# Optimization Methods for Machine Learning
# Lecture 14
# Sparse Convex Optimization

Katya Scheinberg

# Compressed sensing

# A short introduction to *Compressed Sensing*

- An imaging perspective


Scene

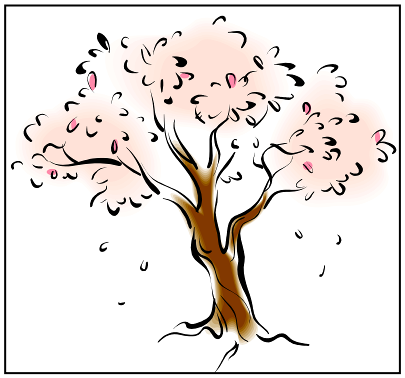
Picture

- Image compression

Why do we compress images?

# Introduction to *Compressed Sensing*

- Images are compressible



Because

- Only certain part of information is important (e.g. objects and their edges)
- Some information is unwanted (e.g. noise)

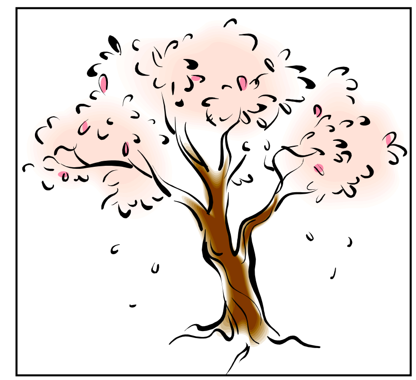
- Image compression
  - Take an input image $u$
  - Pick a good dictionary $\Phi$
  - Find a sparse representation $x$ of $u$ such that $||\Phi x - u||_2$ is small
  - Save $x$

This is traditional compression.

# Introduction to *Compressed Sensing*

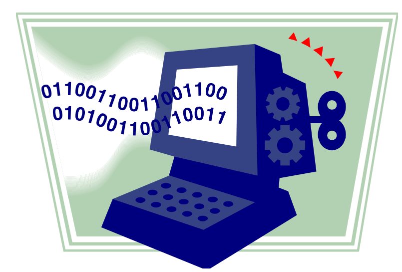
- An imaging perspective

10 Mega Pixels



This is traditional compression.

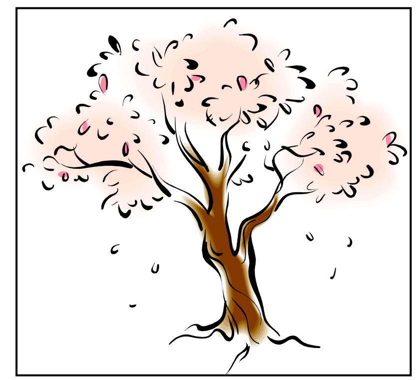
100 Kilobytes

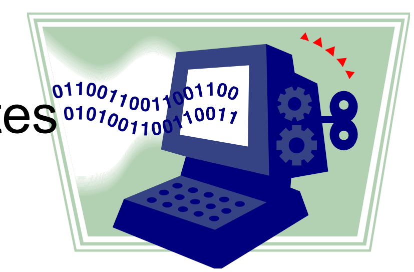


# Introduction to *Compressed Sensing*

- An imaging perspective

$n$: 10 Mega Pixels






$k$:100 Kilobytes

This is traditional compression.

# Introduction to *Compressed Sensing*

- If only 100 kilobytes are saved, why do we need a 10-megapixel camera in the first place?

- Answer: a traditional compression algorithm needs the complete image to compute $\Phi$ and $x$

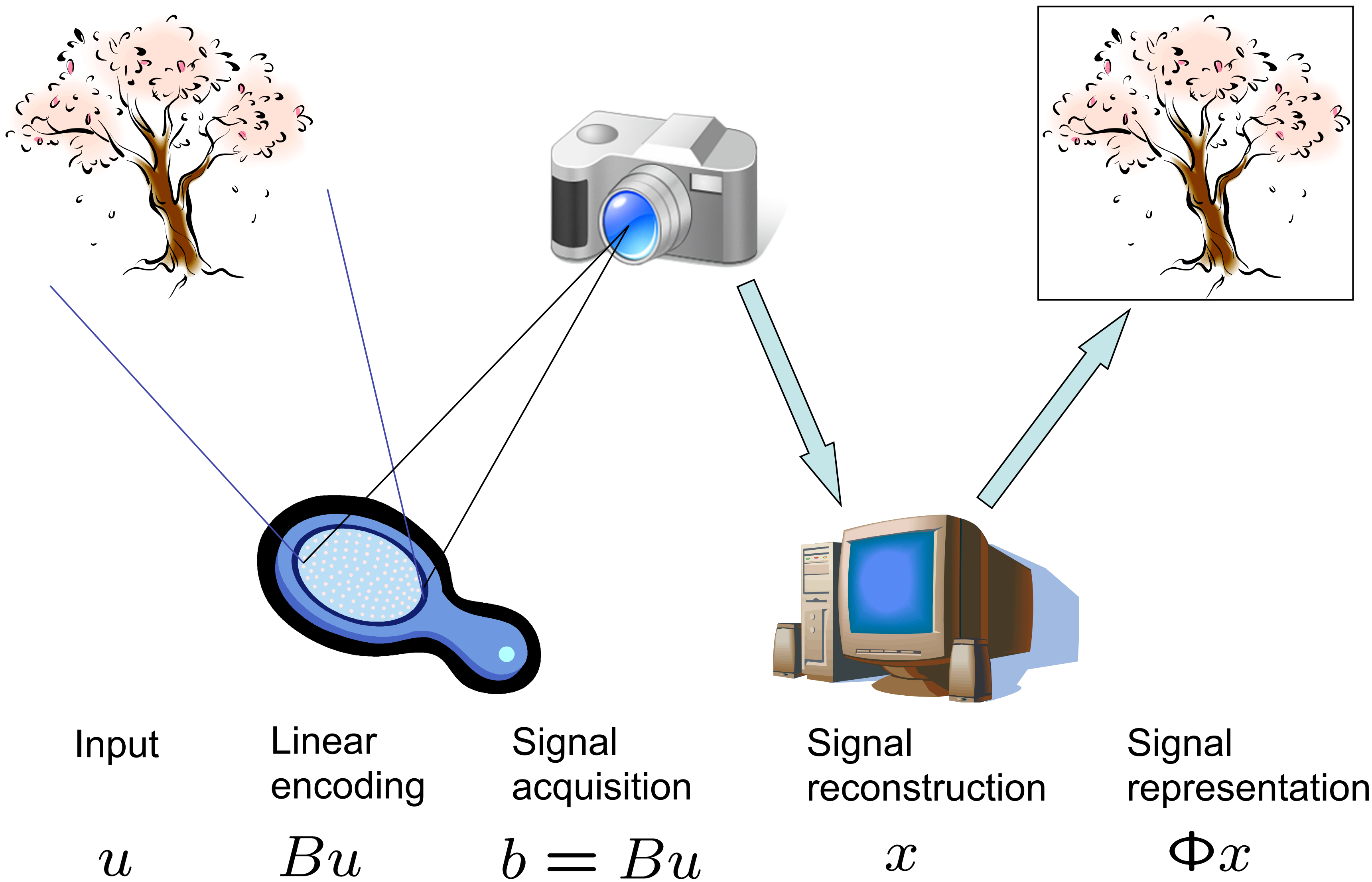
- Can we do better than this?

# Introduction to *Compressed Sensing*

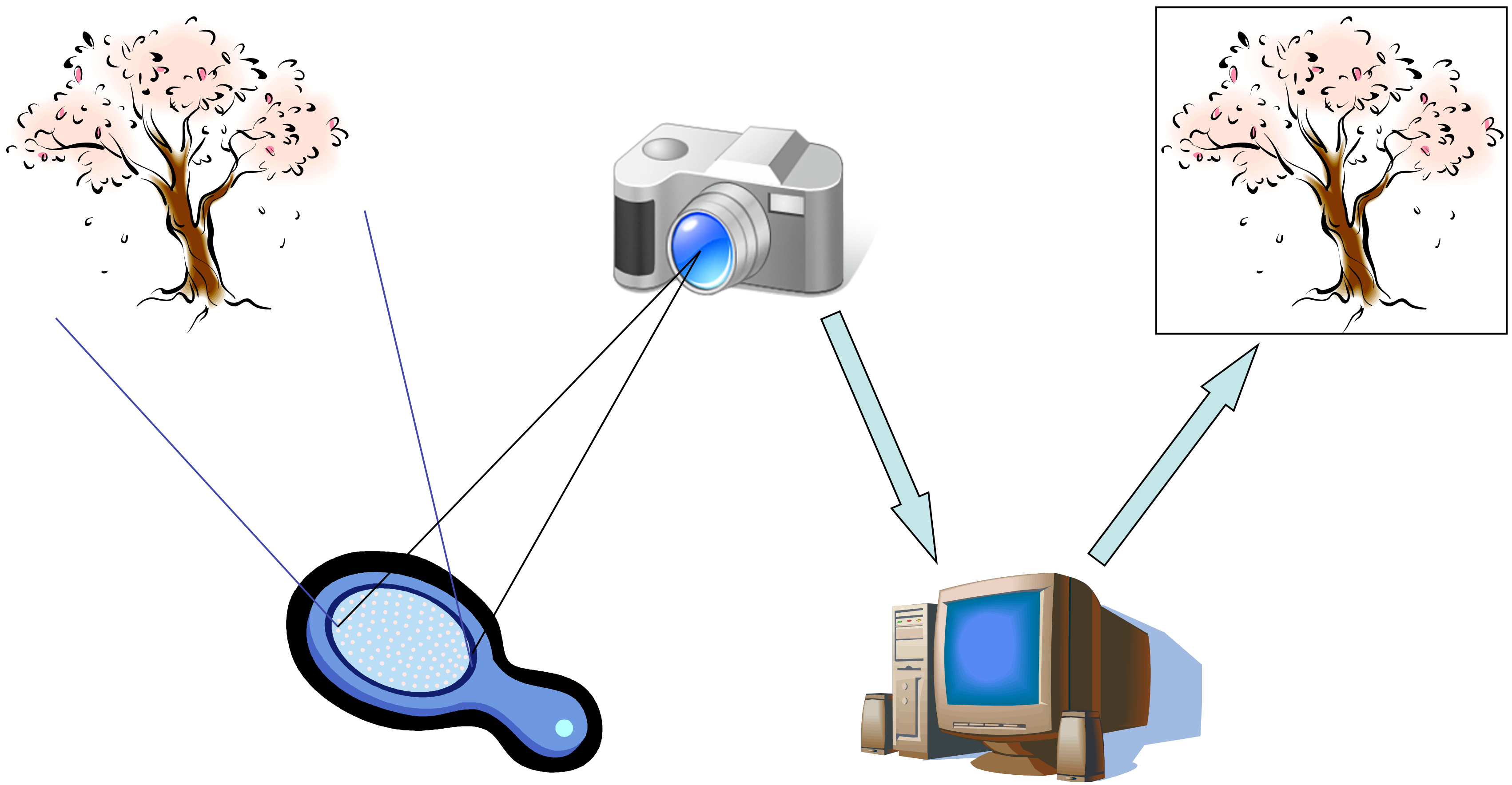
- Let $k$=$||x||_0$, $n$=dim($x$)=dim($u$).

- In compressed sensing based on $l_1$ minimization, the number of measurements is $m$=O($k$ log($n/k$)) (Donoho, Candés-Tao)

# Introduction to *Compressed Sensing*



# Introduction to *Compressed Sensing*



| Input | Linear encoding | Signal acquisition | Signal reconstruction | Signal representation |
|---|---|---|---|---|
| $u$ | $Bu$ | $b = Bu$ | $x$ | $\Phi x$ |

# Introduction to *Compressed Sensing*

- Input: $b$=$Bu$=$B\Phi x$, $A$=$B\Phi$
- Output: $x$

- In compressed sensing, $m$=dim($b$)<<dim($u$)=dim($x$)=$n$

- Therefore, $Ax$ = $b$ is an *underdetermined* system

- Approaches for recovering $x$ (hence the image $u$):
  - Solve min $||x||_0$, subject to $Ax$ = $b$
  - Solve min $||x||_1$, subject to $Ax$ = $b$
  - Other approaches

# Difficulties

- Large scales
- Completely dense data: $A$

However

- Solutions $x$ are expected to be sparse
- The matrices $A$ are often fast transforms

# Recovery by using the $l_1$-norm

Sparse signal reconstruction

$$\begin{aligned} \min \quad & ||x||_0 \\ s.t. \quad & Ax = b. \end{aligned}$$

Sparse signal $x \in \mathbf{R}^n$, matrix $A \in \mathbf{R}^{m \times n}, n >> m$

The system is underdetermined, but if card(x)<m, can recover signal.

The problem is NP-hard in general. Typical relaxation,

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & Ax = b. \end{aligned}$$

## Signal recovery

- Shown by Candes & Tao and Donoho that under certain conditions on matrix A the sparse signal

$$\begin{aligned} \min \quad & ||x||_0 \\ s.t. \quad & Ax = b. \end{aligned}$$

is recovered exactly by solving the convex relaxation

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & Ax = b. \end{aligned}$$

- The matrix property is called “restricted isometry property”

# Restricted Isometry Property

- A vector is said to be s-sparse if it has at most s nonzero entries.
- For a given s the isometric constant $\delta_s$ of a matrix A is the smallest constant such that

$$(1-\delta_s)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta_s)\|x\|_2^2$$

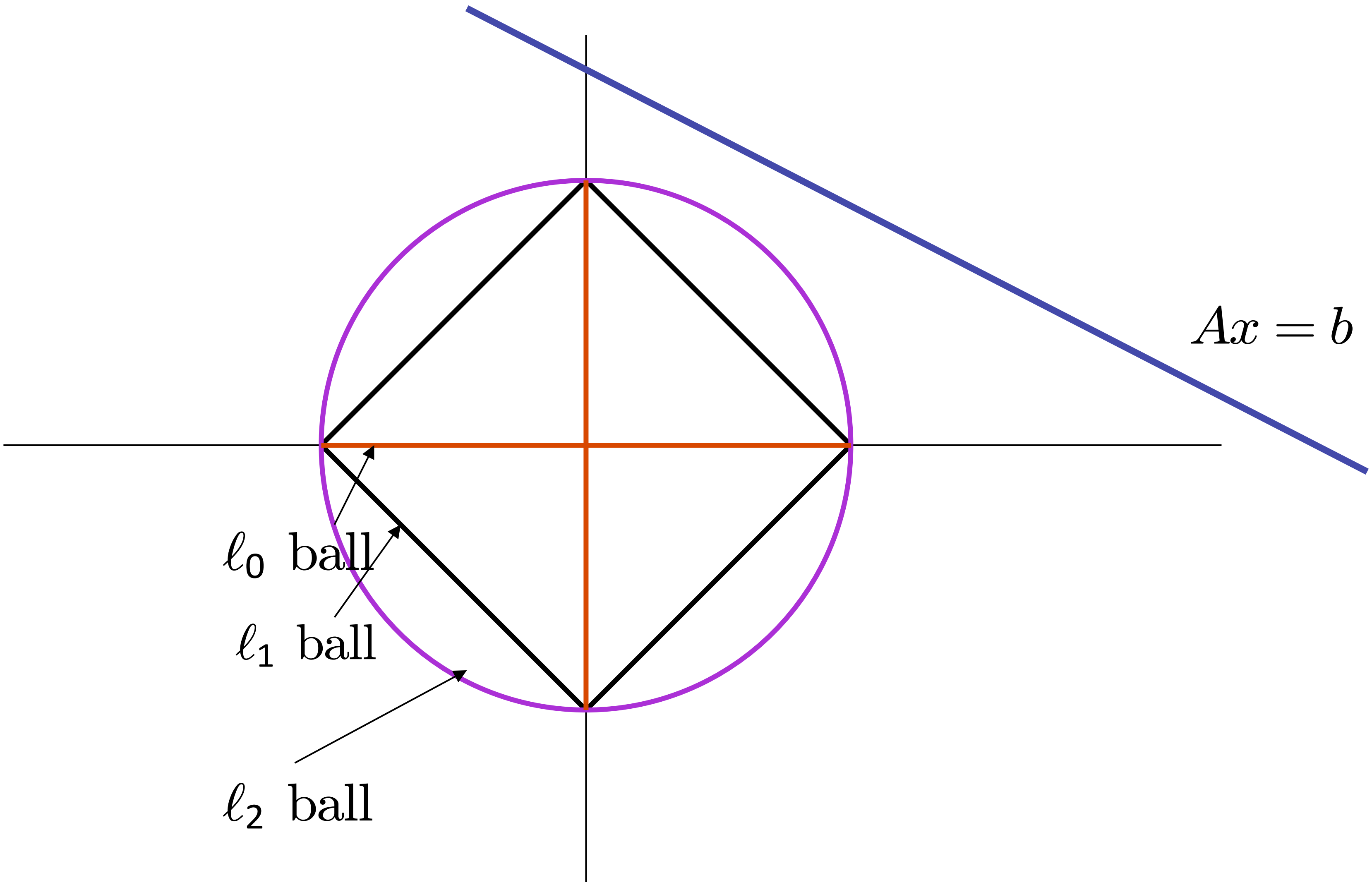
- for any s-sparse $x$ .

Assume that solution $x^*$ to $\min\{\|x\|_0 : Ax = b\}$ is $s$-sparse.

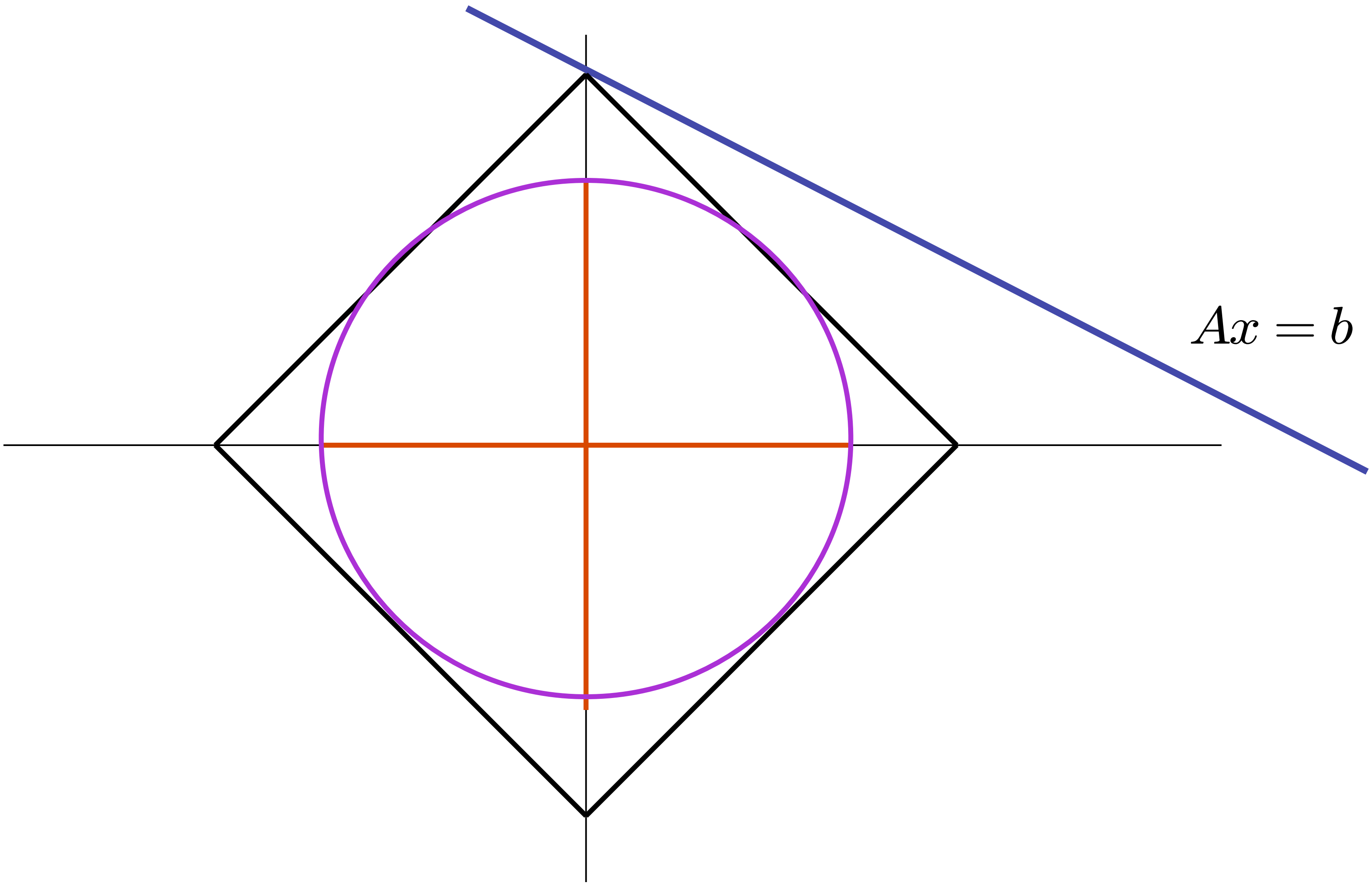
If $\delta_{2s}(A) < 1$ then $x^*$ is the unique solution to $\min\{\|x\|_0 : Ax = b\}$.

If $\delta_{2s}(A) < \sqrt{2} - 1$ then $x^*$ is the solution to $\min\{\|x\|_1 : Ax = b\}$

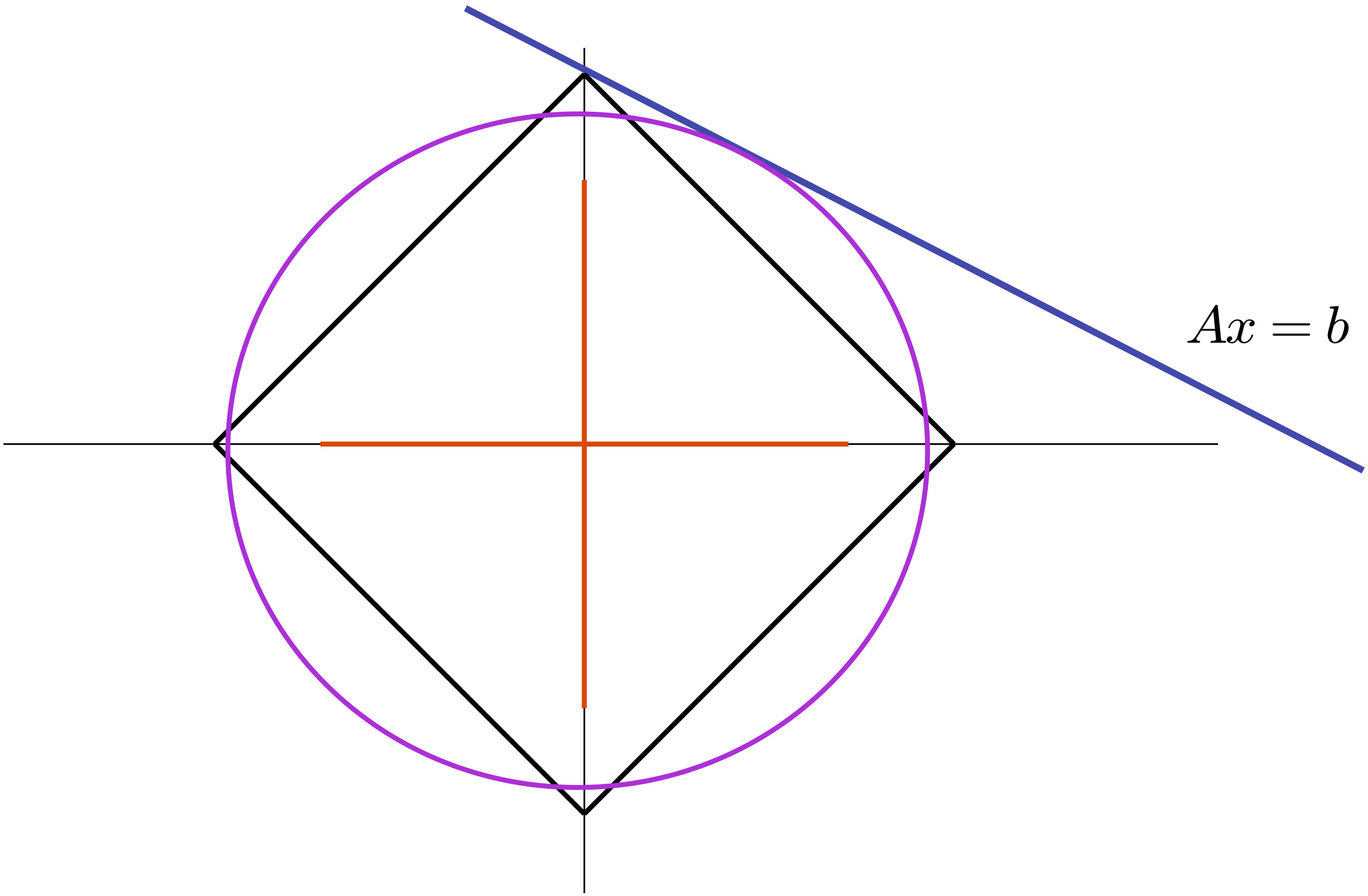
## Why $||\cdot||_1$ norm?



# Why $||\cdot||_1$ norm?

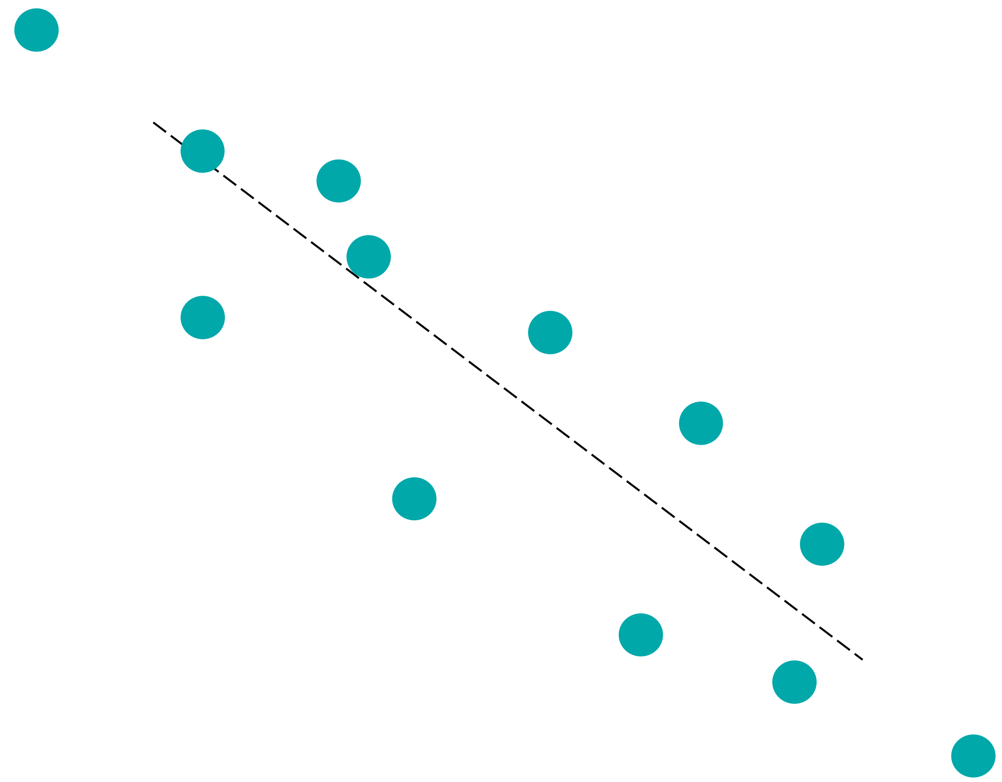


# Why $\|\cdot\|_1$ norm?



# Sparse regularized regression

# Least Squares Linear Regression



# Least squares problem

Standard form of LS problem

$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 \;\Rightarrow\; x = (A^\top A)^{-1} A^\top b$$

Includes solution of a system of linear equations Ax=b.

May be used with additional linear constraints, e.g.

$$\min_{l \leq x \leq u} ||Ax - b||_2^2$$

Ridge regression

$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 + \lambda ||x||_2^2 \;\Rightarrow\; x = (A^\top A + I)^{-1} A^\top b$$

$\lambda$ is the regularization parameter – the trade-off weight.

# Robust least squares regression

Assume matrix $A$ is not known exactly, but each column
$A_i \in B(A_i^0, r) = \{A_i : \|A_i - A_i^0\| \le r\}$
$\Rightarrow A \in \mathcal{A} = B(A_1^0, r) \otimes \ldots \otimes B(A_n^0, r)$.

$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 \Rightarrow \min_{x \in \mathbf{R}^n} \max_{A \in \mathcal{A}} \|Ax - b\|_2^2$$

Less straightforward than for SVM but it is possible to show that the above problem leads to

$$\min_{x \in \mathbf{R}^n} ||A^0 x - b||_2^2 + r||x||_1$$

Another interpretation – feature selection

# Lasso and other formulations

Sparse regularized regression or Lasso:

$$\min \quad \frac{1}{2}||Ax - b||^2 + \lambda ||x||_1$$

Sparse regressor selection

$$\begin{aligned} \min \quad & ||Ax - b|| \\ s.t. \quad & ||x||_1 \leq t. \end{aligned}$$

Noisy signal recovery

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & ||Ax - b|| \leq \epsilon. \end{aligned}$$

## Connection between different formulations

$$\begin{array}{ll} \min & ||Ax - b|| \\ s.t. & ||x||_1 \leq t. \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \min & ||Ax - b||^2 \\ s.t. & ||x||_1 \leq t. \end{array}$$

$$\min \quad \frac{1}{2}||Ax - b|| + \lambda ||x||_1 \quad \not\Longleftrightarrow \quad \min \quad \frac{1}{2}||Ax - b||^2 + \lambda ||x||_1$$

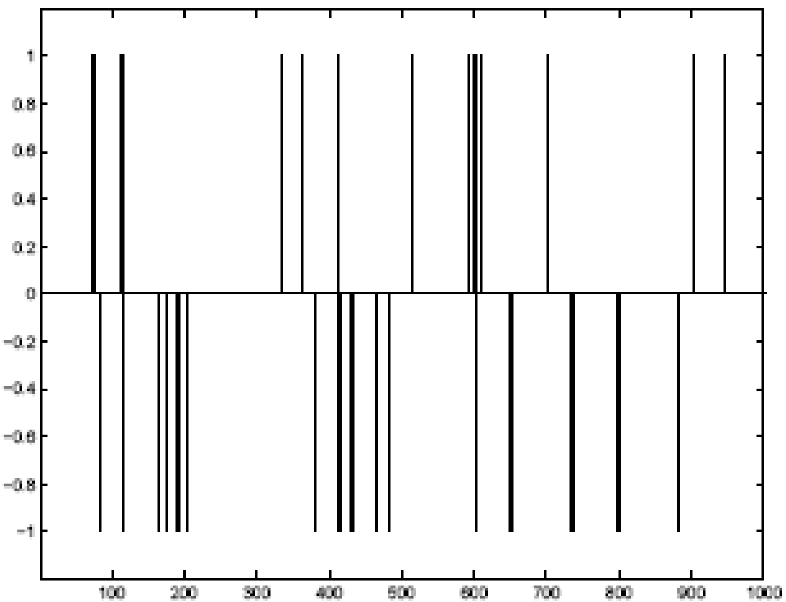
$$\min \quad \frac{1}{2}||Ax - b|| + \lambda ||x||_1 \quad \Longleftrightarrow \quad \begin{array}{ll} \min & ||Ax - b|| \\ s.t. & ||x||_1 \leq t. \end{array}$$

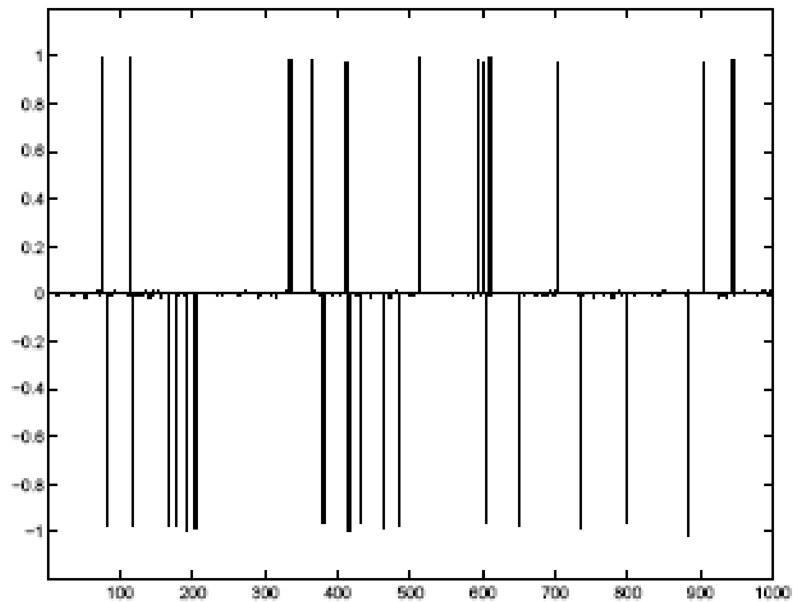
If solution exists

$$\begin{array}{ll} \min & ||x||_1 \\ s.t. & Ax = b \end{array}$$

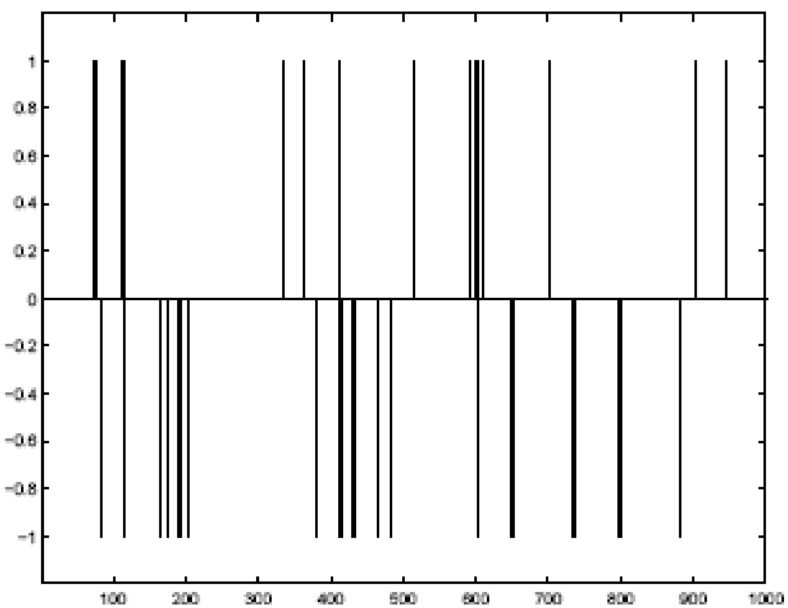
# Example

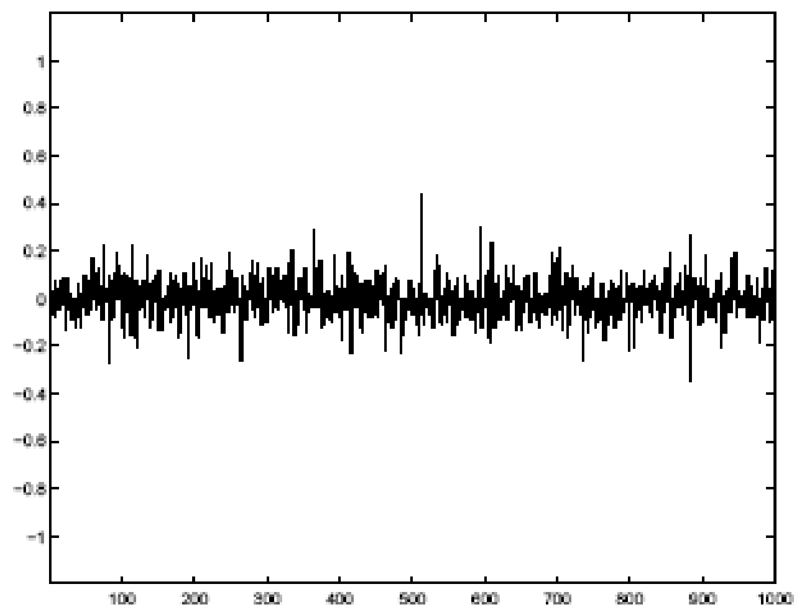
- signal $x \in \mathbf{R}^n$ with $n = 1000$, $\mathbf{card}(x) = 30$
- $m = 200$ (random) noisy measurements: $y = Ax + v$, $v \sim \mathcal{N}(0, \sigma^2 \mathbf{1})$, $A_{ij} \sim \mathcal{N}(0, 1)$
- *left*: original; *right*: $\ell_1$ reconstruction with $\gamma = 10^{-3}$



- $\ell_2$ reconstruction; minimizes $\|Ax - y\|_2 + \gamma\|x\|_2$, where $\gamma = 10^{-3}$
- *left*: original; *right*: $\ell_2$ reconstruction



## Types of convex problems

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & Ax = b \end{aligned}$$

Variable substitution: $x = x' - x'',\ x' \geq 0,\ x'' \geq 0$

$$\begin{aligned} \min \quad & e^\top (x' + x'') \\ \text{s.t.} \quad & A(x' - x'') = b \\ & x' \geq 0, x'' \geq 0 \end{aligned}$$

Linear programming problem

## Types of convex problems

$$\min \quad \frac{1}{2}||Ax - b|| + \lambda||x||_1$$

Variable substitution: $x = x' - x'', \; x' \geq 0, \; x'' \geq 0$

$$\begin{aligned} \min \quad & \frac{1}{2}||A(x' - x'') - b|| + \lambda e^\top (x' + x'') \\ \text{s.t.} \quad & x' \geq 0, x'' \geq 0 \end{aligned}$$

Convex non-smooth objective with linear inequality constraints

# Types of convex problems

Convex QP with linear inequality constraints

$$\begin{aligned} \min \quad & ||Ax - b||^2 \\ s.t. \quad & ||x||_1 \le t. \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \min \quad & ||(Ax' - Ax'' - b||^2 \\ s.t. \quad & e^\top (x', x'') \le t. \\ & x', x'' \ge 0 \end{aligned}$$

SOCP

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & ||Ax - b|| \le \epsilon. \end{aligned}$$

# Optimization approaches

# Lasso

Regularized regression or Lasso:

$$\min \quad \frac{1}{2}||Ax - b||^2 + \lambda ||x||_1$$

$$\begin{aligned} \min \quad & \frac{1}{2}||Ax' - Ax'' - b||^2 + \lambda e^\top (x' + x'') \\ s.t. \quad & x', x'' \geq 0 \end{aligned}$$

Convex QP with nonnegativity constraints

## Standard QP formulation

Reformulate as

$$\min \quad \frac{1}{2}||Mz - b||^2 + \lambda \sum_{i=1}^{n} z_i$$

$$\text{s.t.} \quad z \geq 0 \qquad M = [A, -A]$$

$$\min \quad \frac{1}{2} z^\top M^\top M z - b^\top M z + \lambda \sum_{i=1}^{n} z_i$$

$$\text{s.t.} \quad z \geq 0.$$

How is it different from SVMs dual QP?

## Standard QP formulation

Reformulate as

$$\begin{aligned} \min \quad & \frac{1}{2}||Mz - b||^2 + \lambda \sum_{i=1}^{n} z_i \\ \text{s.t.} \quad & z \geq 0 \qquad M = [A, -A] \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2} z^\top M^\top M z - b^\top M z + \lambda \sum_{i=1}^{n} z_i \\ \text{s.t.} \quad & z \geq 0. \end{aligned}$$

# Standard QP formulation

Reformulate as

$$\begin{aligned} \min \quad & \frac{1}{2}||Mz - b||^2 + \lambda \sum_{i=1}^{n} z_i \\ \text{s.t.} \quad & z \geq 0 \qquad M = [A, -A] \end{aligned}$$

$$\begin{aligned} \min \quad & \frac{1}{2} z^\top M^\top M z - b^\top M z + \lambda \sum_{i=1}^{n} z_i \\ \text{s.t.} \quad & z \geq 0. \end{aligned}$$

Features of this QP

1. $Q=M^TM$, where M is m$\times$n, with n>>m.
2. Forming Q is $O(m^2n)$, factorizing Q+D is $O(m^3)$
3. There are no upper bound constraints.

**IPM complexity is $O(m^3)$ per iteration**

# Dual Problem

$$\min \quad \frac{1}{2}||Ax' - Ax'' - b||^2 + \lambda(x' + x'')$$
$$s.t. \quad x', x'' \geq 0$$

$$\boxed{L(x', x'', s', s'') = \frac{1}{2}||Ax' - Ax'' - b||^2 + \lambda e^\top(x' + x'') - s'^\top x' - s''^\top x''}$$

$$\nabla_{x'} L(x', x'', s', s'') = A^\top(Ax' - Ax'' - b) + \lambda e - s' = 0$$
$$\nabla_{x''} L(x', x'', s', s'') = -A^\top(Ax' - Ax'' - b) + \lambda e - s'' = 0$$
$$s', s'' \geq 0$$

# Dual Problem

Using:

$$(x')^\top A^\top (Ax' - Ax'' - b) + \lambda^\top x' - s'^\top x' = 0$$

$$-(x'')^\top A^\top (Ax' - Ax'' - b) + \lambda^\top x'' - s''^\top x'' = 0$$

$$\max_s \min_x L(x', x'', s', s'') =$$

$$\frac{1}{2}(Ax' - Ax'' - b)^\top (Ax' - Ax'' - b) + \lambda e^\top (x' + x'') - s'^\top x' - s''^\top x'' =$$

$$-\frac{1}{2}(Ax' - Ax'')^\top (Ax' - Ax'') = -\frac{1}{2} x^\top A^\top A x$$

# Lasso

Primal-Dual pair of problems

$$\min \quad \frac{1}{2}||Ax - b||^2 + \lambda||x||_1$$

$$\begin{aligned} \min \quad & \frac{1}{2}x^\top A^\top Ax \\ s.t. \quad & \|A^\top(Ax - b)\|_\infty \leq \lambda \end{aligned}$$

## Optimality Conditions

$$\text{(i)} \quad x_i < 0, \text{ and } (A^\top(Ax - b))_i = \lambda,$$

$$\text{(ii)} \quad x_i > 0, \text{ and } (A^\top(Ax - b))_i = -\lambda,$$

$$\text{(iii)} \quad x_i = 0, \text{ and } -\lambda \leq A^\top(Ax - b)_i \leq \lambda$$

# Coordinate descent

# Coordinate descent

Choose one variable $x_i$ and column $A_i$.
Let $\bar{x}$ and $\bar{A}$ correspond to the fixed part

$$\min_{x_i} \quad \frac{1}{2}||A_i x_i + \bar{A}\bar{x} - b||^2 + \lambda|x_i|$$

## Soft-thresholding operator

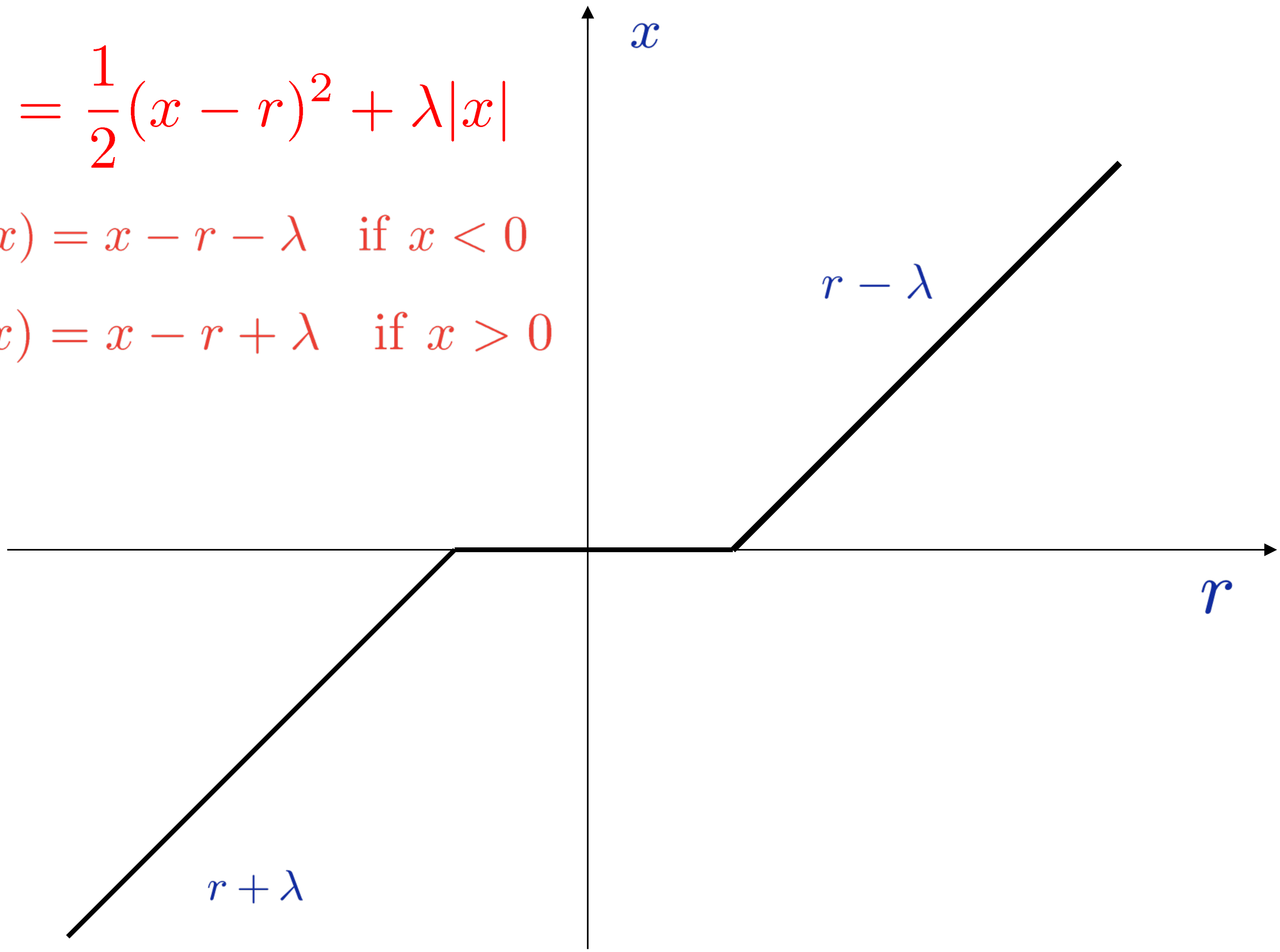
$$\min_{x_i} \frac{1}{2}(x_i - r)^2 + \lambda|x| \rightarrow x_i = \begin{cases} r - \lambda & \text{if } r > \lambda \\ 0 & \text{if } -\lambda \leq r \leq \lambda \\ r + \lambda & \text{if } r < -\lambda \end{cases}$$

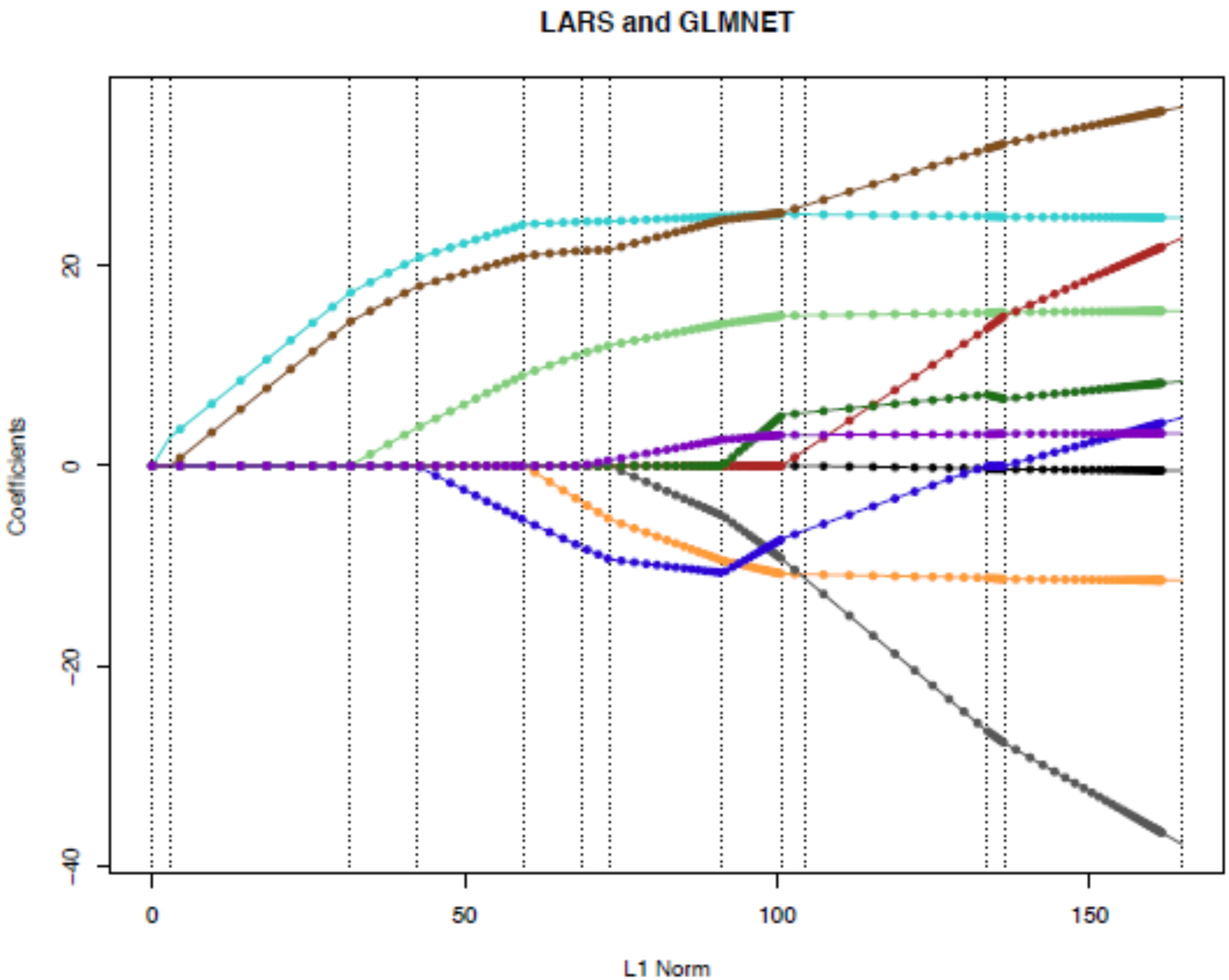
$$r = -{A_i}^\top(\bar{A}\bar{x} - b)/||A_i||^2, \ \lambda \rightarrow \lambda/||A_i||^2$$

$$f(x) = \frac{1}{2}(x-r)^2 + \lambda|x|$$

$$\nabla_x f(x) = x - r - \lambda \quad \text{if } x < 0$$

$$\nabla_x f(x) = x - r + \lambda \quad \text{if } x > 0$$




LARS and GLMNET
Coefficients
L1 Norm
20
0
−20
−40
0
50
100
150