Section 5
Convex Optimisation 1

Convex Combination

convex: unique minimizer.

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 λ_i satisfy $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$.

Convex Sets

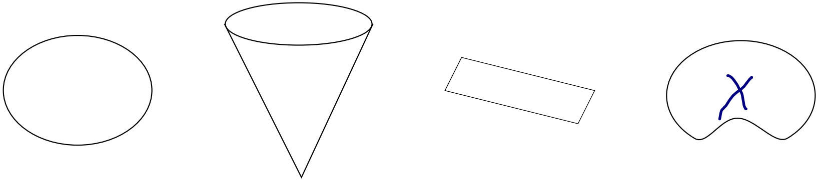
$\forall$ two points in a convex set, the line segment is in the set.

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,

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

Examples



Example of convex sets:

- A *hyperplane* $\mathcal{H} = \left\{ \boldsymbol{x} : \ \boldsymbol{a}^T \boldsymbol{x} = b \right\}$, where $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a} \neq \mathbf{0}$, 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* polyhedron: intersetion of half space.
  $\mathcal{P} = \left\{ \boldsymbol{x} : \ \boldsymbol{a}_j^T \boldsymbol{x} \leq b_j, \ j = 1, \cdots, m, \ \boldsymbol{c}_j^T \boldsymbol{x} = d_j, \ j = 1, \cdots, p \right\}$.
- 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)

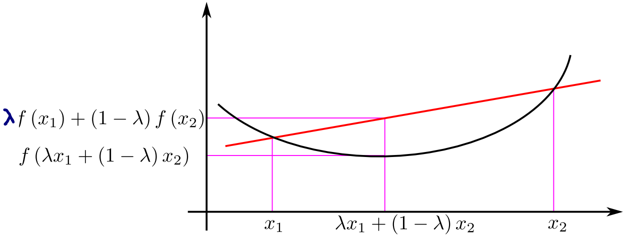
A function $f: \mathbb{R}^n \to \mathbb{R}$ is *convex* if for any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathrm{dom} f \subseteq \mathbb{R}^n$, $\lambda \in [0,1]$, it holds line segment is above the function.

$$\lambda f(\boldsymbol{x}_1) + (1-\lambda) f(\boldsymbol{x}_2) \geq f(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2).$$

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 $\lambda \in (0,1)$. ↳ unique global minimizer (piecewise linear may not be strictly convex)

A Convex Function



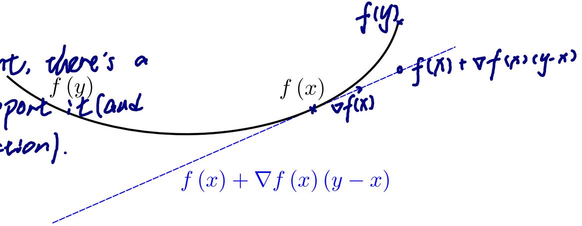
First-Order Condition of Convexity

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 $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$, it holds

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

convex function:
for any given point, there's a hyperspace to support it (and separate the function).



Necessity

f is convex $\Rightarrow \begin{cases} f(y) \geq f(x) + \nabla f(x)^T (y-x) \\ \mathrm{dom}(f) \text{ is convex} \end{cases}$

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,

$(1-t)x + ty = x + t(y-x)$

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

$f(y) \geq \dfrac{f(x+t(y-x)) - (1-t)f(x)}{t}$

Divide both sides by t. It holds

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

$\nabla_{y-x} f(x) = \lim_{t \to 0} \dfrac{f(x+t(y-x)) - f(x)}{t}$ (directional derivative)

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

$\nabla_{y-x} f(x) = \nabla f(x)^T \cdot (y-x)$

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

Sufficiency

$f(y) \geq f(x) + \nabla f(x)^T (y-x) \Rightarrow f(x)$ is convex
dom(f) is convex

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

$$f(\boldsymbol{x}) - f(\boldsymbol{z}) \geq \nabla f(\boldsymbol{z})^T (\boldsymbol{x} - \boldsymbol{z}),$$
$$f(\boldsymbol{y}) - f(\boldsymbol{z}) \geq \nabla f(\boldsymbol{z})^T (\boldsymbol{y} - \boldsymbol{z}).$$

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

$$\lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}) - f(\boldsymbol{z})$$
$$\geq \nabla f(\boldsymbol{z})^T \underbrace{(\lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y} - \boldsymbol{z})}_{0}.$$

By the definition of $\boldsymbol{z}$, the ~~left~~ right side of the inequality is zero. Hence,

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

which proves that f is convex. (assume dom(f) is convex)

Sublevel Sets

Consider the 2-variable case (x is of size 2):

- level set $L_\alpha = \{x \in dom(f): f(x) = \alpha\}$ is a curve (contour).
- sublevel set $C_\alpha = \{x \in dom(f): f(x) \le \alpha\}$ can be a surface.

w.r.t. x

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

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

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

$f(x) \le \alpha \Rightarrow x \in C_\alpha$.

Sublevel Sets of Convex Functions

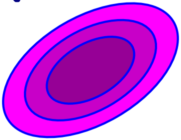
Lemma 5.7

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 $\mathcal{C}_\alpha$, $f(\boldsymbol{x}) \le \alpha$ and $f(\boldsymbol{y}) \le \alpha$. By the convexity of f,

$$\underbrace{f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \le \overbrace{\lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})}^{\text{sublevel set}}}_{\text{convex function}} \le \alpha,$$

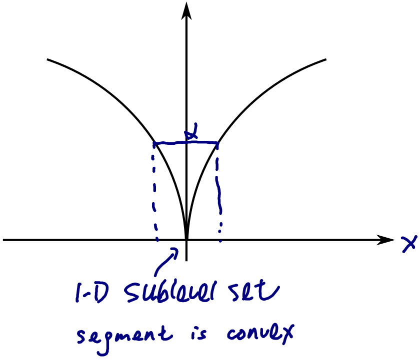
which proves this proposition. $f(\lambda x + (1-\lambda)y) \le \alpha \Rightarrow \lambda x + (1-\lambda)y \in C_\alpha$.



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

$\|x\|_p = (\sum_i x_i^p)^{\frac{1}{p}}$

We've seen ℓ_p-norm in Definition 4.7.

$$\boxed{\begin{array}{l} p(av) = |a|\, p(v) \\ p(u+v) \le p(u) + p(v) \\ p(v) = 0 \Rightarrow v = 0 \end{array}}$$

$p(-v) = |-1|\, p(v) = p(v)$

$2p(v) = p(v) + p(-v) \ge p(v-v) = p(0) = 0 \Rightarrow p(v) \ge 0$

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, $p(\mathbf{0}) = 0$ 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

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

Lemma 5.9

A norm is a convex function.

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

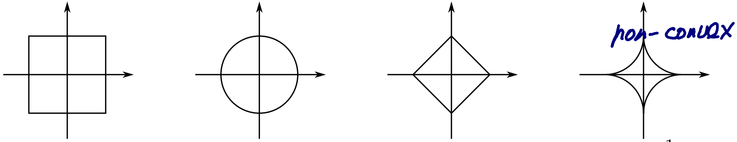
$$\begin{aligned} \|\lambda \boldsymbol{u} + (1-\lambda)\, \boldsymbol{v}\| &\le \|\lambda \boldsymbol{u}\| + \|(1-\lambda)\, \boldsymbol{v}\| \\ &= \lambda \|\boldsymbol{u}\| + (1-\lambda) \|\boldsymbol{v}\|, \end{aligned}$$

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

ℓ_p-Norm

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

Can be verified by using sub-level argument.



$p = \infty$ $\quad$ $p = 2$ $\quad$ $p = 1$ $\quad$ $p = \frac{1}{2}$

Constrained Convex Optimization Problems

A constrained optimization problem of the form

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

$$\text{subject to } h_i(\boldsymbol{x}) \le 0,\ i = 1, \cdots, m,$$

$$\ell_i(\boldsymbol{x}) = 0,\ i = 1, \cdots, r,$$

h(x) convex
x is in sublevel set => x is in convex set.

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).
  - ℓ_i's are affine, i.e., in the form of $\boldsymbol{a}_i^T \boldsymbol{x} + b_i = 0$.
    $\ell_i(\boldsymbol{x}) = 0 \Leftrightarrow \ell_i(\boldsymbol{x}) \le 0$ and $-\ell_i(\boldsymbol{x}) \le 0$.
    Both ℓ_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 $\boldsymbol{x}$ is locally optimal for a convex optimization problem. Then it is also globally optimal.

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

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

Then it is clear that

$$\begin{aligned} f(\boldsymbol{z}) &\overset{\text{convexity}}{\le} (1-\lambda)\, f(\boldsymbol{x}) + \lambda f(\boldsymbol{y}) \overset{\text{local vs global}}{<} f(\boldsymbol{x}), \quad \text{f(z) is no worse than f(x)} \\ h_i(\boldsymbol{z}) &\overset{\text{convexity}}{\le} (1-\lambda)\, h_i(\boldsymbol{x}) + \lambda h_i(\boldsymbol{y}) \overset{\text{upper bound}}{\le} 0,\ i = 0, 1, \cdots, m, \\ \boldsymbol{a}_i^T \boldsymbol{z} &= (1-\lambda)\, \boldsymbol{a}_i^T \boldsymbol{x} + \lambda \boldsymbol{a}_i^T \boldsymbol{y} = b_i,\ i = 1, \cdots, r, \end{aligned}$$

any $\lambda \in (0,1)$ z is feasible

where the inequalities follow from the convexity of the functions f and h_i's. Hence, the point $\boldsymbol{z}$ is feasible and $f(\boldsymbol{z}) < f(\boldsymbol{x})$ for all $\lambda \in (0,1)$. This contradicts with that $\boldsymbol{x}$ is locally optimal and proves the global optimality of $\boldsymbol{x}$.

A Global Optimality Criterion

Theorem 5.11

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

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \underbrace{\nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x})}_{\geq 0 \text{ if } x \text{ is optimal}}, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f).$$

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

x optimal $\Rightarrow \begin{cases} \nabla f(x) = 0, & \text{loose constraints} \\ \nabla f(x) \neq 0, & \text{tight constraints} \end{cases}$

- For an unconstrained convex optimization problem, the sufficient and necessary condition for a globally optimal point $\boldsymbol{x}$ is given by

$$\nabla f(\boldsymbol{x}) = \mathbf{0}.$$

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

$$\nabla f(\boldsymbol{x}) \neq \mathbf{0}.$$

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

Proof

$\nabla f(x)^T(y-x) \geq 0 \Rightarrow x$ optimal

The proof of sufficiency is straightforward. Suppose the inequality holds. Then for all $\boldsymbol{y} \in \mathcal{X}$,

by condition

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

Hence, the point $\boldsymbol{x}$ is globally optimal. by assumption of differentiable

x optimal $\Rightarrow \nabla f(x)^T(y-x) \geq 0$

Conversely, suppose $\boldsymbol{x}$ is optimal, but the inequality does not hold, i.e., for some $\boldsymbol{y} \in \mathcal{X}$ we have

$$\nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) < 0.$$

Consider the point $\boldsymbol{z}(t) = t\boldsymbol{y} + (1-t)\boldsymbol{x},\ t \in [0,1]$. Clearly, $\boldsymbol{z}(t)$ is feasible. Now

chain rule

$$\begin{aligned} \left.\frac{d}{dt} f(\boldsymbol{z}(t))\right|_{t=0} &= \nabla f(\boldsymbol{z}(0)) \cdot \left.\frac{d}{dt} \boldsymbol{z}(t)\right|_{t=0} \\ &= \nabla f(\boldsymbol{x}) \cdot (\boldsymbol{y} - \boldsymbol{x}) < 0, \end{aligned}$$

where the inequality comes from the assumption. It implies that for small positive t, we have $f(\boldsymbol{z}(t)) < f(\boldsymbol{x})$, which contradicts the optimality of $\boldsymbol{x}$. The necessity is therefore proved.

Non-differentiable Functions: Subgradient

non-differentiable + continuous

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 subgradient 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 subdifferential at $\boldsymbol{x}$ and is denoted $\partial f(\boldsymbol{x})$.

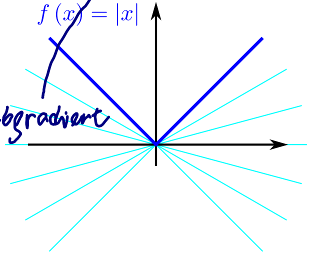
Remark: If f is convex and its subdifferential at $\boldsymbol{x}$ contains exactly one subgradient, then f is differentiable at $\boldsymbol{x}$.

Example

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

subdifferential

subgradient



all hyperplanes support the function.

Section 6
ℓ_1-Minimization

Three Algorithms

- Cyclic Coordinate Descent (CCD)
- Iterative Shrinkage Thresholding (IST)
- Least Angle Regression (LAR)

ℓ_1-Minimization

Want to solve the sparse linear inverse problem:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}.$$

Constrained optimization problem: if we know $\|\boldsymbol{e}\| \leq \epsilon$, subleuel set of convex function is convex. ⇒ convex set x.

all (proper) norms are convex.

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

$\Updownarrow$ equivalent

Unconstrained optimization problem: LASSO

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

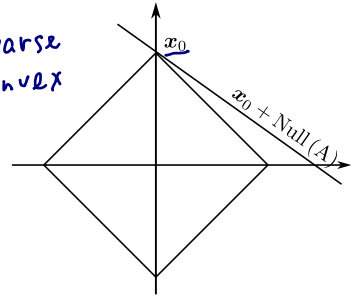
$\exists$ a one-to-one correspondence between ϵ and λ.

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

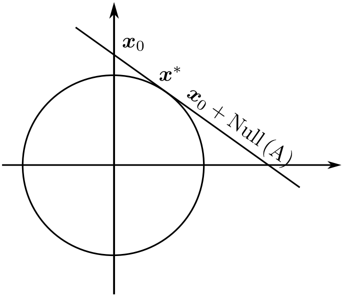
Why ℓ_1-Minimization

A geometric intuition:

l_1:
- sparse
- convex



ℓ_1 tends to give sparse solutions



ℓ_2 tends to give non-sparse solutions

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

Scalar Lasso Problem

$\min_x \frac{1}{2} \|x - y\|_2^2 + \lambda \|x\|_1 \Rightarrow x^\# = \eta_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+$

scale-input scale-output

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

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

$$x^\# - y + \lambda \, \partial_x |x||_{x^\#} = 0,$$

where

$$\partial_x |x| = \begin{cases} 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

$x^\# + \lambda \partial_x |x||_{x^\#} = y$

- suppose $x^\# > 0$

  $x^\# = y - \lambda > 0 \Rightarrow y > \lambda$

- suppose $x^\# < 0$

  $x^\# = y + \lambda < 0 \Rightarrow y < -\lambda$

- suppose $x^\# = 0$

  $0 = y - \lambda \partial_x |x||_{x^\#} \Rightarrow y = k\lambda, \; k \in [-1, 1] \Rightarrow |y| < \lambda$

Scalar Lasso Problem: The Solution

$x^{\#}$ is given by the soft thresholding function:

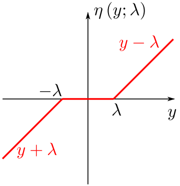
soft thresholding function.

$\eta_\lambda(y) = \text{sign}(y)(|y|-\lambda)_+$

$$x^{\#} = \begin{cases} y - \lambda & \text{if } y > \lambda, \\ 0 & \text{if } |y| \le \lambda, \\ y + \lambda & \text{if } y < -\lambda. \end{cases}$$

$$= \eta(y;\lambda) = \eta_\lambda(y) = \underline{\text{sign}(y)(|y| - \lambda)_+},$$

where $(z)_+ = \max(z, 0)$.



Lasso Problem: Scalar Input Vector Observation

Assume that $\|\boldsymbol{a}\|_2^2 = 1$. Consider the problem

scalar-input vector-output

$$\min_x \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{a}x\|_2^2 + \lambda |x|$$

$$= \frac{1}{2} y^T y - y^T a x + \frac{1}{2} a^T a x^2 + \lambda |x|$$

It optimal solution $x^\#$ is given by

$$\left.\frac{\partial f}{\partial x}\right|_{x^\#} = 0 \Rightarrow \underbrace{a^T a}_{=1} x^\# - y^T a + \lambda \partial_x |x| \big|_{x^\#} = 0$$

$$x^\# = y^T a - \lambda \partial_x |x| \big|_{x^\#}$$

SISO LASSO:
$x^\# = \eta_\lambda(y)$

SIVO LASSO:
$x^\# = \eta_\lambda(y^T a)$

$$x^\# = \begin{cases} \langle \boldsymbol{y}, \boldsymbol{a} \rangle - \lambda & \text{if } \langle \boldsymbol{y}, \boldsymbol{a} \rangle > \lambda, \\ 0 & \text{if } |\langle \boldsymbol{y}, \boldsymbol{a} \rangle| \le \lambda, \\ \langle \boldsymbol{y}, \boldsymbol{a} \rangle + \lambda & \text{if } \langle \boldsymbol{y}, \boldsymbol{a} \rangle < -\lambda. \end{cases}$$

$x^\# > 0 \Rightarrow y^T a > \lambda$

$x^\# = 0 \Rightarrow y^T a = \lambda k, k \in [-1, 1]$
$\Rightarrow |y^T a| \le \lambda$

$x^\# < 0 \Rightarrow y^T a + \lambda < 0$
$\Rightarrow y^T a < -\lambda$

$$= \eta_\lambda \left(\langle \boldsymbol{y}, \boldsymbol{a} \rangle \right).$$

$$= \mathrm{sign}(y^T a)(|y^T a| - \lambda)_+$$

Solving General Lasso: Cyclic Coordinate Descent

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

① set x_j to arbitary value
② start from a certain ordinate. compute the optimal value $x_i^{\#}$.
③ keep $x_i^{\#}$ and go to next coordinate, until converge.

Objective function with respect to x_i:

- $\frac{1}{2} \left\| \boldsymbol{y} - \sum_{j \neq i} \boldsymbol{a}_j x_j - \boldsymbol{a}_i x_i \right\|_2^2 + \lambda \sum_{j \neq i} |x_j| + \lambda |x_i|$
- $\frac{1}{2} \|\boldsymbol{r}_i - \boldsymbol{a}_i x_i\|_2^2 + c + \lambda |x_i|$

fixed

drawback:
· need to run many iterations.
· in each iteration, perform soft thresholding function to each coordination in a sequential way.

Optimal solution for x_i is given by

$$x_i^{\#} = \eta_\lambda \left(\langle \boldsymbol{a}_i, \boldsymbol{r}_i \rangle\right) = \eta_\lambda \left(\left\langle \boldsymbol{a}_i, \boldsymbol{y} - \sum_{j \neq i} \boldsymbol{a}_j \hat{x}_j \right\rangle \right)$$

$$= \eta_\lambda \left(\hat{x}_i + \left\langle \boldsymbol{a}_i, \underbrace{\boldsymbol{y} - \sum_{j} \boldsymbol{a}_j \hat{x}_j}_{\text{residue}} \right\rangle \right)$$

$\langle a_i, a_i \hat{x}_i \rangle = \hat{x}_i$

to update

Three Algorithms

- Cyclic Coordinate Descent (CCD)
- Iterative Shrinkage Thresholding (IST)
- Least Angle Regression (LAR)

The Gradient Descent Method

$f(x) = \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1$
$= \frac{1}{2}x^T A^T A x - y^T A x + \frac{1}{2}y^T y + \lambda\|x\|_1$

$\nabla f(x) = \frac{\partial f}{\partial x} = A^T A x - A^T y + \partial\|x\|_1$
$= -A^T(y - Ax) + \partial\|x\|_1$

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

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

where $t_k > 0$ is a suitable stepsize.

For Lasso problem $f(x) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ which is non-smooth.
Its gradient is given by (see details on page 6-35) $\boxed{\partial\|x\|_1 = \mathrm{sign}(x_i)}$

$$\boxed{\nabla f(x) = -\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) + \partial\|\boldsymbol{x}\|_1}$$

Gradient descent converges very slow.

Gradient Descent Method: Another View

In gradient descent method:

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

$\tilde{f}$: local approximation:

This is equivalent to minimize $\tilde{f}$, ① at each local point x^{k-1}. use gradient info. to construct local approximation.

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

② look global optimal point of $\tilde{f}$ to obtain x_k.

③ iterate.

where

constant   1st order Taylor

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

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

$$= \frac{1}{2t_k}\cdot 2t_k \underline{\nabla f^T(x^{k-1})}\,(x - x^{k-1}) + \frac{1}{2t_k}\left[(x^{k-1})^T x^{k-1} - 2(x^{k-1})^T x + x^T x\right] + c$$

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

$$\left[2t_k \nabla f^T(x^{k-1}) - 2(x^{k-1})^T\right] x$$

Iterative Shrinkage Thresholding (IST)

IST: update the whole vector (coordinates) in x simultaneously, based on local approximation.

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

$$\begin{aligned}
&\tilde{f}(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_1 \\
&:= f\left(\boldsymbol{x}^{k-1}\right) + \left\langle \boldsymbol{x} - \boldsymbol{x}^{k-1}, \nabla f\left(\boldsymbol{x}^{k-1}\right)\right\rangle + \frac{1}{2t_k}\left\|\boldsymbol{x} - \boldsymbol{x}^{k-1}\right\|_2^2 + \lambda \|\boldsymbol{x}\|_1 \\
&= \frac{1}{2t_k}\left\|\boldsymbol{x} - \underbrace{\left(\boldsymbol{x}^{k-1} - t_k \nabla f\left(\boldsymbol{x}^{k-1}\right)\right)}_{\text{known}}\right\|_2^2 + \lambda \|\boldsymbol{x}\|_1 + c \\
&= \sum_i \left[\frac{1}{2t_k}(x_i - z_i)^2 + \lambda |x_i|\right] + c.
\end{aligned}$$

$\nabla f(x^{k-1}) = -A^T(y - Ax^{k-1})$

Therefore,

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

IHT. $x^k = H_s\left(x^{k-1} + \mu A^T(y - Ax^{k-1})\right)$

Three Algorithms

- Cyclic Coordinate Descent (CCD)
- Iterative Shrinkage Thresholding (IST)
- Least Angle Regression (LAR)

Lasso and Sparsity

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

$\lambda = 0$: $\boldsymbol{x}$ is not sparse.

$\lambda \to \infty$: $\boldsymbol{x} = \mathbf{0}$. (most sparse. l-1 norm brings sparsity)

nonzeros in the optimal solution

$\lambda \in [0, \infty)$: one-to-one correspondence between λ and $\|\boldsymbol{x}_\lambda\|_0$, where

regulations ⇔ sparsity

$$\boldsymbol{x}_\lambda := \arg\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1 .$$

Least angle regression: Find a λ to give an x with a specific sparsity.

Piecewise Linearity

the value of each coordinate of x is piecewise linear.

let $x_\lambda = \alpha x_{\lambda_1} + (1-\alpha) x_{\lambda_2}$, to show:

$\underline{\mathrm{sign}(\hat{x}) = \alpha \,\mathrm{sign}(x_{\lambda_1}) + (1-\alpha)\,\mathrm{sign}(x_{\lambda_2})}$

1. $(x_{\lambda_1})_i \neq 0 \Rightarrow \mathrm{sign}((x_{\lambda_1})_i) = \mathrm{sign}((x_{\lambda_2})_i) = \pm 1$
$\Rightarrow \mathrm{sign}((\alpha x_{\lambda_1})_i + ((1-\alpha) x_{\lambda_2})_i) = \pm 1$ convex combination

2. $(x_{\lambda_1})_i = 0$ [sign can be ± 1] choose $\mathrm{sign}((x_{\lambda_1})_i) = \frac{1}{\lambda_1}(A^T(y - A x_{\lambda_1}))_i$
$\mathrm{sign}((x_{\lambda_2})_i) = \frac{1}{\lambda_2}(A^T(y - A x_{\lambda_2}))_i$

Theorem 6.1

$\boldsymbol{x}_\lambda$ is a piecewise linear function of λ.

$\partial \|x\|_1 = \mathrm{sign}(x)$

Proof: $\boldsymbol{x}_\lambda$ is an optimal solution to the Lasso problem if and only if

$\nabla f = -A^T(y - A x_\lambda) + \lambda \,\mathrm{sign}(x_\lambda) = 0$

$$\boxed{\boldsymbol{A}^T (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\lambda) = \lambda \mathrm{sign}(\boldsymbol{x}_\lambda)}.$$

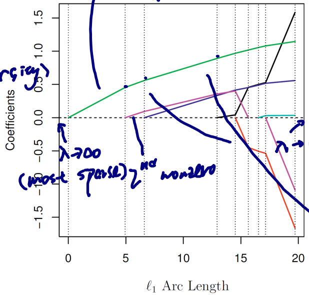
Let $0 < \lambda_1 < \lambda_2$ be two sufficiently close values of λ, so that going from the solution $\boldsymbol{x}_{\lambda_1}$ to $\boldsymbol{x}_{\lambda_2}$ does not require any coordinate of $\boldsymbol{x}_\lambda$ to change its sign. Then it is easy to see that, for all $\lambda = \alpha\lambda_1 + (1-\alpha)\lambda_2$, $\alpha \in [0,1]$, $\hat{\boldsymbol{x}} = \alpha \boldsymbol{x}_{\lambda_1} + (1-\alpha)\boldsymbol{x}_{\lambda_2}$ satisfies the optimality condition. Therefore, $\boldsymbol{x}_\lambda = \hat{\boldsymbol{x}} = \alpha \boldsymbol{x}_{\lambda_1} + (1-\alpha)\boldsymbol{x}_{\lambda_2}$. (linear combination)

Typical Behavior of $\boldsymbol{x}_\lambda$

1st nonzero coefficient

all coefficients are piecewise linear.

(sparsity)



$\lambda \to \infty$ (most sparse)

2nd nonzero

$\lambda \to 0$ (least sparse)

knots: where # coefficients increases.

T. Hastie, et al., Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC, 2015: page 120.

Find the knots, i.e., $\lambda_s > 0$ such that

$$\lim_{\epsilon \to 0^+} \operatorname{sign}(\boldsymbol{x}_{\lambda_s + \epsilon}) \neq \lim_{\epsilon \to 0^+} \operatorname{sign}(\boldsymbol{x}_{\lambda_s - \epsilon})$$

Finding λ_0 (simply by correlation)

Goal: to find λ_0 such that

large λ (sparse) small λ

$$0 = \lim_{\epsilon \to 0^+} \|\boldsymbol{x}_{\lambda_0+\epsilon}\|_0 \neq \lim_{\epsilon \to 0^+} \|\boldsymbol{x}_{\lambda_0-\epsilon}\|_0 = 1.$$

When $\|\boldsymbol{x}_\lambda\|_0 = 1$, let $\mathrm{supp}(\boldsymbol{x}_\lambda) = \{i\}$.
Then the Lasso problem is reduced to $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{a}_i x_i\|_2^2 + \lambda |x_i|$, and

$$x_i^{\#} = (|\boldsymbol{a}_i^T \boldsymbol{y}| - \lambda)_+ \mathrm{sign}(\boldsymbol{a}_i^T \boldsymbol{y}).$$

if $\lambda = \lambda_0 - \varepsilon$, we have exactly 1 entry

This implies that

$$\boxed{\lambda_0 = \max_j |\boldsymbol{a}_j^T \boldsymbol{y}|}.$$

most dominant component

- For a $\lambda > \lambda_0$, $\|\boldsymbol{x}_\lambda\|_0 = 0$
- For a sufficiently small $\epsilon > 0$ and $\lambda \in (\lambda_0 - \epsilon, \lambda_0)$, $\|\boldsymbol{x}_\lambda\|_0 = 1$

Finding the Next Knot (1)

Starting from λ_{s-1}, want to find the next knot $\lambda_s < \lambda_{s-1}$. optimal solution to Lasso problem

- For a $\lambda \in [\lambda_s, \lambda_{s-1}]$, let $\mathcal{I} = \operatorname{supp}(\boldsymbol{x}_\lambda)$ (support won't change) and $\boldsymbol{\delta}_\lambda = \boldsymbol{x}_\lambda - \boldsymbol{x}_{s-1}$:

subgradient = 0

$$\begin{aligned}
\lambda \operatorname{sign}(\boldsymbol{x}_{\lambda,\mathcal{I}}) &= \boldsymbol{A}_{\mathcal{I}}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_\lambda) \\
&= \boldsymbol{A}_{\mathcal{I}}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_{s-1} - \boldsymbol{A}\boldsymbol{\delta}_\lambda) \\
&= \boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{r}_{s-1} - \boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{A}_{\mathcal{I}}\boldsymbol{\delta}_{\lambda,\mathcal{I}}
\end{aligned}$$

previous optimal; assume sign doesn't change

- But $\boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{r}_{s-1} = \lambda_{s-1}\operatorname{sign}(\boldsymbol{x}_{\lambda_{s-1},\mathcal{I}}) = \lambda_{s-1}\operatorname{sign}(\boldsymbol{x}_{\lambda,\mathcal{I}})$.

$$\begin{aligned}
\boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{A}_{\mathcal{I}}\boldsymbol{\delta}_{\lambda,\mathcal{I}} &= \boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{r}_{s-1} - \lambda\operatorname{sign}(\boldsymbol{x}_{\lambda,\mathcal{I}}) \\
&= (\lambda_{s-1} - \lambda)\operatorname{sign}(\boldsymbol{x}_{\lambda,\mathcal{I}}).
\end{aligned}$$

$\delta_{\lambda,I} = (A_I^T A_I)^{-1}(\lambda_{s-1} - \lambda)\operatorname{sign}(x_{\lambda,I})$

$= (A_I^T A_I)^{-1}(\lambda_{s-1} - \lambda)\dfrac{A_I^T r_{s-1}}{\lambda_{s-1}}$

Hence $\boldsymbol{x}_\lambda = \boldsymbol{x}_{s-1} + \boldsymbol{\delta}_\lambda$ where

$$\boldsymbol{\delta}_{\lambda,\mathcal{I}} = \frac{\lambda_{s-1} - \lambda}{\lambda_{s-1}}\left(\boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{A}_{\mathcal{I}}\right)^{-1}\boldsymbol{A}_{\mathcal{I}}^T\boldsymbol{r}_{s-1}.$$

Finding the Next Knot (2)

$$x_\lambda = x_{s-1} + \delta_\lambda = x_{s-1} + \frac{\lambda_{s-1} - \lambda}{\lambda_{s-1}} (A_I^T A_I)^{-1} A_I^T r_{s-1}$$

- Keep track x_λ and $|\langle \boldsymbol{a}_j, \boldsymbol{r}_\lambda \rangle|$ until

$r_\lambda = y - A x_\lambda$

  - Either

(previous zero → nonzero)

$$\max_{j \notin \mathcal{I}} |\langle \boldsymbol{a}_j, \boldsymbol{r}_\lambda \rangle| = \lambda.$$

new knot

Define $i = \arg\max_{j \notin \mathcal{I}} |\langle \boldsymbol{a}_j, \boldsymbol{r}_\lambda \rangle|$ and set $\mathcal{I} = \mathcal{I} \bigcup \{i\}$.

  - Or for some $i \in \mathcal{I}$,

(previous nonzero → zero)

$$(\boldsymbol{x}_\lambda)_i = 0.$$

Set $\mathcal{I} = \mathcal{I} \backslash \{i\}$.

Set λ_s accordingly.

Least Angle Regression

1. $\boldsymbol{r}_0 = \boldsymbol{y}$ and $\boldsymbol{x} = \boldsymbol{0}$.
2. Let $\lambda_0 = \max_j \ |\langle \boldsymbol{a}_j, \boldsymbol{r}_0 \rangle|$, $i = \arg\ \max_j \ |\langle \boldsymbol{a}_j, \boldsymbol{r}_0 \rangle|$ and $\mathcal{I} = \{i\}$.
3. For $s = 1, 2, \cdots$, do
   - 3.1 Find the next knot λ_s.
   - 3.2 Set $\boldsymbol{x}_s = \boldsymbol{x}_{s-1} + \boldsymbol{\delta}_{\lambda_s}$ and $\boldsymbol{r}_s = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}_s$.

Return the sequence $\{\lambda_s, \boldsymbol{x}_s\}$, $s = 0, 1, 2, \cdots$

resolution sparse level

Stable Recovery of Exact Sparse Signals

Theorem 6.2

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

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

LARS. optimum in terms of magnitude. (constant error (ℓ₁))

where the constant C_S depends only on δ_{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 6.3

Suppose that $\boldsymbol{x}_0$ *is an an arbitrary vector in* $\mathbb{R}^n$ *and let* $\boldsymbol{x}_{0,S}$ *be the truncated vector corresponding to the* S *largest values of* $\boldsymbol{x}_0$ *(in absolute value). When the matrix* $\boldsymbol{A}$ *satisfies RIP, the solution* $\boldsymbol{x}^\#$ *obeys*

$$\left\| \boldsymbol{x}^\# - \boldsymbol{x}_0 \right\|_2 \leq C_{1,S} \cdot \epsilon + C_{2,S} \cdot \frac{\left\| \boldsymbol{x}_0 - \boldsymbol{x}_{0,S} \right\|_1}{\sqrt{S}}.$$

extra error

x_0 is not sparse; x^* is LARS to $x_{0,S}$

No algorithm performs fundamentally better than ℓ_1-min.

exactly by sparse assumption

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)

$\|Ah\|_2 \le 2\varepsilon$

$S \quad \lambda \quad \varepsilon$

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

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

Tube constraint:

triangle inequality

$$\|\boldsymbol{A}\boldsymbol{h}\|_2 = \left\|\boldsymbol{A}\boldsymbol{x}^\# - \boldsymbol{A}\boldsymbol{x}_0\right\|_2 \le \underbrace{\left\|\boldsymbol{A}\boldsymbol{x}^\# - \boldsymbol{y}\right\|_2}_{\le \varepsilon:\ x^\# \text{ is feasible, satisfy the constraint}} + \underbrace{\|\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{y}\|_2}_{\le \varepsilon:\ \text{model}} \le 2\epsilon.$$

$h \triangleq x^\# - x_0$

error between
l-1 minimizer
and ground truth

Analysis for Exact Sparse Signals (2)

$\|h_{T_0^c}\|_1 \le \|h_{T_0}\|_1$

Cone constraint: Let $\boldsymbol{x}^{\#} = \boldsymbol{x}_0 + \boldsymbol{h}$. Then

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

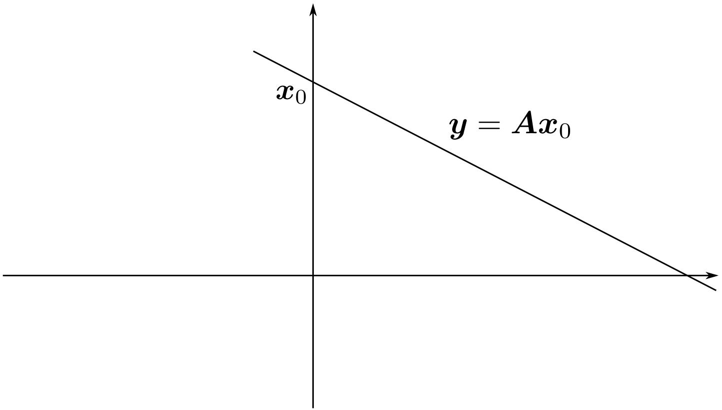
Proof:

$x^{\#}$ is global minimum in terms of ℓ_1 norm.

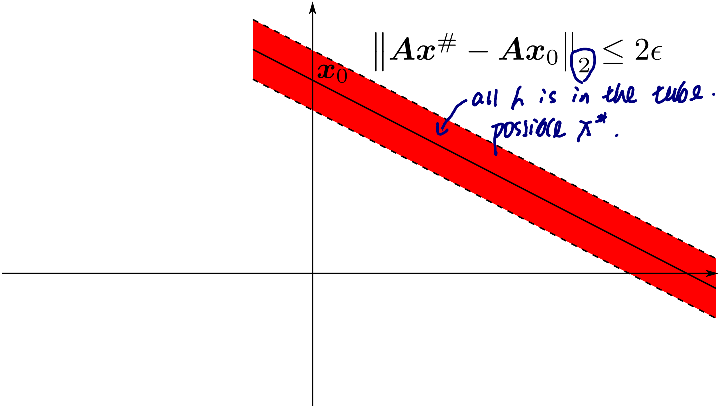
ℓ_1 norm can be divided. as sum of absolute subvectors.

$$\begin{aligned}
\|\boldsymbol{x}_0\|_1 &\ge \left\|\boldsymbol{x}^{\#}\right\|_1 = \|\boldsymbol{x}_0 + \boldsymbol{h}\|_1 \\
&= \left\|(\boldsymbol{x}_0 + \boldsymbol{h})_{\mathcal{T}_0}\right\|_1 + \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1 \\
&\overset{\text{triangle ineq.}}{\ge} \|\boldsymbol{x}_0\|_1 - \|\boldsymbol{h}_{\mathcal{T}_0}\|_1 + \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1 .
\end{aligned}$$

Geometric Interpretation

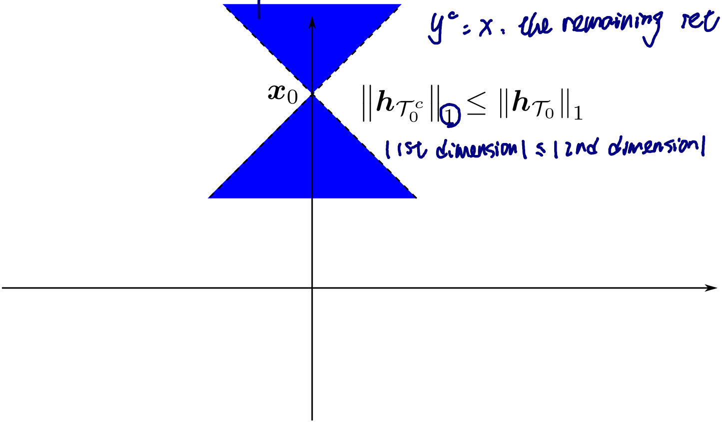


Geometric Interpretation



Geometric Interpretation

possible x^*



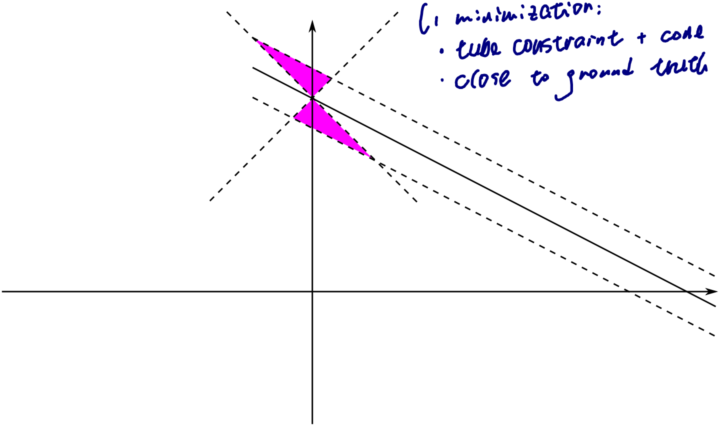
y. true support set

y^c = x. the remaining set

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

|1st dimension| ≤ |2nd dimension|

Geometric Interpretation

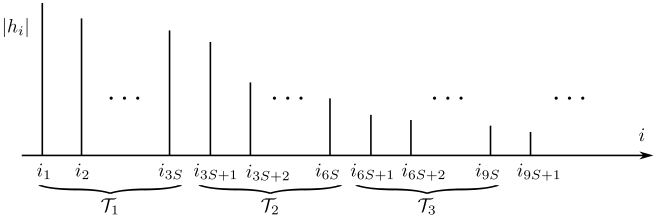


ℓ_1 minimization:
- tube constraint + cone constraint
- close to ground truth

Proof

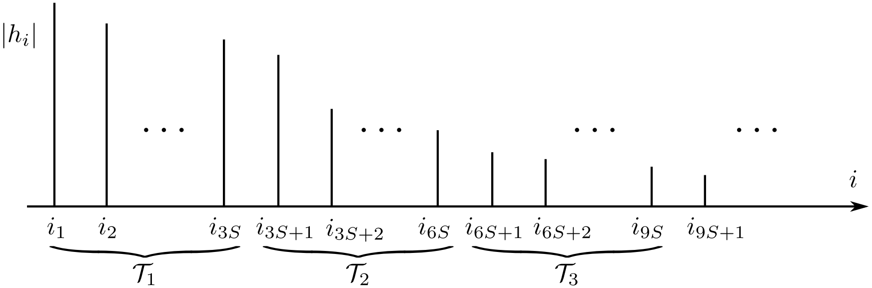
Since $\|\boldsymbol{Ah}\|_2 \le 2\epsilon$, want to show $\|\boldsymbol{h}\|_2 \approx \|\boldsymbol{Ah}\|_2$.
(This is not true in general. For example $\boldsymbol{Ah} = \boldsymbol{0}$ but $\|\boldsymbol{h}\|_2$ can be ∞)

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 $\mathcal{T}_j = \{n_\ell,\ (j-1)\,M + 1 \le \ell \le jM\}$.
Hence $\mathcal{T}_1$ contains the indices of the M largest entries (in magnitude) of $\boldsymbol{h}_{\mathcal{T}_0^c}$, $\mathcal{T}_2$ contains the indices of the next M largest entries (in magnitude) of $\boldsymbol{h}_{\mathcal{T}_0^c}$.



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 ℓ_2-norm of $\boldsymbol{h}$ concentrates on $\mathcal{T}_{01} = \mathcal{T}_0 \bigcup \mathcal{T}_1$.

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

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

$$\begin{aligned}
\left\|\boldsymbol{h}_{\mathcal{T}_{01}^c}\right\|_2^2 &\leq \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1^2 \sum_{k=M+1}^{N} \frac{1}{k^2} \\
&\overset{(a)}{\leq} \left\|\boldsymbol{h}_{\mathcal{T}_0^c}\right\|_1^2 / M \overset{(b)}{\leq} \frac{\|\boldsymbol{h}_{\mathcal{T}_0}\|_1^2}{M} \\
&\overset{(c)}{\leq} \frac{\|\boldsymbol{h}_{\mathcal{T}_0}\|_2^2 \cdot |\mathcal{T}_0|}{M} \leq \rho \|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2^2 ,
\end{aligned}$$

where (a) holds as $\sum_{k=M+1}^{N} 1/k^2 \leq 1/M$, (b) is from the ℓ_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.

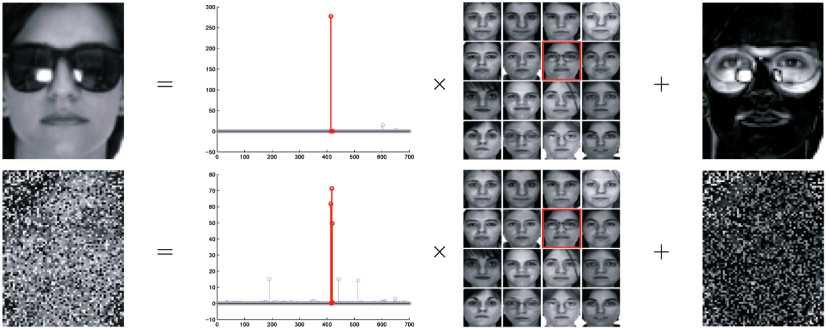
Proof: Step 3

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{h}\|_2 &= \left\| \boldsymbol{A}_{\mathcal{T}_{01}} \boldsymbol{h}_{\mathcal{T}_{01}} + \sum_{j\ge 2} \boldsymbol{A}_{\mathcal{T}_j} \boldsymbol{h}_{\mathcal{T}_j} \right\|_2 \ge \|\boldsymbol{A}_{\mathcal{T}_{01}} \boldsymbol{h}_{\mathcal{T}_{01}}\|_2 - \left\| \sum_{j\ge 2} \boldsymbol{A}_{\mathcal{T}_j} \boldsymbol{h}_{\mathcal{T}_j} \right\|_2 \\
&\ge \|\boldsymbol{A}_{\mathcal{T}_{01}} \boldsymbol{h}_{\mathcal{T}_{01}}\|_2 - \sum_{j\ge 2} \left\| \boldsymbol{A}_{\mathcal{T}_j} \boldsymbol{h}_{\mathcal{T}_j} \right\|_2 \\
&\ge \sqrt{1-\delta_{|\mathcal{T}_0|+M}} \, \|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2 - \sqrt{1+\delta_M} \sum_{j\ge 2} \left\| \boldsymbol{h}_{\mathcal{T}_j} \right\|_2 \\
&\ge \underbrace{\left(\sqrt{1-\delta_{4S}} - \sqrt{\rho}\sqrt{1+\delta_{4S}} \right)}_{C_{4S}} \|\boldsymbol{h}_{\mathcal{T}_{01}}\|_2 .
\end{aligned}$$

Hence,

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

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



The Setup

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

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 $\mathbb{R}^m$.

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

Or equivalently

if y is from subject i, it should be sparsely represented by linear combination of the corresponding training set.

$$\boldsymbol{y} \approx [\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \cdots, \boldsymbol{\Phi}_C]\, \boldsymbol{c} = \overset{m\times n\ n\times 1}{\boldsymbol{\Phi}\boldsymbol{c}} \in \mathbb{R}^m$$

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

The ℓ_1-minimisation formulation for face recognition:

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

Robust Face Recognition

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

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

can be sparse or not.

where $\boldsymbol{e}$ 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:

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

Or

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

Gradient Computation

Definition 6.4 (Gradient)

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

Example 6.5

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

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

  $\frac{d}{dx} a^T x = \frac{d}{dx} x^T a = a$

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

  $\frac{d}{dx} x^T B x = (B + B^T) x$

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

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

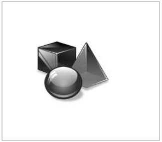
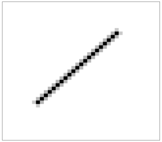
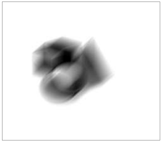
Section 7
Low Rank Matrix Recovery

Netflix Problem

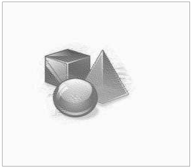
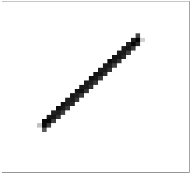
	Black Swan	Titanic	True Grit	The King's Speech
J. Cameron		★★★★★	★★★☆☆	
C. Eastwood	★★★★☆		★★★★★	
P. Jackson		★★★☆☆		★★★★☆
Roman Polanski	★★★★★			★★★★★

Blind Deconvolution [Ahmed, Recht, and Romberg, 2013]

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

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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_K \geq 0$.

low-rank
vs
sparsity

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

Low Rank Matrix Recovery

$\left[\begin{matrix} - \\ - \\ \vdots \end{matrix}\right]\left[\begin{matrix} | & | & \cdots \end{matrix}\right] = \left[\begin{matrix} \circ & & \\ & \circ & \\ & & \ddots \end{matrix}\right]$

$tr(A^T A) = \sum_{ij} A_{ij}^2 = \|A\|_F^2 = \langle A, A \rangle$

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 $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_L$:

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

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

The low-rank matrix recovery problem is given by

$$\min_{\boldsymbol{X}} \|\overset{L\times 1}{\boldsymbol{y}} - \mathcal{A}(\overset{m\times n}{\boldsymbol{X}})\|_2^2 \quad \text{s.t. } \operatorname{rank}(\boldsymbol{X}) \le R.$$

Example 7.2

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

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

$$\mathcal{A}: \quad \mathbb{R}^{m\times n} \to \mathbb{R}^{L}$$
$$\boldsymbol{X} \mapsto \boldsymbol{y} = \overset{L\times mn}{\boldsymbol{A}} \text{vect}\overset{mn\times 1}{(\boldsymbol{X})},$$

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

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

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.

unconstrained

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

bilinear optimization:
- auxiliary variable (# target ↗)
- convex (remove rank constraint)

Alternating projection:

fix L, optimize R

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

fix R, optimize L

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

Alternating Projection (2)

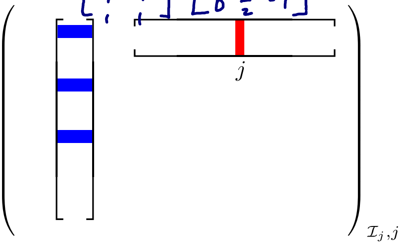
$$L \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad R^T \quad \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 3 \end{bmatrix}$$

Details on fixing $\boldsymbol{L}$ and updating $\boldsymbol{R}$:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & ? \\ 1 & ? & 3 \\ ? & 1 & 1 \end{bmatrix}$$

$$L \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \qquad R^T \begin{bmatrix} 1 & \frac{1}{2} & 2 \\ 0 & \frac{1}{2} & -1 \end{bmatrix}$$

$$\begin{array}{c} j \\ \begin{bmatrix} & 1 & \\ & ? & \\ & 3 & \\ \cdots & ? & \cdots \\ & 5 & \\ & ? & \\ & \vdots & \end{bmatrix}_{\mathcal{I}_j,j} \end{array} =$$



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

Nuclear Norm Minimization

Define the nuclear norm

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

which is the ℓ_1-norm of the singular value vector.

Constrained optimization problem:

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

Unconstrained optimization problem:

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

ℓ_1-norm and Nuclear Norm

ℓ_1-norm

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

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

Nuclear norm

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

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

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

Soft Thresholding Function

$\sum \min_{x_i} \left[\frac{1}{2}(x_i - z_i)^2 + \lambda |x_i| \right] \Rightarrow x_i = \eta_\lambda(z_i)$

$\partial \|x\|_1 = \sum_i \operatorname{sign}(x_i) e_i$

$\partial \|X\|_* = \sum_i \operatorname{sign}(\sigma_i) u_i v_i^T$

ℓ_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

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

induction

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

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

ISTA

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

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

$\frac{1}{2}\|y-A\cdot \underbrace{\mathrm{vec}(X)}_{x}\|_2^2=\frac{1}{2}\langle y-Ax, y-Ax\rangle$

$=\frac{1}{2}\left[x^TA^TAx-2y^TAx+y^Ty\right]$

linear operator $\mathcal{A}$ $\Longleftrightarrow$ matrix A

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

$\frac{\partial}{\partial x}=A^TAx-A^Ty=-A(y-Ax)$

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

t_k: 1 in most cases

Iterative Hard Thresholding Algorithm

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

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

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

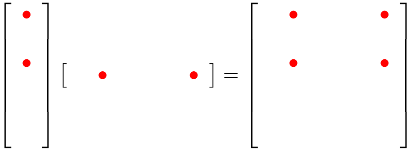
$$\boldsymbol{X}^k = H_{\underbrace{R,\boldsymbol{\sigma}}_{\text{largest } R \text{ singular values}}}\left(\boldsymbol{X}^{k-1}+t_k\mathcal{A}^*\left(\boldsymbol{y}-\mathcal{A}\left(\boldsymbol{X}^{k-1}\right)\right)\right).$$

Comments on Performance Guarantees

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

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

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



  - 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

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

blurred image

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

This bilinear problem is difficult to solve.

- Scaling ambiguity.

Blind Deconvolution: The Idea

$\begin{bmatrix} s[0] \\ s[1] \\ s[2] \\ \vdots \\ s[N] \end{bmatrix} [h[0]\ h[1] \cdots h[L]]$

sum of anti-diagonal elements

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

$y = \mathcal{A}(s \cdot h^T) = \mathcal{A}(X)$

$\text{rank}(X) = 1$

(components)

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

solve vectors

solve matrix with low rank

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

SVD for vectors. (components)