# COMP 545: Advanced topics in optimization
# From simple to complex ML systems

Lecture 5

# Overview

- In the past lectures, we:
    - Talked a little bit about **general smooth optimization** problems
    - This included some non-convex optimization, but mostly convex
    - The discussion was quite abstract (no particular application)

# Overview

- In the past lectures, we:
  - Talked a little bit about **general smooth optimization** problems
  - This included some non-convex optimization, but mostly convex
  - The discussion was quite abstract (no particular application)
- For the next 2-3 lectures, we will consider (possibly) the simplest non-convex setting: **sparse model selection**
  - We will provide motivation, background and alternative solutions
  - We will focus on how we can **provably and efficiently solve** such problems

# Overview

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{C} \end{aligned}$$

# Overview

We will consider convex objectives..

$$\min_{x} \quad f(x)$$

$$\text{s.t.} \quad x \in \mathcal{C}$$

..over non-convex constraints

# Overview

We will consider convex objectives..

$$\min_{x} \quad f(x)$$

..over non-convex constraints

$$\text{s.t.} \quad x \in \mathcal{C}$$

- We will focus on the cases of (structured) sparsity and low-rankness

(But I'm open to other alternatives as we proceed)

# Sparse linear regression

(Not again man!..)

- Generative model: $y_i = a_i^\top x^\star + w_i$
  - $a_i \in \mathbb{R}^p$ : features
  - $y_i \in \mathbb{R}$ : responses
  - $w_i \in \mathbb{R}$ : additive noise

# Sparse linear regression

(Not again man!..)

- Generative model: $y_i = a_i^\top x^\star + w_i$
  - $a_i \in \mathbb{R}^p$ : features
  - $y_i \in \mathbb{R}$ : responses
  - $w_i \in \mathbb{R}$ : additive noise
- Generative prior: $x^\star \in \mathbb{R}^p$ is $k$-sparse: $\|x^\star\|_0 = k,\ k \ll p$

# Sparse linear regression

(Not again man!..)

- Generative model: $y_i = a_i^\top x^\star + w_i$
  - $a_i \in \mathbb{R}^p$ : features
  - $y_i \in \mathbb{R}$ : responses
  - $w_i \in \mathbb{R}$ : additive noise
- Generative prior: $x^\star \in \mathbb{R}^p$ is $k$-sparse: $\|x^\star\|_0 = k,\ k \ll p$
- Assuming data set $\{y_i, a_i\}_{i=1}^n$, $n < p$, find $x^\star \in \mathbb{R}^p$

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \le k \end{aligned}$$

# Sparse linear regression

(Not again man!..)

- Generative model: $y_i = a_i^\top x^\star + w_i$
  - $a_i \in \mathbb{R}^p$ : features
  - $y_i \in \mathbb{R}$ : responses
  - $w_i \in \mathbb{R}$ : additive noise
- Generative prior: $x^\star \in \mathbb{R}^p$ is $k$-sparse: $\|x^\star\|_0 = k,\ k \ll p$
- Assuming data set $\{y_i, a_i\}_{i=1}^n$, $n < p$, find $x^\star \in \mathbb{R}^p$

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \le k \end{aligned}$$

- Any suggestions how to solve this?

# Sparse linear regression

(Not again man!..)

# Sparse linear regression

(Not again man!..)

– Solution #1: convexification + proj. gradient descent

**LASSO**

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2$$
$$\text{s.t.} \quad \|x\|_1 \leq \lambda$$

$$\longrightarrow \quad x_{t+1} = \Pi_{\|\cdot\|_1 \leq \lambda}\left(x_t - \eta \nabla f(x_t)\right)$$

(Pros & Cons?)

# Sparse linear regression

(Not again man!..)

- Solution #1: convexification + proj. gradient descent

**LASSO**

$$\min_{x\in\mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \qquad \text{s.t.} \quad \|x\|_1 \le \lambda$$

$$\longrightarrow \quad x_{t+1} = \Pi_{\|\cdot\|_1\le\lambda}\left(x_t - \eta\nabla f(x_t)\right)$$

(Pros & Cons?)

- Solution #2: convexification + **proximal** gradient descent

**Basis pursuit (denoising)**

$$\min_{x\in\mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 + \rho\|x\|_1 \quad \longrightarrow \quad x_{t+1} = \mathrm{Prox}_{\rho\|\cdot\|_1}\left(x_t - \eta\nabla f(x_t)\right)$$

(Pros & Cons?)

# Sparse linear regression

(Not again man!..)

- Solution #1: convexification + proj. gradient descent

**LASSO**

$$\min_{x\in\mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \qquad \text{s.t.} \quad \|x\|_1 \le \lambda$$

$$\longrightarrow \quad x_{t+1} = \Pi_{\|\cdot\|_1 \le \lambda}\left(x_t - \eta \nabla f(x_t)\right)$$

(Pros & Cons?)

- Solution #2: convexification + **proximal** gradient descent

**Basis pursuit (denoising)**

$$\min_{x\in\mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 + \rho\|x\|_1 \quad \longrightarrow \quad x_{t+1} = \mathrm{Prox}_{\rho\|\cdot\|_1}\left(x_t - \eta \nabla f(x_t)\right)$$

(Pros & Cons?)

- Solution #3: keep non-convexity + **non-convex projected gradient descent**

**Hard-thresholding**

$$\min_{x\in\mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \qquad \text{s.t.} \quad \|x\|_0 \le k$$

$$\longrightarrow \quad x_{t+1} = H_k\left(x_t - \eta \nabla f(x_t)\right)$$

(Pros & Cons?)

# But before we proceed..

- Some questions:

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \leq k \end{aligned}$$

# But before we proceed..

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \leq k \end{aligned}$$

- Some questions:
  - Q: "How easy it is to solve $\ell_0$-pseudo norm problems?"
  - A: "Sparsity makes problems exponentially hard to solve"

(This assumes the most general case)

# But before we proceed..

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \frac{1}{2}\|y - Ax\|_2^2$$
$$\text{s.t.} \quad \|x\|_0 \leq k$$

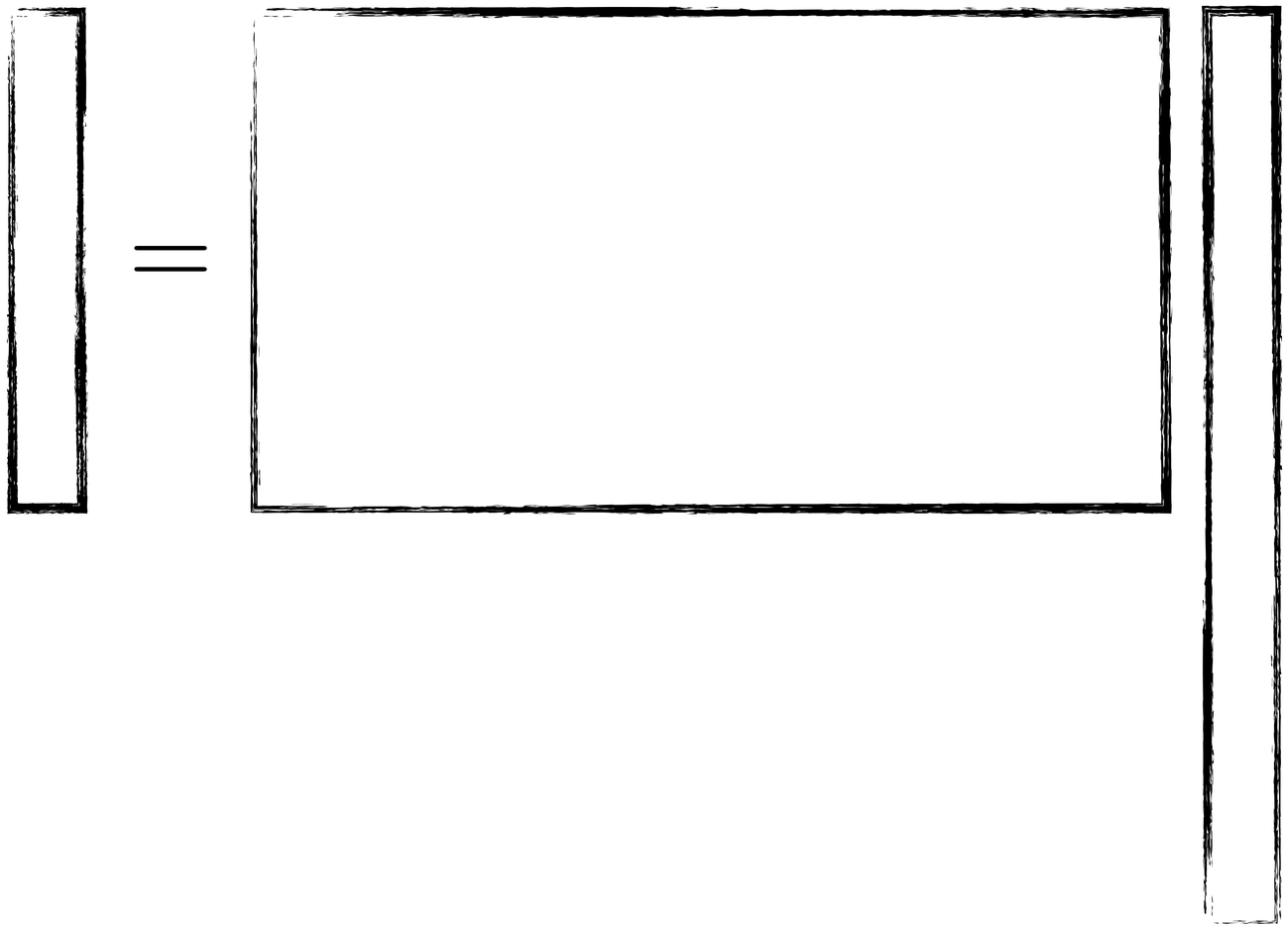
- Some questions:
  - Q: "How easy it is to solve $\ell_0$-pseudo norm problems?"
  - A: "Sparsity makes problems exponentially hard to solve"
    (This assumes the most general case)
  - Q: "But isn't the problem underdetermined? $(n \ll p)$"
  - A: "Yes, without any constraints, the problem has infinite solutions"

# But before we proceed..

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \tfrac{1}{2}\|y - Ax\|_2^2$$
$$\text{s.t.} \quad \|x\|_0 \leq k$$

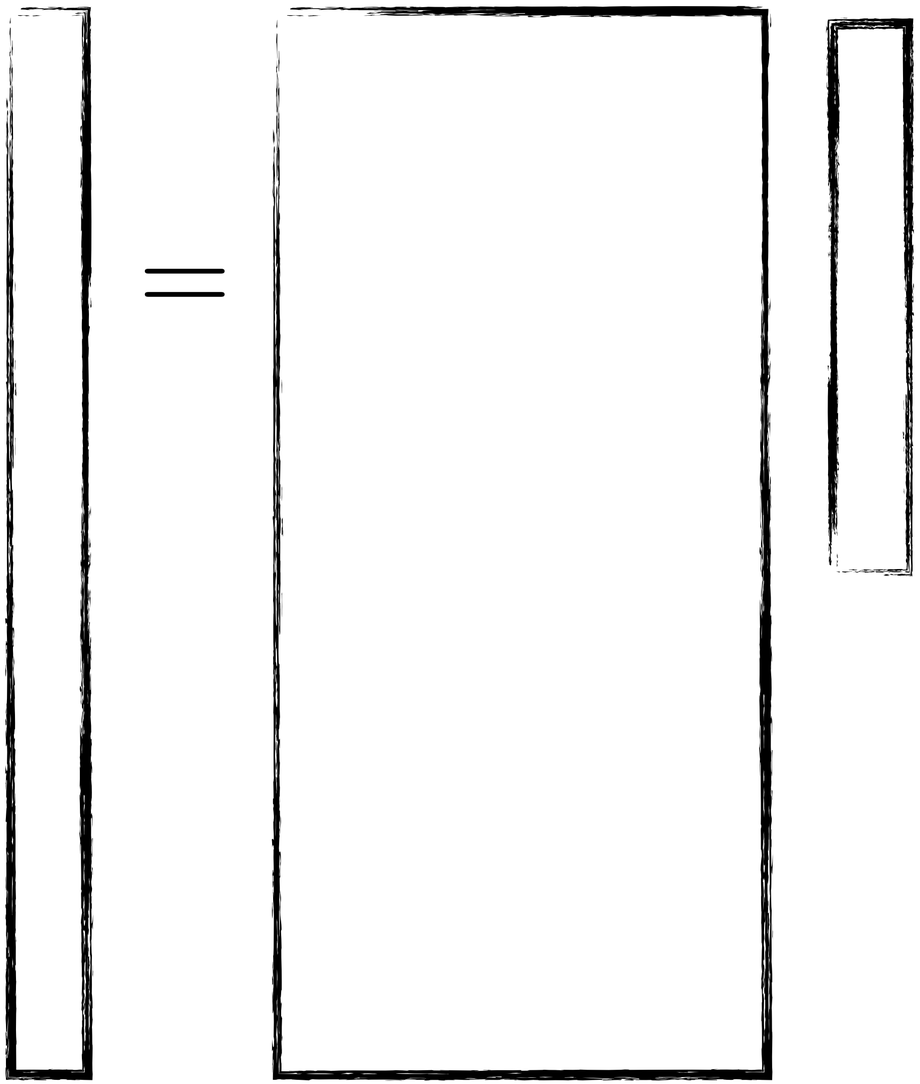
- Some questions:
  - Q: "How easy it is to solve $\ell_0$-pseudo norm problems?"
  - A: "Sparsity makes problems exponentially hard to solve"
    (This assumes the most general case)
  - Q: "But isn't the problem underdetermined? $(n \ll p)$"
  - A: "Yes, without any constraints, the problem has infinite solutions"
  - Q: "Why then do we have hopes solving this problem?"
  - A: "Under assumptions on $A$, and the relation between $(n, p, k)$, we Will see that on average this problem can be solved in polynomial time"

# Over- vs. under-parameterized



Over-parameterized

# Over- vs. under-parameterized



Under-parameterized

# Iterative hard thresholding (IHT)

(It is just projected gradient descent on sparsity constraints)

# Iterative hard thresholding (IHT)

(It is just projected gradient descent on sparsity constraints)

- IHT:

$$x_{t+1} = H_k\left(x_t - \eta \nabla f(x_t)\right)$$

where $H_k(z) \in \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \|x - z\|_2^2$ s.t. $\|x\|_0 \leq k$

# Iterative hard thresholding (IHT)

(It is just projected gradient descent on sparsity constraints)

- IHT:

$$x_{t+1} = H_k\left(x_t - \eta \nabla f(x_t)\right)$$

where $H_k(z) \in \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \|x - z\|_2^2$ s.t. $\|x\|_0 \leq k$

(Have we seen this before?)

# Iterative hard thresholding (IHT)

(It is just projected gradient descent on sparsity constraints)

- IHT:

$$x_{t+1} = H_k\left(x_t - \eta \nabla f(x_t)\right)$$

where $H_k(z) \in \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \|x - z\|_2^2$ s.t. $\|x\|_0 \leq k$

(Have we seen this before?)

(If yes, how we solve it?)

# Iterative hard thresholding (IHT)

(It is just projected gradient descent on sparsity constraints)

- IHT:

$$x_{t+1} = H_k\left(x_t - \eta \nabla f(x_t)\right)$$

where $H_k(z) \in \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \|x - z\|_2^2$ s.t. $\|x\|_0 \leq k$

(Have we seen this before?)

(If yes, how we solve it?)

- Now, imagine yourself implementing this.. What are the hyper-parameters?

# Iterative hard thresholding (IHT)

(It is just projected gradient descent on sparsity constraints)

- IHT:

$$x_{t+1} = H_k\left(x_t - \eta \nabla f(x_t)\right)$$

where $H_k(z) \in \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \|x - z\|_2^2$ s.t. $\|x\|_0 \leq k$

(Have we seen this before?)

(If yes, how we solve it?)

- Now, imagine yourself implementing this.. What are the hyper-parameters?
  - "How do we set the step size?"
  - "How do we select the initial point? (it is non-convex after all)"
  - "What if we don't know the sparsity level?"
  - "Are there any other tricks we can pull-off?"

# But, still, wait a minute..

- We already mentioned that the problem is hard to solve

(But maybe not true for ALL instances of the feature matrix + responses)

# But, still, wait a minute..

- We already mentioned that the problem is hard to solve

(But maybe not true for ALL instances of the feature matrix + responses)

- Imagine that $A = I$

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & f(x) := \tfrac{1}{2}\|y - Ax\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \leq k \end{aligned}$$

# But, still, wait a minute..

- We already mentioned that the problem is hard to solve

(But maybe not true for ALL instances of the feature matrix + responses)

- Imagine that $A = I$

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \frac{1}{2}\|y - Ax\|_2^2 \qquad \text{s.t.} \quad \|x\|_0 \leq k$$

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \frac{1}{2}\|y - x\|_2^2 \qquad \text{s.t.} \quad \|x\|_0 \leq k$$

(But we know how to solve this exactly and efficiently!)

# But, still, wait a minute..

- We already mentioned that the problem is hard to solve

(But maybe not true for ALL instances of the feature matrix + responses)

- Imagine that $A = I$

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \frac{1}{2}\|y - Ax\|_2^2 \qquad \text{s.t.} \quad \|x\|_0 \leq k$$

$$\longrightarrow$$

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \frac{1}{2}\|y - x\|_2^2 \qquad \text{s.t.} \quad \|x\|_0 \leq k$$

(But we know how to solve this exactly and efficiently!)

- Property of $I$: isometry

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|I(x_1 - x_2)\|_2^2 \leq (1-\delta)\|x_1 - x_2\|_2^2, \quad \text{for some } \delta \in [0,1], \ \forall x_1, x_2 \in \mathbb{R}^p$$

(Interpretation?)

# But, still, wait a minute..

- We don't care about the geometry of all vectors; only that of sparse vectors

# But, still, wait a minute..

- We don't care about the geometry of all vectors; only that of sparse vectors
- Restricted isometry property: for a matrix $A \in \mathbb{R}^{n \times p}$, we have:

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

(Interpretation?)

# But, still, wait a minute..

- We don't care about the geometry of all vectors; only that of sparse vectors
- Restricted isometry property: for a matrix $A \in \mathbb{R}^{n \times p}$, we have:

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

(Interpretation?)

- Other properties in literature: Nullspace property,
  restricted eigenvalue property
- How can we use this property in proving convergence of IHT?

Whiteboard

# But, still, wait a minute..

- We don't care about the geometry of all vectors; only that of sparse vectors
- Restricted isometry property: for a matrix $A \in \mathbb{R}^{n \times p}$, we have:

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

(Interpretation?)

- Other properties in literature: Nullspace property,
  restricted eigenvalue property
- How can we use this property in proving convergence of IHT?

Whiteboard

- We get linear convergence to the global optimum!

# How does it perform in practice?

Demo

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- It turns out that matrices that satisfy:

$$\mathbb{P}_{A \sim \mathcal{D}^{n \times p}} \left[ \left| \|Ax\|_2^2 - \|x\|_2^2 \right| > \epsilon \cdot \|x\|_2^2 \right] \leq 2e^{-\Omega(n)}$$

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- It turns out that matrices that satisfy:

Probability of being outside the interval..
$(1-\epsilon), (1+\epsilon)$

$$\mathbb{P}_{A \sim \mathcal{D}^{n \times p}}\left[\left|\|Ax\|_2^2 - \|x\|_2^2\right| > \epsilon \cdot \|x\|_2^2\right] \leq 2e^{-\Omega(n)}$$

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- It turns out that matrices that satisfy:

Probability of being outside the interval..
$(1-\epsilon), (1+\epsilon)$

$$\mathbb{P}_{A \sim \mathcal{D}^{n \times p}}\left[\left|\|Ax\|_2^2 - \|x\|_2^2\right| > \epsilon \cdot \|x\|_2^2\right] \leq 2e^{-\Omega(n)}$$

..is small

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- It turns out that matrices that satisfy:

Probability of being outside the interval..
$(1-\epsilon), (1+\epsilon)$

$$\mathbb{P}_{A \sim \mathcal{D}^{n \times p}} \left[ \left| \|Ax\|_2^2 - \|x\|_2^2 \right| > \epsilon \cdot \|x\|_2^2 \right] \leq 2e^{-\Omega(n)}$$

..is small

also satisfy RIP with probability $1 - 2e^{-\Omega(n)}$ whenever $n \geq \Omega\left(\frac{k}{\delta^2} \log \frac{p}{k}\right)$

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- It turns out that matrices that satisfy:

Probability of being outside the interval..
$(1-\epsilon), (1+\epsilon)$

$$\mathbb{P}_{A \sim \mathcal{D}^{n \times p}} \left[ \left| \|Ax\|_2^2 - \|x\|_2^2 \right| > \epsilon \cdot \|x\|_2^2 \right] \leq 2e^{-\Omega(n)}$$

..is small

also satisfy RIP with probability $1 - 2e^{-\Omega(n)}$ whenever $n \geq \Omega\left(\frac{k}{\delta^2} \log \frac{p}{k}\right)$

- What type of $A \in \mathbb{R}^{n \times p}$ satisfy this: Gaussian, Bernoulli, ..

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- It turns out that matrices that satisfy:

Probability of being outside the interval..
$(1-\epsilon), (1+\epsilon)$

$$\mathbb{P}_{A \sim \mathcal{D}^{n \times p}} \left[ \left| \|Ax\|_2^2 - \|x\|_2^2 \right| > \epsilon \cdot \|x\|_2^2 \right] \leq 2e^{-\Omega(n)}$$

..is small

also satisfy RIP with probability $1 - 2e^{-\Omega(n)}$ whenever $n \geq \Omega\left(\frac{k}{\delta^2} \log \frac{p}{k}\right)$

- What type of $A \in \mathbb{R}^{n \times p}$ satisfy this: Gaussian, Bernoulli, ..

(In our case, we generate it as Gaussian so with high probability we are fine)

# What did go wrong here?

- Q: "Are we certain that $A \in \mathbb{R}^{n \times p}$ satisfies RIP?"
- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

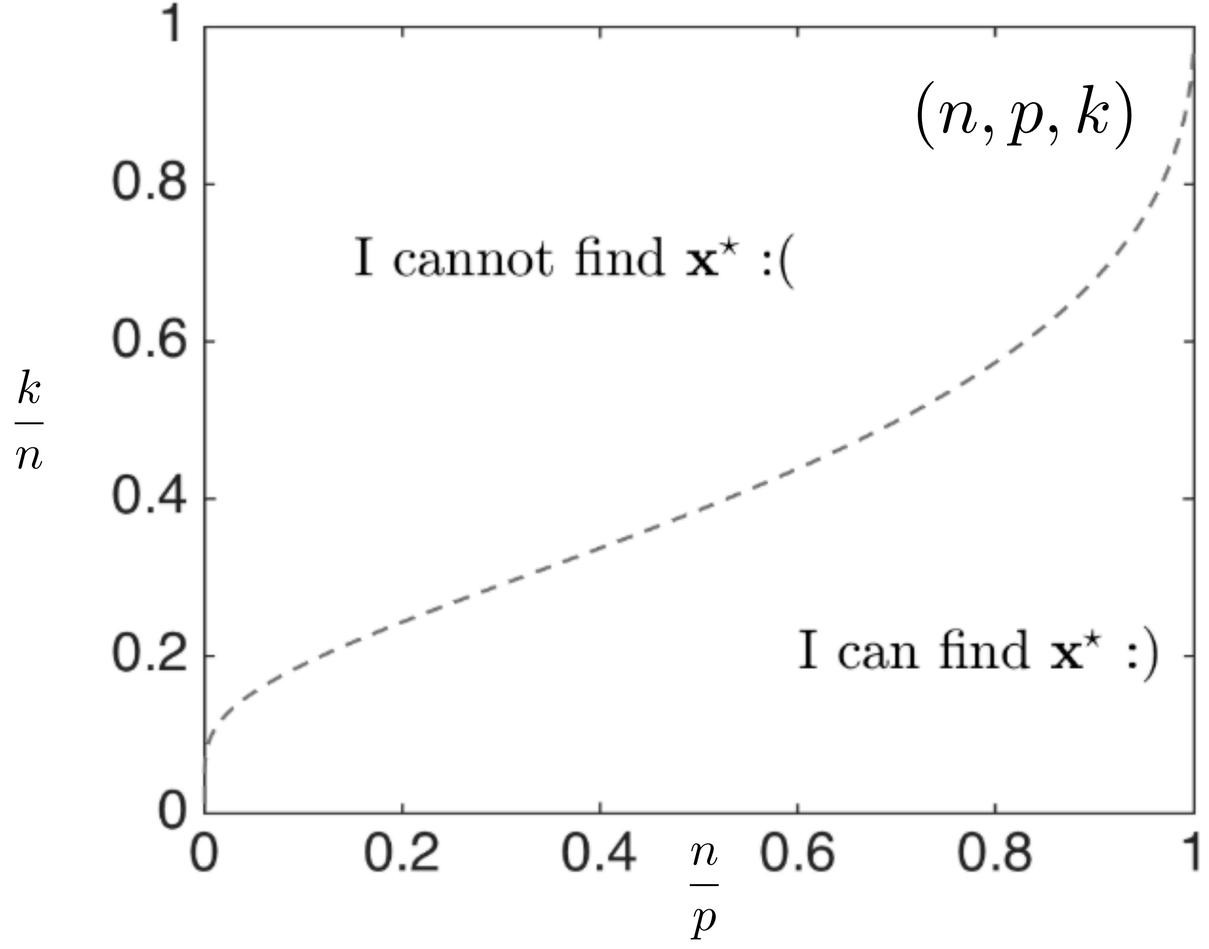
Note: Checking whether a fixed matrix actually satisfies RIP is NP-hard..

# What did go wrong here?

- Q: "Maybe the problem is not easily solvable?"

# What did go wrong here?

- Q: "Maybe the problem is not easily solvable?"
- The notion of phase transition:

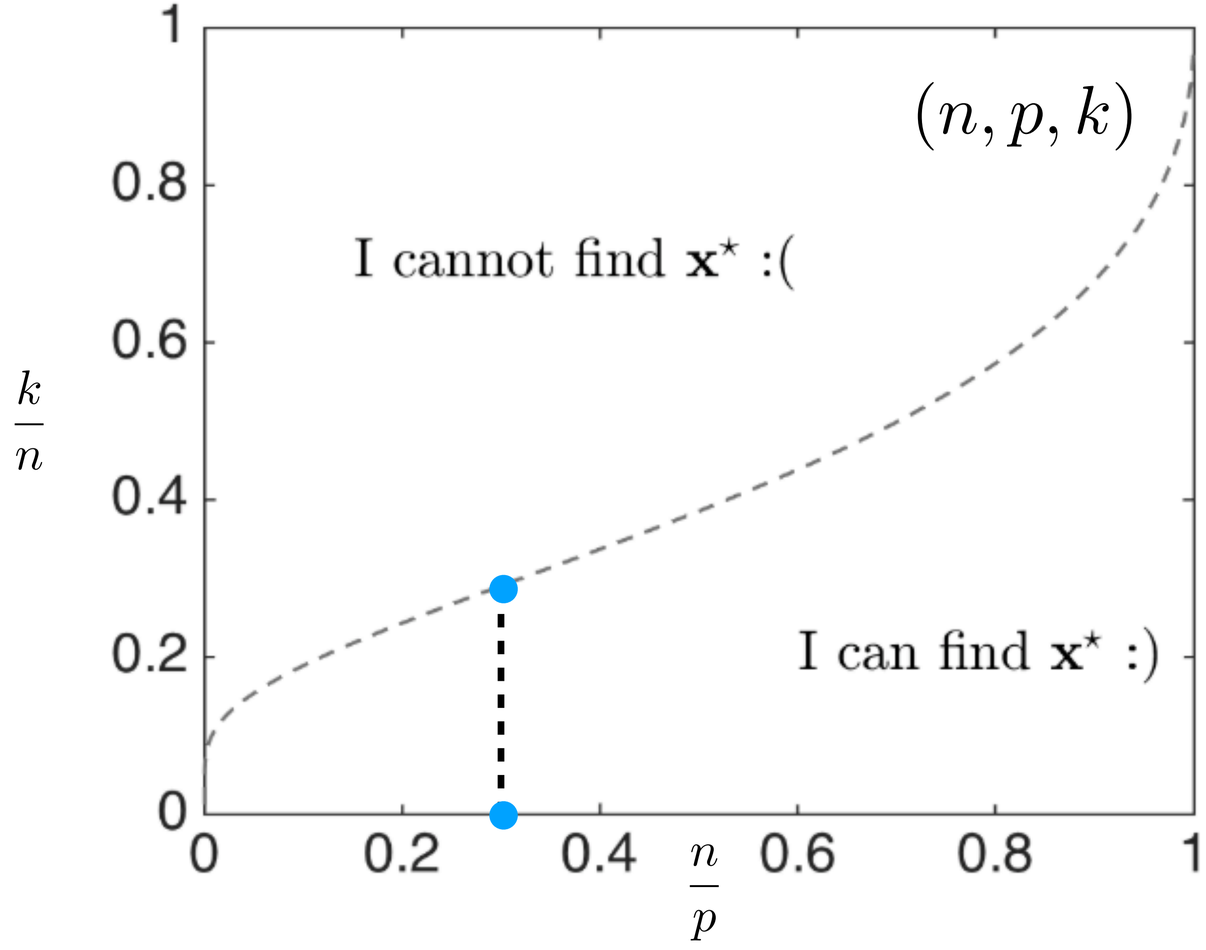



$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

# What did go wrong here?

- Q: "Maybe the problem is not easily solvable?"
- The notion of phase transition:

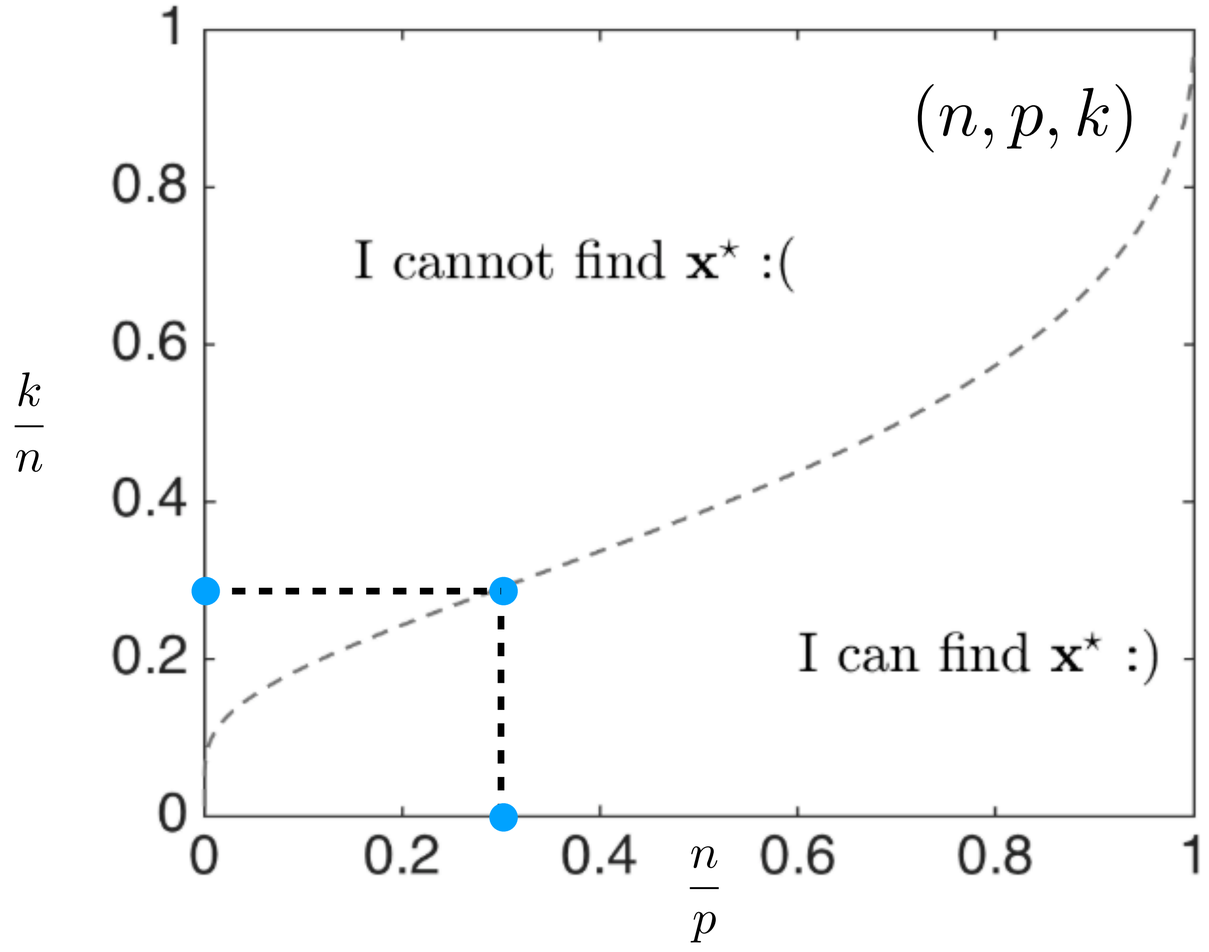



$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

# What did go wrong here?

- Q: "Maybe the problem is not easily solvable?"
- The notion of phase transition:

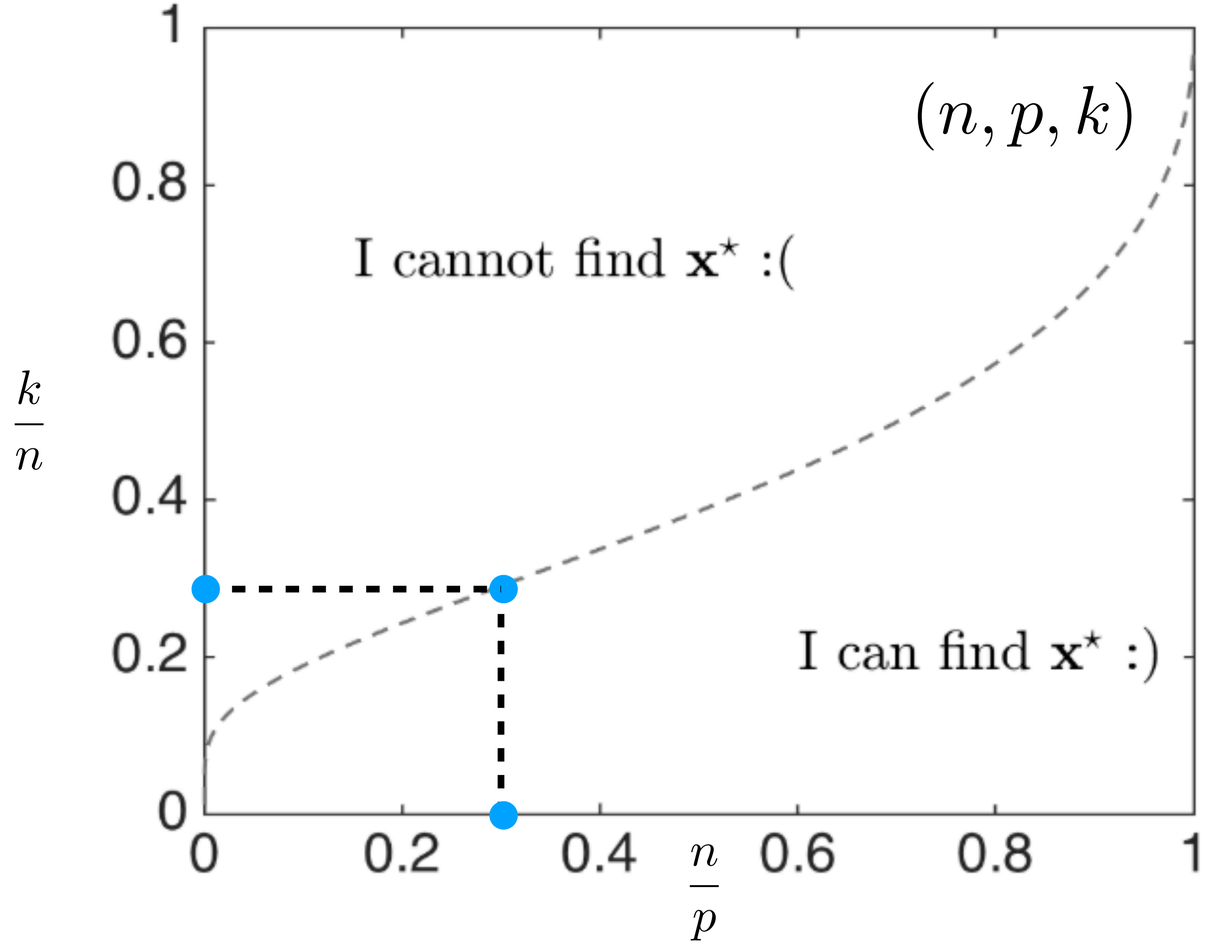



$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

# What did go wrong here?

- Q: "Maybe the problem is not easily solvable?"
- The notion of phase transition:



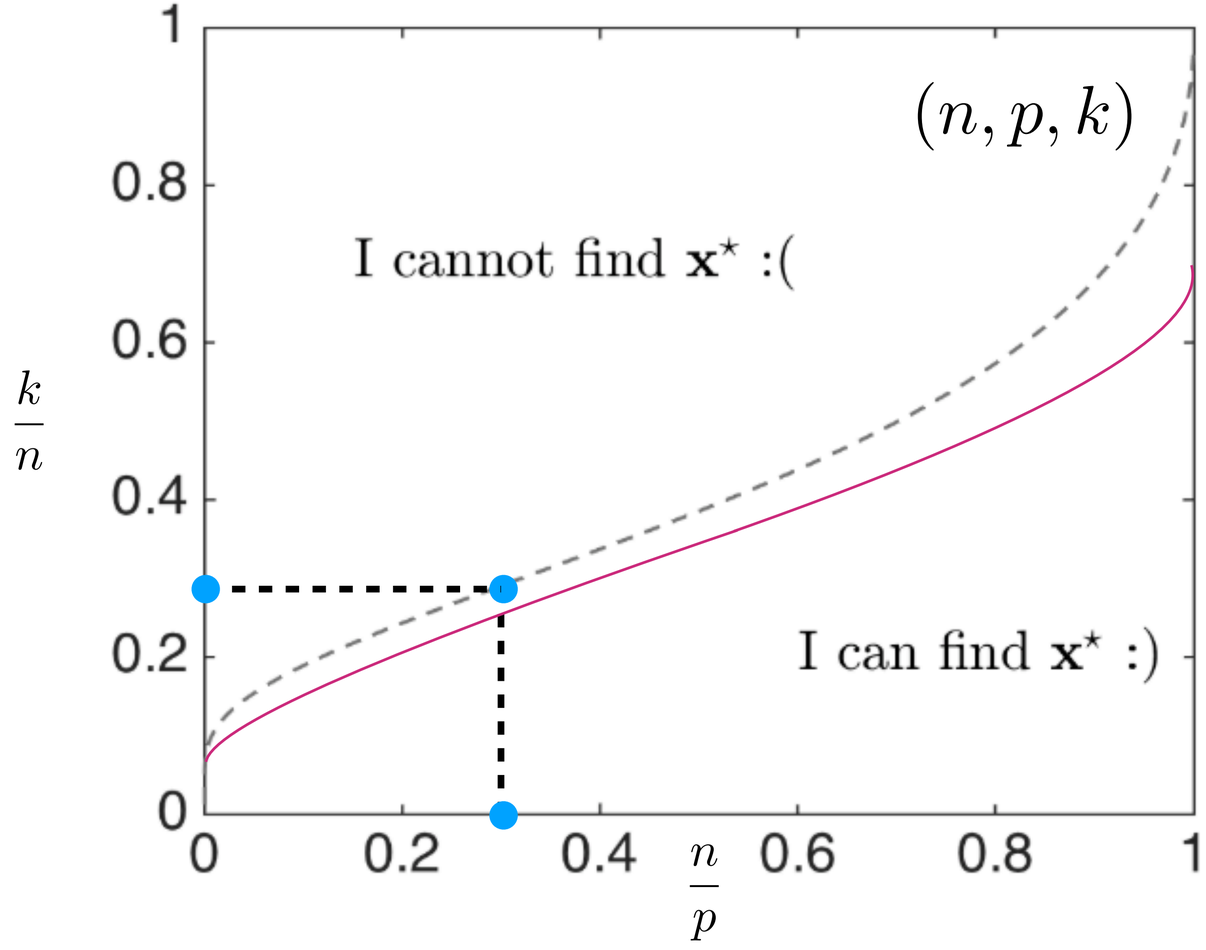

$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

Maybe the sparsity level is too much?

# What did go wrong here?

- Q: "Maybe the problem is not easily solvable?"
- The notion of phase transition:




$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

Maybe the sparsity level is too much?

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?

Whiteboard

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?
- Thus, we propose:

$$\eta = \frac{1}{1+\delta}$$

(In literature section, there is work that validates theoretically this selection)

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?
- Thus, we propose:

$$\eta = \frac{1}{1+\delta}$$

(In literature section, there is work that validates theoretically this selection)

- But, is this practical?

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?
- Thus, we propose:

$$\eta = \frac{1}{1+\delta}$$

(In literature section, there is work that validates theoretically this selection)

- But, is this practical? Generally, no!

(Value of $\delta$ is NP-hard to find)

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?
- What about find a more practical step-size selection?

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?
- What about find a more practical step-size selection?

Whiteboard

# What did go wrong here?

- Q: "How can we then improve its performance?"
- Via hyper-parameter tuning! Can we tune the step size via theory?
- What about find a more practical step-size selection?

Whiteboard

Demo

# Exact line search

- What we performed is exact line search:

"Given the direction we want to move towards, find the best step size such that we minimize the objective function"

# Exact line search

- What we performed is exact line search:

"Given the direction we want to move towards, find the best step size such that we minimize the objective function"

- I.e.,

$$\eta = \arg\min_{\eta \in \mathbb{R}_+} f\left(x_t - \eta \nabla_{Q_t} f(x_t)\right)$$

# Exact line search

- What we performed is exact line search:

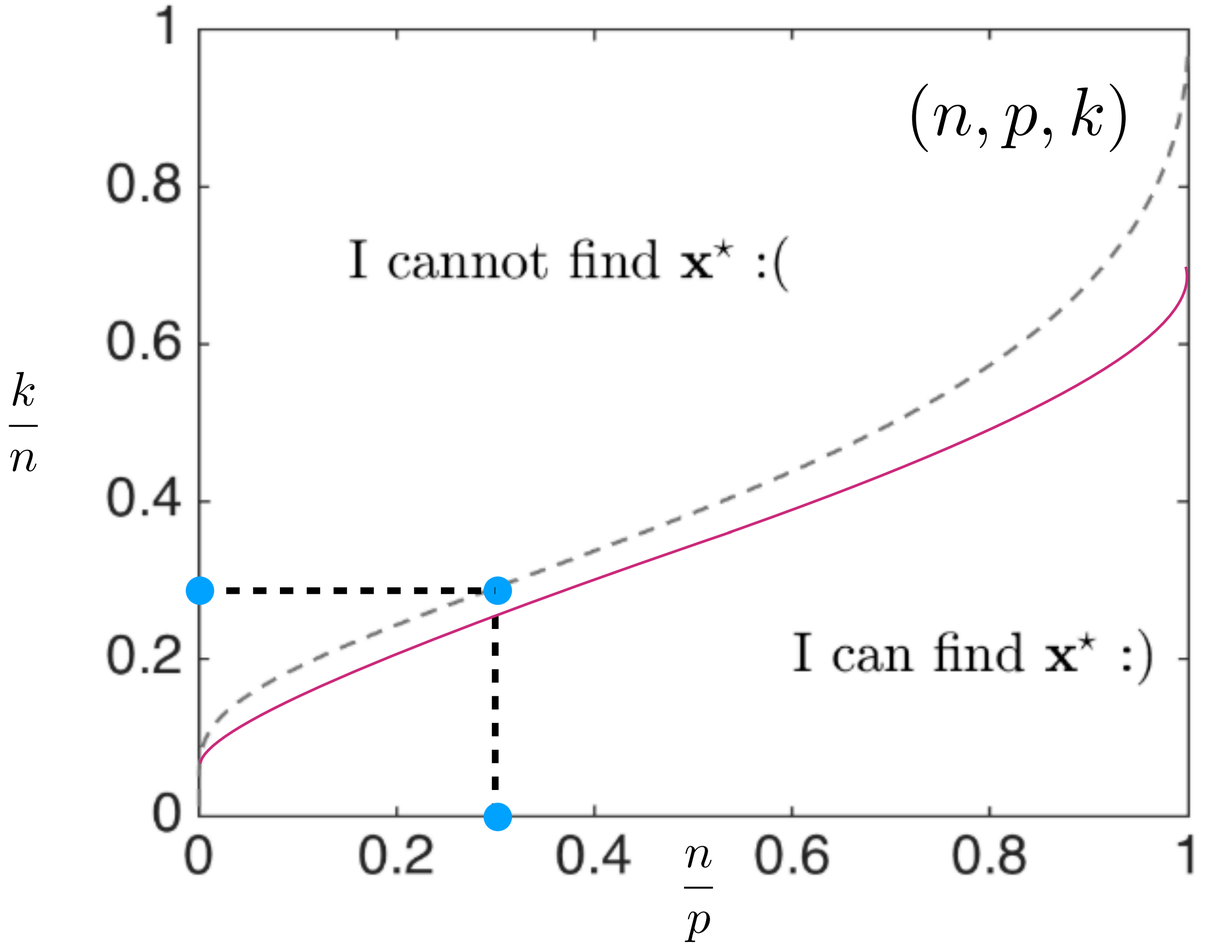
"Given the direction we want to move towards, find the best step size such that we minimize the objective function"

- I.e.,

$$\eta = \arg\min_{\eta \in \mathbb{R}_+} f\left(x_t - \eta \nabla_{Q_t} f(x_t)\right)$$

- Q: "Great! Why don't we use that all the time?"
- A: "Because, moving beyond least squares, solving this might be as difficult as the original problem"

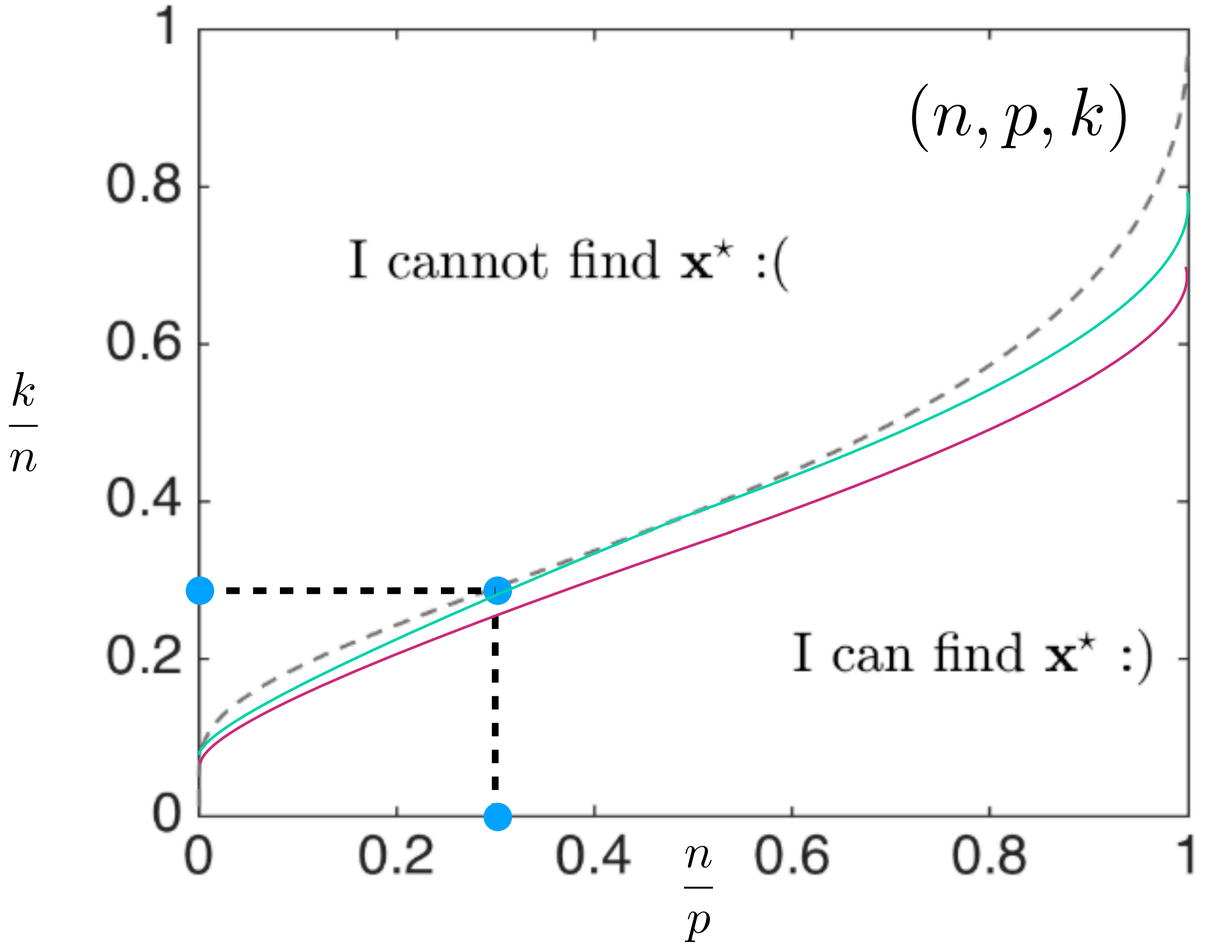
# Phase transition update




$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

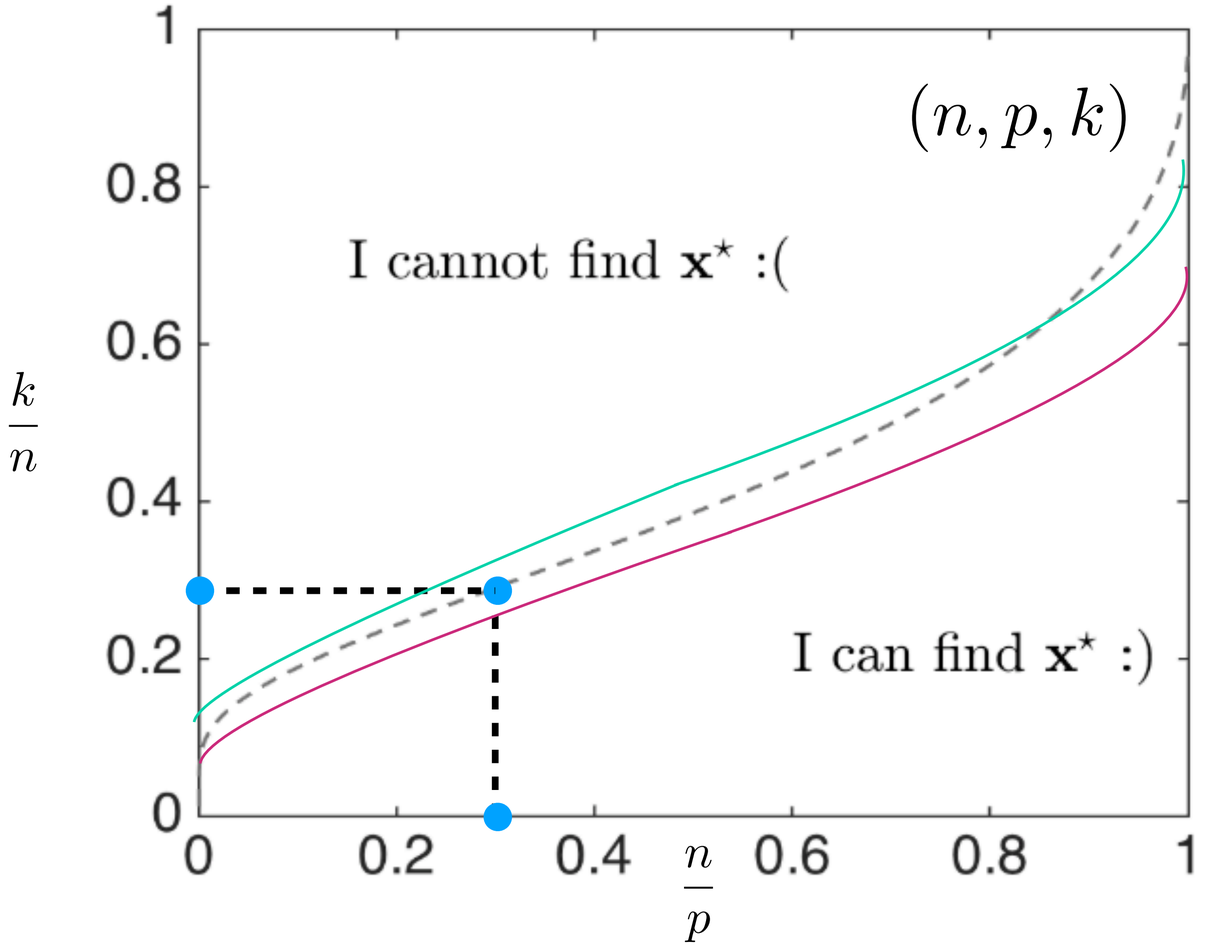
# Phase transition update




$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

# Phase transition update




$$\min_{x \in \mathbb{R}^p} \quad \|x\|_1$$

$$\text{s.t.} \quad y = Ax$$

(Some of these methods can be found in the Review part)

# But does this step size selection work in theory?

Whiteboard

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"

(Demo)

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"
- A: "We pay for the "energy" we leave out"

(Demo)

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"
- A: "We pay for the "energy" we leave out"
- Q: "What happens if we overshoot sparsity level?" (Demo)

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"
- A: "We pay for the "energy" we leave out"
- Q: "What happens if we overshoot sparsity level?" (Demo)

# What about the sparsity level $k$?

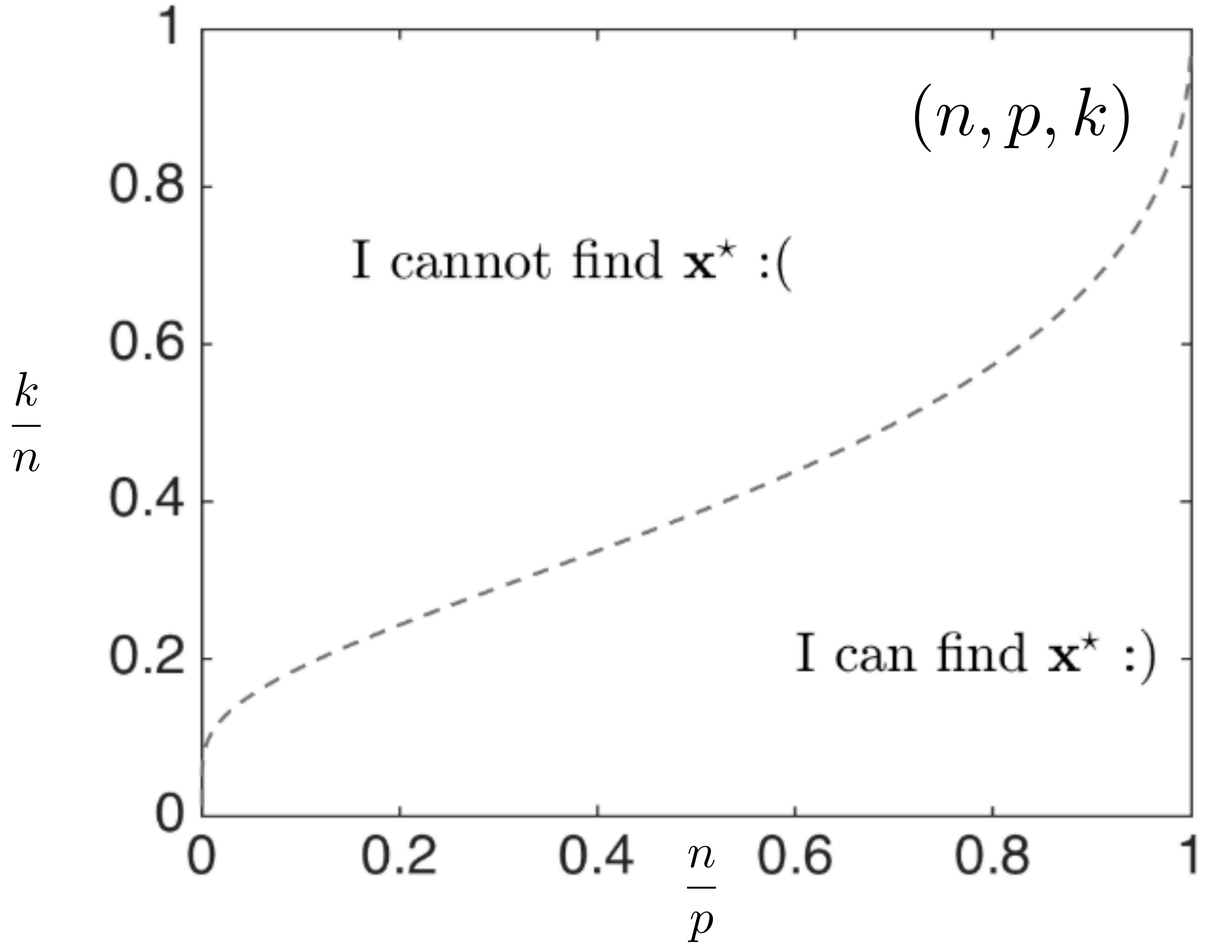
- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"
- A: "We pay for the "energy" we leave out"
- Q: "What happens if we overshoot sparsity level?" (Demo)
- A: "We actually get denser and denser solutions"

# What about the sparsity level $k$?

- There is no clear and provable tweak that suggests how to set $k$
- Q: "What if we set $k = p$?"
- A: "Umm.. Think harder :)"
- Q: "What happens if we undershoot sparsity level?"
- A: "We pay for the "energy" we leave out"
- Q: "What happens if we overshoot sparsity level?" (Demo)
- A: "We actually get denser and denser solutions"
- Q: "Is there any non-provable tweaks?"
- A: "Problem-dependent strategies"

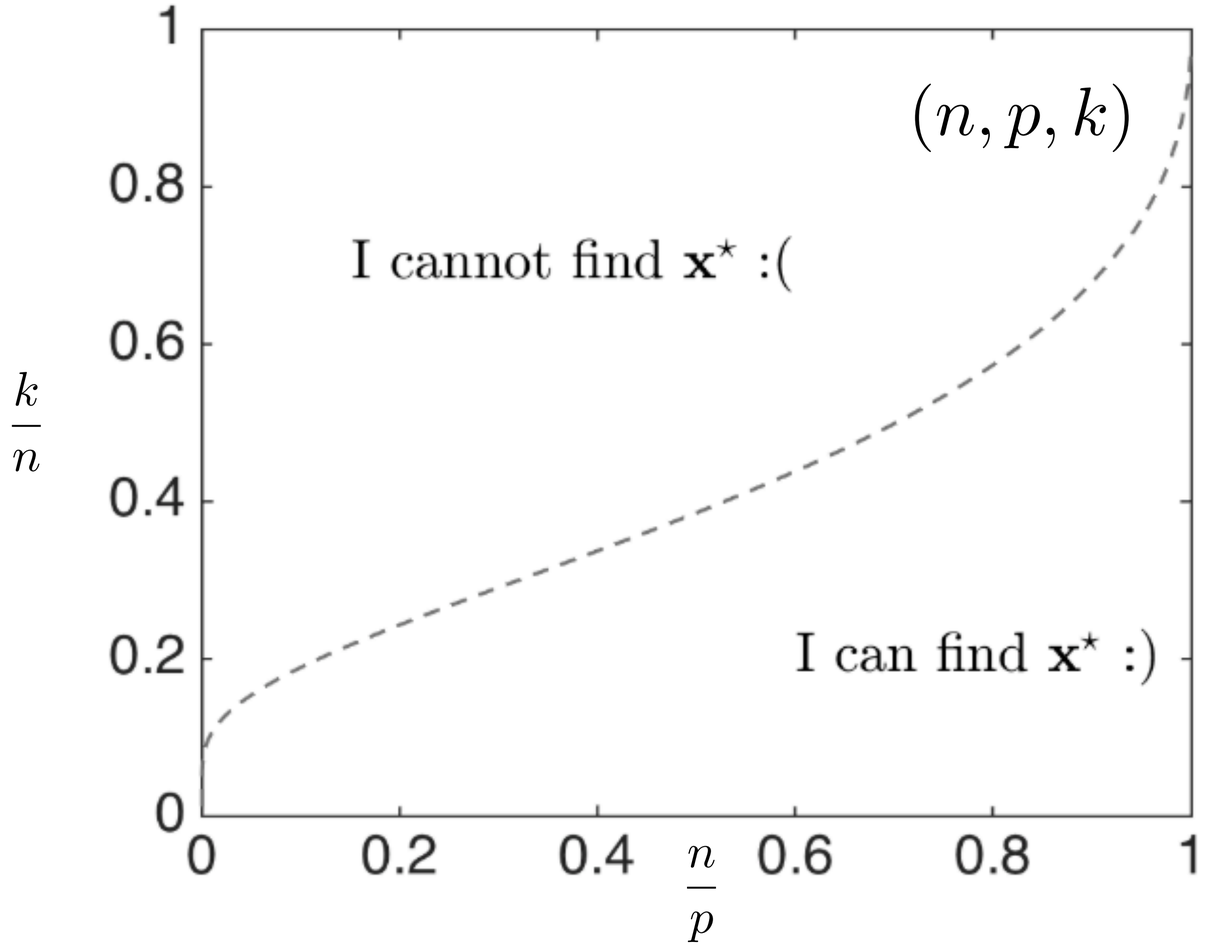
# Heuristic for $k$ on linear regression

(More like an upper bound for sparsity level)



# Heuristic for $k$ on linear regression

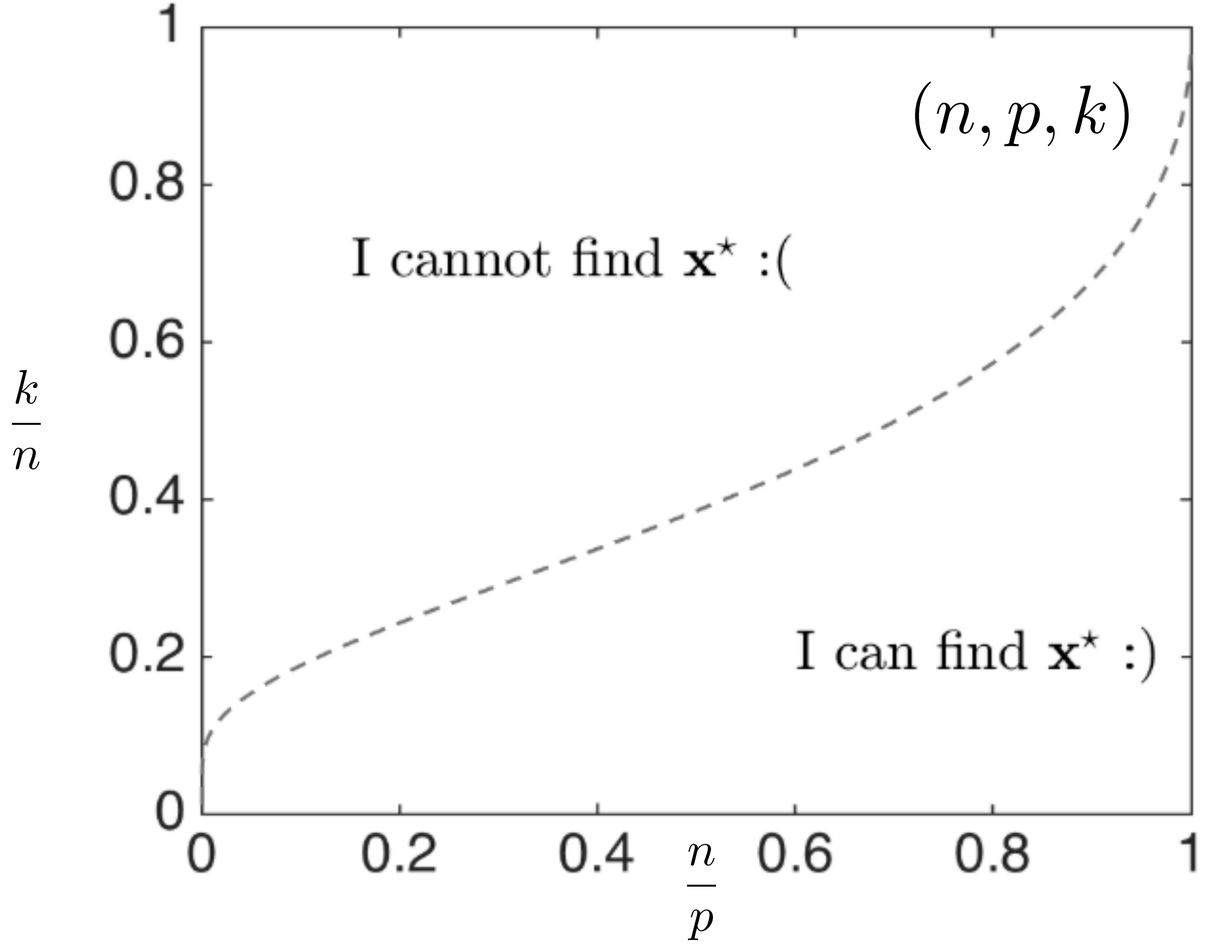
(More like an upper bound for sparsity level)




– We know $(n, p)$

# Heuristic for $k$ on linear regression

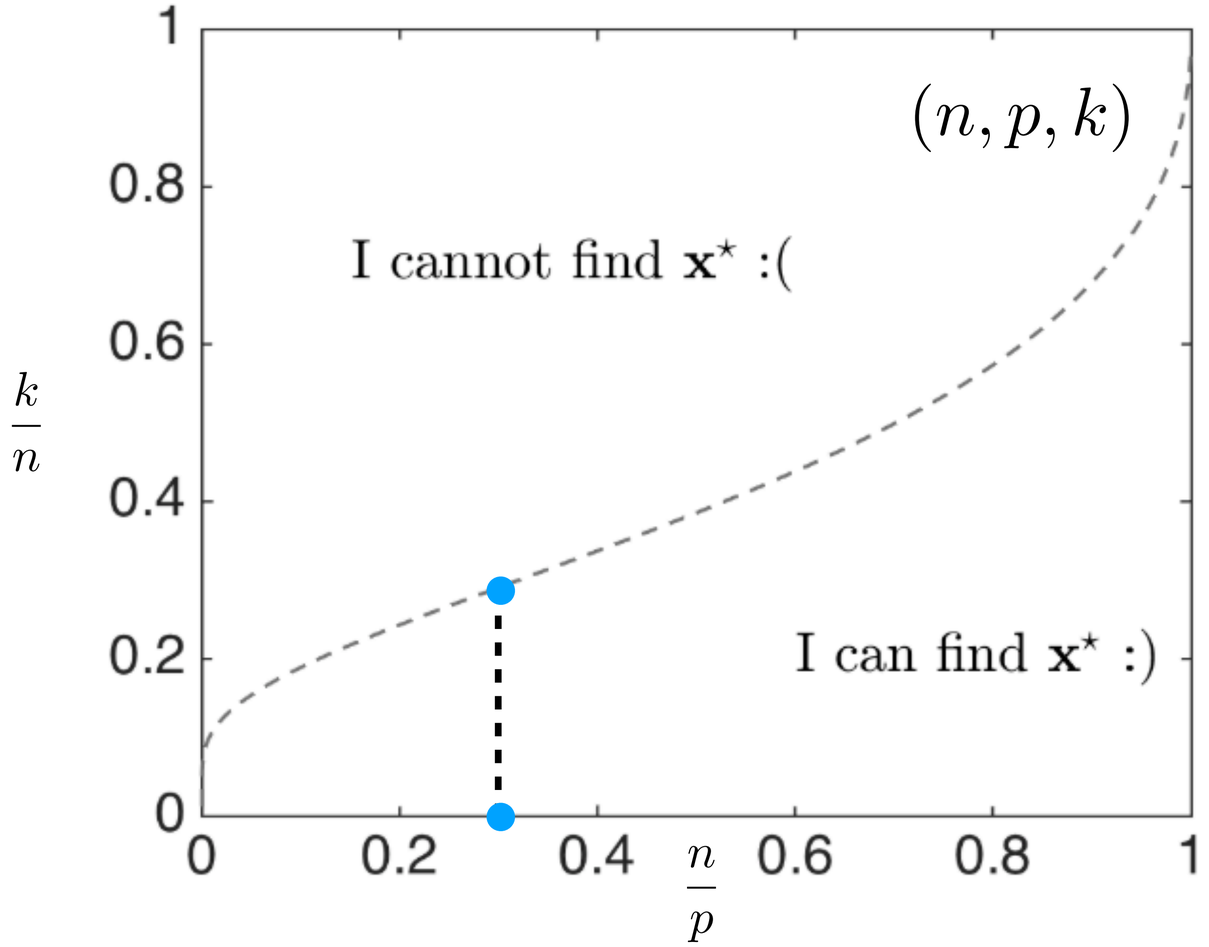
(More like an upper bound for sparsity level)




- We know $(n, p)$
- Map these values in the phase transition plot

# Heuristic for $k$ on linear regression

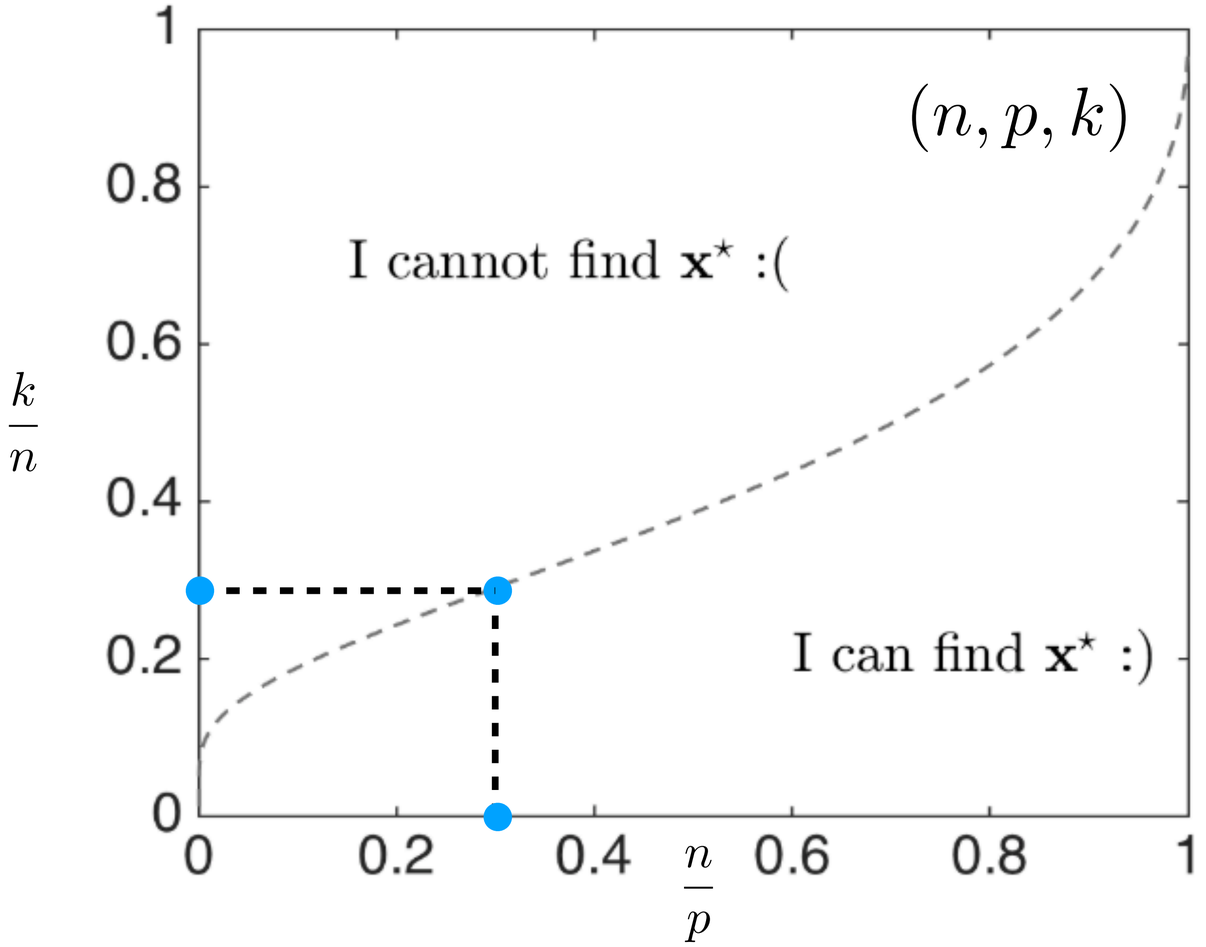
(More like an upper bound for sparsity level)




- We know $(n, p)$
- Map these values in the phase transition plot

# Heuristic for $k$ on linear regression

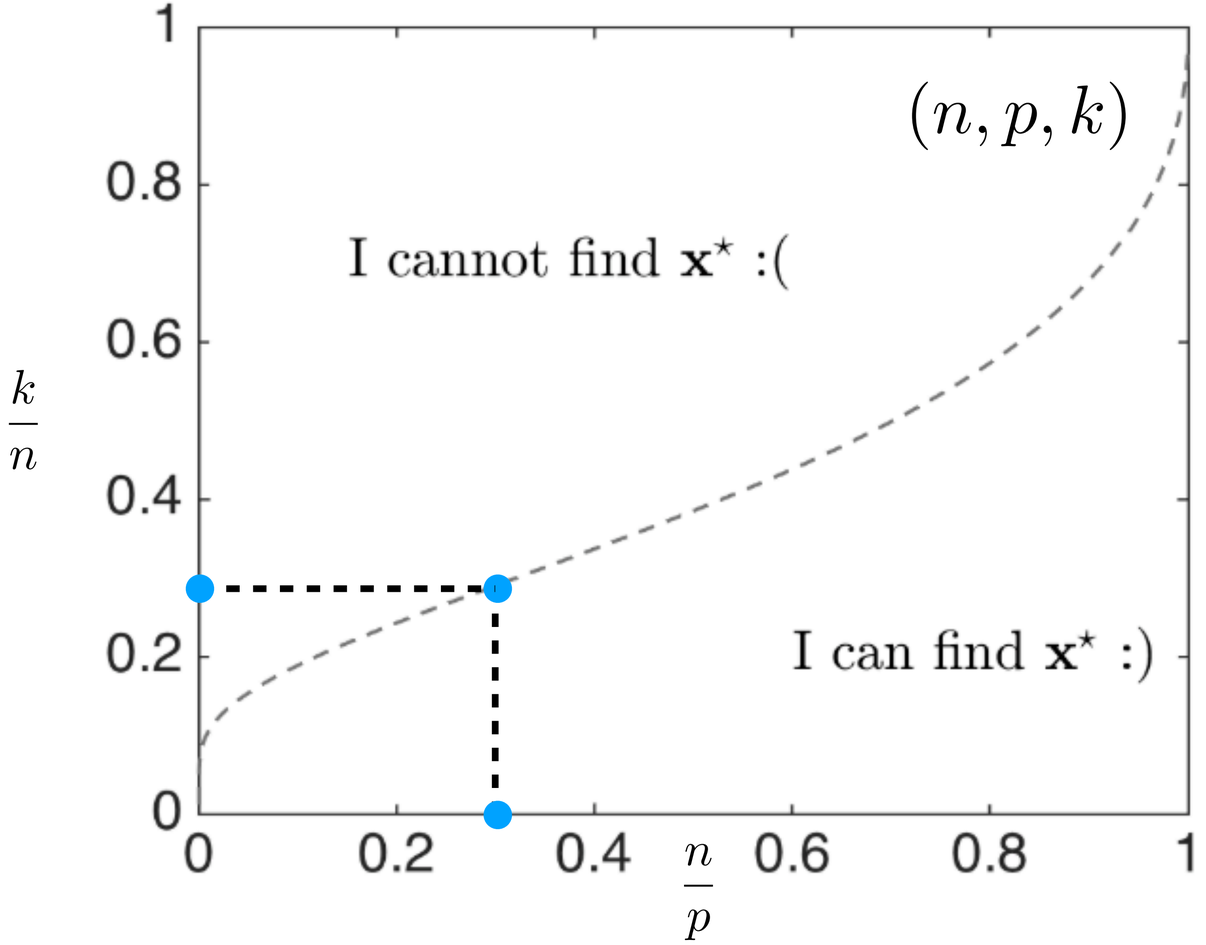
(More like an upper bound for sparsity level)




- We know $(n, p)$
- Map these values in the phase transition plot

# Heuristic for $k$ on linear regression

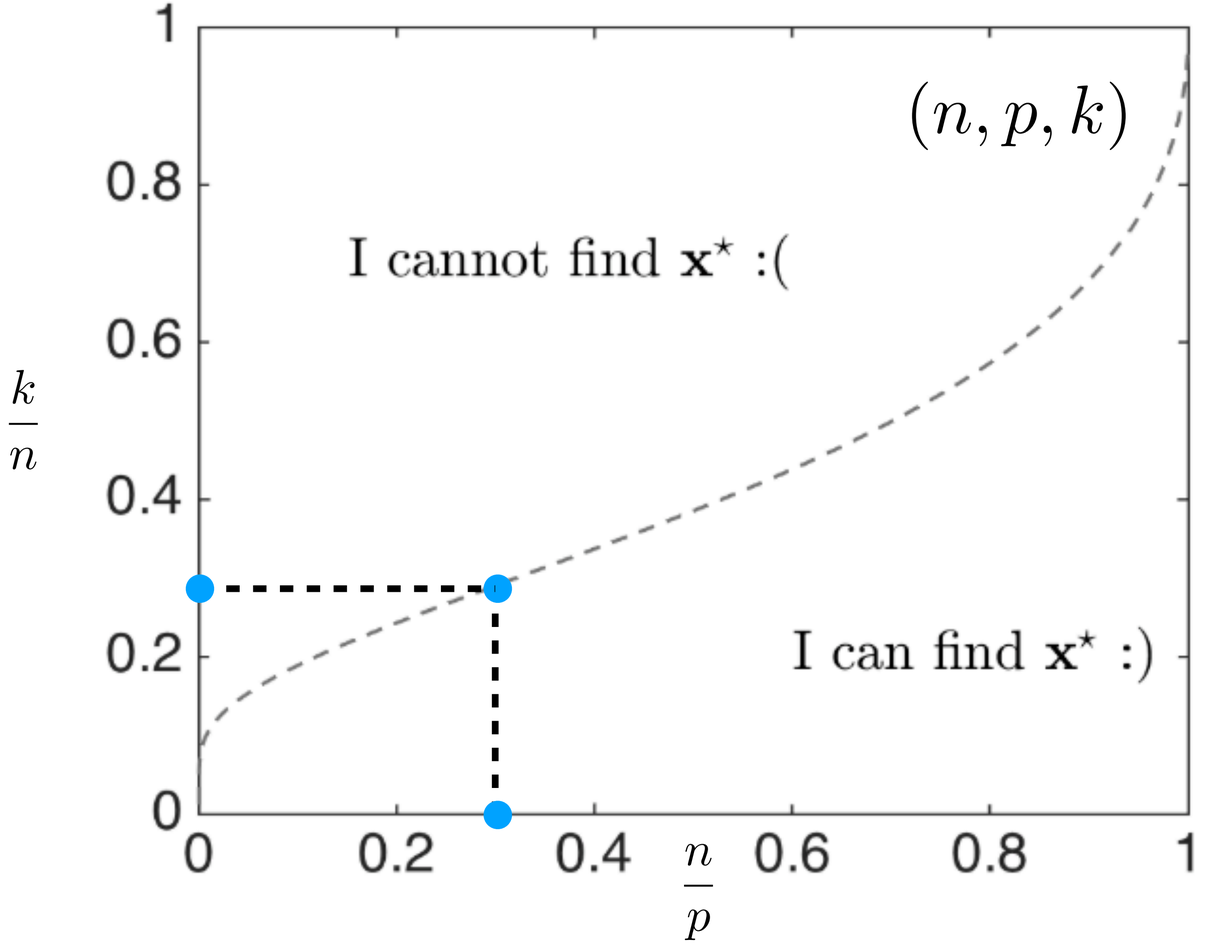
(More like an upper bound for sparsity level)




- We know $(n, p)$
- Map these values in the phase transition plot
- The curve indicates what levels might work

# Heuristic for $k$ on linear regression

(More like an upper bound for sparsity level)




- We know $(n, p)$
- Map these values in the phase transition plot
- The curve indicates what levels might work
- Run algorithm, take results, interpret: if it works, good; otherwise, do binary search over sparsity level

"All these sound interesting.. but do they extend to other objectives? And how are they related with what we discussed so far?"

(Lipschitz gradient continuity, strong convexity, Hessians, etc..)

# A different view of RIP

- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

# A different view of RIP

- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- Simplify for $2k$ −sparse

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \|x\|_2^2, \quad \forall\ 2k - \text{sparse } x \in \mathbb{R}^p$$

# A different view of RIP

- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- Simplify for $2k$ −sparse

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \|x\|_2^2, \quad \forall\ 2k - \text{sparse } x \in \mathbb{R}^p$$

- What is the Hessian of the objective? $\nabla^2 f(\cdot) = A^\top A$

# A different view of RIP

- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- Simplify for $2k$ −sparse

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \|x\|_2^2, \quad \forall\ 2k - \text{sparse } x \in \mathbb{R}^p$$

- What is the Hessian of the objective? $\nabla^2 f(\cdot) = A^\top A$

- Rewriting RIP:

$$(1-\delta)\|x\|_2^2 \leq x^\top A^\top A x \leq (1+\delta)\|x\|_2^2 \Rightarrow (1-\delta)I \preceq A^\top A \preceq (1+\delta)I$$

# A different view of RIP

- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- Simplify for $2k$ −sparse

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \|x\|_2^2, \quad \forall\ 2k - \text{sparse } x \in \mathbb{R}^p$$

- What is the Hessian of the objective? $\nabla^2 f(\cdot) = A^\top A$

- Rewriting RIP:

$$(1-\delta)\|x\|_2^2 \leq x^\top A^\top A x \leq (1+\delta)\|x\|_2^2 \Rightarrow \boxed{(1-\delta)I \preceq A^\top A \preceq (1+\delta)I}$$

# A different view of RIP

- Reminder

$$(1-\delta)\|x_1 - x_2\|_2^2 \leq \|A(x_1 - x_2)\|_2^2 \leq (1+\delta)\|x_1 - x_2\|_2^2,$$
$$\text{for some } \delta \in (0,1),\ \forall\ k\text{-sparse } x_1, x_2 \in \mathbb{R}^p$$

- Simplify for $2k$ −sparse

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \|x\|_2^2, \quad \forall\ 2k - \text{sparse } x \in \mathbb{R}^p$$

- What is the Hessian of the objective? $\nabla^2 f(\cdot) = A^\top A$

- Rewriting RIP:

$$(1-\delta)\|x\|_2^2 \leq x^\top A^\top A x \leq (1+\delta)\|x\|_2^2 \Rightarrow \boxed{(1-\delta)I \preceq A^\top A \preceq (1+\delta)I}$$

- When objective has Lipschitz continuous gradients and is strongly convex:

$$\mu I \preceq \nabla^2 f(x) \preceq LI$$

# Restricted smoothness and strong convexity

- "Restricted": the properties hold over a subset of $\mathbb{R}^p$

  Let's say $x \in \mathcal{C} \subseteq \mathbb{R}^p$

# Restricted smoothness and strong convexity

- "Restricted": the properties hold over a subset of $\mathbb{R}^p$

  Let's say $x \in \mathcal{C} \subseteq \mathbb{R}^p$

- Restricted smoothness (Lipschitz gradient continuity)

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathcal{C}$$

(Note that this $L$ is different from that of general smoothness)

# Restricted smoothness and strong convexity

- "Restricted": the properties hold over a subset of $\mathbb{R}^p$

  Let's say $x \in \mathcal{C} \subseteq \mathbb{R}^p$

- Restricted smoothness (Lipschitz gradient continuity)

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathcal{C}$$

(Note that this $L$ is different from that of general smoothness)

- Restricted strong convexity

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathcal{C}$$

(Note that this $\mu$ is different from that of general strong convexity)

# Examples

- Sparse logistic regression

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \leq k \end{aligned}$$

$y_i \in \{\pm 1\}$

# Examples

- Sparse logistic regression

$\in \{\pm 1\}$

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \leq k \end{aligned}$$

(where sparsity gets in the picture)

Satisfies restricted strong convexity with constant:

$$\mu = (\gamma_k + \lambda), \text{ where } \gamma_k = \lambda_{\min}\left(A^\top \Lambda A, k\right)$$

# Examples

- Sparse logistic regression

$\in \{\pm 1\}$

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i a_i^\top x\right)\right) + \frac{\lambda}{2}\|x\|_2^2 \\ \text{s.t.} \quad & \|x\|_0 \leq k \end{aligned}$$

(where sparsity gets in the picture)

Satisfies restricted strong convexity with constant:

$$\mu = (\gamma_k + \lambda), \text{ where } \gamma_k = \lambda_{\min}\left(A^\top \Lambda A, k\right)$$

and restricted smoothness with constant:

$$L = (\lambda_{\max}\left(A^\top A, k\right) + \lambda)$$

(For more information, see "Gradient hard thresholding pursuit")

# Examples

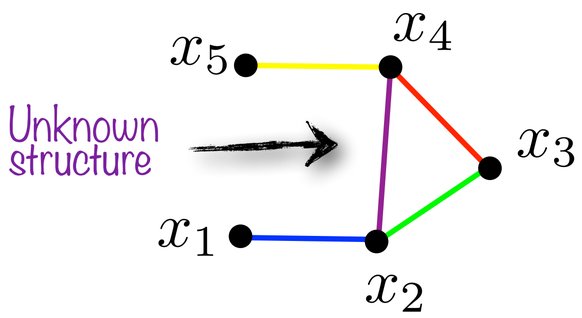
- Graphical model selection (under Gaussian assumptions)

Whiteboard

# Examples

- Graphical model selection (under Gaussian assumptions)

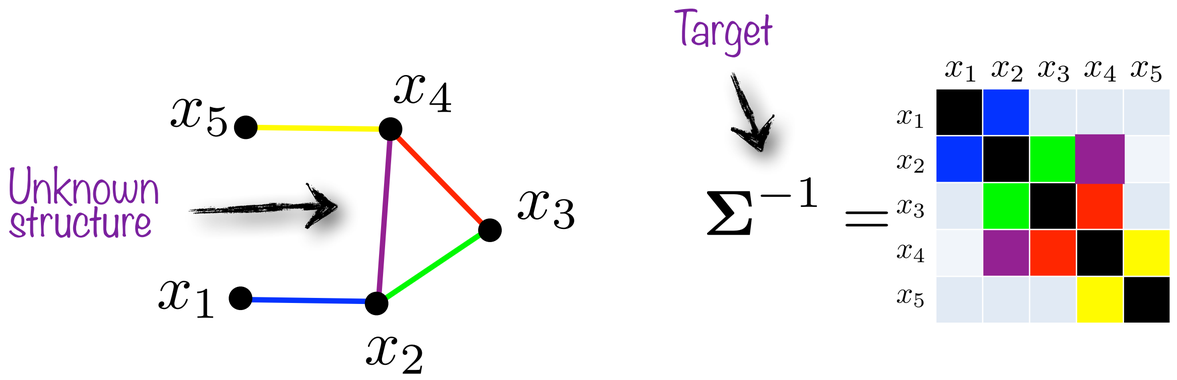
Whiteboard



# Examples

- Graphical model selection (under Gaussian assumptions)

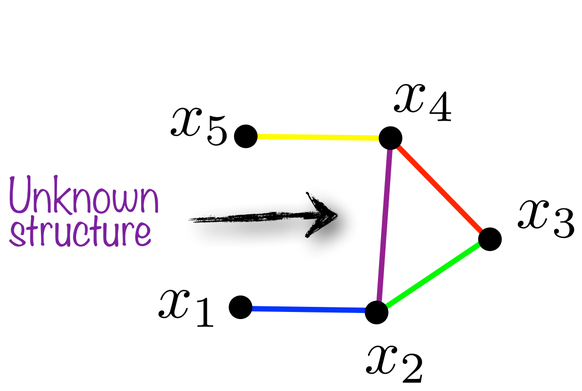
Whiteboard

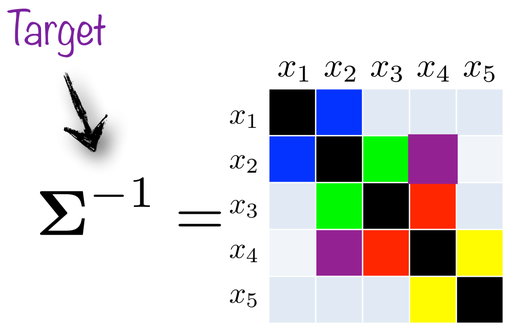


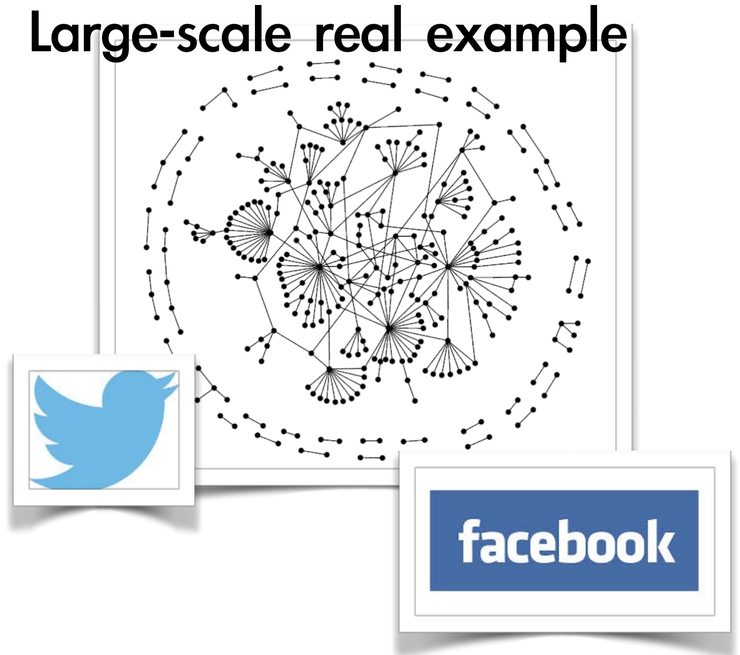
# Examples

- Graphical model selection (under Gaussian assumptions)

Whiteboard



# Examples

- Graphical model selection (under Gaussian assumptions)
  - Given a data set $\mathcal{D}$, drawn from a joint pdf with unknown covariance $\mathbf{\Sigma}$, the aim is to learn a sparse matrix $\mathbf{\Theta}$ that approximates $\mathbf{\Sigma}^{-1}$.

    Input: sample covariance $\widehat{\mathbf{\Sigma}}$ calculated usually from limited samples

# Examples

- Graphical model selection (under Gaussian assumptions)
  - Given a data set $\mathcal{D}$, drawn from a joint pdf with unknown covariance $\mathbf{\Sigma}$, the aim is to learn a sparse matrix $\mathbf{\Theta}$ that approximates $\mathbf{\Sigma}^{-1}$.

Input: sample covariance $\widehat{\mathbf{\Sigma}}$ calculated usually from limited samples

**Optimization problem 1**

Self-concordant function

Convex

$$\min_{\Theta \in \mathbb{R}^{p \times p}} \; -\log\det(\Theta) + \mathrm{Tr}\left(\widehat{\Sigma}\Theta\right) + \rho\|\Theta\|_1$$

# Examples

- Graphical model selection (under Gaussian assumptions)
  - Given a data set $\mathcal{D}$, drawn from a joint pdf with unknown covariance $\mathbf{\Sigma}$, the aim is to learn a sparse matrix $\mathbf{\Theta}$ that approximates $\mathbf{\Sigma}^{-1}$.

Input: sample covariance $\widehat{\mathbf{\Sigma}}$ calculated usually from limited samples

**Optimization problem 1**

Self-concordant function

Convex

$$\min_{\Theta \in \mathbb{R}^{p \times p}} \quad -\log\det(\Theta) + \mathrm{Tr}\left(\widehat{\Sigma}\Theta\right) + \rho\|\Theta\|_1$$

**Optimization problem 2**

Non-convex

$$\min_{\Theta \in \mathbb{R}^{p \times p}} \quad -\log\det(\Theta) + \mathrm{Tr}\left(\widehat{\Sigma}\Theta\right)$$

$$\text{s.t.} \quad \|\Theta\|_0 \leq k$$

# Beyond plain sparsity

- Our discussion so far holds for discrete structures beyond sparsity:

  Block-sparsity, overlapping block-sparsity, dispersive models, tree sparsity, graph-sparsity, etc..

# Beyond plain sparsity

- Our discussion so far holds for discrete structures beyond sparsity:

  Block-sparsity, overlapping block-sparsity, dispersive models, tree sparsity, graph-sparsity, etc..

- As long as the projection onto the combinatorial constraint can be computed efficiently:

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & \|x - y\|_2^2 \\ \text{s.t.} \quad & x \in \mathcal{C} \end{aligned}$$

- Various extensions include inexact projections, greedy approaches, and there are connections with (sub/super)modular optimization

# Interlude: Statistics in Data Science

- We will use the example of RIP

  (The goal is to highlight that we cannot have data science without using a diverse set of tools; IMHO, ML is not an research area "fallen from the sky": it is just a cool combination of optimization, statistics, coding and linear algebra)

- Disclaimer: this is not a complete introduction to concentration inequalities

  (How many would be interested in learning about concentration inequalities (as a course)? )

# Papers to review – due next Tuesday

(Select one of the following papers)

- "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples", Needell et al., 2009.

- "Compressed sensing using generative models", Bora et al., 2017.

- "A Nearly-Linear Time Framework for Graph-Structured Sparsity", Hegde et al., 2015.

- "Deep image prior", Ulyanov et al., 2018.

# Conclusion

- This lecture considers **sparse model selection** in Data Science applications
- While there are rigorous and efficient methods also in the convex domain we followed the **non-convex path** of hard thresholding methods
- We discussed some global convergence guarantees, and highlighted the importance of hyper-parameter tuning

# Next lecture

- We will consider the case of **low-rank recovery**, natural extension of sparsity - there, we have different ways to exploit non-convexity