# Section 5
# Convex Optimisation 1

# Convex Combination

## Definition 5.1

A *convex combination* is a linear combination of points where all coefficients are non-negative and sum to 1.

More specifically, let $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_\ell \in \mathbb{R}^n$. A convex combination of these points is of the form

$$\sum_{i=1}^{\ell} \lambda_i \boldsymbol{x}_i,$$

where the real coefficients $\lambda_i$ satisfy $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$.
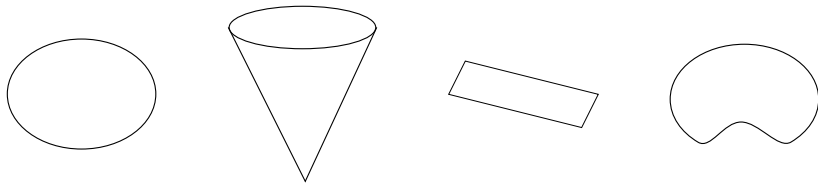
# Convex Sets

## Definition 5.2

A set $\mathcal{X}$ is a *convex set* if and only if the convex combination of any two points in the set belongs to the set.
That is,

$$\mathcal{X} \subseteq \mathbb{R}^n \text{ is convex} \iff \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \ \lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2 \in \mathcal{X}, \ \forall \lambda \in [0, 1].$$

# Examples



Example of convex sets:

- A *hyperplane* $\mathcal{H} = \left\{ \boldsymbol{x} : \boldsymbol{a}^T \boldsymbol{x} = b \right\}$, where $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a} \neq \boldsymbol{0}$, and $b \in \mathbb{R}$.
- A *halfspace* $\mathcal{H}_+ = \left\{ \boldsymbol{x} : \boldsymbol{a}^T \boldsymbol{x} \leq b \right\}$, where $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a} \neq 0$, and $b \in \mathbb{R}$.
- A *polyhedron*
  $\mathcal{P} = \left\{ \boldsymbol{x} : \boldsymbol{a}_j^T \boldsymbol{x} \leq b_j, \ j = 1, \cdots, m, \ \boldsymbol{c}_j^T \boldsymbol{x} = d_j, \ j = 1, \cdots, p \right\}$.
- Intersections of convex sets are convex.

# Convex Functions

## Definition 5.3

The *domain* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as the set of the points where the function $f$ is finite, i.e.,
$$\mathrm{dom} f = \{\boldsymbol{x} \in \mathbb{R}^n : |f(\boldsymbol{x})| < \infty\}.$$

Example: $\mathrm{dom} \log x = \mathbb{R}^+$.
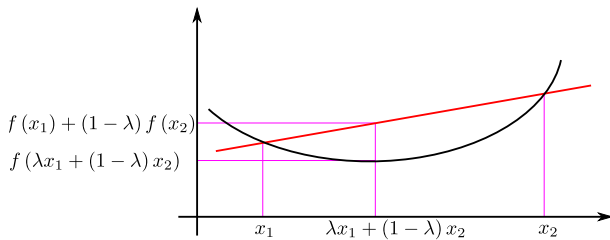
## Definition 5.4 (Convex functions)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathrm{dom} f \subseteq \mathbb{R}^n$, $\lambda \in [0, 1]$, it holds
$$\lambda f(\boldsymbol{x}_1) + (1 - \lambda) f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1 - \lambda) \boldsymbol{x}_2).$$

This definition implies that $\mathrm{dom} f$ is convex. However, in this lecture notes, we usually assume $\mathrm{dom}\, f = \mathbb{R}^n$ for simplicity.
A function $f$ is *strictly convex* if strict inequality holds whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and $\lambda \in (0, 1)$.
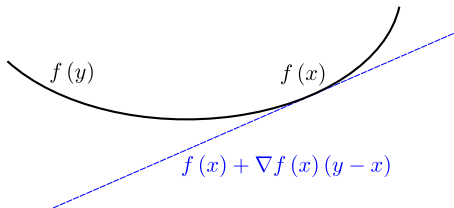
# A Convex Function

# First-Order Condition of Convexity

## Theorem 5.5

*Suppose a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then it is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$, it holds*

$$f\left(\boldsymbol{y}\right) \geq f\left(\boldsymbol{x}\right) + \nabla f\left(\boldsymbol{x}\right)^T \left(\boldsymbol{y} - \boldsymbol{x}\right). \tag{12}$$



$f\left(y\right)$    $f\left(x\right)$

$f\left(x\right) + \nabla f\left(x\right)\left(y - x\right)$

# Necessity

Assume first that $f$ is convex and $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}\,(f)$. Since $\mathrm{dom}\,(f)$ is convex, $\boldsymbol{x} + t\,(\boldsymbol{y} - \boldsymbol{x}) \in \mathrm{dom}\,(f)$ for all $0 < t \leq 1$. By convexity of $f$,

$$f\left(\boldsymbol{x} + t\,(\boldsymbol{y} - \boldsymbol{x})\right) \leq (1 - t)\, f\,(\boldsymbol{x}) + t f\,(\boldsymbol{y}).$$

Divide both sides by $t$. It holds

$$f\,(\boldsymbol{y}) \geq f\,(\boldsymbol{x}) + \frac{f\left(\boldsymbol{x} + t\,(\boldsymbol{y} - \boldsymbol{x})\right) - f\,(\boldsymbol{x})}{t}.$$

Take the limit as $t \to 0$ yields (12).

# Sufficiency

To show the other direction (sufficiency), assume that (12) holds. Choose any $x \neq y$ and $\lambda \in [0, 1]$. Let $z = \lambda x + (1 - \lambda) y$. Applying (12) twice yields

$$f(x) \geq f(z) + \nabla f(z)^T (x - z),$$
$$f(y) \geq f(z) + \nabla f(z)^T (y - z).$$

Multiply the first inequality by $\lambda$ and the second by $1 - \lambda$, and then add them together. It holds

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(z)$$
$$+ \lambda \nabla f(z)^T (x - z) + (1 - \lambda) \nabla f(z)^T (y - z).$$

Now note that $x - z = (1 - \lambda)(x - y)$ and $y - z = -\lambda(x - y)$. One obtains

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(z),$$

which proves that $f$ is convex.

# Sublevel Sets

## Definition 5.6 (Sublevel Sets, a.k.a. Lower Contour Sets)

The $\alpha$-*sublevel set* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$\mathcal{C}_\alpha = \{\boldsymbol{x} \in \mathrm{dom}\,(f) : \ f(\boldsymbol{x}) \leq \alpha\}.$$

# Sublevel Sets of Convex Functions
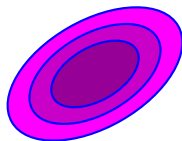
## Lemma 5.7

*Sublevel sets of a convex function $f$ are convex.*

Proof: We shall show that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}_\alpha$, it holds $\lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y} \in \mathcal{C}_\alpha$ for all $\lambda \in [0, 1]$. By the definition of $\mathcal{C}_\alpha$, $f(\boldsymbol{x}) \leq \alpha$ and $f(\boldsymbol{y}) \leq \alpha$. By the convexity of $f$,

$$f(\lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}) \leq \alpha,$$
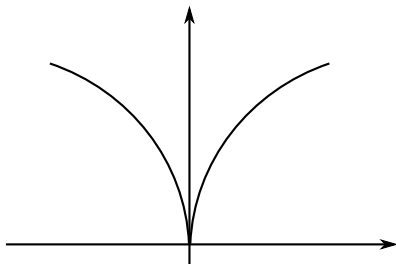
which proves this proposition.

# Sublevel Sets

The converse of Lemma 5.7 is not true.
That sublevel sets of a function $f$ are convex does not imply that $f$ is convex.

# Norm

We've seen $\ell_p$-norm in Definition 4.7.

## Definition 5.8

Given a vector space $\mathcal{V}$ over the field $\mathbb{F}$ of complex (real) numbers, a norm on $\mathcal{V}$ is a function $p : \mathcal{V} \to \mathbb{R}$ with the following properties:
For all $a \in \mathbb{F}$ and all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}$,

1. $p(a\boldsymbol{v}) = |a|\, p(\boldsymbol{v})$, (absolute scalability)
2. $p(\boldsymbol{u} + \boldsymbol{v}) \le p(\boldsymbol{u}) + p(\boldsymbol{v})$, (triangle inequality)
3. if $p(\boldsymbol{v}) = 0$ then $\boldsymbol{v}$ is the zero vector. (separates points)

Positivity follows: By the first axiom, $p(\boldsymbol{0}) = 0$ and $p(-\boldsymbol{v}) = p(\boldsymbol{v})$.
Then by triangle inequality,
$$0 \le p(\boldsymbol{v}) + p(-\boldsymbol{v}) = 2p(\boldsymbol{v}) \;\Rightarrow\; 0 \le p(\boldsymbol{v}).$$

# Convexity of a Norm

## Lemma 5.9

*A norm is a convex function.*

Proof: For any given $u, v \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, it holds that

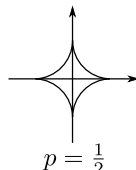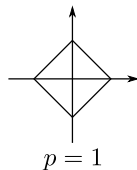$$\|\lambda u + (1 - \lambda) v\| \leq \|\lambda u\| + \|(1 - \lambda) v\|$$
$$= \lambda \|u\| + (1 - \lambda) \|v\|,$$

where we have used the triangle inequality and the absolute scalability. This establishes the convexity of the norm.

# $\ell_p$-Norm

In Definition 4.7, it mentioned that $\ell_p$-norm is a proper norm iff $p \geq 1$.

Can be verified by using sub-level argument.



$p = \infty$      $p = 2$      $p = 1$      $p = \frac{1}{2}$
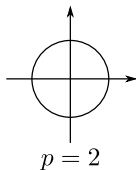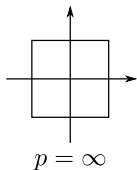
# Constrained Convex Optimization Problems

A constrained optimization problem of the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{subject to } h_i(\boldsymbol{x}) \leq 0, \ i = 1, \cdots, m,$$
$$\ell_i(\boldsymbol{x}) = 0, \ i = 1, \cdots, r,$$

is convex if

- the objective function $f_0$ is convex, and
- the feasible set is convex.
    - $h_i$'s are convex (consequence of Lemma 5.7).
    - $\ell_i$'s are affine, i.e., in the form of $\boldsymbol{a}_i^T \boldsymbol{x} + b_i = 0$.
      $\ell_i(\boldsymbol{x}) = 0 \Leftrightarrow \ell_i(\boldsymbol{x}) \leq 0$ and $-\ell_i(\boldsymbol{x}) \leq 0$.
      Both $\ell_i$ and $-\ell_i$ need to be convex $\Rightarrow \ell_i$ is affine.

# Local Optimality and Global Optimality

## Theorem 5.10

*Suppose that a feasible point $x$ is locally optimal for a convex optimization problem. Then it is also globally optimal.*

Proof: Suppose that $x$ is not globally optimal, i.e., there exists a feasible $y \neq x$ such that $f(y) < f(x)$. Consider a point $z$ on the line segment between $x$ and $y$, i.e.,

$$z = (1 - \lambda) x + \lambda y, \ \lambda \in (0, 1).$$

Then it is clear that

$$f(z) \leq (1 - \lambda) f(x) + \lambda f(y) < f(x),$$
$$h_i(z) \leq (1 - \lambda) h_i(x) + \lambda h_i(y) \leq 0, \ i = 0, 1, \cdots, m,$$
$$a_i^T z = (1 - \lambda) a_i^T x + \lambda a_i^T y = b_i, \ i = 1, \cdots, r,$$

where the inequalities follow from the convexity of the functions $f$ and $h_i$'s. Hence, the point $z$ is feasible and $f(z) < f(x)$ for all $\lambda \in (0, 1)$. This contradicts with that $x$ is not locally optimal and proves the global optimality of $x$.

# A Global Optimality Criterion

## Theorem 5.11

Suppose that the objective $f_0$ in a convex optimization problem is differentiable, i.e.,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}), \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f).$$

Let $\mathcal{X}$ denote the feasible set

$$\mathcal{X} = \left\{ \boldsymbol{x} : \ h_i(\boldsymbol{x}) \leq 0, \ i = 1, \cdots, m, \ \boldsymbol{a}_i^T \boldsymbol{x} = b_i, \ i = 1, \cdots, r \right\}.$$

Then an $\boldsymbol{x} \in \mathcal{X}$ is optimal if and only if

$$\nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) \geq 0, \ \forall \boldsymbol{y} \in \mathcal{X}.$$

# Consequence of Theorem 5.11

▶ For an unconstrained convex optimization problem, the sufficient and necessary condition for a globally optimal point $x$ is given by

$$\nabla f\left(x\right) = 0.$$

▶ In a constrained convex optimization problem, it may happen that

$$\nabla f\left(x\right) \neq 0.$$

This implies that $x$ is at the boundary of the feasible set. (This is actually linked to KKT conditions and will be discussed later.)

# Proof

The proof of sufficiency is straightforward. Suppose the inequality holds. Then for all $y \in \mathcal{X}$,
$$f\left(y\right) \geq f\left(x\right) + \nabla f\left(x\right)^T \left(y - x\right) \geq f\left(x\right).$$
Hence, the point $x$ is globally optimal.

Conversely, suppose $x$ is optimal, but the inequality does not hold, i.e., for some $y \in \mathcal{X}$ we have
$$\nabla f\left(x\right)^T \left(y - x\right) < 0.$$
Consider the point $z\left(t\right) = ty + \left(1 - t\right)x$, $t \in \left[0, 1\right]$. Clearly, $z\left(t\right)$ is feasible. Now
$$\begin{aligned}\frac{d}{dt}\left.f\left(z\left(t\right)\right)\right|_{t=0} &= \nabla f\left(z\left(0\right)\right) \cdot \left.\frac{d}{dt}z\left(t\right)\right|_{t=0}\\ &= \nabla f\left(x\right) \cdot \left(y - x\right) < 0,\end{aligned}$$
where the inequality comes from the assumption. It implies that for small positive $t$, we have $f\left(z\left(t\right)\right) < f\left(x\right)$, which contradicts the optimality of $x$. The necessity is therefore proved.

# Non-differentiable Functions: Subgradient

## Definition 5.12

If $f : \mathcal{U} \to \mathbb{R}$ is a convex function defined on a convex open set $\mathcal{U} \subset \mathbb{R}^n$, a vector $\boldsymbol{v} \in \mathbb{R}^n$ is called a <span style="color:red">subgradient</span> at a point $\boldsymbol{x} \in \mathcal{U}$ if

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \boldsymbol{v}^T (\boldsymbol{y} - \boldsymbol{x}), \ \forall \boldsymbol{y} \in \mathcal{U}.$$

The set of all subgradients at $\boldsymbol{x}$ is called the <span style="color:red">subdifferential</span> at $\boldsymbol{x}$ and is denoted $\partial f(\boldsymbol{x})$.

Remark: If $f$ is convex and its subdifferential at $\boldsymbol{x}$ contains exactly one subgradient, then $f$ is differentiable at $\boldsymbol{x}$.

# Example

$$f(x) = |x| \Rightarrow \partial f = \begin{cases} 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$



$f(x) = |x|$
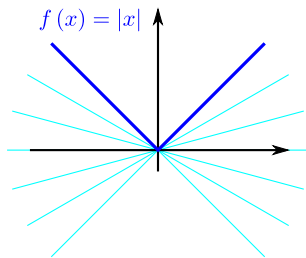
# Section 6
# $\ell_1$-Minimization

# $\ell_1$-Minimization

Want to solve the sparse linear inverse problem:
$$\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{e}.$$

Constrained optimization problem: if we know $\|\boldsymbol{e}\| \leq \epsilon$,
$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \text{ subject to } \|\boldsymbol{Ax} - \boldsymbol{y}\|_2 \leq \epsilon.$$

Unconstrained optimization problem: LASSO
$$\min_{\boldsymbol{x}} \tfrac{1}{2} \|\boldsymbol{Ax} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{x}\|_1.$$

$\exists$ a one-to-one correspondence between $\epsilon$ and $\lambda$.

- $\lambda \to 0$ implies $\epsilon \to 0$.
- $\lambda \to \infty$ implies $\epsilon \to \infty$.

# Why $\ell_1$-Minimization

A geometric intuition:
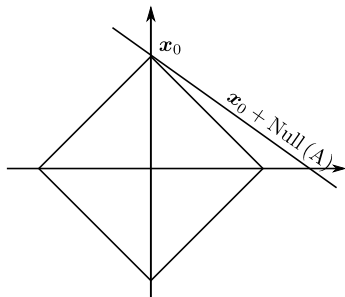


$\ell_1$ tends to give sparse solutions
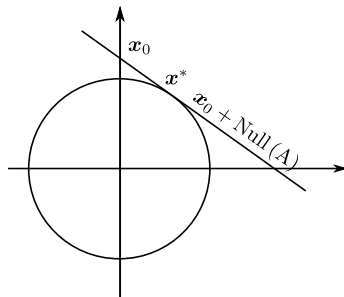
$\ell_2$ tends to give non-sparse solutions

Feasible solution for $\boldsymbol{y} = \boldsymbol{Ax}$: $\boldsymbol{x} \in \mathcal{X} = \boldsymbol{x}_0 + \mathcal{N}ull(\boldsymbol{A})$.

# Solve the Lasso Problem: Scalar Case

$$\min_x \underbrace{\frac{1}{2}(x-y)^2 + \lambda|x|}_{f(x)}.$$

The minimum of $f(x)$ is achieved at $x^{\#}$ s.t. $\frac{d}{dx}f(x^{\#}) = 0$, i.e.,

$$x^{\#} - y + \lambda \left.\frac{d|x|}{dx}\right|_{x^{\#}} = 0, \text{ where } \frac{d|x|}{dx} = \begin{cases} 1 & \text{if } x > 0, \\ [-1,1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

Or equivalently, $x^{\#}$ is given by the soft thresholding function

$$x^{\#} = \eta(y; \lambda) = \begin{cases} y - \lambda & \text{if } y \geq \lambda, \\ 0 & \text{if } -\lambda < y < \lambda, \\ y + \lambda & \text{if } y \leq -\lambda. \end{cases}$$
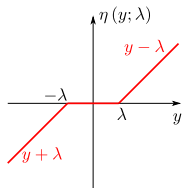
# Vector Case: the Gradient Descent Method

Gradient descent method: To solve $\min_{\boldsymbol{x}}\ f\left(\boldsymbol{x}\right),$
one iteratively updates

$$\boldsymbol{x}^k = \boldsymbol{x}^{k-1} - t_k \nabla f\left(\boldsymbol{x}^{k-1}\right),$$

where $t_k > 0$ is a suitable stepsize.

For Lasso problem $f\left(x\right) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1$ which is non-smooth.
Its gradient is given by (see details on page 6-22)

$$-\boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right) + \partial\left\|\boldsymbol{x}\right\|_1.$$

Gradient descent converges very slow.

# Gradient Descent Method: Another View

In gradient descent method:

$$\boldsymbol{x}^k = \boldsymbol{x}^{k-1} - t_k \nabla f\left(\boldsymbol{x}^{k-1}\right).$$

This is equivalent to minimize $\tilde{f}$,

$$\boldsymbol{x}_k = \arg \min_{\boldsymbol{x}} \ \tilde{f}\left(\boldsymbol{x}\right)$$

where

$$\tilde{f}\left(\boldsymbol{x}\right) := f\left(\boldsymbol{x}^{k-1}\right) + \left\langle \boldsymbol{x} - \boldsymbol{x}^{k-1}, \nabla f\left(\boldsymbol{x}^{k-1}\right)\right\rangle + \frac{1}{2t_k}\left\|\boldsymbol{x} - \boldsymbol{x}^{k-1}\right\|_2^2$$

$$= \frac{1}{2t_k}\left\|\boldsymbol{x} - \left(\boldsymbol{x}^{k-1} - t_k \nabla f\left(\boldsymbol{x}^{k-1}\right)\right)\right\|_2^2 + c.$$

# Iterative Shrinkage Thresholding (IST)

To solve $\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_1$, we apply the proximal regularization:
$$\boldsymbol{x}^k = \arg \min_{\boldsymbol{x}} \tilde{f}(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_1$$
where
$$\tilde{f}(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_1$$
$$:= f(\boldsymbol{x}^{k-1}) + \langle \boldsymbol{x} - \boldsymbol{x}^{k-1}, \nabla f(\boldsymbol{x}^{k-1}) \rangle + \frac{1}{2t_k} \|\boldsymbol{x} - \boldsymbol{x}^{k-1}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$$
$$= \frac{1}{2t_k} \|\boldsymbol{x} - (\boldsymbol{x}^{k-1} - t_k \nabla f(\boldsymbol{x}^{k-1}))\|_2^2 + \lambda \|\boldsymbol{x}\|_1 + c$$
$$= \sum_i \left[ \frac{1}{2t_k} (x_i - z_i)^2 + \lambda |x_i| \right] + c.$$

Therefore,
$$\boldsymbol{x}^k = \eta \left( \boldsymbol{x}^{k-1} + t_k \boldsymbol{A}^T (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{k-1}) ; \lambda t_k \right).$$

# Stable Recovery of Exact Sparse Signals

## Theorem 6.1

*Let $S$ be such that $\delta_{4S} \leq \frac{1}{2}$. Then for any signal $\boldsymbol{x}_0$ supported on $\mathcal{T}_0$ with $|\mathcal{T}_0| \leq S$ and any perturbation $\boldsymbol{e}$ with $\|\boldsymbol{e}\|_2 \leq \epsilon$, the solution $\boldsymbol{x}^{\#}$ obeys*

$$\left\| \boldsymbol{x}^{\#} - \boldsymbol{x}_0 \right\|_2 \leq C_S \cdot \epsilon,$$

*where the constant $C_S$ depends only on $\delta_{4S}$.*

## Typical value of $C_S$

$$C_S \approx \begin{cases} 8.82 & \text{for } \delta_{4S} = \frac{1}{5}, \\ 10.47 & \text{for } \delta_{4S} = \frac{1}{4}. \end{cases}$$

# Stable Recovery of Approximately Sparse Signals

## Theorem 6.2

*Suppose that $x_0$ is an an arbitrary vector in $\mathbb{R}^n$ and let $x_{0,S}$ be the truncated vector corresponding to the $S$ largest values of $x_0$ (in absolute value). When the matrix $A$ satisfies RIP, the solution $x^{\#}$ obeys*

$$\left\| x^{\#} - x_0 \right\|_2 \leq C_{1,S} \cdot \epsilon + C_{2,S} \cdot \frac{\|x_0 - x_{0,S}\|_1}{\sqrt{S}}.$$

No algorithm performs fundamentally better than $\ell_1$-min.

## Typical values

$C_{1,S} \approx 12.04$ and $C_{2,S} \approx 8.77$ for $\delta_{4S} = \frac{1}{5}$.

# Analysis for Exact Sparse Signals (1)

Assume that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$, $\|\boldsymbol{x}\|_0 \leq S$, and $\|\boldsymbol{w}\|_2 \leq \epsilon$.
Cast the recovery problem as

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \text{ subject to } \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \epsilon.$$

Tube constraint:

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 = \left\|\boldsymbol{A}\boldsymbol{x}^{\#} - \boldsymbol{A}\boldsymbol{x}_0\right\|_2 \leq \left\|\boldsymbol{A}\boldsymbol{x}^{\#} - \boldsymbol{y}\right\|_2 + \|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{y}\|_2 \leq 2\epsilon.$$

# Analysis for Exact Sparse Signals (2)

<span style="color:red">Cone constraint</span>: Let $\boldsymbol{x}^{\#} = \boldsymbol{x}_0 + \boldsymbol{h}$. Then

$$\left\| \boldsymbol{h}_{\mathcal{T}_0^c} \right\|_1 \leq \left\| \boldsymbol{h}_{\mathcal{T}_0} \right\|_1 .$$

<span style="color:cyan">Proof</span>:

$$
\begin{aligned}
\|\boldsymbol{x}_0\|_1 &\geq \left\| \boldsymbol{x}^{\#} \right\|_1 = \|\boldsymbol{x}_0 + \boldsymbol{h}\|_1 \\
&= \left\| (\boldsymbol{x}_0 + \boldsymbol{h})_{\mathcal{T}_0} \right\|_1 + \left\| \boldsymbol{h}_{\mathcal{T}_0^c} \right\|_1 \\
&\geq \|\boldsymbol{x}_0\|_1 - \left\| \boldsymbol{h}_{\mathcal{T}_0} \right\|_1 + \left\| \boldsymbol{h}_{\mathcal{T}_0^c} \right\|_1 .
\end{aligned}
$$
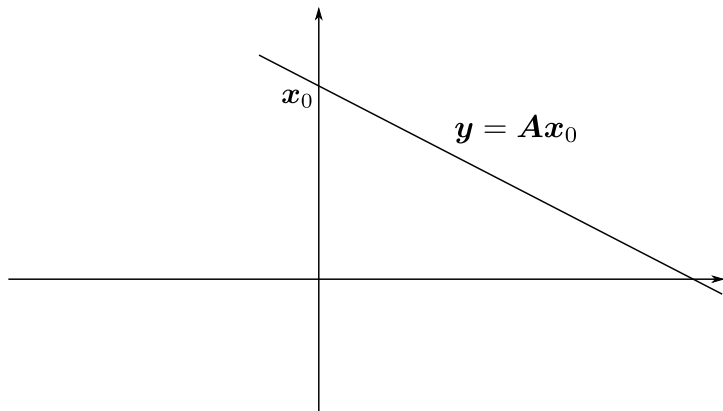
# Geometric Interpretation

# Geometric Interpretation



$$\left\| \boldsymbol{A}\boldsymbol{x}^{\#} - \boldsymbol{A}\boldsymbol{x}_0 \right\|_2 \leq 2\epsilon$$

$\boldsymbol{x}_0$

# Geometric Interpretation



$$\|\boldsymbol{h}_{\mathcal{T}_0^c}\|_1 \leq \|\boldsymbol{h}_{\mathcal{T}_0}\|_1$$

with $\boldsymbol{x}_0$ marked at the center.
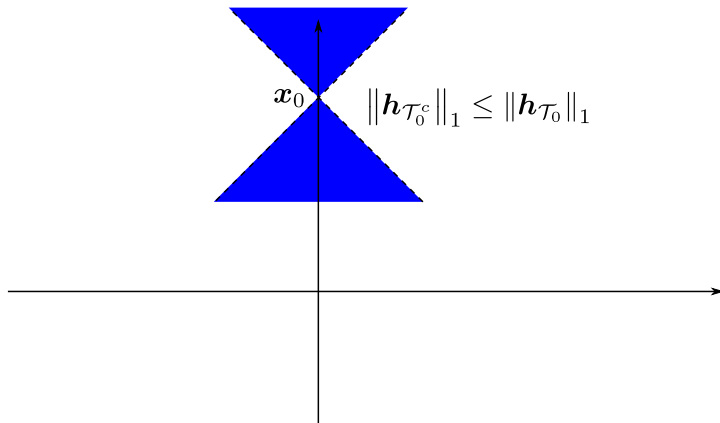
# Geometric Interpretation
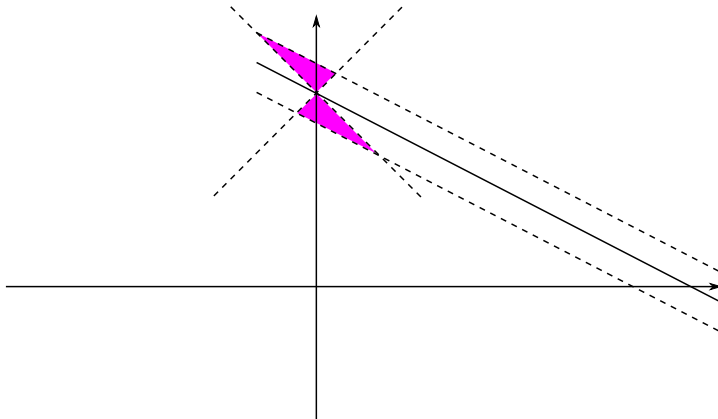
# Proof

Since $\|Ah\|_2 \leq 2\epsilon$, want to show $\|h\|_2 \approx \|Ah\|_2$.
(This is not true in general. For example $Ah = 0$ but $\|h\|_2$ can be $\infty$)

Divide $\mathcal{T}_0^c$ into subsets of size $M$ ($M = 3|\mathcal{T}_0|$).
List the entries in $\mathcal{T}_0^c$ as $n_1, \cdots, n_{N-|\mathcal{T}_0|}$ in decreasing order of their magnitudes.
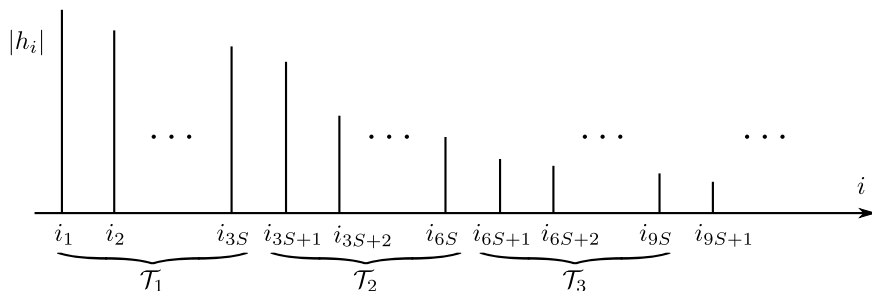Set $\mathcal{T}_j = \{n_\ell, \ (j-1)M + 1 \leq \ell \leq jM\}$.
Hence $\mathcal{T}_1$ contains the indices of the $M$ largest entries (in magnitude) of $h_{\mathcal{T}_0^c}$, $\mathcal{T}_2$ contains the indices of the next $M$ largest entries (in magnitude) of $h_{\mathcal{T}_0^c}$.
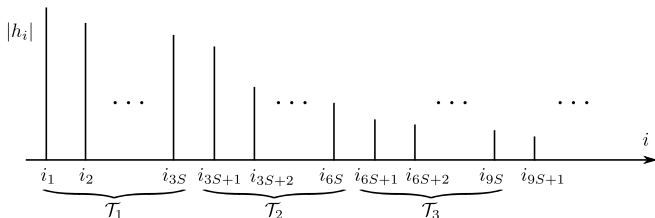


Define $\rho = |\mathcal{T}_0|/M$ ($\rho = 1/3$ when $M = 3|\mathcal{T}_0|$).

# Some Observations



▶ The $k^{th}$-largest value of $\boldsymbol{h}_{\mathcal{T}_0^c}$ obeys

$$\left|\boldsymbol{h}_{\mathcal{T}_0^c}(k)\right| \leq \frac{\sum_{\ell=1}^k \left|\boldsymbol{h}_{\mathcal{T}_0^c}(\ell)\right|}{k} \leq \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1 / k.$$

▶

$$\left|\boldsymbol{h}_{\mathcal{T}_{j+1}}(k)\right| \leq \frac{\left\|\boldsymbol{h}_{\mathcal{T}_j}\right\|_1}{M}.$$

# Proof: Step 1

The $\ell_2$-norm of $\boldsymbol{h}$ concentrates on $\mathcal{T}_{01} = \mathcal{T}_0 \bigcup \mathcal{T}_1$.

$$\|\boldsymbol{h}\|_2^2 = \|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2^2 + \|\boldsymbol{h}_{\mathcal{T}_{01}^c}\|_2^2 \leq (1 + \rho) \|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2^2.$$

**Proof**: From $\left|\boldsymbol{h}_{\mathcal{T}_0^c}\right|_{(k)} \leq \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1 / k$, it holds

$$\left\|\boldsymbol{h}_{\mathcal{T}_{01}^c}\right\|_2^2 \leq \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1^2 \sum_{k=M+1}^{N} \frac{1}{k^2}$$

$$\overset{(a)}{\leq} \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1^2 / M \overset{(b)}{\leq} \frac{\left\|\boldsymbol{h}_{\mathcal{T}_0}\right\|_1^2}{M}$$

$$\overset{(c)}{\leq} \frac{\left\|\boldsymbol{h}_{\mathcal{T}_0}\right\|_2^2 \cdot |\mathcal{T}_0|}{M} \leq \rho \left\|\boldsymbol{h}_{\mathcal{T}_{01}}\right\|_2^2,$$

where $(a)$ holds as $\sum_{k=M+1}^{N} 1/k^2 \leq 1/M$, $(b)$ is from the $\ell_1$-cone constraint, and $(c)$ comes from the Cauchy-Schwartz inequality.

# Proof: Step 2 - A Technical Result

$$\sum_{j \geq 2} \left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_2 \leq \sqrt{\rho} \cdot \left\| \boldsymbol{h}_{\mathcal{T}_0} \right\|_2 .$$

Proof: By construction $\left| \boldsymbol{h}_{\mathcal{T}_{j+1}} (k) \right| \leq \left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_1 / M$. Then

$$\left\| \boldsymbol{h}_{\mathcal{T}_{j+1}} \right\|_2^2 = \sum_{k \in \mathcal{T}_{j+1}} \left| \boldsymbol{h}_{\mathcal{T}_{j+1}} (k) \right|^2 \leq M \cdot \frac{\left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_1^2}{M^2} = \frac{\left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_1^2}{M}.$$

Hence,

$$\sum_{j \geq 2} \left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_2 \leq \sum_{j \geq 2} \left\| \boldsymbol{h}_{\mathcal{T}_{j-1}} \right\|_1 / \sqrt{M} \overset{(a)}{=} \sum_{j \geq 1} \left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_1 / \sqrt{M} = \left\| \boldsymbol{h}_{\mathcal{T}_0^c} \right\|_1 / \sqrt{M}$$

$$\overset{(b)}{\leq} \left\| \boldsymbol{h}_{\mathcal{T}_0} \right\|_1 / \sqrt{M} \overset{(c)}{\leq} \sqrt{\frac{|\mathcal{T}_0|}{M}} \left\| \boldsymbol{h}_{\mathcal{T}_0} \right\|_2 = \sqrt{\rho} \left\| \boldsymbol{h}_{\mathcal{T}_0} \right\|_2 ,$$

where $(a)$ uses the variable change $j' = j - 1$, $(b)$ and $(c)$ follow from the cone constraint and the Cauchy-Schwartz inequality respectively.

# Proof: Step 3

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 = \left\|\boldsymbol{A}_{\mathcal{T}_{01}}\boldsymbol{h}_{\mathcal{T}_{01}} + \sum_{j\geq 2}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_{\mathcal{T}_j}\right\|_2 \geq \|\boldsymbol{A}_{\mathcal{T}_{01}}\boldsymbol{h}_{\mathcal{T}_{01}}\|_2 - \left\|\sum_{j\geq 2}\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_{\mathcal{T}_j}\right\|_2$$

$$\geq \|\boldsymbol{A}_{\mathcal{T}_{01}}\boldsymbol{h}_{\mathcal{T}_{01}}\|_2 - \sum_{j\geq 2}\left\|\boldsymbol{A}_{\mathcal{T}_j}\boldsymbol{h}_{\mathcal{T}_j}\right\|_2$$

$$\geq \sqrt{1-\delta_{|\mathcal{T}_0|+M}}\,\|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2 - \sqrt{1+\delta_M}\sum_{j\geq 2}\left\|\boldsymbol{h}_{\mathcal{T}_j}\right\|_2$$

$$\geq \underbrace{\left(\sqrt{1-\delta_{4S}} - \sqrt{\rho}\sqrt{1+\delta_{4S}}\right)}_{C_{4S}}\|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2.$$
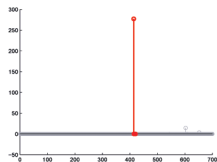
Hence,

$$\|\boldsymbol{h}\|_2 \leq \sqrt{1+\rho}\,\|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2 \leq \frac{\sqrt{1+\rho}}{C_{4S}}\,\|\boldsymbol{A}\boldsymbol{h}\|_2 \leq \frac{\sqrt{1+\rho}}{C_{4S}}\cdot 2\epsilon.$$

# Face Recognition with Block Occlusion [Wright et al., 2009]
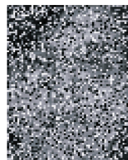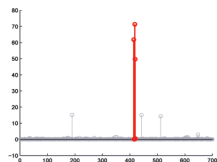
# The Setup

- A set of training samples $\{\phi_i, l_i\}$
    - $\phi_i \in \mathbb{R}^m$ is the vector representation of the images.
    - $l_i \in \{1, 2, \cdots, C\}$ label for the $C$ subjects.
- Test sample $y$

Assumption:

- For simplicity, assume a good face alignment.

# Face Recognition via Sparse Linear Regression

Sufficiently many images of the same subject $i$ form a low-dimensional linear subspace in $\mathbb{R}^m$.

$$\boldsymbol{y} \approx \sum_{\{j|l_j=i\}} \boldsymbol{\phi}_j c_j =: \boldsymbol{\Phi}_i \boldsymbol{c}_i.$$

Or equivalently

$$\boldsymbol{y} \approx [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \cdots, \boldsymbol{\Phi}_C]\, \boldsymbol{c} = \boldsymbol{\Phi}\boldsymbol{c} \in \mathbb{R}^m$$

$$\text{where } \boldsymbol{c} = \left[\cdots, \boldsymbol{0}^T, \boldsymbol{c}_i^T, \boldsymbol{0}^T, \cdots\right]^T.$$

The $\ell_1$-minimisation formulation for face recognition:

$$\min \|\boldsymbol{c}\|_1 \quad \text{s.t. } \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{c}\|_2 \leq \epsilon.$$

# Robust Face Recognition

When we have corruption and occlusion $y \not\approx \mathbf{\Phi x}$. Instead

$$y \approx \mathbf{\Phi} c + e,$$

where $e$ is an unknown error vector whose entries can be very large.

Assumption: only a fraction of pixels is corrupted ($\geq 70\%$ in some cases).

Robust face recognition formulation:

$$\min \; \|c\|_1 + \|e\|_1 \quad \text{s.t. } y = \mathbf{\Phi} c + e.$$

Or

$$\min \; \|w\|_1 \quad \text{s.t. } y = [\mathbf{\Phi}, \mathbf{I}] \, w.$$

# Gradient Computation

## Definition 6.3 (Gradient)

$$\nabla f\left(\boldsymbol{x}\right) := \left[\frac{d}{dx_1}f, \cdots, \frac{d}{dx_n}f\right]^T.$$

## Example 6.4

Let $f\left(\boldsymbol{x}\right) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right\|_2^2$. Then $\nabla f = -\boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right)$.

▶
$$\frac{d}{d\boldsymbol{x}}\boldsymbol{a}^T\boldsymbol{x} = \frac{d}{d\boldsymbol{x}}\boldsymbol{x}^T\boldsymbol{a} = \boldsymbol{a}.$$

▶
$$\frac{d}{d\boldsymbol{x}}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = 2\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}.$$

▶ $f\left(\boldsymbol{x}\right) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}$,

$$\frac{d}{d\boldsymbol{x}}f = \boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^T\boldsymbol{y} = -\boldsymbol{A}^T\left(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\right).$$

# Section 7
# Low Rank Matrix Recovery

# Netflix Problem

|  | Black Swan | Titanic | True Grit | The King's Speech |
|---|---|---|---|---|
| J. Cameron | | ★★★★★ | ★★★☆☆ | |
| C. Eastwood | ★★★★☆ | | ★★★★★ | |
| P. Jackson | | ★★★☆☆ | | ★★★★☆ |
| Roman Polanski | ★★★★★ | | | ★★★★★ |

# Blind Deconvolution

$$\boldsymbol{y} = \boldsymbol{s} \star \boldsymbol{h} : \; y\left[n\right] = \sum_{\ell=0}^{L} s\left[n - \ell\right] h\left[\ell\right].$$
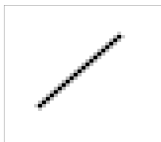


After deblurring:
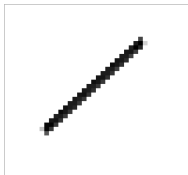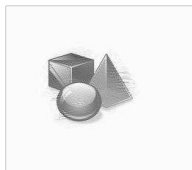
# Low Rank Matrices and Approximations

Consider a matrix $\boldsymbol{X}_0 \in \mathbb{R}^{m \times n}$ with its SVD
$$\boldsymbol{X}_0 = \sum_{k=1}^{\min(m,n)} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^T,$$
where $K = \min(m, n)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K \geq 0$.

## Theorem 7.1 (The Eckart-Young Theorem)

*The best low-rank approximation of $\boldsymbol{X}_0$, i.e.,*

$$\min_{\boldsymbol{X}} \ \|\boldsymbol{X} - \boldsymbol{X}_0\|_F^2 \quad \text{s.t. } \operatorname{rank}(\boldsymbol{X}) = R,$$

*is given by simply truncating the SVD*

$$\hat{\boldsymbol{X}} = \sum_{k=1}^{R} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^T.$$

Remark: $\|\boldsymbol{X}\|_F^2 = \sum_{i,j} X_{i,j}^2 = \|\operatorname{vec}(\boldsymbol{X})\|_2^2$.

# Low Rank Matrix Recovery

Let $\mathcal{A}: \mathbb{R}^{m \times n} \to \mathbb{R}^L$ is a linear measurement operator that takes $L$ inner products with predefined matrices $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_L$:

$$\mathcal{A}: \quad \mathbb{R}^{m \times n} \to \mathbb{R}^L$$

$$\boldsymbol{X}_0 \mapsto y_l = \langle \boldsymbol{X}_0, \boldsymbol{A}_l \rangle = \operatorname{trace}\left(\boldsymbol{A}_l^T \boldsymbol{X}_0\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} X_0\left[i, j\right] A_l\left[i, j\right].$$

The low-rank matrix recovery problem is given by

$$\min_{\boldsymbol{X}} \|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{X}\right)\|_2^2 \quad \text{s.t. } \operatorname{rank}\left(\boldsymbol{X}\right) \leq R.$$

## Example 7.2

In the Netflix problem, $\boldsymbol{A}_l\left[i, j\right] = 1$ and $\boldsymbol{A}_l\left[s, t\right] = 0$ for all $[s, t] \neq [i, j]$.

# Another Look at the Linear Operator $\mathcal{A}$

$$\mathcal{A}: \quad \mathbb{R}^{m \times n} \to \mathbb{R}^L$$
$$\boldsymbol{X} \mapsto \boldsymbol{y} = \boldsymbol{A}\text{vect}\left(\boldsymbol{X}\right),$$

where $\boldsymbol{A} \in \mathbb{R}^{L \times (m \cdot n)}$.

$$\begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & \ddots & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & & 1 \end{bmatrix}$$

# Alternating Projection

To solve
$$\min_{\boldsymbol{X}} \ \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 \quad \text{s.t. rank}(\boldsymbol{X}) \leq R$$
is the same as to look for an $\boldsymbol{L} \in \mathbb{R}^{m \times R}$ and a $\boldsymbol{R} \in \mathbb{R}^{n \times R}$ s.t.

$$\min_{\boldsymbol{L}, \boldsymbol{R}} \ \left\|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{L}\boldsymbol{R}^T\right)\right\|_2^2.$$

Alternating projection:

$$\boldsymbol{R}_{k+1} = \arg \min_{\boldsymbol{R}} \ \left\|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{L}_k\boldsymbol{R}^T\right)\right\|_2^2,$$

$$\boldsymbol{L}_{k+1} = \arg \min_{\boldsymbol{L}} \ \left\|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{L}\boldsymbol{R}_{k+1}^T\right)\right\|_2^2.$$

Details on fixing $L$ and updating $R$:



$$\begin{bmatrix} 1 \\ 3 \\ 5 \\ \vdots \end{bmatrix} = \boldsymbol{X}_0\left[\mathcal{I}_j, j\right] = \boldsymbol{L}_{\mathcal{I}_j,:} \boldsymbol{R}_{j,:}^T$$

# Nuclear Norm Minimization

Define the nuclear norm

$$\|\boldsymbol{X}\|_* = \sum_{k=1}^{\min(m,n)} \sigma_i,$$

which is the $\ell_1$-norm of the singular value vector.

Constrained optimization problem:

$$\min_{\boldsymbol{X}} \ \|\boldsymbol{X}\|_* \quad \text{s.t. } \|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{X}\right)\|_2^2 \le \epsilon.$$

Unconstrained optimization problem:

$$\min_{\boldsymbol{X}} \ \frac{1}{2}\|\boldsymbol{y} - \mathcal{A}\left(\boldsymbol{X}\right)\|_2^2 + \lambda \|\boldsymbol{X}\|_* .$$

# $\ell_1$-norm and Nuclear Norm

## $\ell_1$-norm

Write $\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}_i$ where $\boldsymbol{e}_i$ is the $i^{\text{th}}$ natural basis vector.
$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$.

$$\partial \|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} \text{sign}\,(x_i)\,\boldsymbol{e}_i = \{\boldsymbol{v}: \ v_i = \text{sign}\,(x_i)\}.$$

## Nuclear norm

$\boldsymbol{X} = \sum_{i=1}^{\min(m,n)} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ and $\|\boldsymbol{X}\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i$.

$$\partial \|\boldsymbol{X}\|_* = \sum_{i=1}^{\min(m,n)} \text{sign}\,(\sigma_i)\,\boldsymbol{u}_i \boldsymbol{v}_i^T$$

$$= \left\{ \boldsymbol{U}_r \boldsymbol{V}_r^T + \boldsymbol{U}_{m-r} \boldsymbol{T} \boldsymbol{V}_{n-r}^T : \ \boldsymbol{T} \in \mathbb{R}^{(m-r)\times(n-r)}, \ \sigma\,(\boldsymbol{T}) \leq 1 \right\}.$$

# Soft Thresholding Function

$\ell_1$-norm minimization with given $\boldsymbol{z} \in \mathbb{R}^n$

Let $\hat{\boldsymbol{x}} = \arg \min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$. Then

$$\hat{\boldsymbol{x}} = \sum_i \eta\left(z_i; \lambda\right) \boldsymbol{e}_i \quad \text{where } \eta\left(z_i; \lambda\right) = \text{sign}\left(z_i\right) \max\left(0, |z_i| - \lambda\right).$$

Nuclear norm minimization with given $\boldsymbol{Z} \in \mathbb{R}^{m \times n}$

Let $\hat{\boldsymbol{X}} = \arg \min_{\boldsymbol{X}} \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2 + \lambda \|\boldsymbol{X}\|_*$. Then

$$\hat{\boldsymbol{X}} = \sum_{i=1}^{\min(m,n)} \eta\left(\sigma_i; \lambda\right) \boldsymbol{u}_i \boldsymbol{v}_i^T \quad \text{where } \eta\left(\sigma_i; \lambda\right) = \text{sign}\left(\sigma_i\right) \max\left(0, |\sigma_i| - \lambda\right).$$

# ISTA

$\min \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1$

- $\frac{\partial}{\partial \boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = -\boldsymbol{A}^T (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$.
- $f = \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \Rightarrow \frac{1}{2t_k} \left\| \boldsymbol{x} - \left( \boldsymbol{x}^{k-1} - t_k \nabla f \right) \right\|_2^2$.
- 
$$\boldsymbol{x}^k = \eta \left( \boldsymbol{x}^{k-1} + t_k \boldsymbol{A}^T \left( \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{k-1} \right); \lambda t_k \right).$$

$\min \frac{1}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 + \lambda \|\boldsymbol{X}\|_*$

- $\frac{\partial}{\partial \boldsymbol{X}} \frac{1}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 = -\mathcal{A}^* (\boldsymbol{y} - \mathcal{A}(\boldsymbol{x}))$.
- $f = \frac{1}{2} \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X})\|_2^2 \Rightarrow \frac{1}{2t_k} \left\| \boldsymbol{X} - \left( \boldsymbol{X}^{k-1} - t_k \nabla f \right) \right\|_F^2$.
- 
$$\boldsymbol{X}^k = \eta_{\boldsymbol{\sigma}} \left( \boldsymbol{X}^{k-1} + t_k \mathcal{A}^* \left( \boldsymbol{y} - \mathcal{A} \left( \boldsymbol{X}^{k-1} \right) \right); \lambda t_k \right).$$

# Iterative Hard Thresholding Algorithm

$$\min \tfrac{1}{2} \left\| \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \right\|_2^2 \quad \text{s.t. } \left\| \boldsymbol{x} \right\|_0 \leq S$$

$$\boldsymbol{x}^k = H_S \left( \boldsymbol{x}^{k-1} + \mu_k \boldsymbol{A}^T \left( \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{k-1} \right) \right).$$

$$\min \tfrac{1}{2} \left\| \boldsymbol{y} - \mathcal{A}\left( \boldsymbol{X} \right) \right\|_2^2 \quad \text{s.t. rank} \left( \boldsymbol{X} \right) \leq R$$

$$\boldsymbol{X}^k = H_{R,\boldsymbol{\sigma}} \left( \boldsymbol{X}^{k-1} + t_k \mathcal{A}^* \left( \boldsymbol{y} - \mathcal{A}\left( \boldsymbol{X}^{k-1} \right) \right) \right).$$

# Comments on Performance Guarantees

▶ When $\mathcal{A}(\cdot)$ is a Gaussian random 'projection', RIP condition will hold with high probability:

$$1 - \delta \leq \|\mathcal{A}(\boldsymbol{X})\|_2^2 \leq 1 + \delta, \quad \forall \boldsymbol{X} \text{ s.t. } \text{rank}(\boldsymbol{X}) \leq R.$$

▶ For matrix completion: difficult when $\boldsymbol{X}$ is low-rank and sparse.

$$\begin{bmatrix} \bullet \\ \\ \bullet \\ \\ \\ \end{bmatrix} \begin{bmatrix} & \bullet & & \bullet & \end{bmatrix} = \begin{bmatrix} \bullet & & \bullet \\ \\ \bullet & & \bullet \\ \\ \\ \end{bmatrix}$$

  ▶ Want coherence constant small:

$$\mu(\boldsymbol{U}) := \frac{N}{R} \max_{1 \leq i \leq N} \|\mathcal{P}_{\boldsymbol{U}} \boldsymbol{e}_i\|_2^2 = O(1).$$

# Blind Deconvolution: The Problem

Given a convolution of two signals

$$y[n] = \sum_{\ell=0}^{L} s[n-\ell]\, h[\ell],$$

what are $x[n]$ and $h[n]$?

This bilinear problem is difficult to solve.

▶ Scaling ambiguity.

$$sh^T = \begin{matrix} \\ \\ y[0] \\ y[1] \\ y[2] \\ y[3] \\ y[4] \\ y[5] \\ \vdots \end{matrix} \begin{bmatrix} s[-2]h[0] & s[-2]h[1] & s[-2]h[2] \\ s[-1]h[0] & s[-1]h[1] & s[-1]h[2] \\ s[0]h[0] & s[0]h[1] & s[0]h[2] \\ s[1]h[0] & s[1]h[1] & s[1]h[2] \\ s[2]h[0] & s[2]h[1] & s[2]h[2] \\ s[3]h[0] & s[3]h[1] & s[3]h[2] \\ s[4]h[0] & s[4]h[1] & s[4]h[2] \\ s[5]h[0] & s[5]h[1] & s[5]h[2] \\ s[6]h[0] & s[6]h[1] & s[6]h[2] \\ \vdots & \vdots & \vdots \end{bmatrix}$$

Each entries of $y = x \star h$ is a sum along a skew diagonal of the rank-1 matrix $xh^T$:

$$\min \|X\|_* \text{ s.t. } y = \mathcal{A}(X).$$