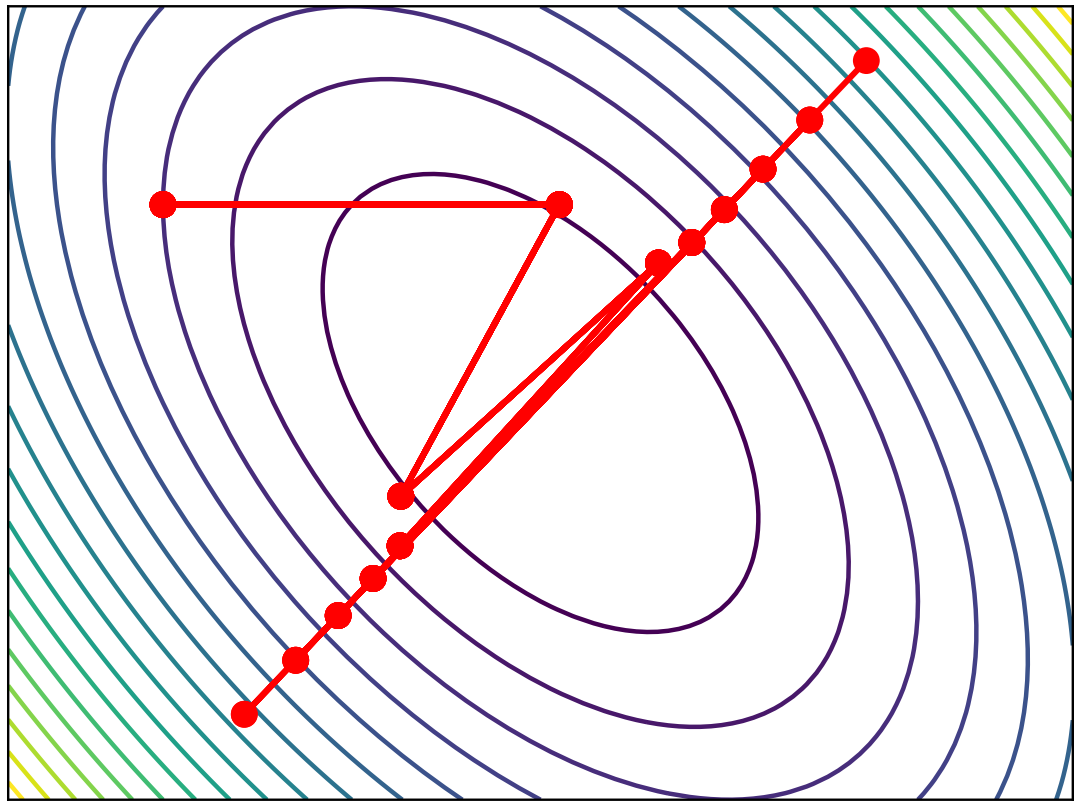
# Divergence of a gradient method



# Divergence of a gradient method



# Divergence of a gradient method



# Stability guarantees (convex)

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* $\mathbb{E}[\|f'(x^\star; S)\|^2] \leq \sigma^2$. *Then*

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

# Stability guarantees (convex)

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* $\mathbb{E}[\|f'(x^\star; S)\|^2] \leq \sigma^2$. *Then*

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

## Theorem (Asi & D. 18)

*Under the same assumptions,*

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

# Stability guarantees (convex)

Use any model with $f_x(y;s) \geq \inf_z f(z;s)$ (i.e. good lower bound)

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

## Theorem (Asi & D. 19)

*Assume $\mathcal{X}^\star = \operatorname{argmin}_{x \in \mathcal{X}} F(x)$ is non-empty and there exists $p < \infty$ such that*

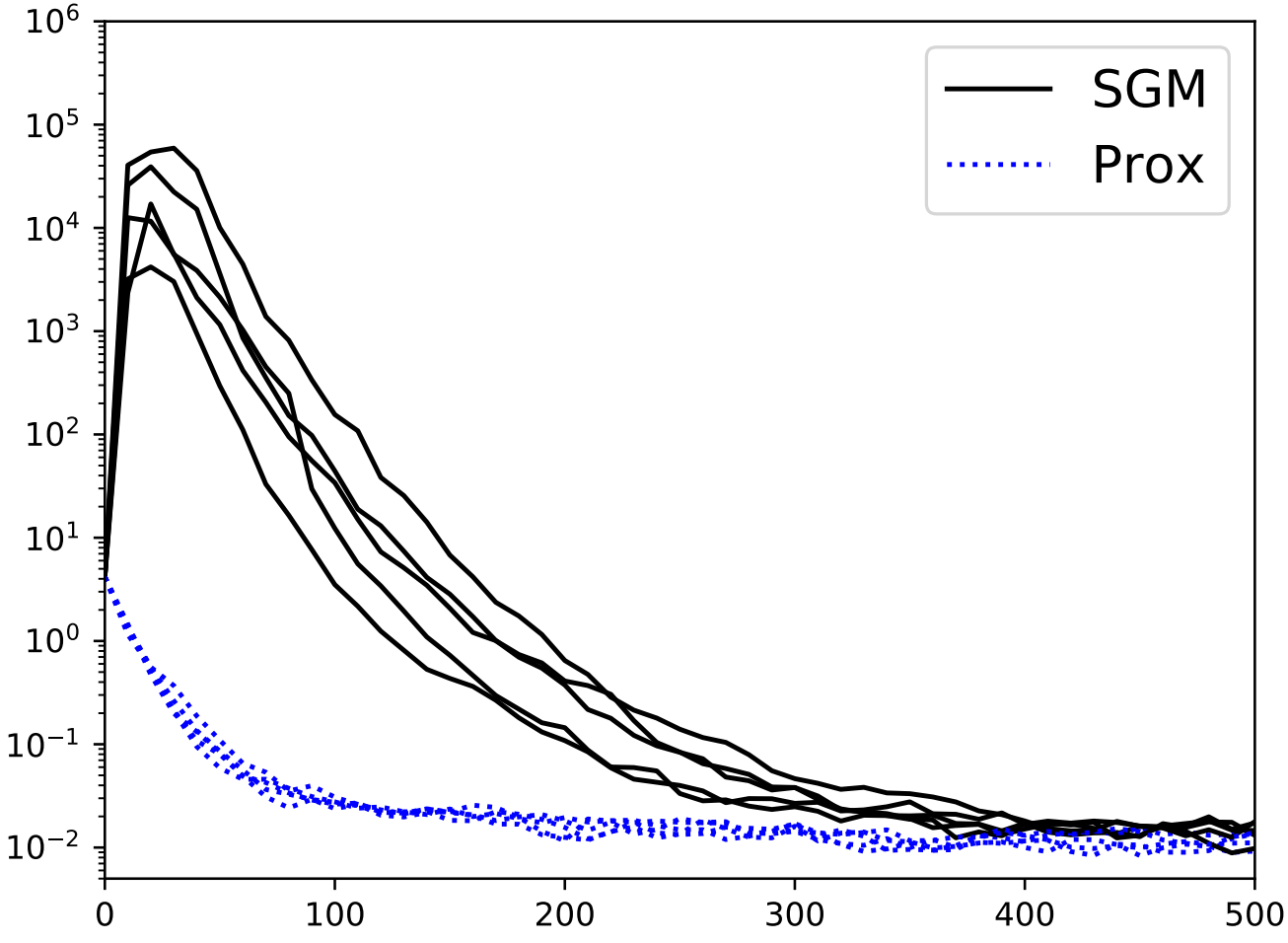
$$\mathbb{E}[\|f'(x;S)\|^2] \leq C(1 + \operatorname{dist}(x, \mathcal{X}^\star)^p).$$

*Then*

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

# Example behaviors

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



# 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,*

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

# New asymptotic analysis (convex case)

## 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,*

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

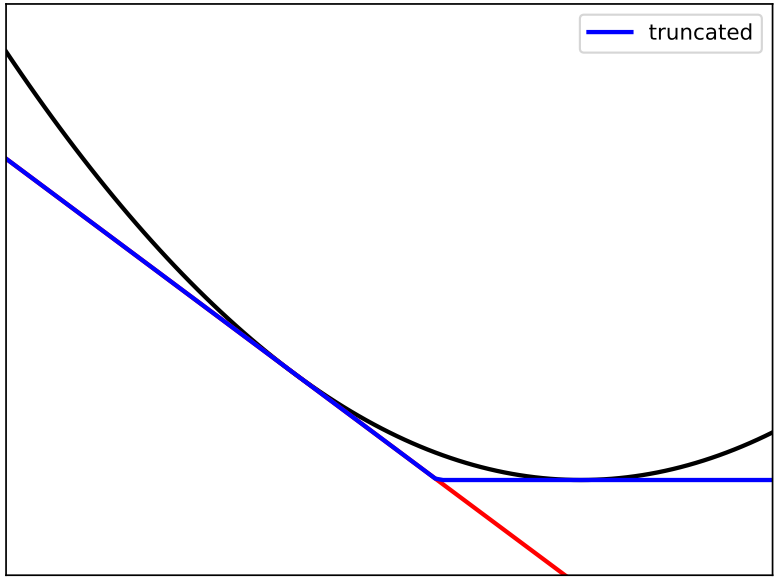
# New asymptotic analysis (convex case)

## 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 19]
- Key insight: subgradients of $f_{x_k}(\cdot; S_k)$ close to $\nabla f(x_k; S_k)$



# Convergence to stationarity in weakly convex cases

Convergence requires *Moreau envelope* [Davis & Drusvyatskiy 18]

$$F_\lambda(x) := \inf_{y \in X} \left\{ F(y) + \frac{\lambda}{2} \|y - x\|_2^2 \right\},$$

Important properties:

- Proximal mapping:

$$x^\lambda := \text{prox}_{F/\lambda}(x) := \operatorname*{argmin}_{y \in X} \left\{ F(y) + \frac{\lambda}{2} \|y - x\|_2^2 \right\}$$

  satisfies

$$\nabla F_\lambda(x) = \lambda(x - x^\lambda)$$

- Near stationarity and decrease:

$$F(x^\lambda) \leq F(x) \quad \text{and} \quad \text{dist}(0, \partial F(x^\lambda)) \leq \|\nabla F_\lambda(x)\|_2$$

# Convergence to stationarity in weakly convex cases

Convergence requires *Moreau envelope* [Davis & Drusvyatskiy 18]

$$F_\lambda(x) := \inf_{y \in X} \left\{ F(y) + \frac{\lambda}{2} \|y - x\|_2^2 \right\},$$

Important properties:

- Proximal mapping:

$$x^\lambda := \text{prox}_{F/\lambda}(x) := \operatorname*{argmin}_{y \in X} \left\{ F(y) + \frac{\lambda}{2} \|y - x\|_2^2 \right\}$$

  satisfies

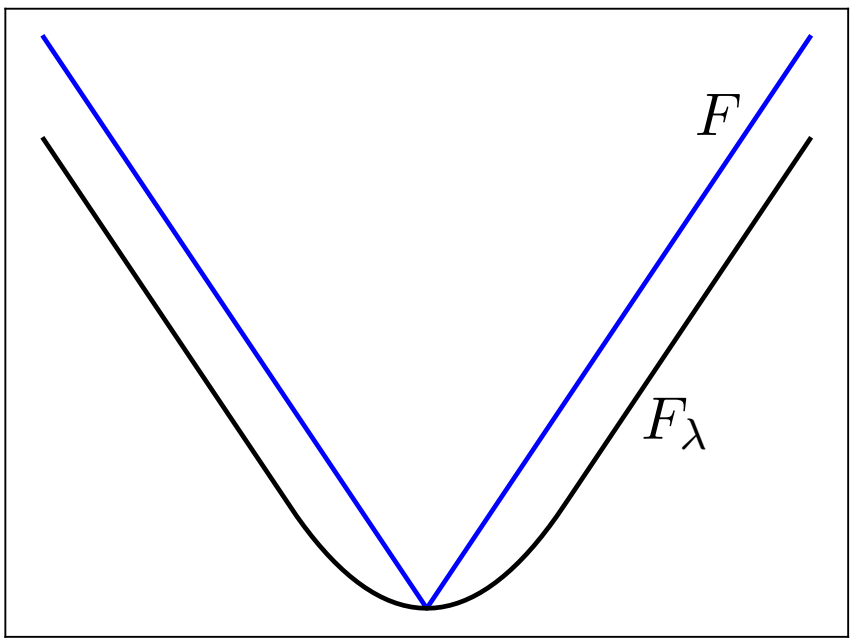
$$\nabla F_\lambda(x) = \lambda(x - x^\lambda)$$

- Near stationarity and decrease:

$$F(x^\lambda) \leq F(x) \quad \text{and} \quad \text{dist}(0, \partial F(x^\lambda)) \leq \|\nabla F_\lambda(x)\|_2$$

Convergence: Say iterates $x_k$ converge if $\nabla F_\lambda(x_k) \to 0$

# Moreau envelope of the absolute value




For $F(x) = |x|$,

$$F_\lambda(x) = \begin{cases} \frac{\lambda}{2}x^2 & \text{if } |x| \le \lambda^{-1} \\ |x| - \frac{1}{2\lambda} & \text{if } |x| > \lambda^{-1} \end{cases}$$

- $F_\lambda'(x) = \lambda x$
- $|F_\lambda'(x)| = \lambda \operatorname{dist}(x, 0)$
- prox step $x^\lambda = 0$ if $|x| \le 1/\lambda$

# Convergence in weakly convex cases

Use regularized stochastic-proximal point method,

$$x_{k+1} = \underset{x \in X}{\operatorname{argmin}} \left\{ f(x; S_k) + \frac{\rho(S_k)}{2} \|x - x_k\|_2^2 + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}.$$

## Theorem (Asi & D. 19)

*Let random $f$ be $\rho(s)$ weakly convex with $\mathbb{E}[\rho^2(S)] < \infty$. With proximal-point iteration, iterates $x_k$ satisfy $F_\lambda(x_k) \overset{a.s.}{\to} G$ and*

$$\sum_{k=1}^{\infty} \alpha_k \|\nabla F_\lambda(x_k)\|_2^2 < \infty.$$

# Convergence in weakly convex cases

Use regularized stochastic-proximal point method,

$$x_{k+1} = \underset{x \in X}{\operatorname{argmin}} \left\{ f(x; S_k) + \frac{\rho(S_k)}{2} \|x - x_k\|_2^2 + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}.$$

## Theorem (Asi & D. 19)

*Let random $f$ be $\rho(s)$ weakly convex with $\mathbb{E}[\rho^2(S)] < \infty$. With proximal-point iteration, iterates $x_k$ satisfy $F_\lambda(x_k) \overset{a.s.}{\to} G$ and*

$$\sum_{k=1}^{\infty} \alpha_k \|\nabla F_\lambda(x_k)\|_2^2 < \infty.$$

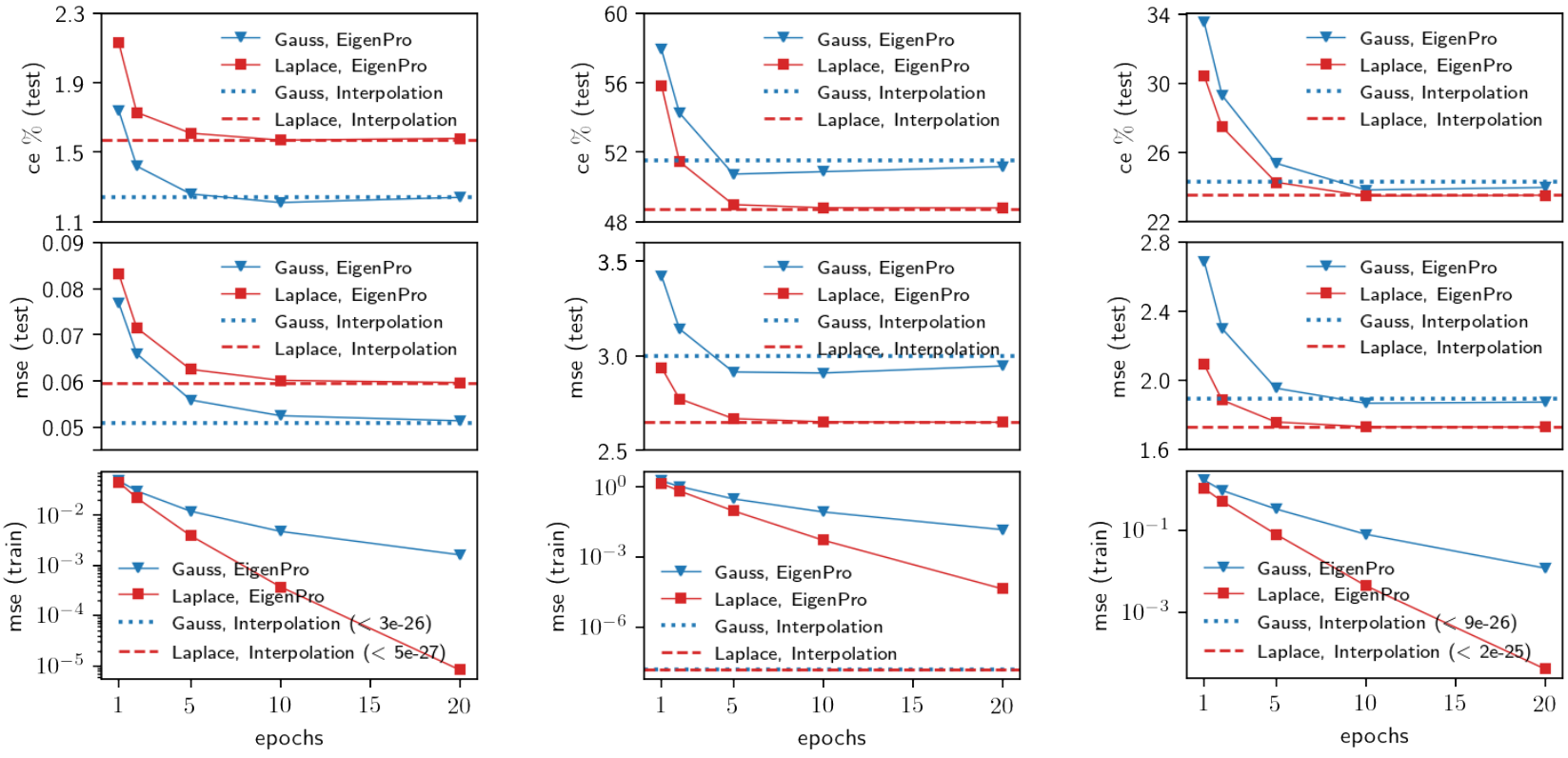
## Proposition (Asi & D. 19)

*If iterates $x_k$ remain bounded and image of stationary points has measure zero,*

$$\nabla F_\lambda(x_k) \overset{a.s.}{\to} 0.$$

# 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?

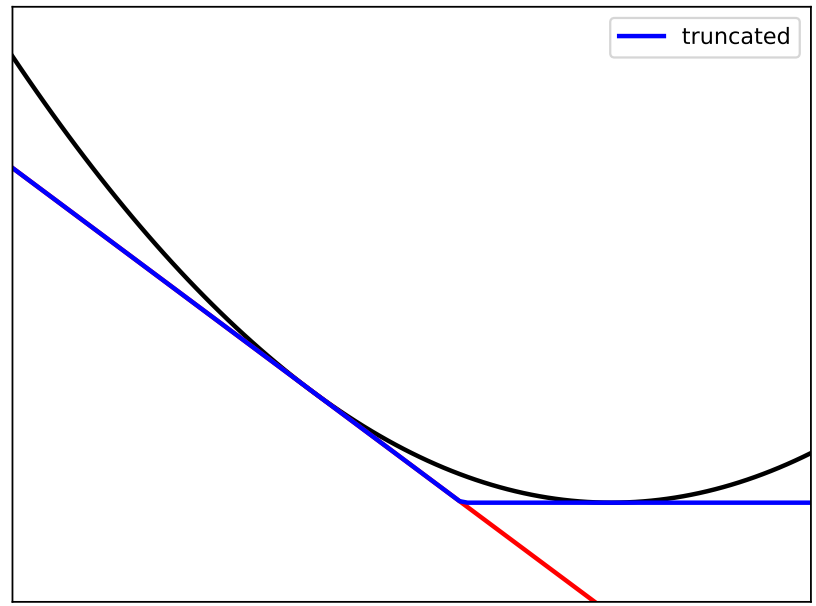
$$\underset{x}{\text{minimize}} \;\; F(x) := \mathbb{E}[f(x;S)] = \int f(x;s)dP(s)$$

**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

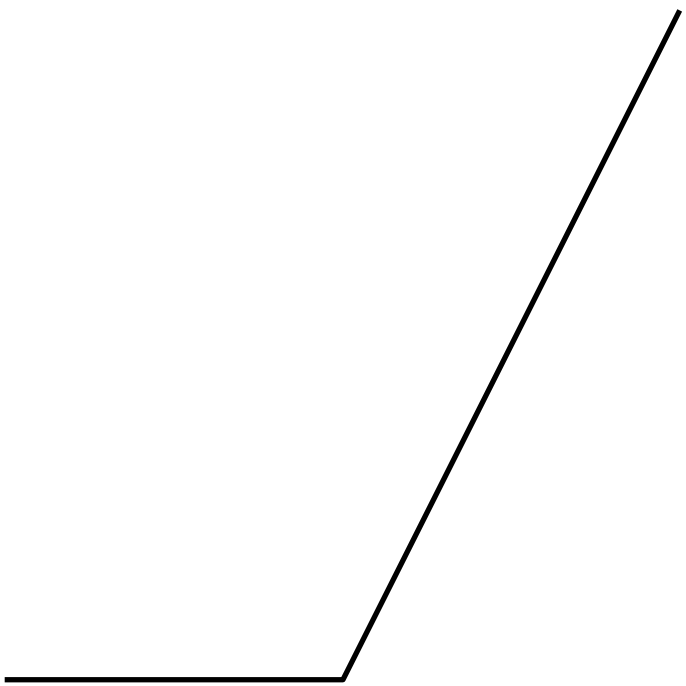
# Sharp problems

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



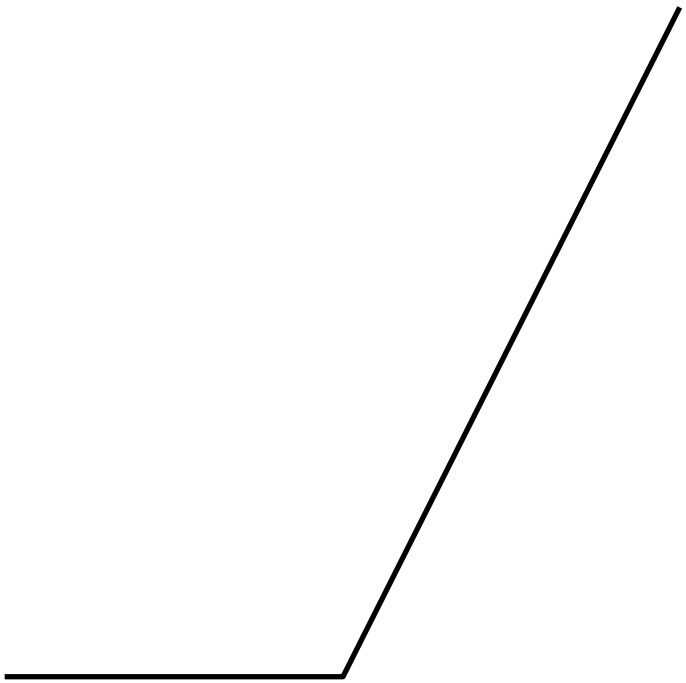
# Sharp problems

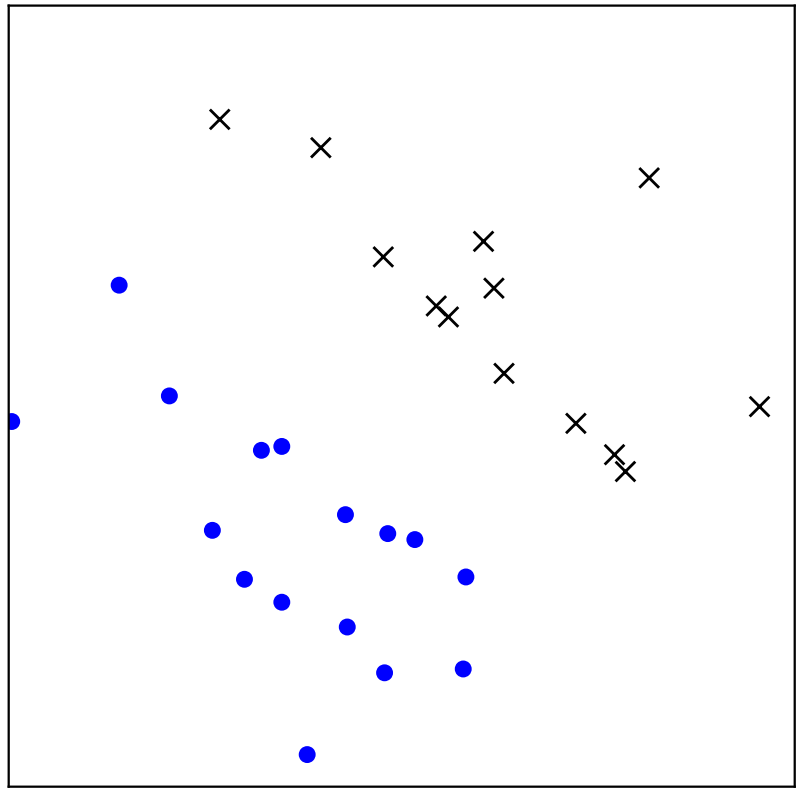
**Definition:** An objective $F$ is *sharp* if

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

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

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





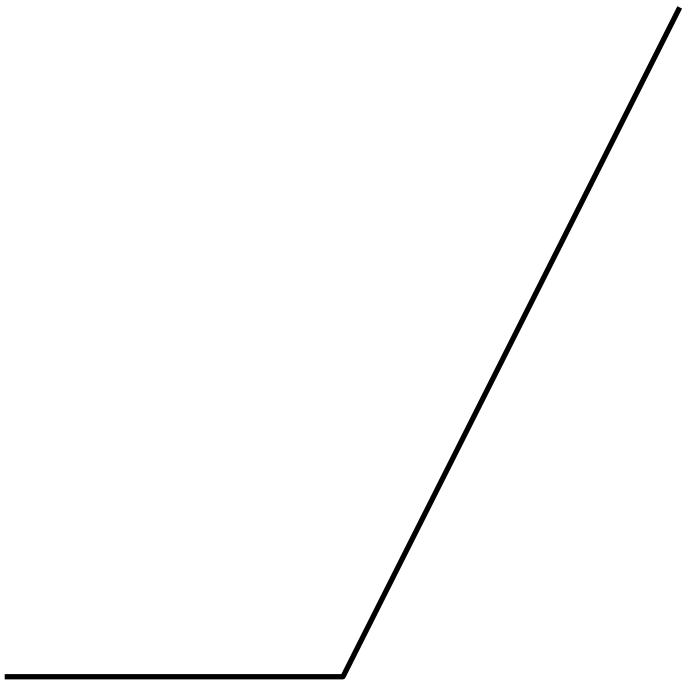
# Sharp problems

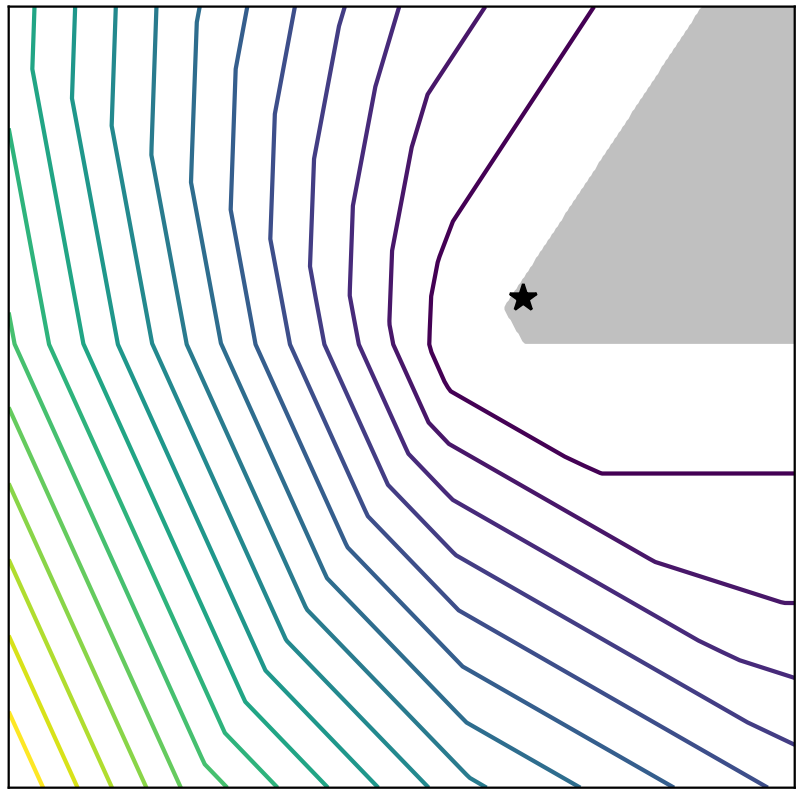
**Definition:** An objective $F$ is *sharp* if

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

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. If $F$ is convex,*

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

# Sharp weakly 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]

- Phase retrieval $F(x) = \frac{1}{m} \left\| (Ax)^2 - (Ax^\star)^2 \right\|_1$
- Blind deconvolution [Charisopoulos et al. 19]

# Sharp weakly 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]

- Phase retrieval $F(x) = \frac{1}{m} \left\| (Ax)^2 - (Ax^\star)^2 \right\|_1$
- Blind deconvolution [Charisopoulos et al. 19]

Theorem (Asi & D. 19)

*Let $F$ have sharp growth and be easy. There exists $c \in (0, 1)$ such that on the event $x_k \to X^\star$,*

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

# Outline

# Methods

## Iterate

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

# 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 ($f \geq 0$):

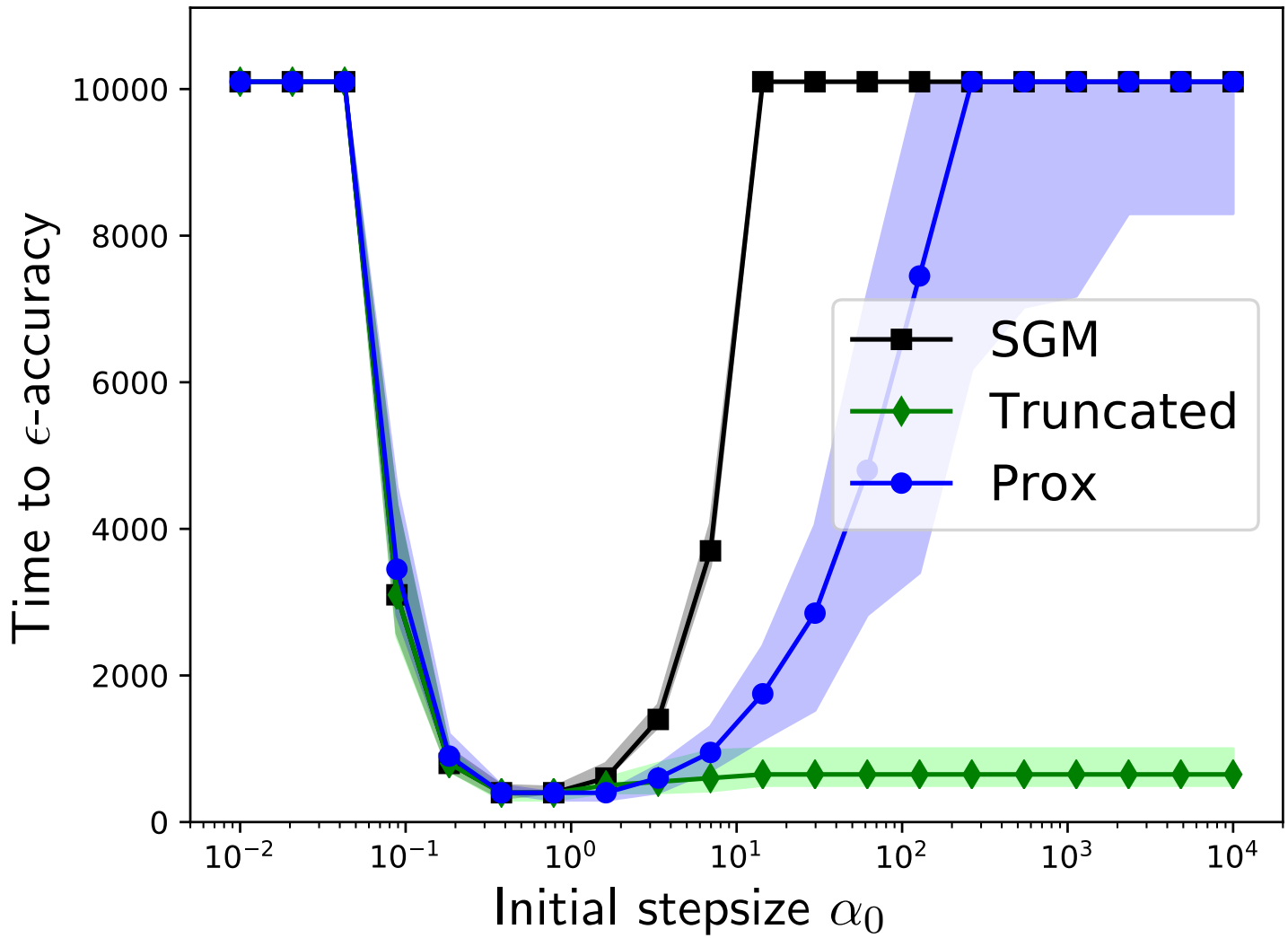
$$f_{x_k}(x; S_k) = \left[ f(x_k; S_k) + \langle f'(x_k; S_k), x - x_k \rangle \right]_+$$

- (Stochastic) proximal point

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

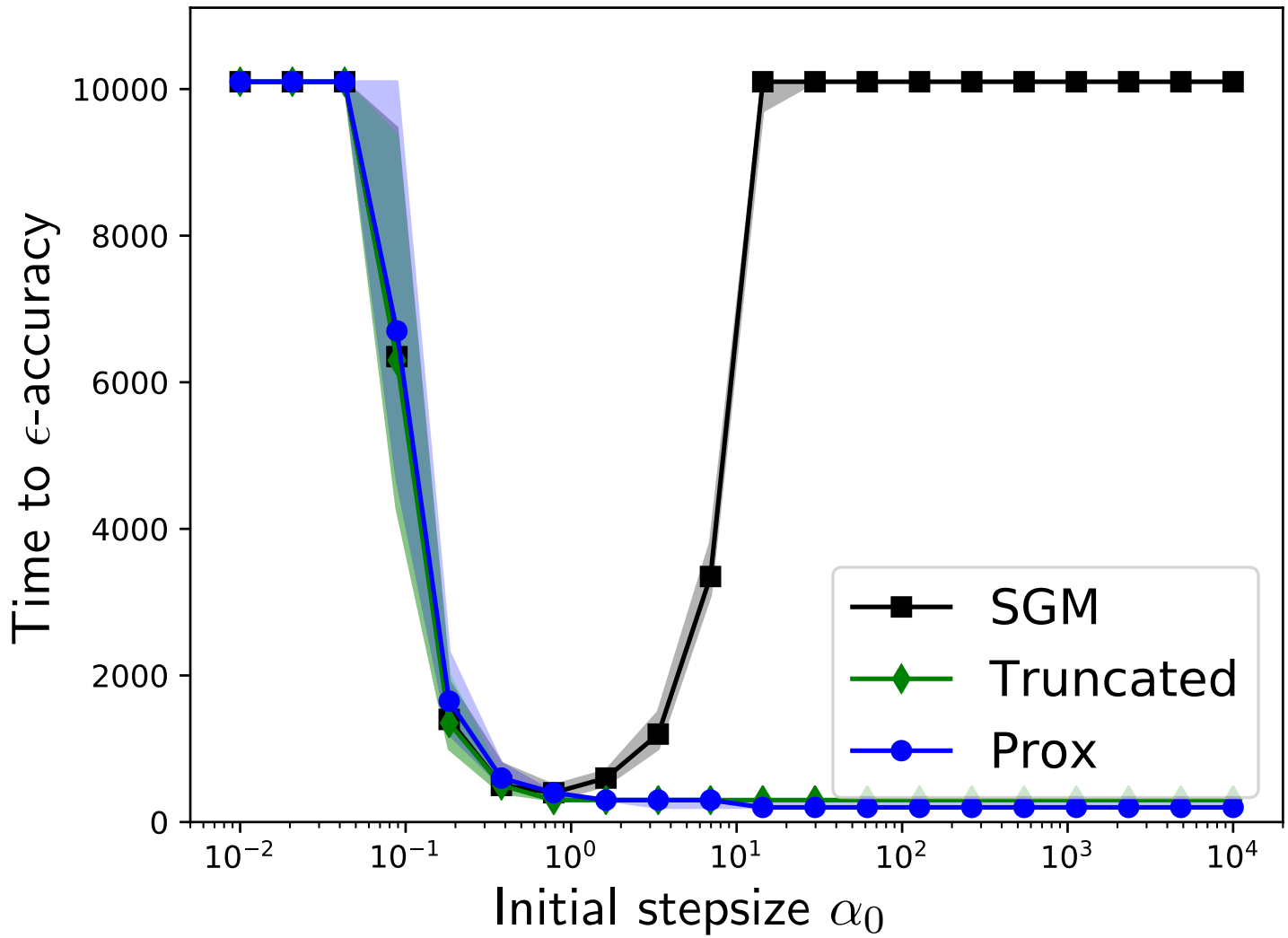
# Linear regression with low noise

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

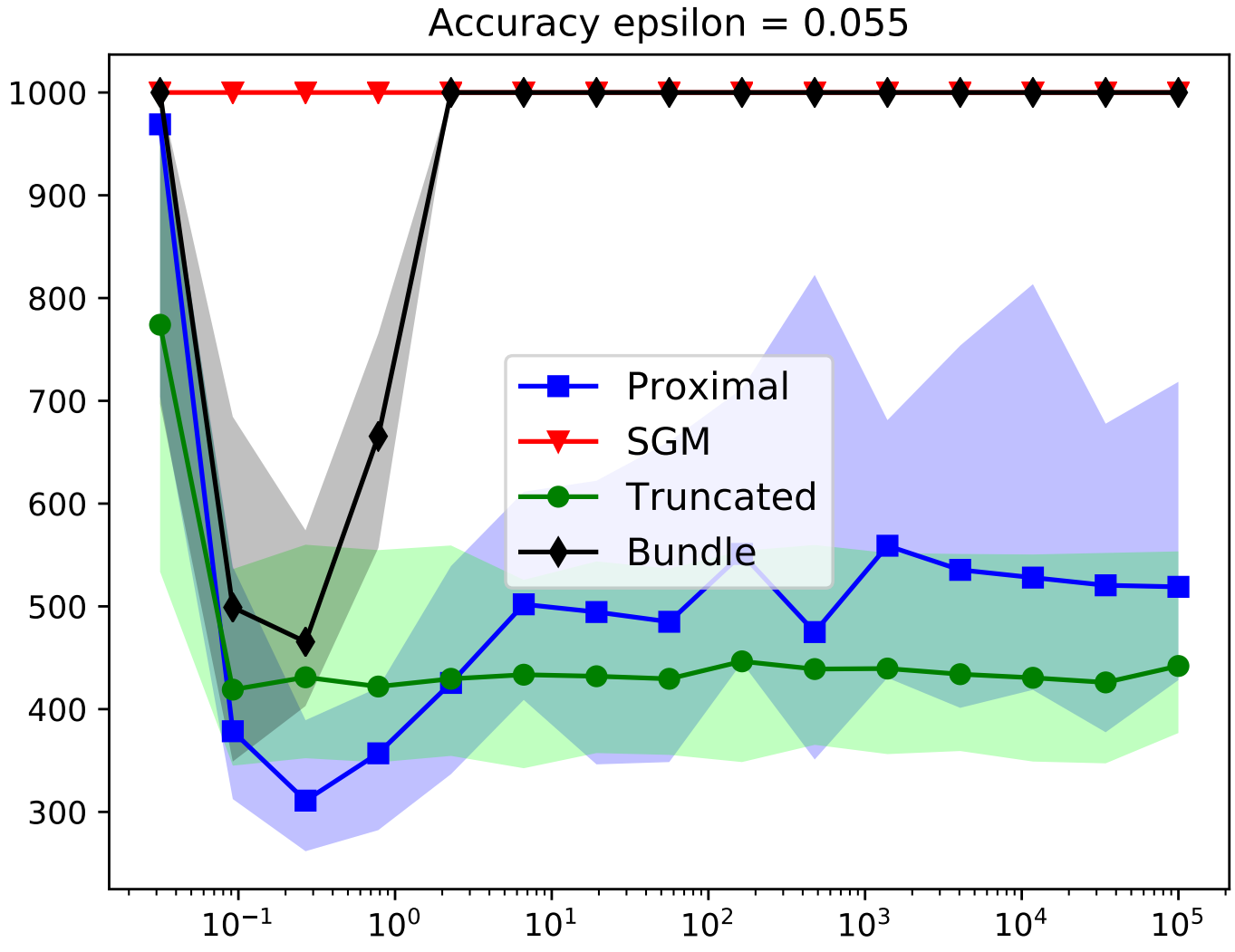


# Linear regression with no noise

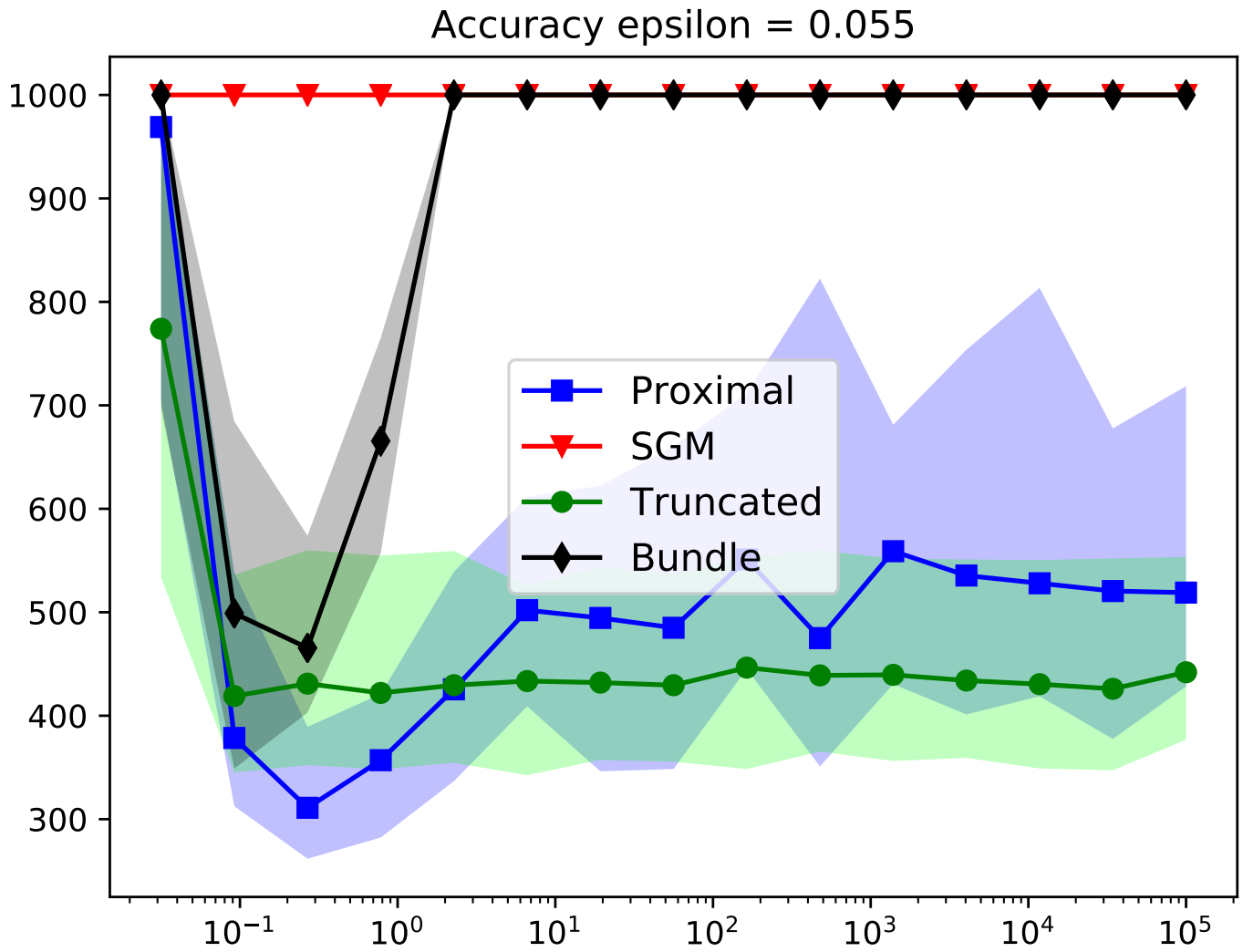
$$F(x) = \frac{1}{2m} \sum_{i=1}^{m} (a_i^T x - b_i)^2$$



# Linear regression with "poor" conditioning



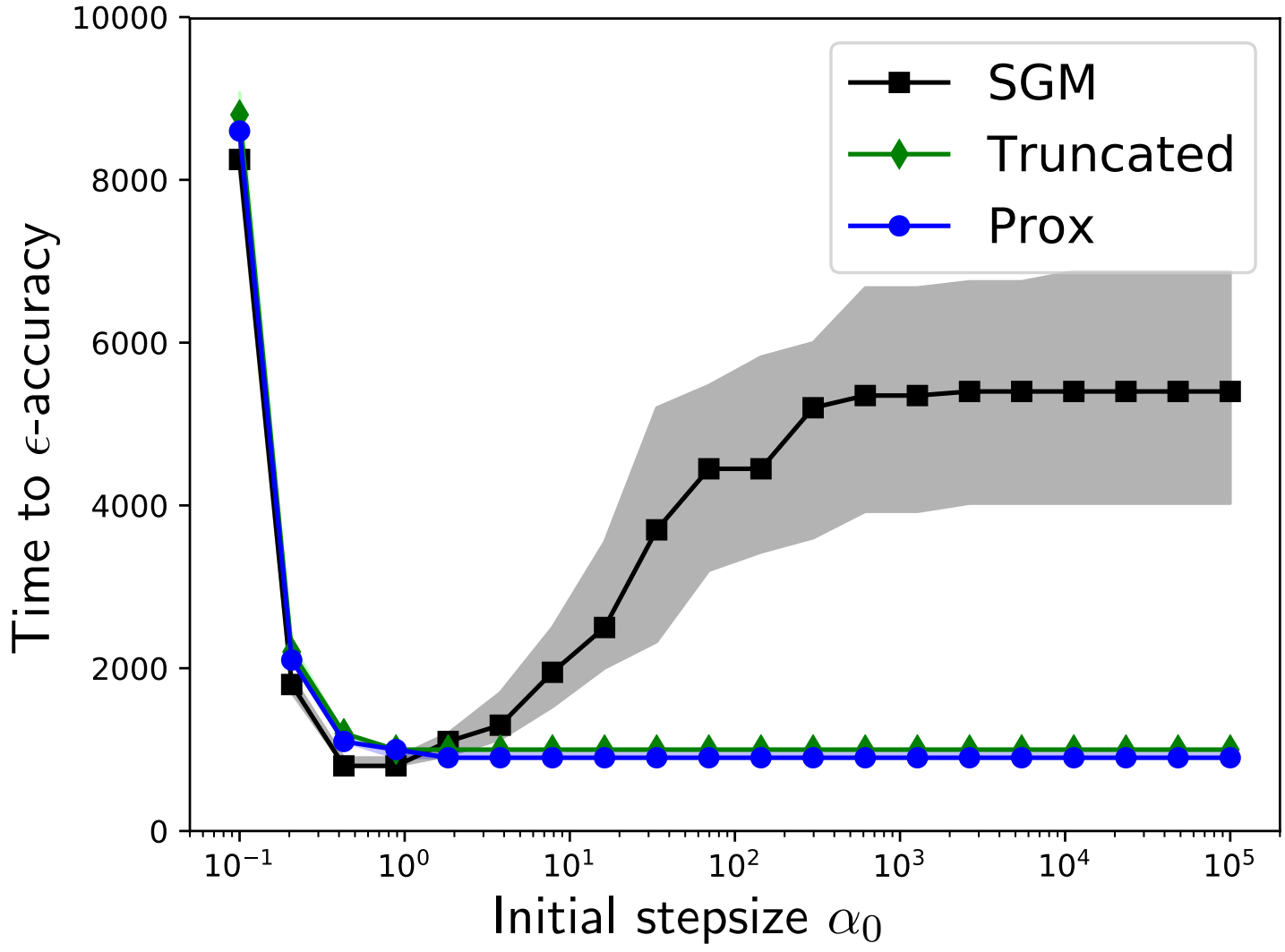
# Linear regression with "poor" conditioning




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

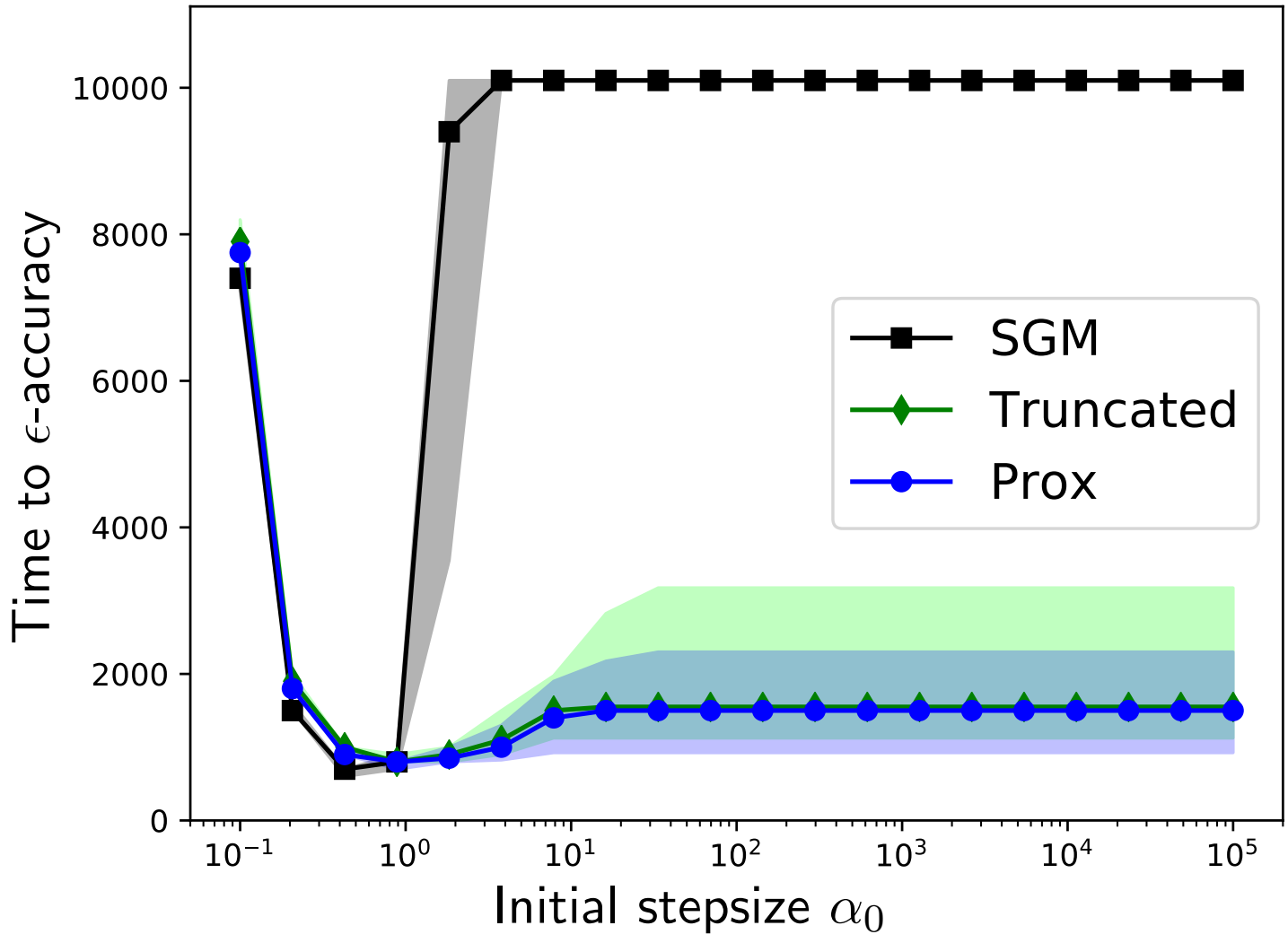
# Multiclass hinge loss: no noise

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



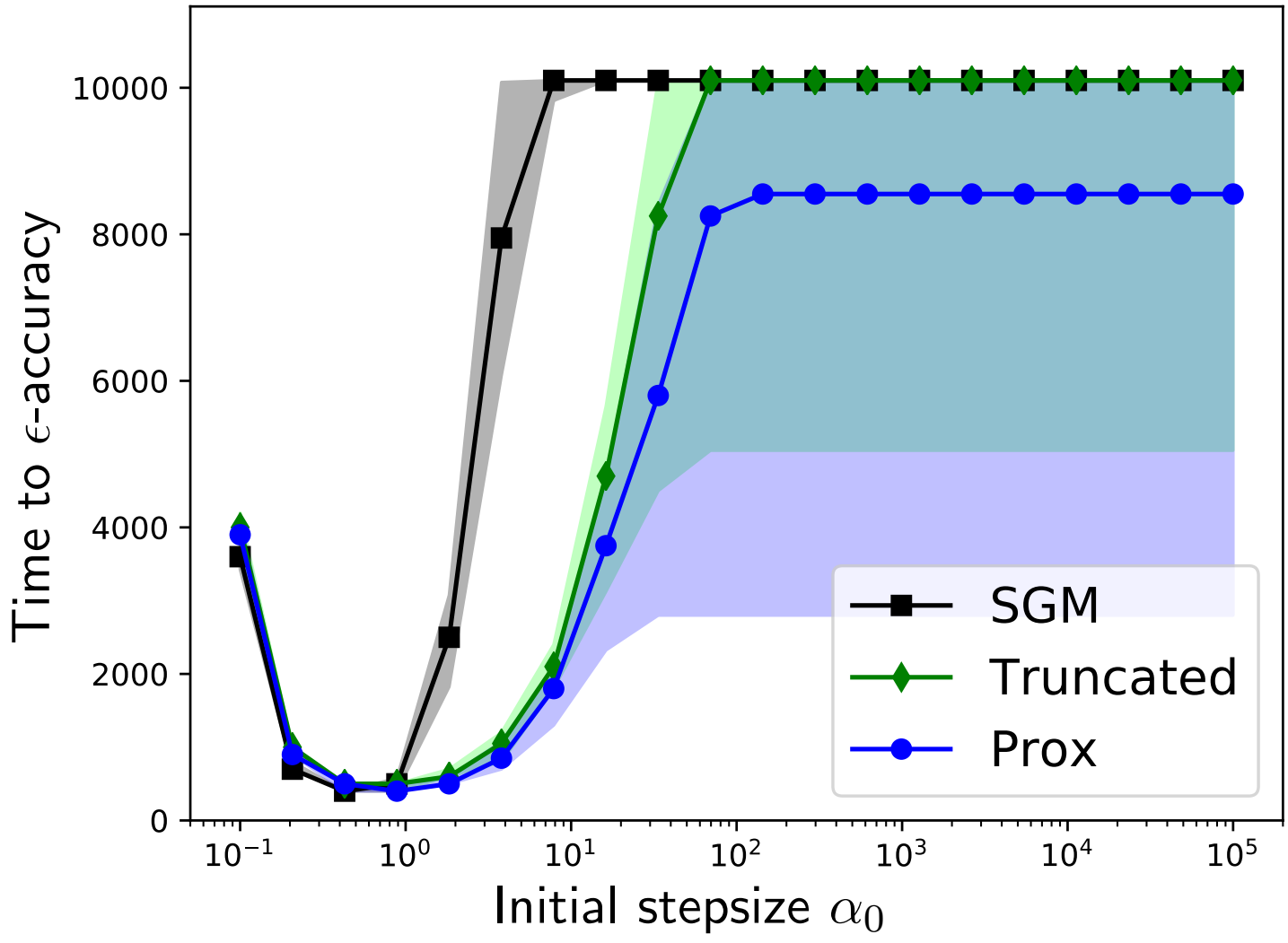
# Multiclass hinge loss: small label flipping

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

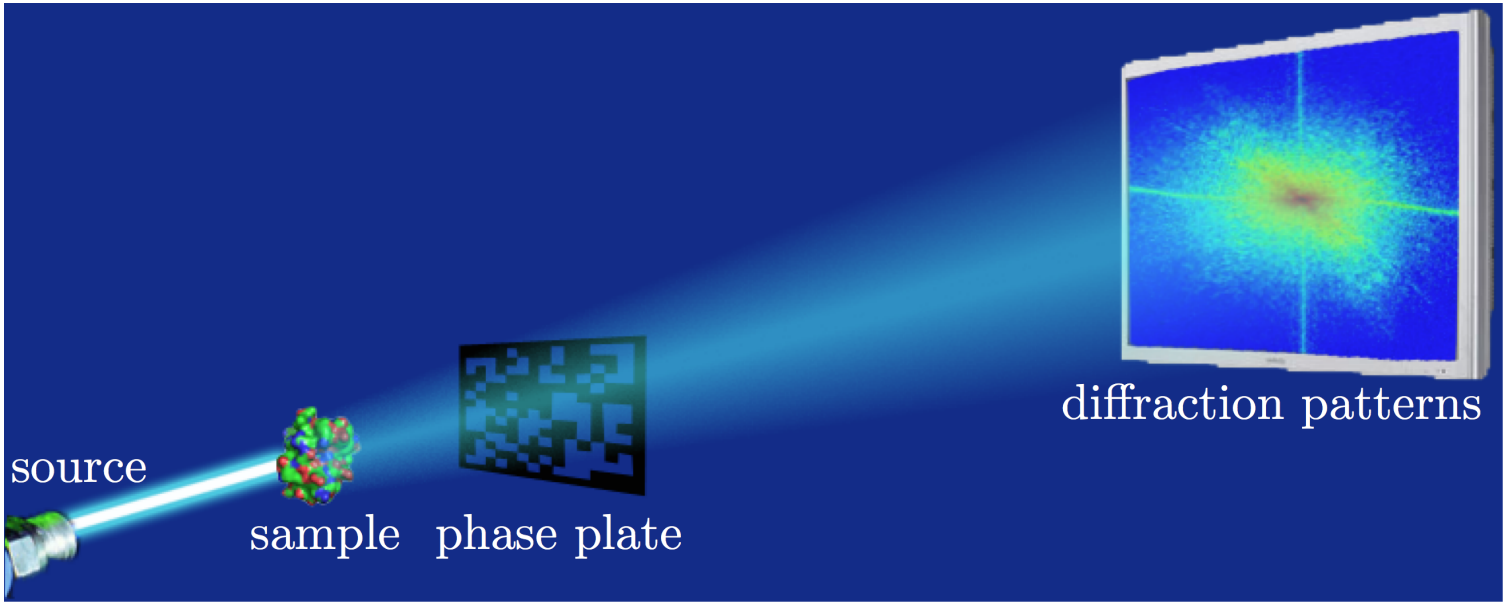


# Multiclass hinge loss: substantial label flipping

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

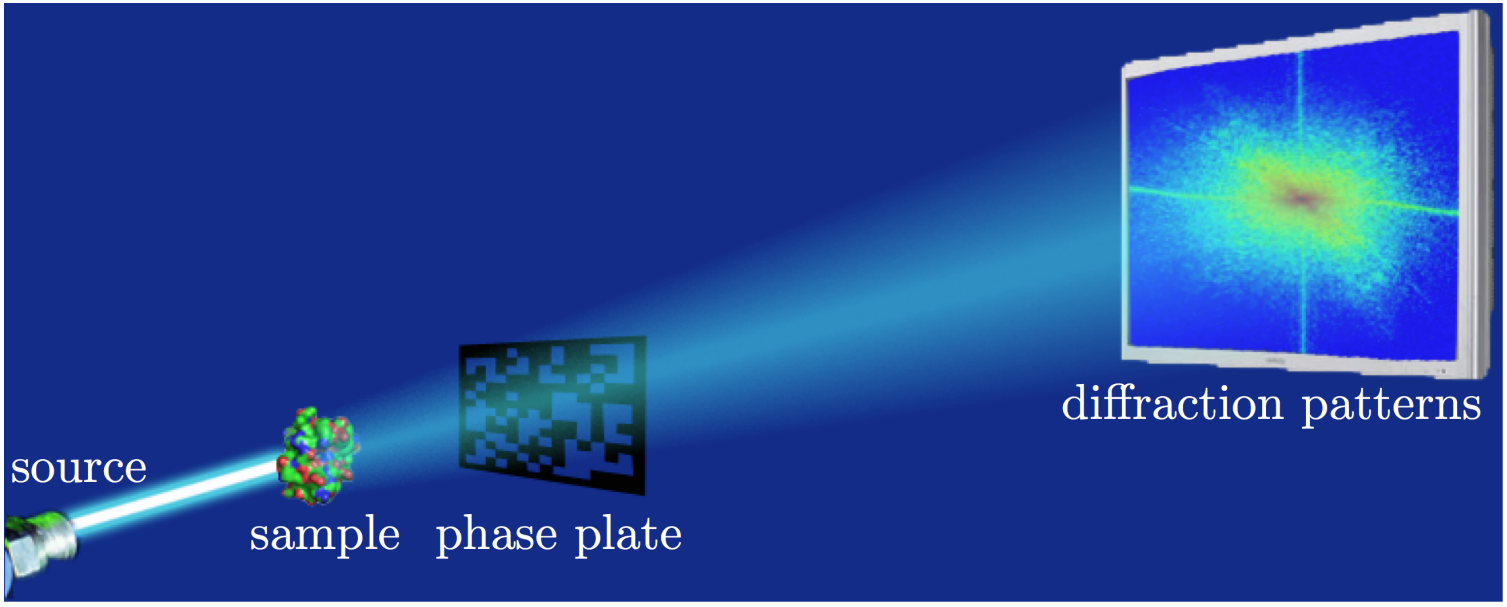


# (Robust) Phase retrieval




[Candès, Li, Soltanolkotabi 15]

# (Robust) Phase retrieval




[Candès, Li, Soltanolkotabi 15]

Observations (usually)

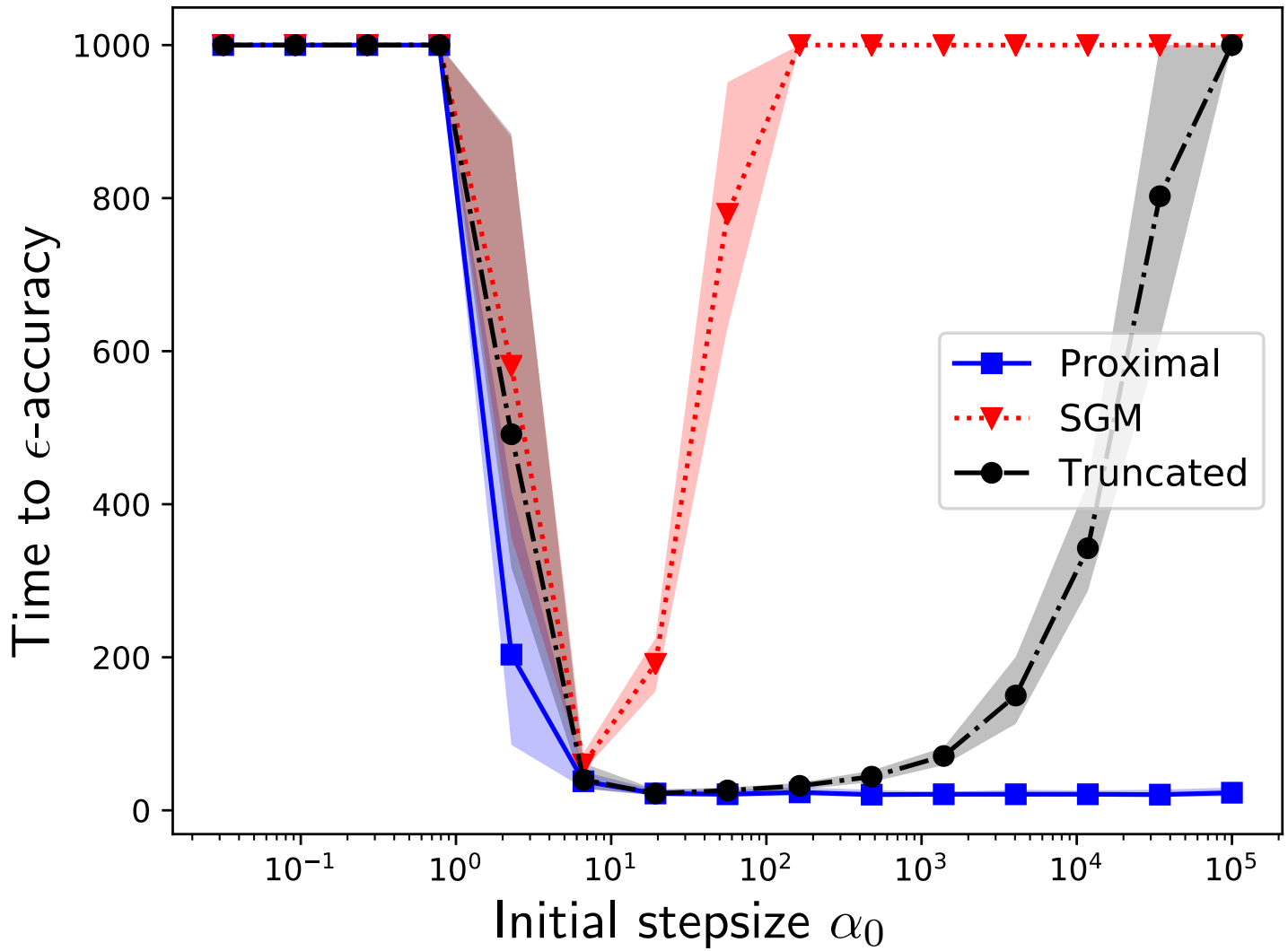
$$b_i = \langle a_i, x^\star \rangle^2$$

yield objective

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

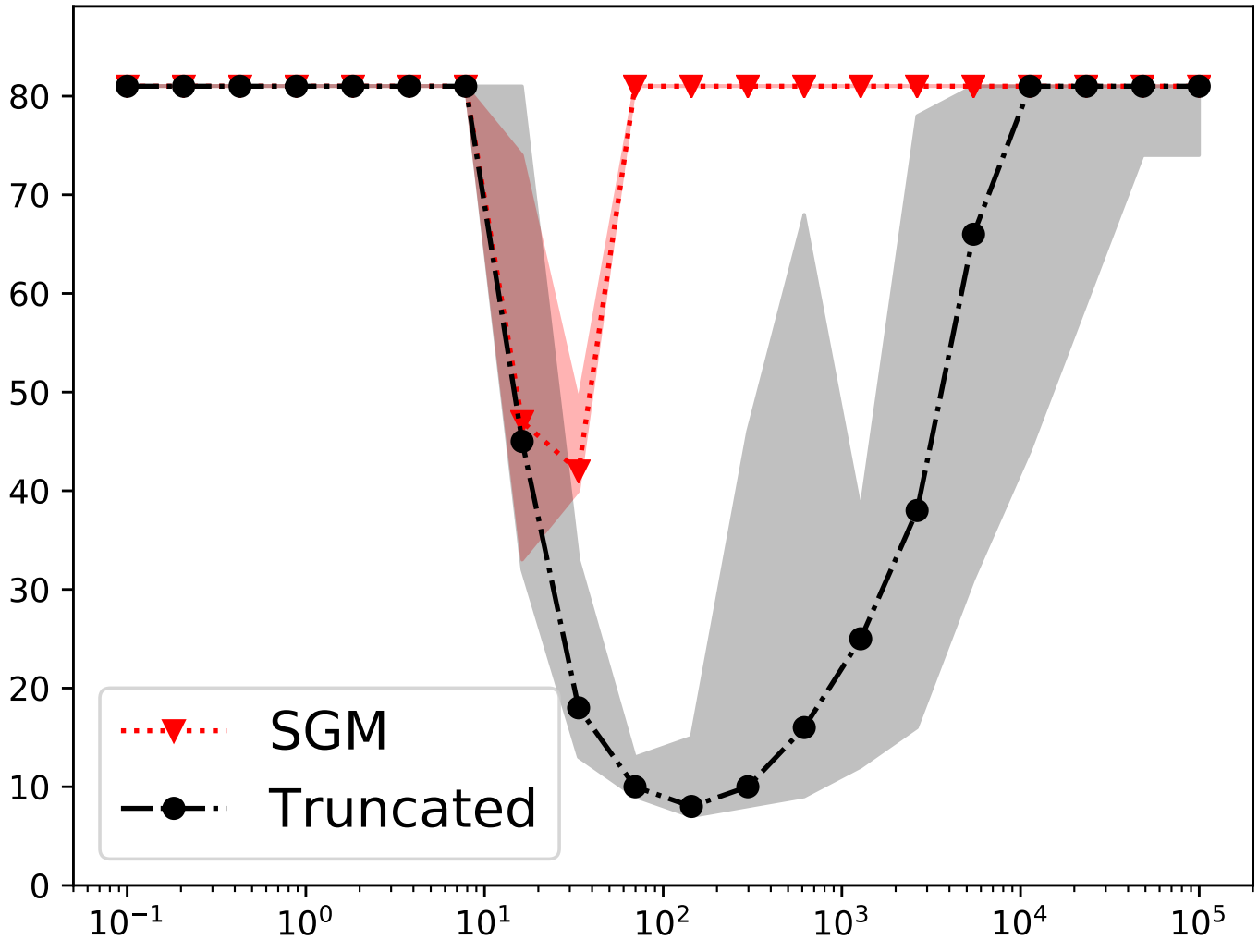
# Phase retrieval without noise

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



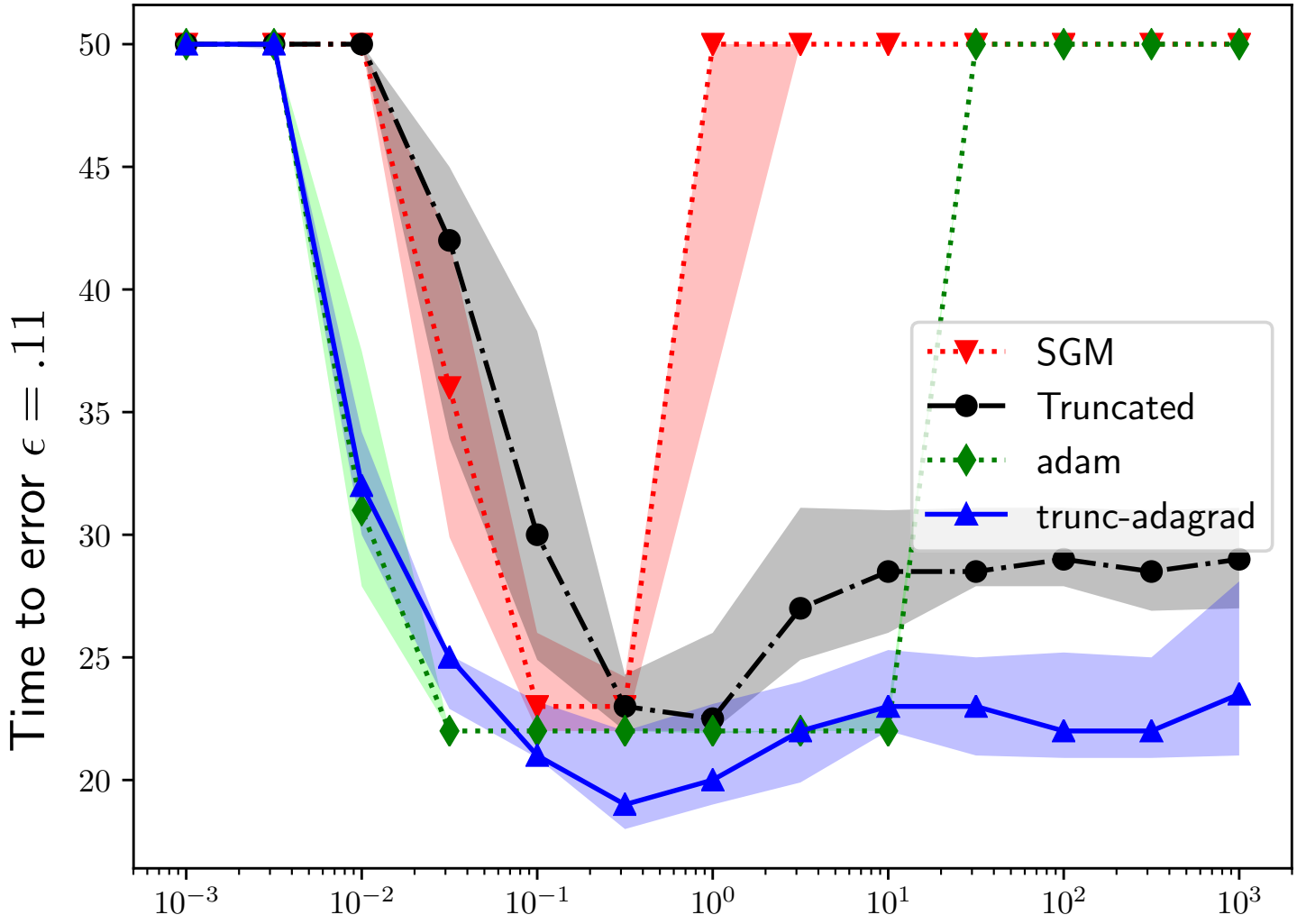
# Matrix completion without noise

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



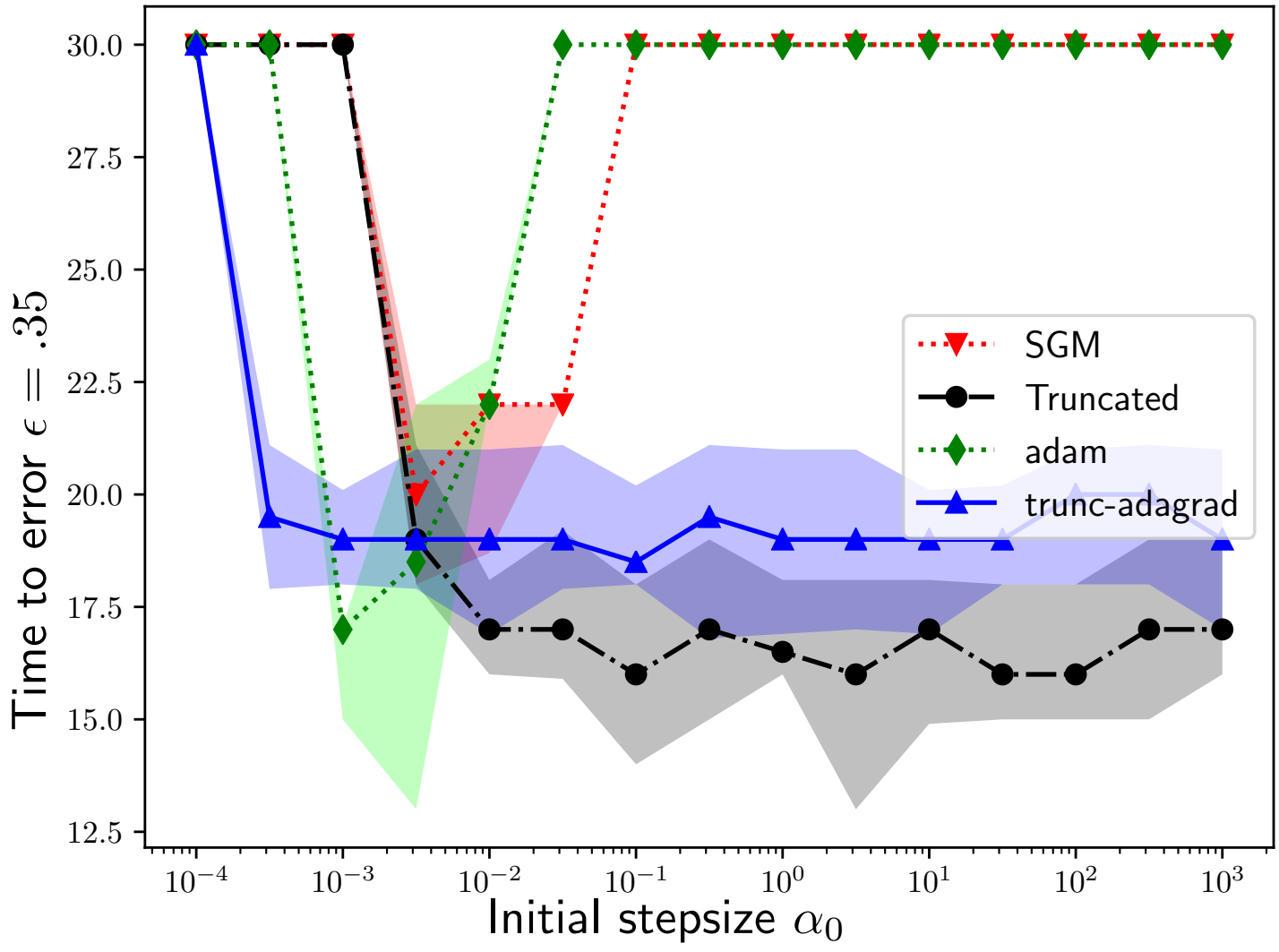
# Deep learning experiments

CIFAR 10 Dataset: 10 class image classification



# Deep learning experiment: dog recognition

Stanford Dogs: 120 class dog breed classification



# 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**

- Parallelism?

- The importance of better models in stochastic optimization. arXiv:1903.08619
- Stochastic (Approximate) Proximal Point Methods: Convergence, Optimality, and Adaptivity. arXiv:1810.05633