Solving composite optimization problems, with applications to phase retrieval

John Duchi (based on joint work with Feng Ruan)

Outline

Composite optimization problems

Methods for composite optimization

Application: robust phase retrieval

Experimental evaluation

Large scale composite optimization?

What I hope to accomplish today

- Investigate problem structures that are not *quite* convex but still amenable to elegant solution approaches
- Show how we can leverage stochastic structure to turn hard non-convex problems into "easy" ones
  [Keshavan, Montanari, Oh 10; Loh & Wainwright 12]
- Consider large scale versions of these problems

Composite optimization problems

The problem:

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

where

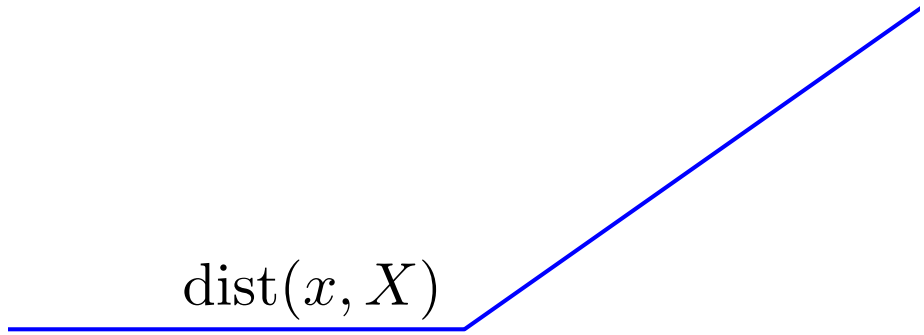
$$h : \mathbb{R}^m \to \mathbb{R} \text{ is convex} \quad \text{and} \quad c : \mathbb{R}^n \to \mathbb{R}^m \text{ is smooth}$$

Motivation: the exact penalty

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \quad x \in X$$

equivalent (for all large enough λ) to

$$\underset{x}{\text{minimize}} \;\; f(x) + \lambda \operatorname{dist}(x, X)$$

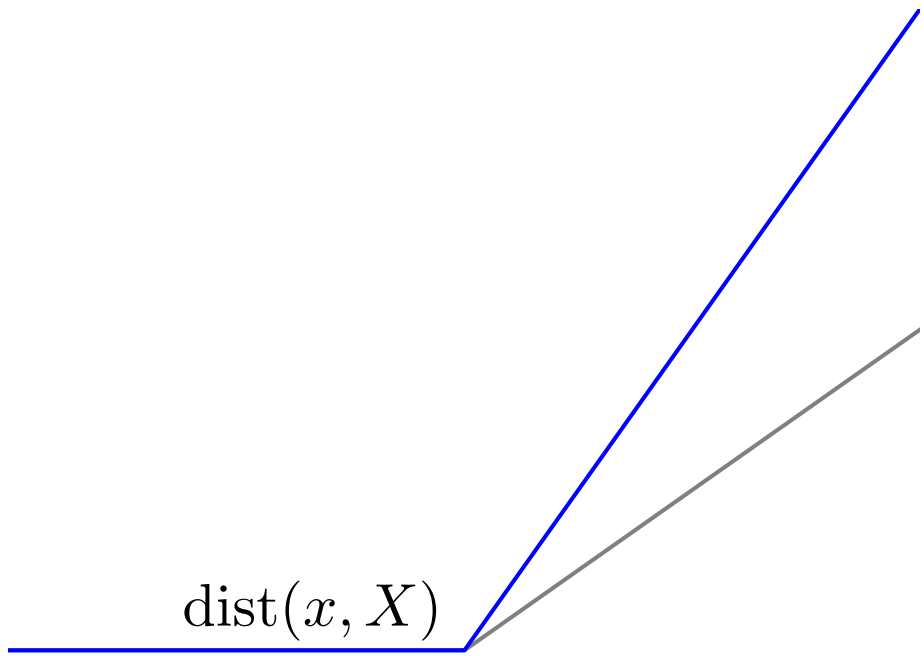


Motivation: the exact penalty

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \quad x \in X$$

equivalent (for all large enough λ) to

$$\underset{x}{\text{minimize}} \;\; f(x) + \lambda \operatorname{dist}(x, X)$$

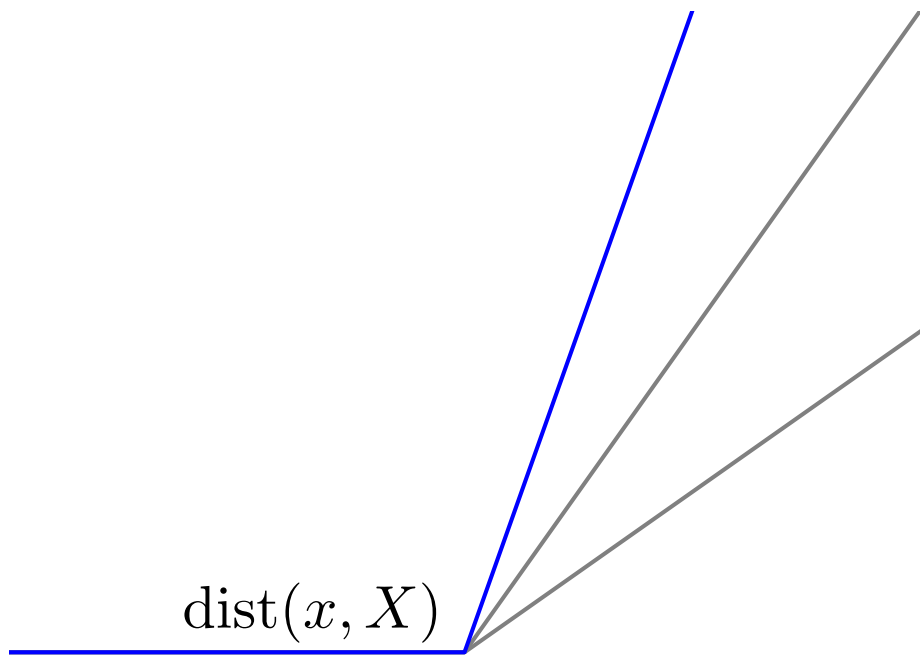


Motivation: the exact penalty

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \quad x \in X$$

equivalent (for all large enough λ) to

$$\underset{x}{\text{minimize}} \;\; f(x) + \lambda \operatorname{dist}(x, X)$$



Motivation: the exact penalty

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \quad c(x) = 0$$

equivalent to (for all large enough λ)

$$\underset{x}{\text{minimize}} \;\; f(x) + \lambda \, \|c(x)\|$$

[Fletcher & Watson 80, 82; Burke 85]

Motivation: the exact penalty

$$\underset{x}{\text{minimize}} \;\; f(x) \quad \text{subject to} \quad c(x) = 0$$

equivalent to (for all large enough λ)

$$\underset{x}{\text{minimize}} \;\; f(x) + \underbrace{\lambda \, \|c(x)\|}_{=h(c(x))}$$

where

$$h(z) = \lambda \, \|z\|$$

[Fletcher & Watson 80, 82; Burke 85]

Motivation: nonlinear measurements and modeling

- Have true signal $x^\star \in \mathbb{R}^n$ and measurement vectors $a_i \in \mathbb{R}^n$

Motivation: nonlinear measurements and modeling

- Have true signal $x^\star \in \mathbb{R}^n$ and measurement vectors $a_i \in \mathbb{R}^n$
- Observe nonlinear measurements

$$b_i = \phi(\langle a_i, x^\star \rangle) + \xi_i, \quad i = 1, \ldots, m$$

for $\phi(\cdot)$ a nonlinear function but *smooth* function

An objective:

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

Motivation: nonlinear measurements and modeling

- Have true signal $x^\star \in \mathbb{R}^n$ and measurement vectors $a_i \in \mathbb{R}^n$
- Observe nonlinear measurements

$$b_i = \phi(\langle a_i, x^\star \rangle) + \xi_i, \quad i = 1, \ldots, m$$

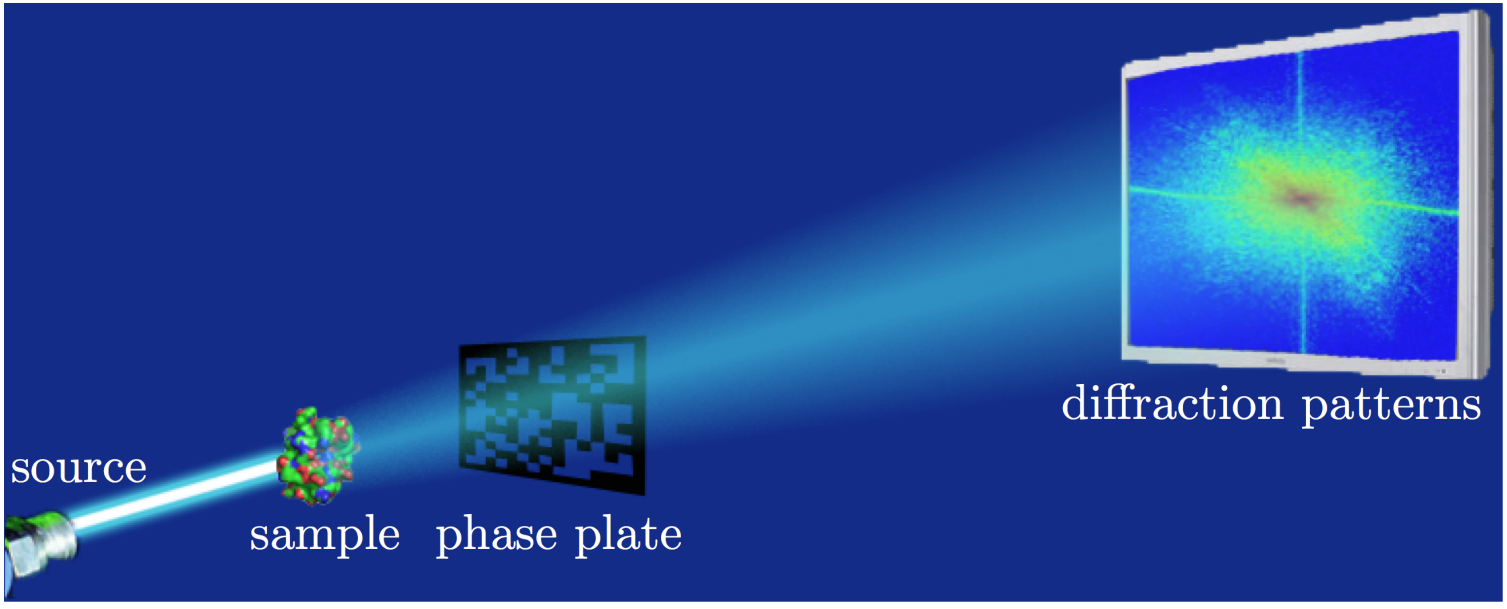
for $\phi(\cdot)$ a nonlinear function but *smooth* function

An objective:

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

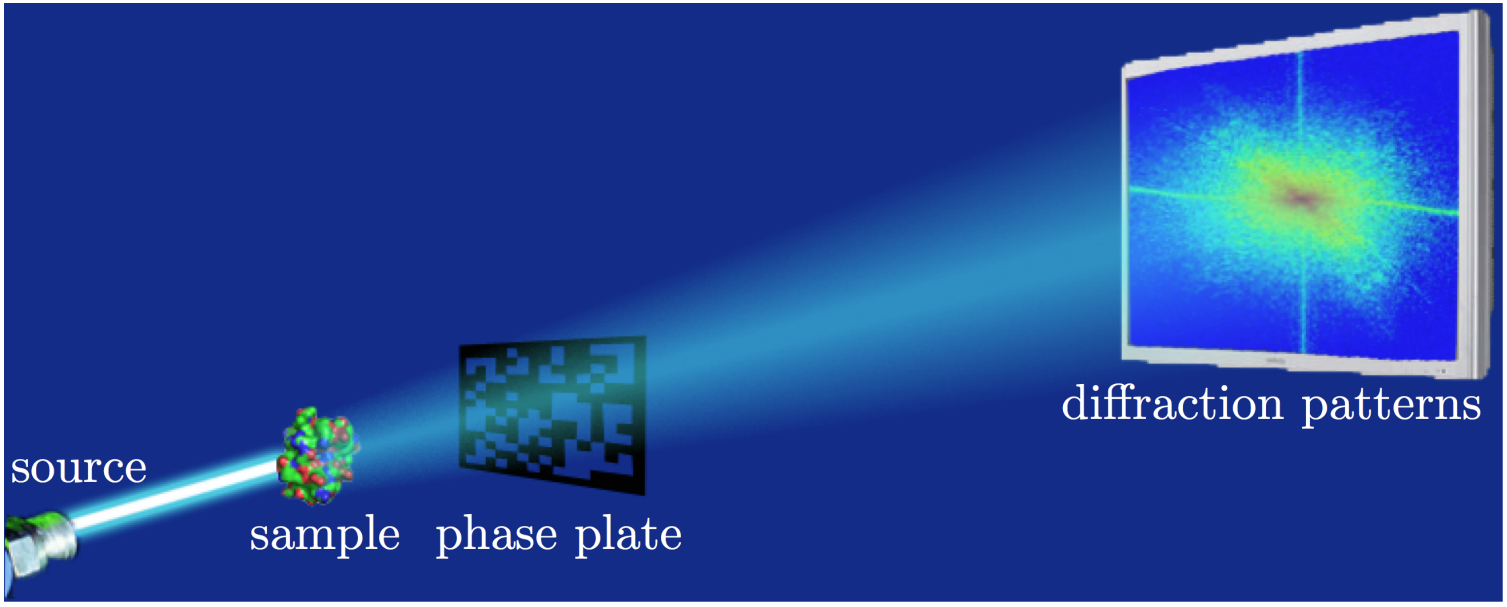
Nonlinear least squares [Nocedal & Wright 06; Plan & Vershynin 15; Oymak & Soltanolkotabi 16]

(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

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

Optimization methods

How do we solve optimization problems?

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

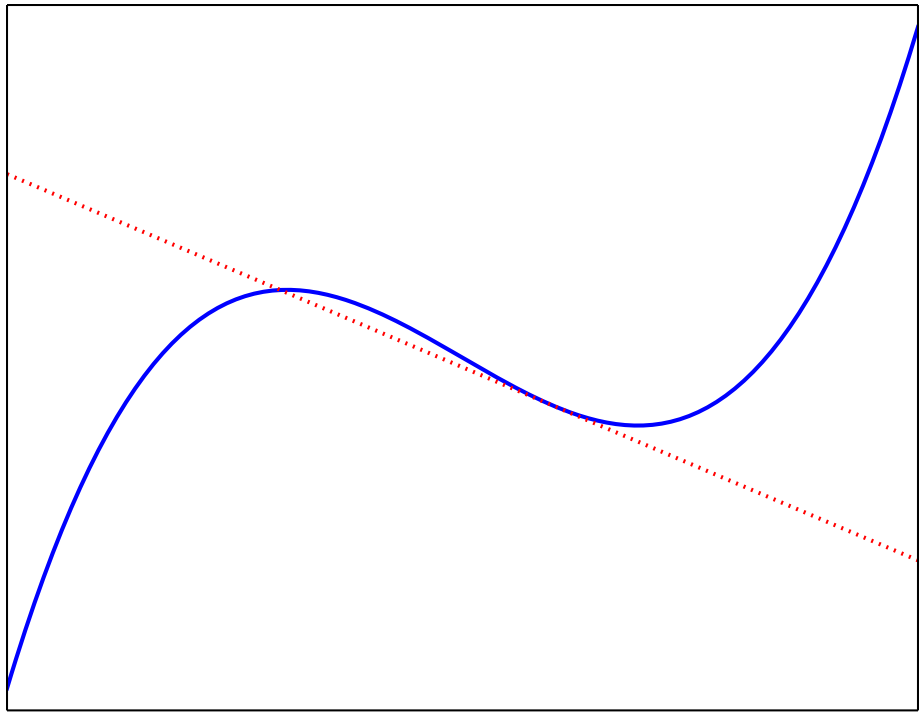
Optimization methods

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

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



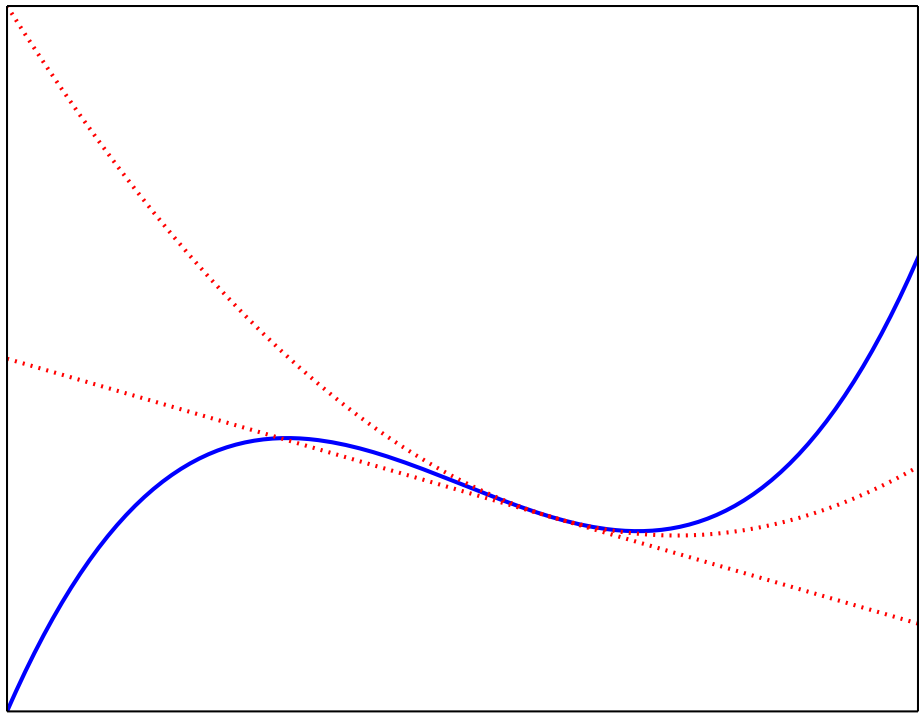
Optimization methods

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

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



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

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

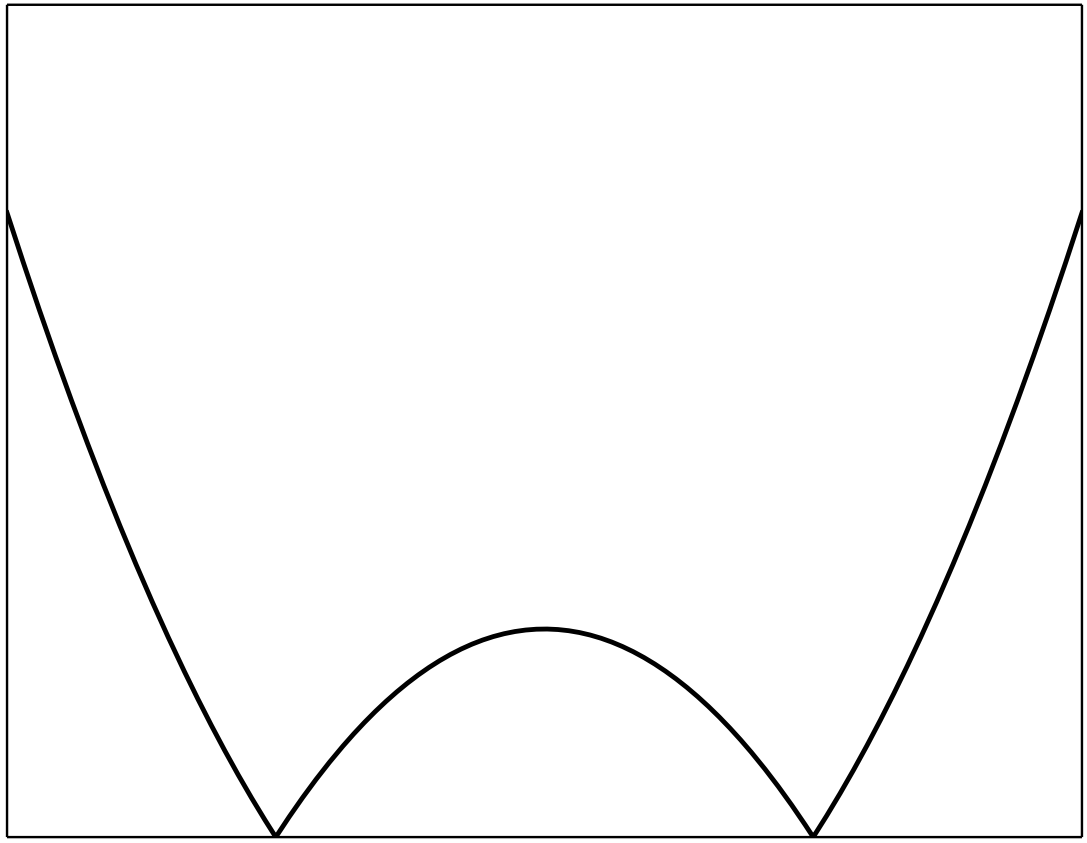
[Burke 85; Drusvyatskiy, Ioffe, Lewis 16]

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

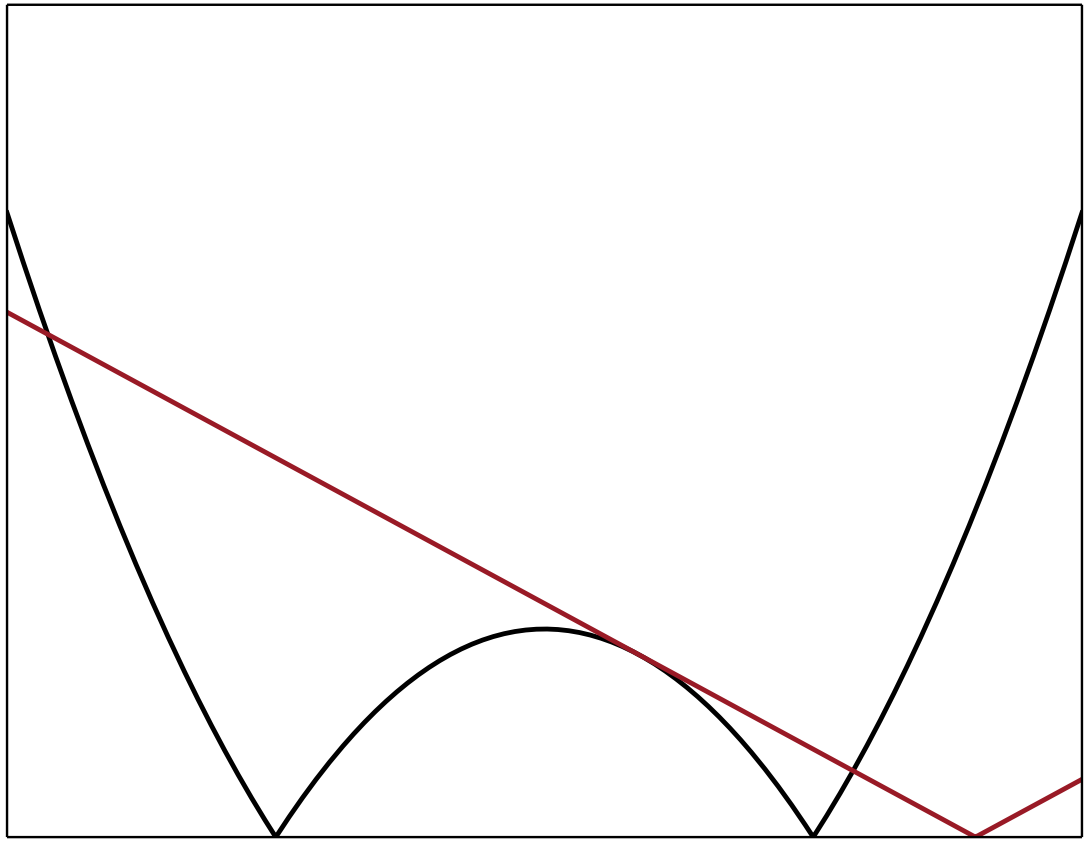


Modeling composite problems

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

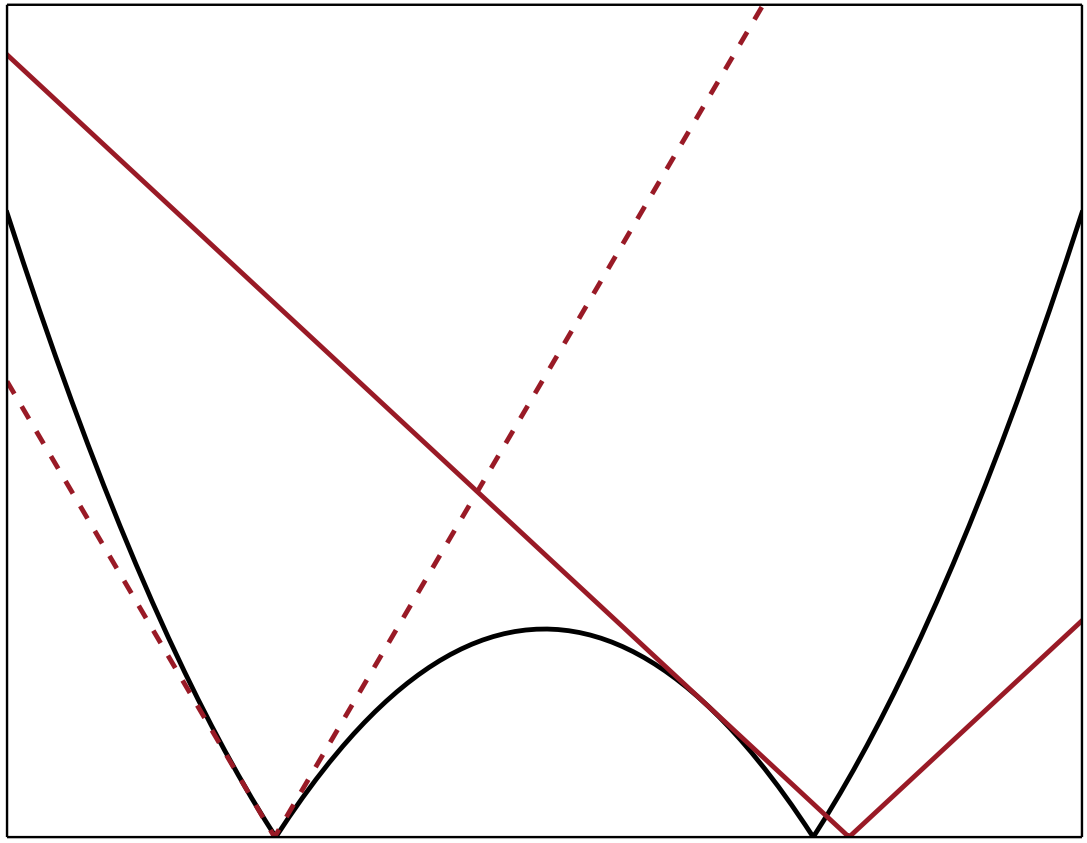


Modeling composite problems

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

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

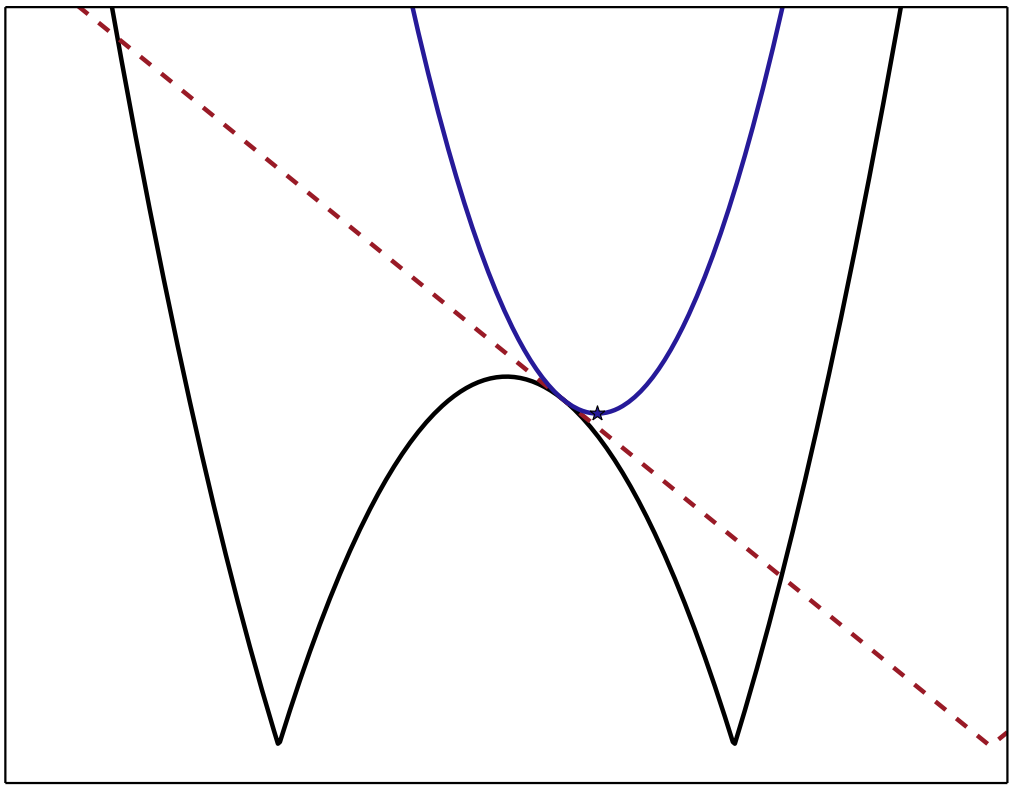
$$= \operatorname*{argmin}_{x \in X} \left\{ h\left(c(x_k) + \nabla c(x_k)^T (x - x_k)\right) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

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

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

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

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

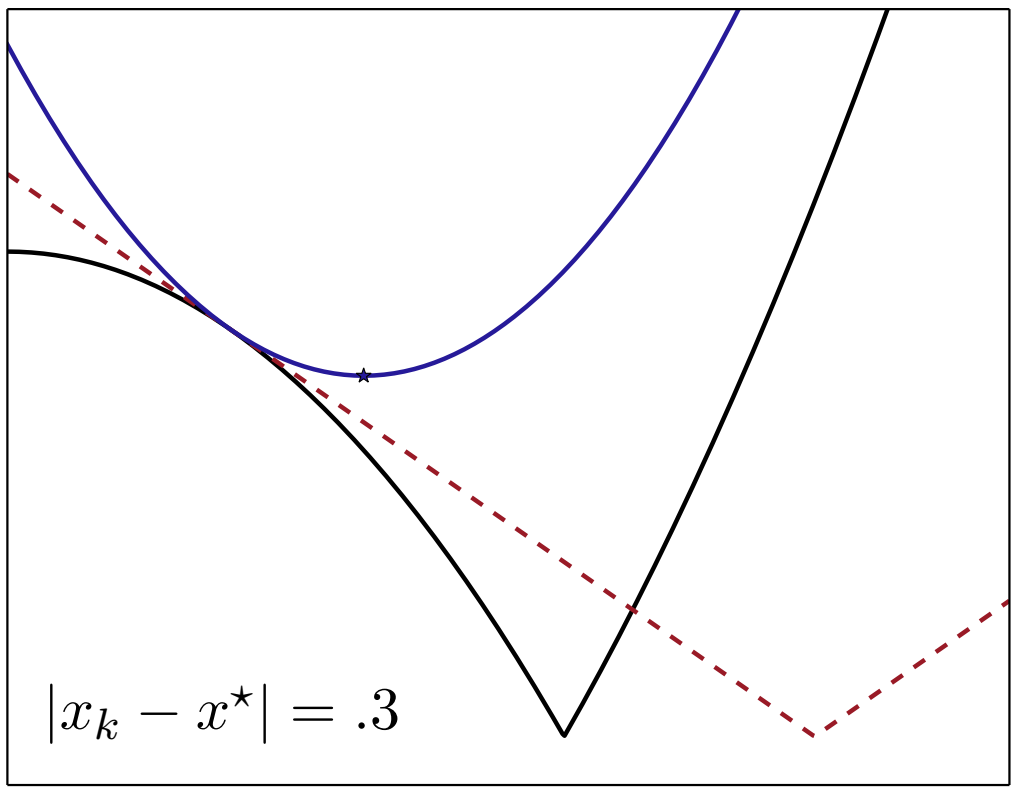


The prox-linear method [Burke, Drusvyatskiy et al.]

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

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

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

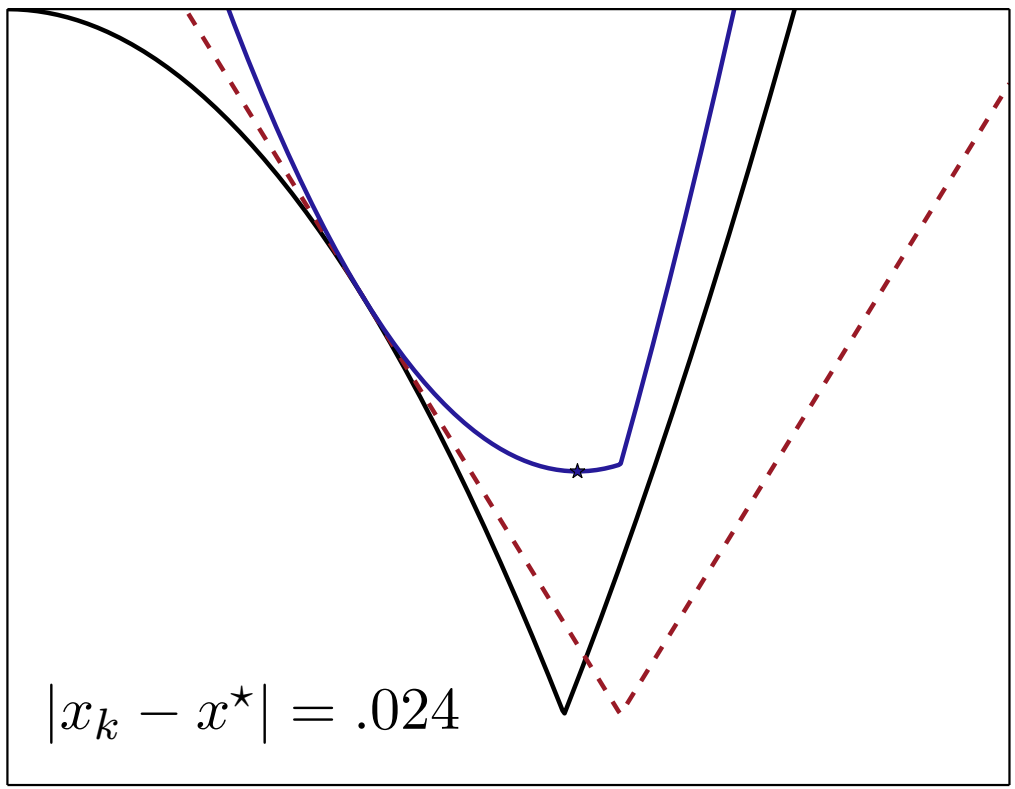


The prox-linear method [Burke, Drusvyatskiy et al.]

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

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

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

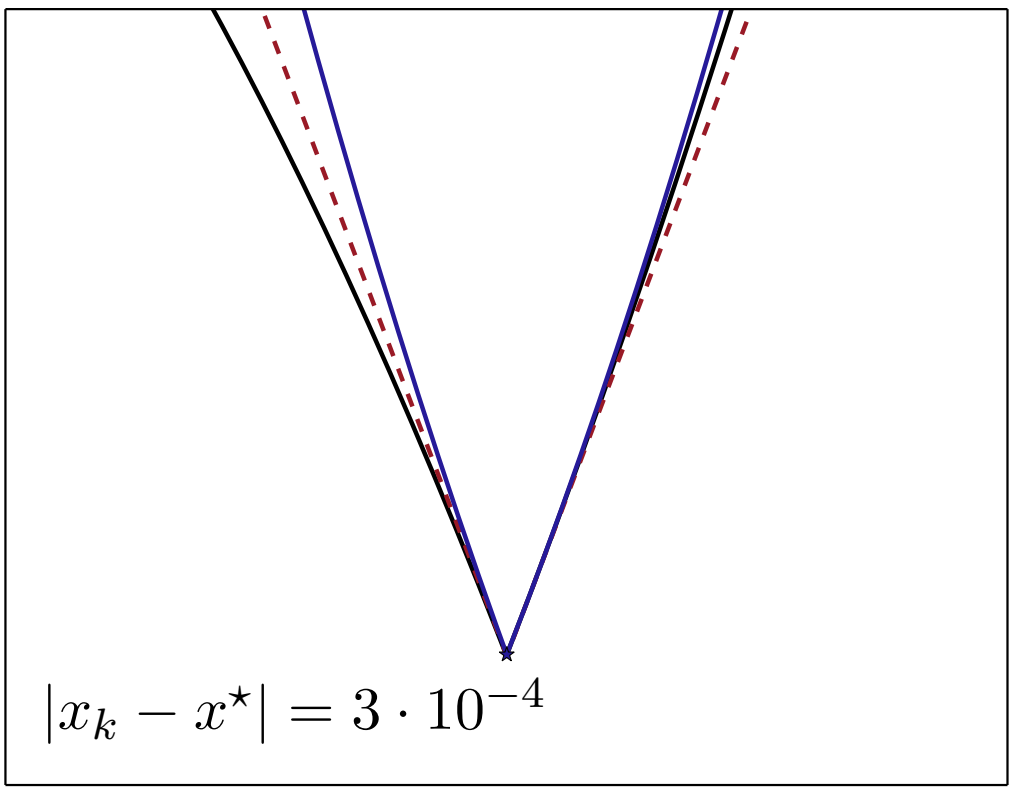


The prox-linear method [Burke, Drusvyatskiy et al.]

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

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

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

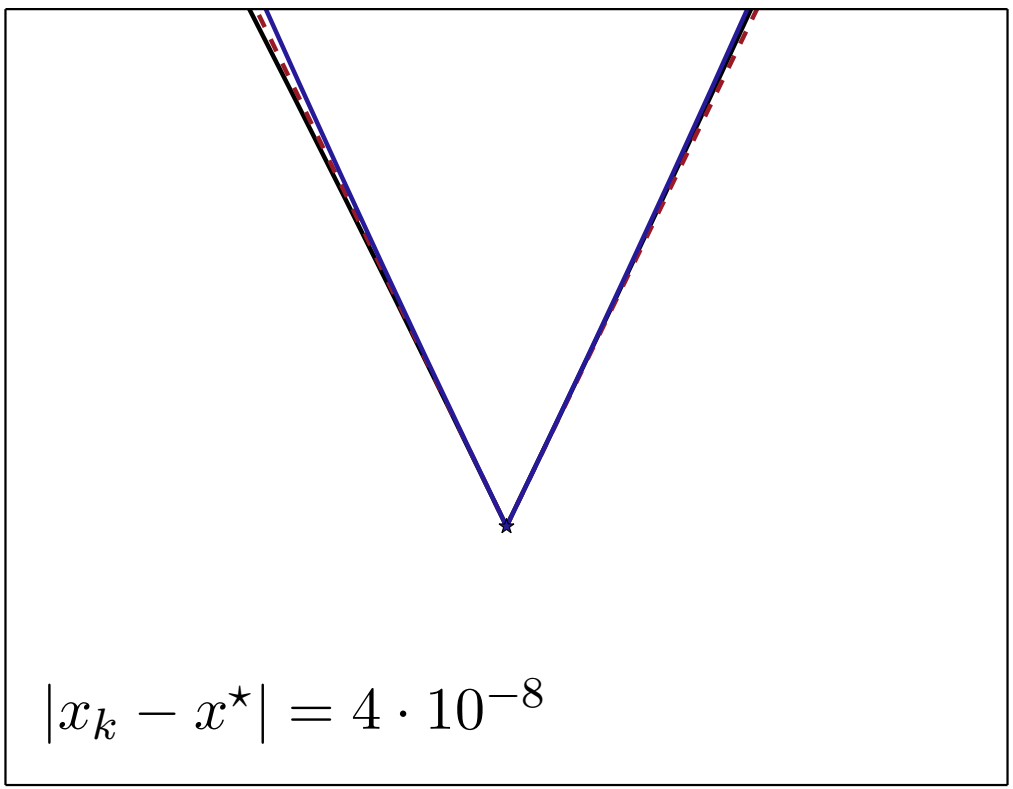


The prox-linear method [Burke, Drusvyatskiy et al.]

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

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

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



Robust phase retrieval problems

A nice application for these composite methods

Robust phase retrieval problems

Data model: true signal $x^\star \in \mathbb{R}^n$, for $p_{\rm fail} < \frac{1}{2}$ observe

$$b_i = \langle a_i, x^\star \rangle^2 + \xi_i \quad \text{where} \quad \xi_i = \begin{cases} 0 & \text{w.p. } \geq 1 - p_{\rm fail} \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

Robust phase retrieval problems

Data model: true signal $x^\star \in \mathbb{R}^n$, for $p_{\text{fail}} < \frac{1}{2}$ observe

$$b_i = \langle a_i, x^\star \rangle^2 + \xi_i \quad \text{where} \quad \xi_i = \begin{cases} 0 & \text{w.p.} \ \ \geq 1 - p_{\text{fail}} \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

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 $x^\star \in \mathbb{R}^n$, for $p_{\text{fail}} < \frac{1}{2}$ observe

$$b_i = \langle a_i, x^\star \rangle^2 + \xi_i \quad \text{where} \quad \xi_i = \begin{cases} 0 & \text{w.p.} \ \ \geq 1 - p_{\text{fail}} \\ \text{arbitrary} & \text{otherwise} \end{cases}$$

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,

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

A convergence theorem

Three key ingredients.

(1) Stability: $f(x) - f(x^\star) \geq \lambda \left\|x - x^\star\right\|_2 \left\|x + x^\star\right\|_2$

(2) Close models: $|f_x(y) - f(y)| \leq \frac{1}{m} \left\|\left|A^T A\right|\right\|_{\mathrm{op}} \|x - y\|_2^2$

(3) A good initialization

A convergence theorem

Three key ingredients.

(1) Stability: $f(x) - f(x^\star) \geq \lambda \left\| x - x^\star \right\|_2 \left\| x + x^\star \right\|_2$

(2) Close models: $|f_x(y) - f(y)| \leq \frac{1}{m} \left\| A^T A \right\|_{\text{op}} \|x - y\|_2^2$

(3) A good initialization

- Measurement matrix $A = [a_1 \ \cdots \ a_m]^T \in \mathbb{R}^{m \times n}$ and

$$\frac{1}{m} A^T A = \frac{1}{m} \sum_{i=1}^{m} a_i a_i^T$$

- Convex model f_x of f at x defined by

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

A convergence theorem

Three key ingredients.

(1) Stability: $f(x) - f(x^\star) \geq \lambda \left\|x - x^\star\right\|_2 \left\|x + x^\star\right\|_2$

(2) Close models: $|f_x(y) - f(y)| \leq \frac{1}{m} \left\|\left|A^T A\right|\right\|_{\text{op}} \|x - y\|_2^2$

(3) A good initialization

- Measurement matrix $A = [a_1 \ \cdots \ a_m]^T \in \mathbb{R}^{m \times n}$ and

$$\frac{1}{m} A^T A = \frac{1}{m} \sum_{i=1}^m a_i a_i^T$$

- Convex model f_x of f at x defined by

$$f_x(y) = \frac{1}{m} \sum_{i=1}^m \left| \langle a_i, x \rangle^2 + 2 \langle a_i, x \rangle \langle a_i, y - x \rangle \right|$$

A convergence theorem

Three key ingredients.

(1) Stability: $f(x) - f(x^\star) \geq \lambda \left\|x - x^\star\right\|_2 \left\|x + x^\star\right\|_2$

(2) Close models: $|f_x(y) - f(y)| \leq \frac{1}{m} \left\|\left|A^T A\right|\right\|_{\mathrm{op}} \|x - y\|_2^2$

(3) A good initialization

Theorem (D. & Ruan 17)

Define $\mathrm{dist}(x, x^\star) = \min\{\|x - x^\star\|_2, \|x + x^\star\|_2\}$. *Let* x_k *be generated by the prox-linear method and* $L = \frac{1}{m} \left\|\left|A^T A\right|\right\|_{\mathrm{op}}$. *Then*

$$\mathrm{dist}(x_k, x^\star) \leq \left(\frac{2L}{\lambda} \mathrm{dist}(x_0, x^\star)\right)^{2^k}.$$

Unpacking the convergence theorem

Theorem (D. & Ruan 17)

Define $\operatorname{dist}(x, x^\star) = \min\{\|x - x^\star\|_2, \|x + x^\star\|_2\}$. *Let* x_k *be generated by the prox-linear method and* $L = \frac{1}{m}\left\|\left|A^T A\right|\right\|_{\text{op}}$. *Then*

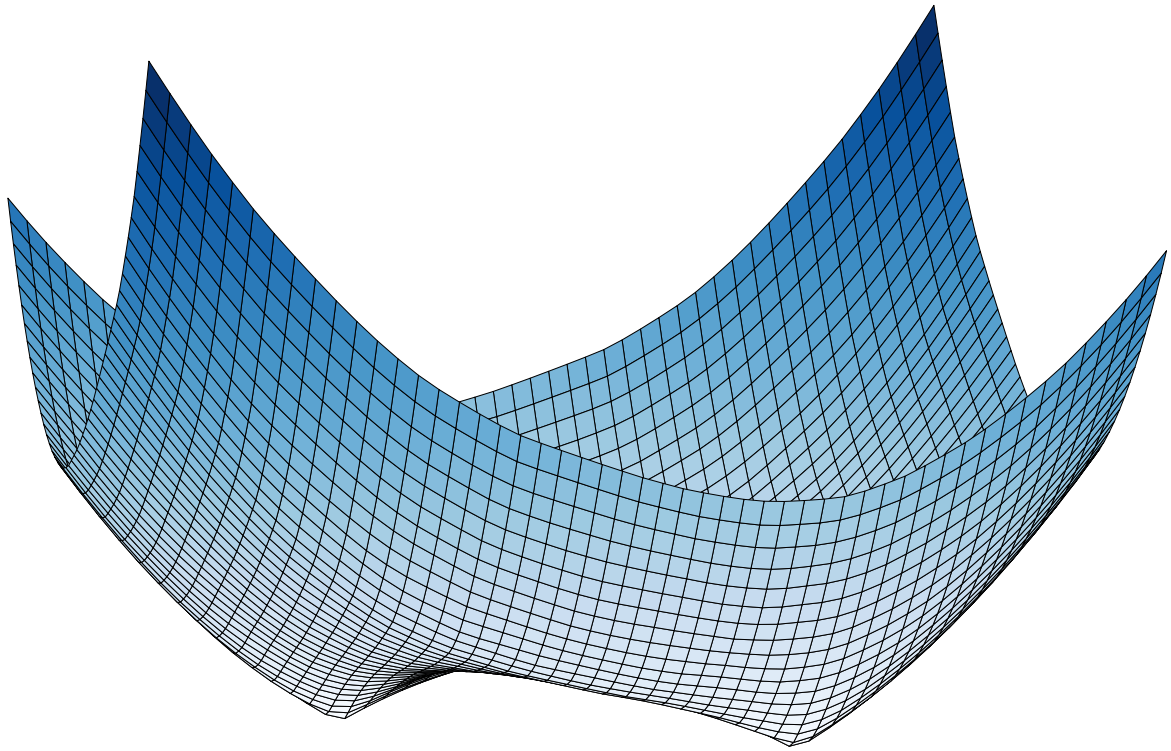
$$\operatorname{dist}(x_k, x^\star) \leq \left(\frac{2L}{\lambda}\operatorname{dist}(x_0, x^\star)\right)^{2^k}.$$

- Quadratic convergence: for all intents and purposes, 6 iterations
- Requires solving explicit convex optimization problems (quadratic programs) with *no* tuning parameters

Ingredients in convergence: stability

1. Stability: (cf. Eldar and Mendelson 14)

$$f(x) - f(x^\star) \geq \lambda \|x - x^\star\|_2 \|x + x^\star\|_2$$



Ingredients in convergence: stability

1. Stability: (cf. Eldar and Mendelson 14)

$$f(x) - f(x^\star) \geq \lambda \|x - x^\star\|_2 \|x + x^\star\|_2$$

What is necessary?

Proposition (D. & Ruan 17)

Assume uniformity condition: for all $u, v \in \mathbb{R}^n$ *and* $a \sim P$

$$P(|u^T a a^T v| \geq \epsilon_0 \|u\|_2 \|v\|_2) \geq c > 0.$$

Then f *is* $\frac{1}{2}\epsilon_0$*-stable with probability at least* $1 - e^{-cm}$.

Ingredients in convergence: stability

1. Stability: (cf. Eldar and Mendelson 14)

$$f(x) - f(x^\star) \geq \lambda \|x - x^\star\|_2 \|x + x^\star\|_2$$

What is necessary?

Proposition (D. & Ruan 17)

Assume uniformity condition: for all $u, v \in \mathbb{R}^n$ *and* $a \sim P$

$$P(|u^T a a^T v| \geq \epsilon_0 \|u\|_2 \|v\|_2) \geq c > 0.$$

Then f *is* $\frac{1}{2}\epsilon_0$*-stable with probability at least* $1 - e^{-cm}$.

(Gaussians satisfy this)

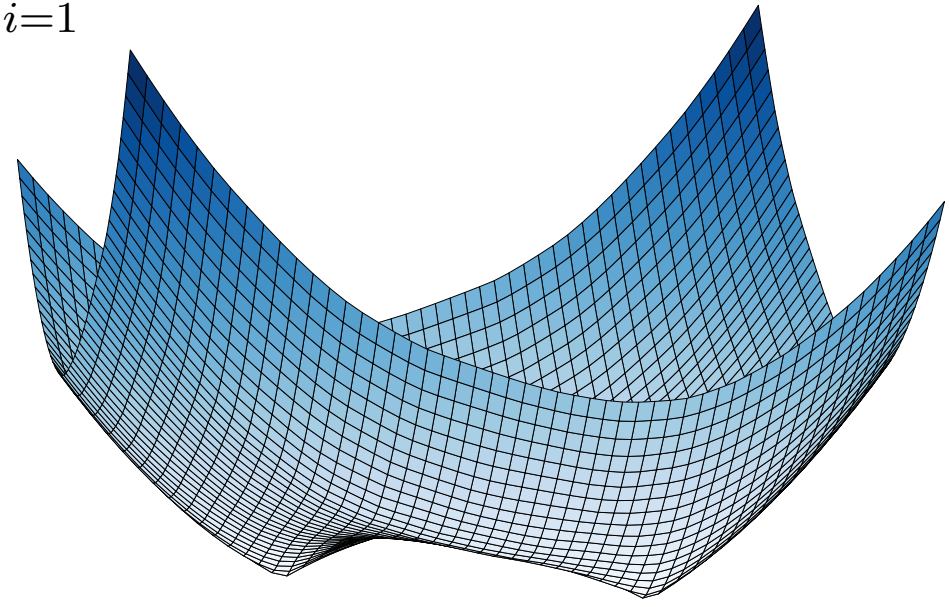
Ingredients in convergence: stability

Growth condition (stability):

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

and under random a_i with uniform enough support,

$$f(x) = \frac{1}{m}\sum_{i=1}^m \left|(x - x^\star)^T a_i a_i^T (x + x^\star)\right| \gtrsim \|x - x^\star\|_2 \|x + x^\star\|_2$$



Ingredients in convergence

2. Approximation: need $\frac{1}{m}\left\|\left\|A^TA\right\|\right\|_{\text{op}} = O(1)$

What is necessary?

Proposition (Vershynin 11)

If the measurement vectors a_i are sub-Gaussian, then

$$\frac{1}{m}\left\|\left\|A^TA\right\|\right\|_{\text{op}} \leq O(1)\cdot\sqrt{\frac{n}{m}+t} \quad \textit{w.p.} \ \geq 1-e^{-mt^2}.$$

Ingredients in convergence

2. Approximation: need $\frac{1}{m}\left\|\left\|A^TA\right\|\right\|_{\text{op}} = O(1)$

What is necessary?

Proposition (Vershynin 11)

If the measurement vectors a_i are sub-Gaussian, then

$$\frac{1}{m}\left\|\left\|A^TA\right\|\right\|_{\text{op}} \leq O(1)\cdot\sqrt{\frac{n}{m}+t} \quad \textit{w.p.} \ \geq 1-e^{-mt^2}.$$

Heavy-tailed data gets $\frac{1}{m}\left\|\left\|A^TA\right\|\right\|_{\text{op}} = O(1)$ with reasonable probability for m a bit larger

Ingredients in convergence: spectral initialization

Insight: [Wang, Giannakis, Eldar 16] Most vectors $a_i \in \mathbb{R}^n$ are orthogonal to $x^\star$

Ingredients in convergence: spectral initialization

Insight: [Wang, Giannakis, Eldar 16] Most vectors $a_i \in \mathbb{R}^n$ are orthogonal to $x^\star$

$$X^{\text{init}} := \sum_{i: b_i \leq \text{median}(b)} a_i a_i^T$$

satisfies

$$X^{\text{init}} \approx \mathbb{E}[a_i a_i^T] - c d^\star d^{\star T} \quad \text{where} \quad d^\star = x^\star / \|x^\star\|_2$$

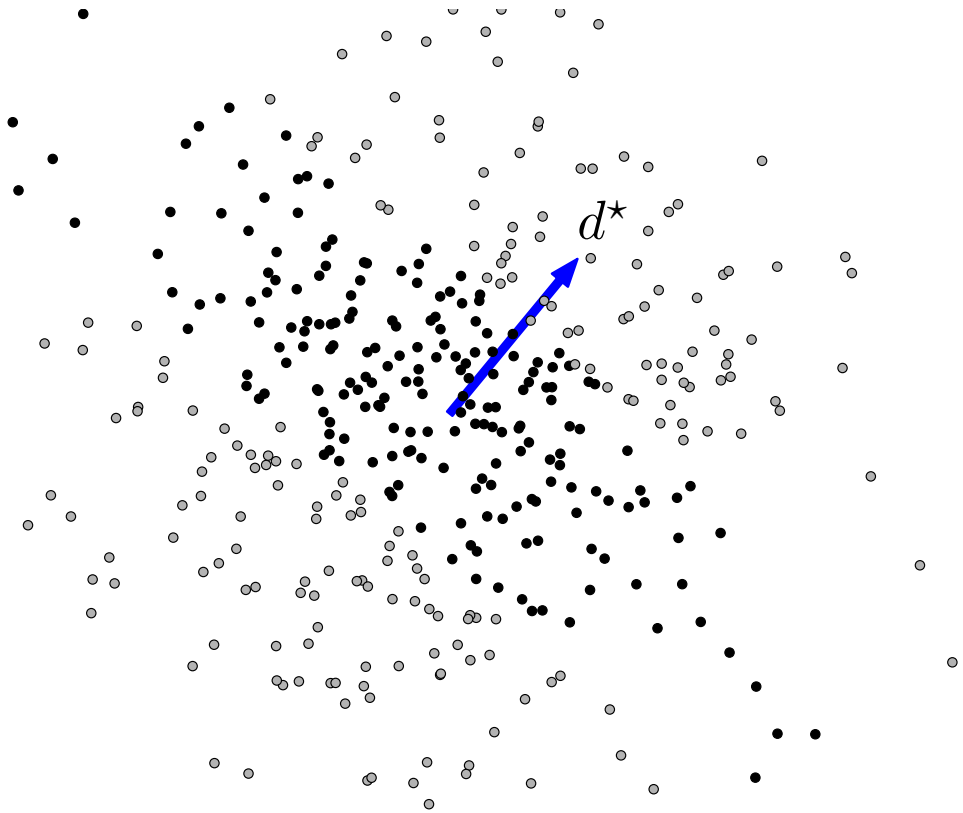
Ingredients in convergence: spectral initialization

Insight: [Wang, Giannakis, Eldar 16] Most vectors $a_i \in \mathbb{R}^n$ are orthogonal to $x^\star$

$$X^{\text{init}} := \sum_{i:b_i \le \text{median}(b)} a_i a_i^T$$

satisfies

$$X^{\text{init}} \approx \mathbb{E}[a_i a_i^T] - c d^\star d^{\star T} \quad \text{where} \quad d^\star = x^\star / \|x^\star\|_2$$



Ingredients in convergence: spectral initialization

3. Initialization: We need $\mathrm{dist}(x_0, x^\star) \lesssim \frac{1}{2} \|x^\star\|_2$

Ingredients in convergence: spectral initialization

3. Initialization: We need $\operatorname{dist}(x_0, x^\star) \lesssim \frac{1}{2} \|x^\star\|_2$

Estimate direction $\widehat{d} \approx x^\star / \|x^\star\|_2$ and radius $\widehat{r}$ by

$$X^{\text{init}} := \sum_{i : b_i \leq \text{median}(b)} a_i a_i^T \quad \text{and} \quad \widehat{d} = \operatorname*{argmin}_{d \in \mathbb{S}^{n-1}} \left\{ d^T X^{\text{init}} d \right\}$$

$$\widehat{r} := \left(\frac{1}{m} \sum_{i=1}^{m} b_i^2 \right)^{\frac{1}{2}} \approx \|x^\star\|_2$$

Ingredients in convergence: spectral initialization

3. Initialization: We need $\operatorname{dist}(x_0, x^\star) \lesssim \frac{1}{2} \|x^\star\|_2$

Estimate direction $\widehat{d} \approx x^\star / \|x^\star\|_2$ and radius $\widehat{r}$ by

$$X^{\text{init}} := \sum_{i : b_i \le \text{median}(b)} a_i a_i^T \quad \text{and} \quad \widehat{d} = \operatorname*{argmin}_{d \in \mathbb{S}^{n-1}} \left\{ d^T X^{\text{init}} d \right\}$$

$$\widehat{r} := \left(\frac{1}{m} \sum_{i=1}^{m} b_i^2 \right)^{\frac{1}{2}} \approx \|x^\star\|_2$$

Proposition (D. & Ruan 17)

Under appropriate orthogonality conditions, $x_0 = \widehat{r}\widehat{d}$ *satisfies*

$$\operatorname{dist}(x_0, x^\star) \lesssim \sqrt{\frac{n}{m} + t}$$

with probability at least $1 - e^{-mt^2}$

Take-home result

- Stability: measurements a_i are uniform enough in direction
- Closeness: a_i are sub-Gaussian or normalized
- Sufficient conditions for initialization: for $v \in \mathbb{S}^n$,

$$\mathbb{E}[a_i a_i^T \mid \langle a_i, v\rangle^2 \le \|v\|_2^2] = I_n - cvv^T + E$$

  where $c > 0$ and E is a small error
- Measurement failure probability $p_{\text{fail}} \le \frac{1}{4}$

Theorem (D. & Ruan 17)

If these conditions hold and $m/n \gtrsim 1$, then the spectral initialization succeeds and iterates x_k of prox-linear algorithm satisfy

$$\operatorname{dist}(x_k, x_0) \le (O(1) \cdot \operatorname{dist}(x_0, x^\star))^{2^k}$$

Experiments

1. Random (Gaussian) measurements
2. Adversarially chosen outliers
3. Real images

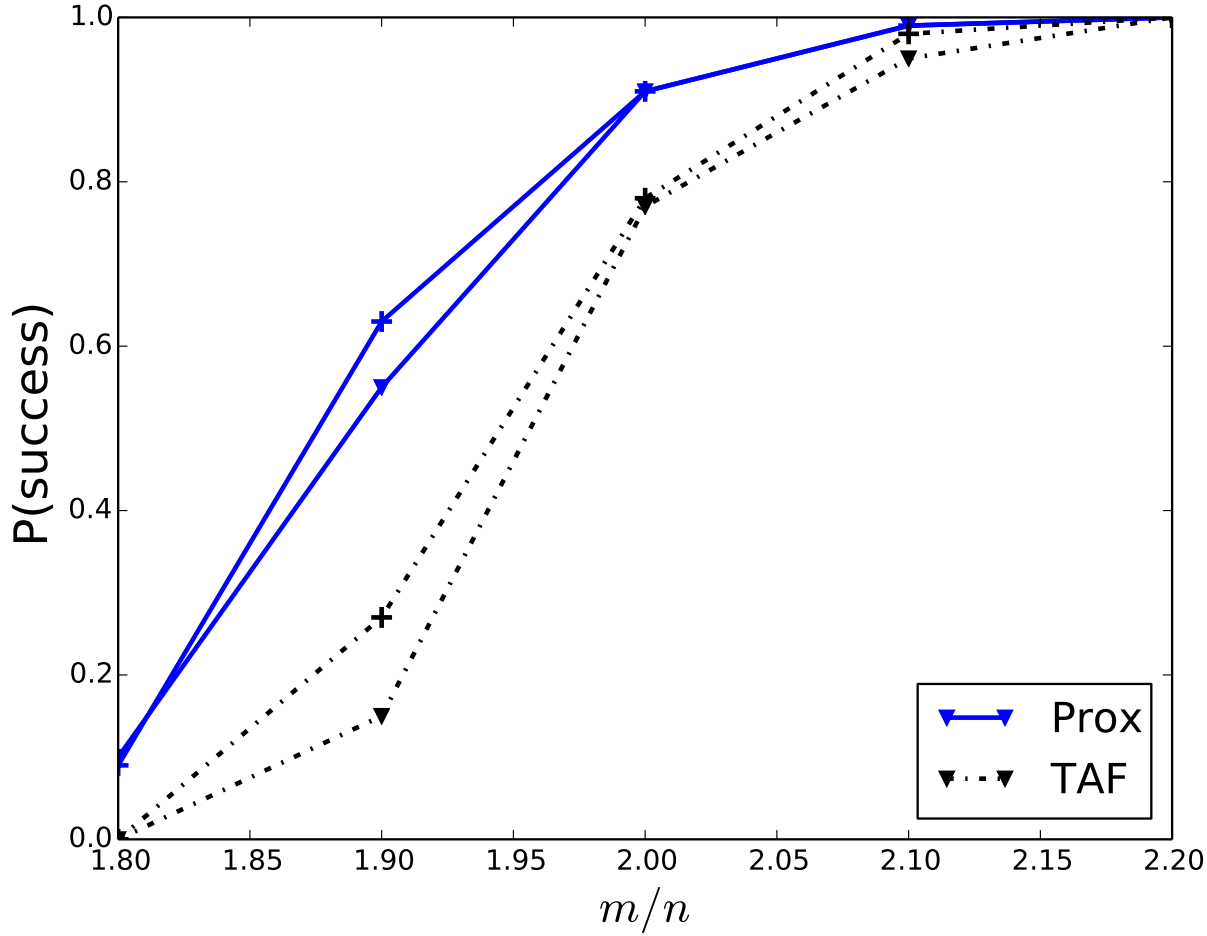
Experiment 1: random Gaussian measurements

- Data generation: dimension $n = 3000$,

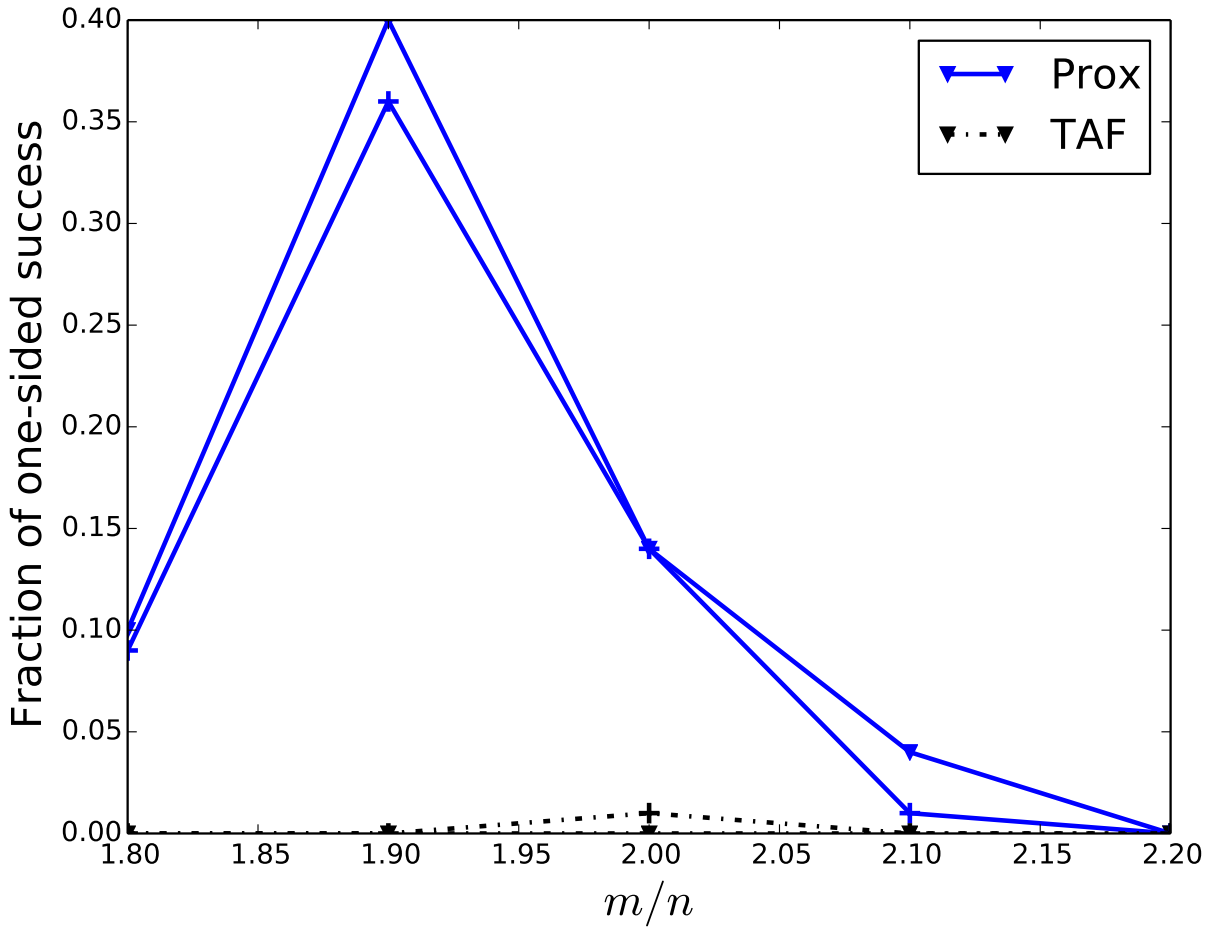
$$a_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I_n) \quad \text{and} \quad b_i = \langle a_i, x^\star \rangle^2$$

- Compare to Wang, Giannakis, Eldar's Truncated Amplitude Flow (best performing non-convex approach)
- Look at success probability against m/n (note that $m \geq 2n - 1$ is necessary for injectivity)

Experiment 1: random Gaussian measurements



Experiment 1: random Gaussian measurements



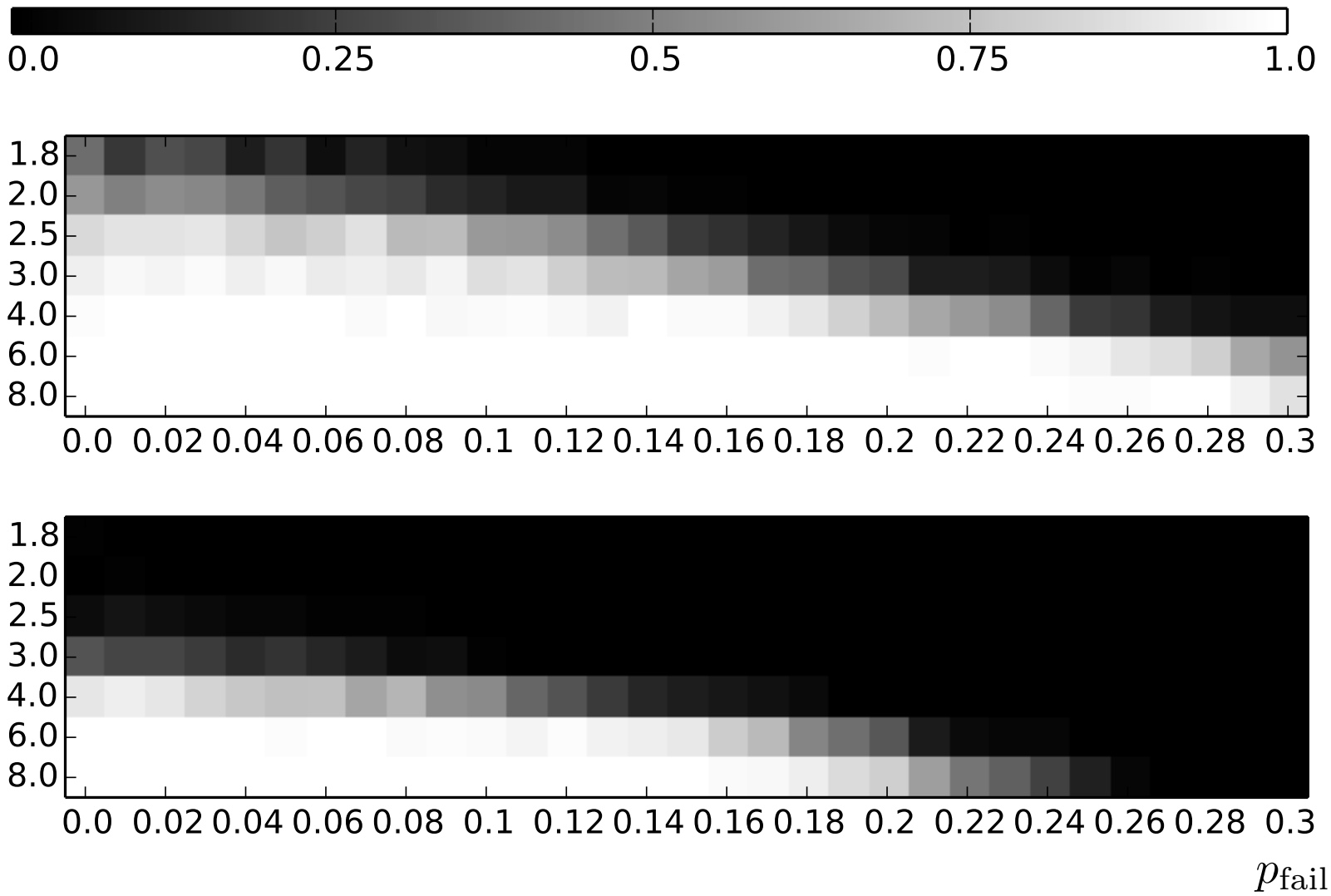
Experiment 2: corrupted measurements

- 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 2: corrupted measurements



Experiment 3: digit recovery

- Data generation: handwritten 16×16 grayscale digits, sensing matrix

$$A = \begin{bmatrix} H_n S_1 \\ H_n S_2 \\ H_n S_3 \end{bmatrix} \in \mathbb{R}^{3n \times n}$$

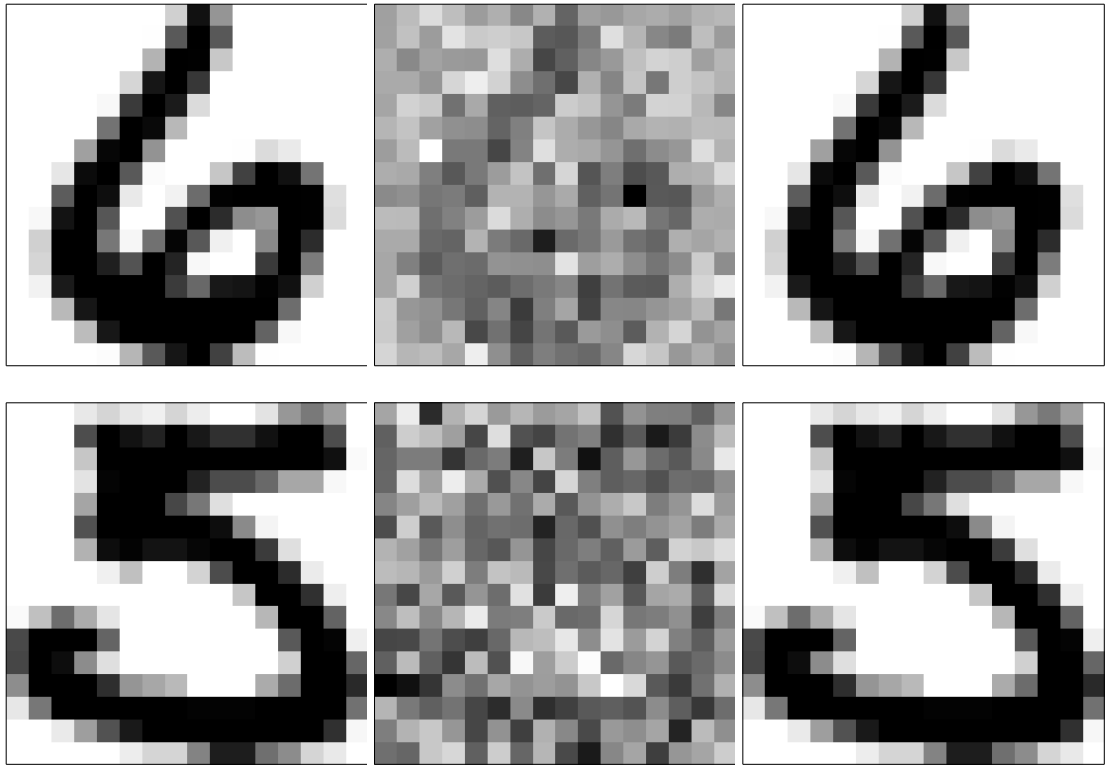
  where $n = 256$, S_l are diagonal random sign matrices, H_n is Hadamard transform matrix

- Observe

$$b = (Ax^\star)^2 + \xi \quad \text{where} \quad \xi_i = \begin{cases} 0 & \text{w.p. } 1 - p_{\text{fail}} \\ \text{Cauchy} & \text{otherwise} \end{cases}$$

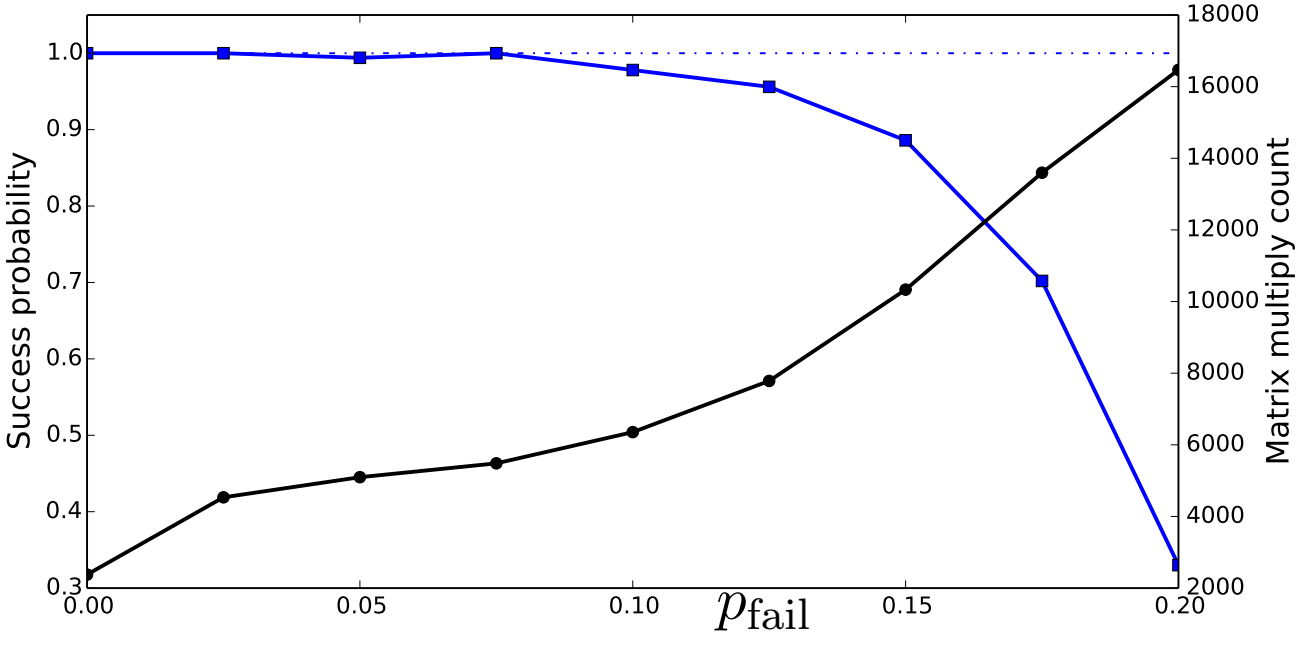
- Other non-convex approaches designed for Gaussian data; unclear how to parameterize them

Experiment 3: digit recovery



Left: true image. Middle: spectral initialization. Right: solution.

Experiment 3: digit recovery



Performance of composite optimization scheme versus failure probability

Experiment 4: real images

Signal size $n = 2^{22}$, measurements $m = 3 \cdot 2^{24}$

Experiment 4: real images

Signal size $n = 2^{22}$, measurements $m = 3 \cdot 2^{24}$

Composite optimization at scale

Question: What if we have composite problems with a *really big* sample?

Composite optimization at scale

Question: What if we have composite problems with a *really big* sample?

- Typical stochastic optimization setup,

$$f(x) = \mathbb{E}[F(x; S)] \quad \text{where} \quad F(x; S) = h(c(x; S); S)$$

Composite optimization at scale

Question: What if we have composite problems with a *really big* sample?

- Typical stochastic optimization setup,

$$f(x) = \mathbb{E}[F(x; S)] \quad \text{where} \quad F(x; S) = h(c(x; S); S)$$

- Example: large scale (robust) nonlinear regression

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

A stochastic composite method

- Define (random) convex approximation

$$F_x(y; s) = h(c(x; s) + \nabla c(x; s)^T (y - x); s)$$

A stochastic composite method

- Define (random) convex approximation

$$F_x(y;s) = h(\underbrace{c(x;s) + \nabla c(x;s)^T(y-x)}_{\approx c(y;s)};s)$$

A stochastic composite method

- Define (random) convex approximation

$$F_x(y;s) = h(\underbrace{c(x;s) + \nabla c(x;s)^T(y-x)}_{\approx c(y;s)};s)$$

- Then iterate for $k = 1, 2, \ldots$

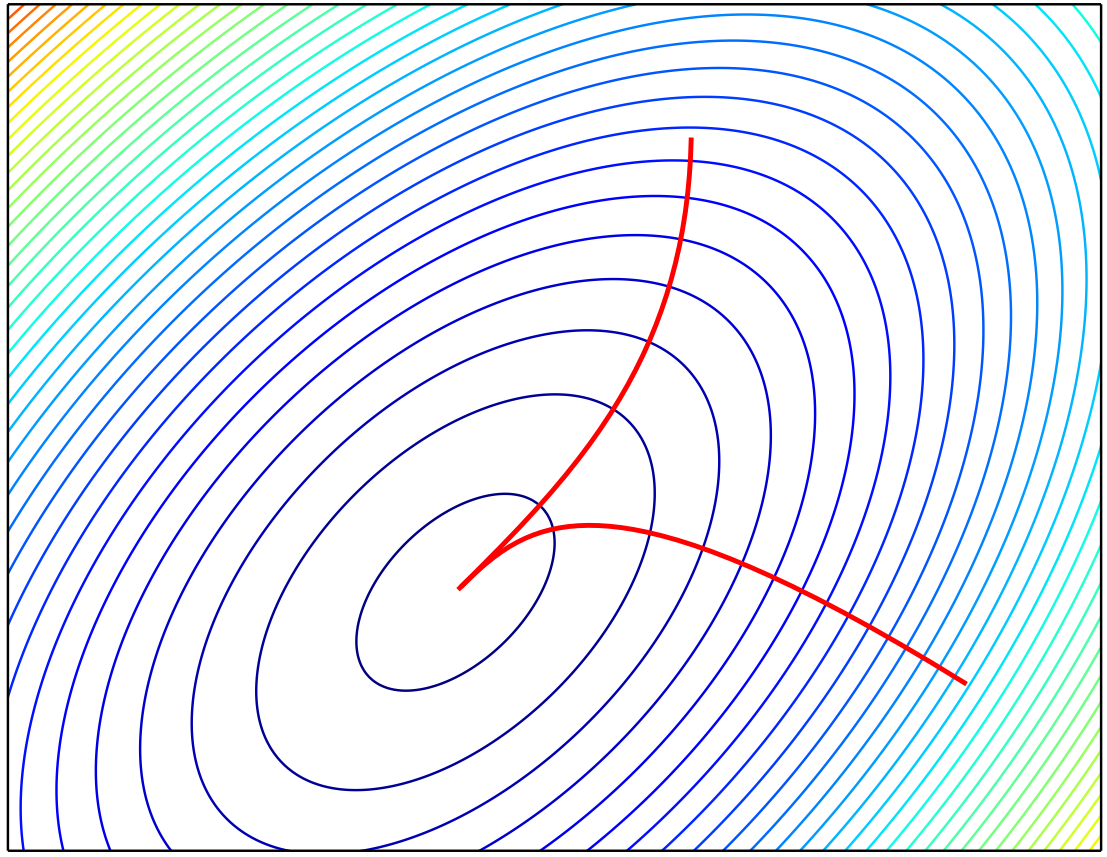
$$S_k \overset{\text{iid}}{\sim} P$$

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

Understanding convergence behavior

Ordinary differential equations (gradient flow):

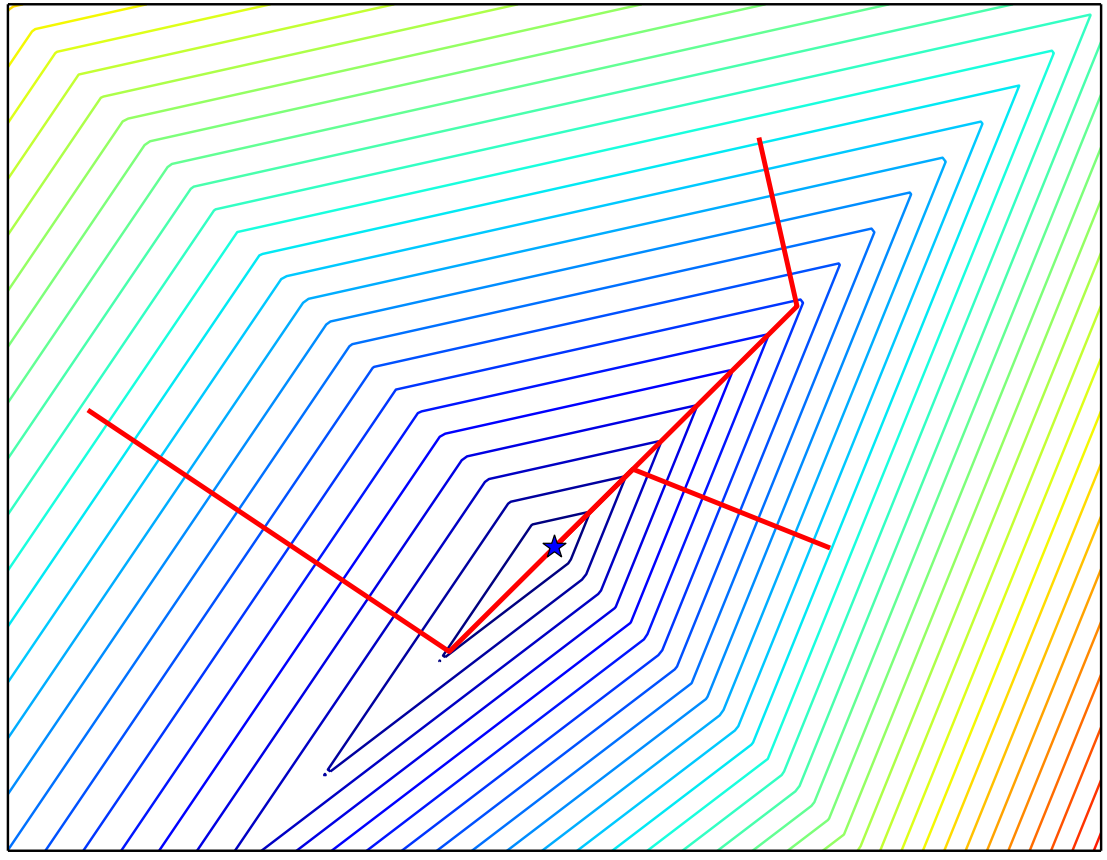
$$\dot{x} = -\nabla f(x) \quad \text{i.e.} \quad \frac{d}{dt}x(t) = -\nabla f(x(t))$$



Understanding convergence behavior

Ordinary differential *inclusions* (subgradient flow):

$$\dot{x} \in -\partial f(x) \quad \text{i.e.} \quad \frac{d}{dt}x(t) \in -\partial f(x(t))$$



The differential inclusion

For stochastic function

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

the *generalized* subgradient (for non-convex, non-smooth) is [D. & Ruan 17]

$$\partial f(x) = \int \nabla c(x;s)\partial h(c(x;s);s)dP(s)$$

Theorem (D. & Ruan 17)

For stochastic composite problem, the subdifferential inclusion $\dot{x} \in -\partial f(x)$ *has a unique trajectory for all time and*

$$f(x(t)) - f(x(0)) \leq -\int_0^t \|\partial f(x(\tau))\|^2 \, d\tau.$$

It also has limit points and they are stationary.

The limiting differential inclusion

Recall our iteration

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

Optimality conditions: using $F_x(y; s) = h(c(x; s) + \nabla c(x; s)^T (y - x))$,

The limiting differential inclusion

Recall our iteration

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

Optimality conditions: using $F_x(y; s) = h(c(x; s) + \nabla c(x; s)^T (y - x))$,

$$0 \in \nabla c(x_k; s) \partial h(c(x_k; s) + \nabla c(x_k; s)^T (x_{k+1} - x_k)) + \frac{1}{\alpha_k} [x_{k+1} - x_k]$$

The limiting differential inclusion

Recall our iteration

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

Optimality conditions: using $F_x(y; s) = h(c(x; s) + \nabla c(x; s)^T (y - x))$,

$$0 \in \nabla c(x_k; s) \partial h(\underbrace{c(x_k; s) + \nabla c(x_k; s)^T (x_{k+1} - x_k)}_{= c(x_k;s) \pm O(\|x_k - x_{k+1}\|^2)}) + \frac{1}{\alpha_k} [x_{k+1} - x_k]$$

The limiting differential inclusion

Recall our iteration

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

Optimality conditions: using $F_x(y; s) = h(c(x; s) + \nabla c(x; s)^T (y - x))$,

$$0 \in \nabla c(x_k; s) \partial h(\underbrace{c(x_k; s) + \nabla c(x_k; s)^T (x_{k+1} - x_k)}_{= c(x_k;s) \pm O(\|x_k - x_{k+1}\|^2)}) + \frac{1}{\alpha_k} [x_{k+1} - x_k]$$

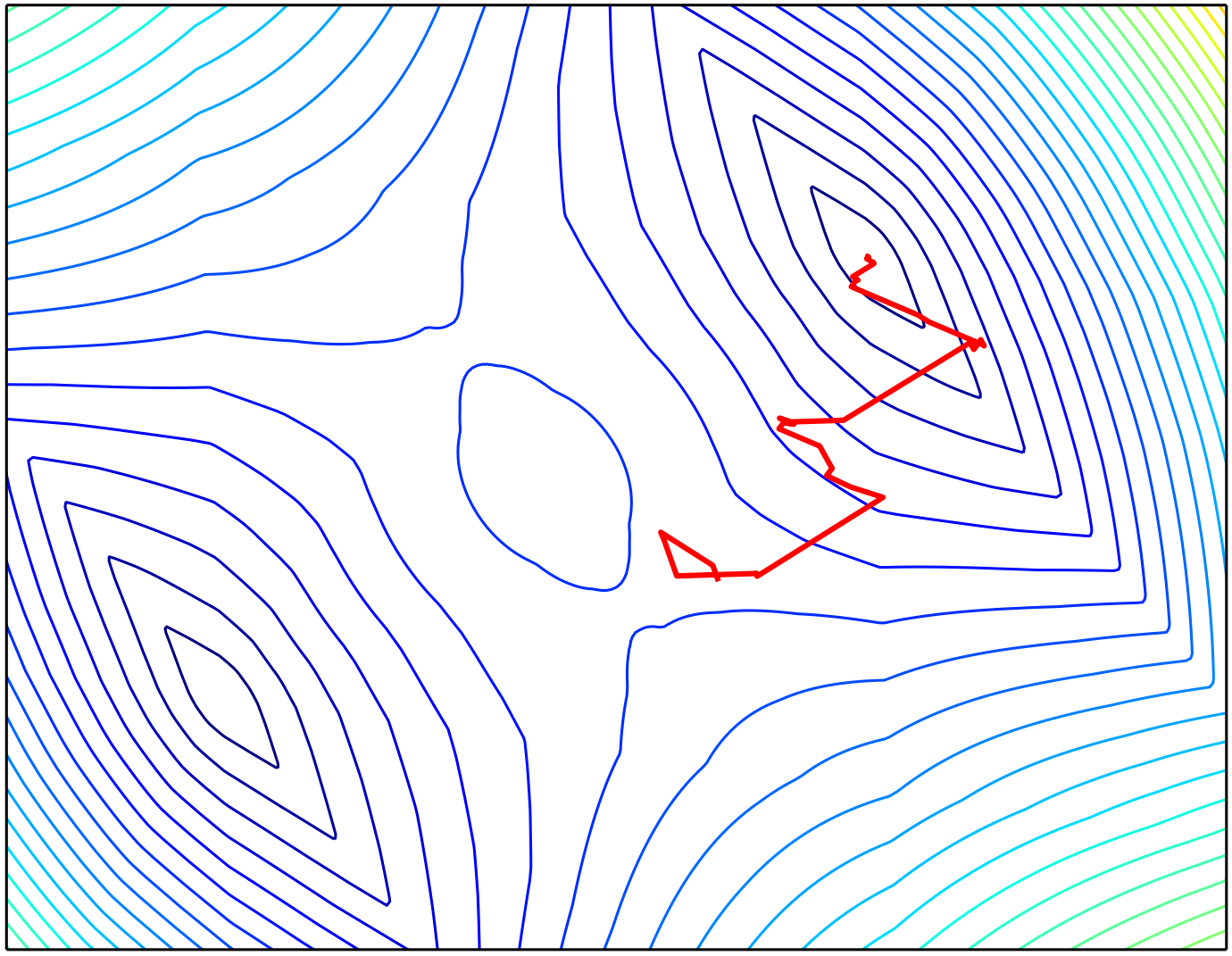
i.e.

$$\frac{1}{\alpha_k} [x_{k+1} - x_k] \in -\nabla c(x_k; s) \partial h(c(x_k; s); s) + \text{subgradient mess} + \text{Noise}$$

$$= -\partial f(x_k) + \text{subgradient mess} + \text{Noise}$$

Graphical example

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



A convergence guarantee

Consider the stochatsic composite optimization problem

$$\underset{x\in X}{\text{minimize}} \;\; f(x) := \mathbb{E}[F(x;S)] \;\; \text{where} \;\; F(x;s) = h(c(x;s);s).$$

Use the iteration

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

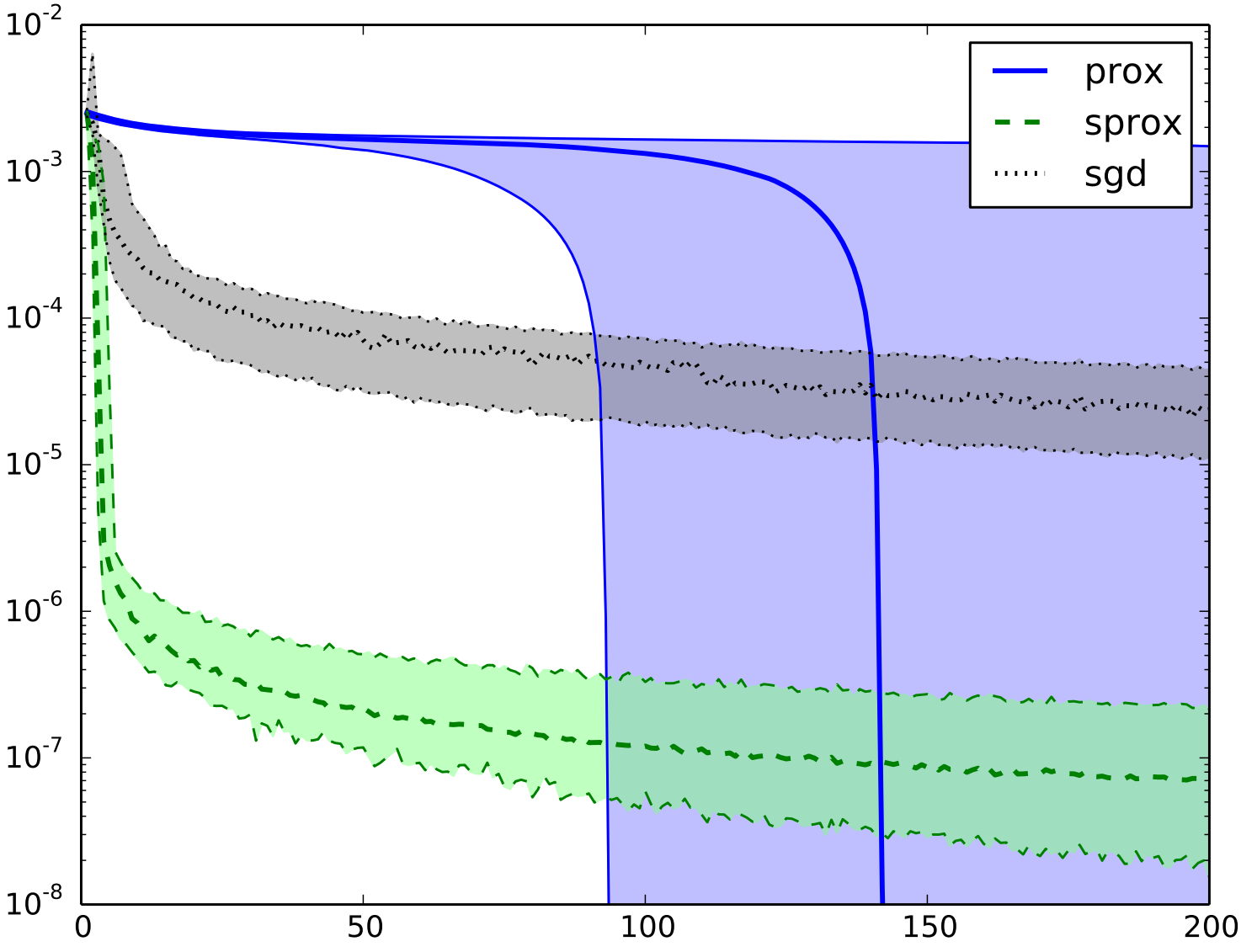
Theorem (D. & Ruan 17)

Assume X is compact and $\sum_{k=1}^{\infty} \alpha_k = \infty$, $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. Then the sequence $\{x_k\}$ satisfies

(1) *$f(x_k)$ converges*

(2) *All cluster points of x_k are stationary*

Experiment: noiseless phase retrieval



Conclusions

1. Broadly interesting structures for *non-convex* problems that are still approximable
2. Statistical modeling allows solution of non-trivial, non-smooth, non-convex problems
3. Large scale efficient methods still important

Conclusions

1. Broadly interesting structures for *non-convex* problems that are still approximable
2. Statistical modeling allows solution of non-trivial, non-smooth, non-convex problems
3. Large scale efficient methods still important

References

- *Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval* arXiv:1705.02356
- *Stochastic Methods for Composite Optimization Problems* arXiv:1703.08570