Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 4: Concise signal models and compressive sensing

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2021)

License Information for Mathematics of Data Slides

Outline

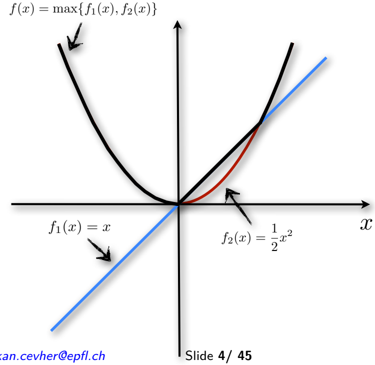
- Deficiency of smooth models
- Sparsity and compressive sensing
- Atomic norms
- Non-smooth minimization via Subgradient descent

Non-smooth minimization: A simple example

What if we simultaneously want $f_1(x), f_2(x), \ldots, f_k(x)$ to be small?

A natural approach in some cases: Minimize $f(x) = \max\{f_1(x), \ldots, f_k(x)\}$

- *The good news*: If each $f_i(x)$ is convex, then $f(x)$ is convex
- *The bad (!) news*: Even if each $f_i(x)$ is smooth, $f(x)$ may be non-smooth
  - e.g., $f(x) = \max\{x, x^2\}$



A statistical learning motivation for non-smooth optimization

Linear Regression

Consider the classical linear regression problem:

$$\mathbf{b} = \mathbf{A}\mathbf{x}^{\natural} + \mathbf{w}$$

with $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ are known, $\mathbf{x}^{\natural}$ is unknown, and $\mathbf{w}$ is noise. Assume *for now* that $n \geq p$ (more later).

A statistical learning motivation for non-smooth optimization

Linear Regression

Consider the classical linear regression problem:

$$\mathbf{b} = \mathbf{A}\mathbf{x}^{\natural} + \mathbf{w}$$

with $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$ are known, $\mathbf{x}^{\natural}$ is unknown, and $\mathbf{w}$ is noise. Assume *for now* that $n \geq p$ (more later).

- *Standard approach:* Least squares: $\mathbf{x}^{\star}_{\text{LS}} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$
  - Convex, smooth, and an *explicit solution*: $\mathbf{x}^{\star}_{\text{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \mathbf{A}^{\dagger}\mathbf{b}$
- *Alternative approach:* Least absolute value deviation: $\mathbf{x}^{\star} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$
  - The advantage: Improved robustness against outliers (i.e., less sensitive to high noise values)
  - The bad (!) news: A *non-differentiable* objective function

Our main motivating example this lecture: The case $n \ll p$

Deficiency of smooth models

Recall the practical performance of an estimator $\mathbf{x}^\star$.

Practical performance

Denote the numerical approximation at time t by $\mathbf{x}^t$. The practical performance is determined by

$$\left\|\mathbf{x}^t - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\left\|\mathbf{x}^t - \mathbf{x}^\star\right\|_2}_{\text{numerical error}} + \underbrace{\left\|\mathbf{x}^\star - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} .$$

Remarks:

- *Non-smooth* estimators of $\mathbf{x}^\natural$ can help *reduce the statistical error*.
- This improvement *may* require higher computational costs.

Example: Least-squares estimation in the linear model

∘ Recall the linear model and the LS estimator.

LS estimation in the linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{n \times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ denotes the unknown noise. The LS estimator for $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b}$ is defined as

$$\mathbf{x}^\star_{\mathsf{LS}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\}.$$

Remarks:

∘ If $\mathbf{A}$ has full column rank, $\mathbf{x}^\star_{\mathsf{LS}} = \mathbf{A}^\dagger \mathbf{b}$ is uniquely defined.

∘ *When $n < p$*, $\mathbf{A}$ cannot have full column rank, and hence $\mathbf{x}^\star_{\mathsf{LS}} \in \left\{ \mathbf{A}^\dagger \mathbf{b} + \mathbf{h} : \mathbf{h} \in \mathrm{null}\,(\mathbf{A}) \right\}$.

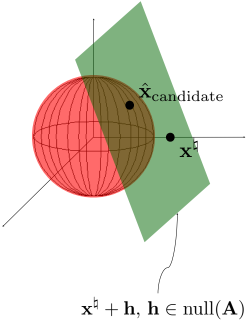
Observation:

∘ The estimation error $\left\| \mathbf{x}^\star_{\mathsf{LS}} - \mathbf{x}^\natural \right\|_2$ can be *arbitrarily large*!

A candidate solution

Continuing the LS example:

- There exist infinitely many $\mathbf{x}$'s such that $\mathbf{b} = \mathbf{A}\mathbf{x}$
- Suppose that $\mathbf{w} = 0$ (i.e. no noise). Let us just choose the one $\hat{\mathbf{x}}_{\text{candidate}}$ with the smallest norm $\|\mathbf{x}\|_2$.



Observation: ◦ Unfortunately, *this still fails when $n < p$*

A candidate solution contd.

Proposition ([7])

Suppose that $\mathbf{A} \in \mathbb{R}^{n \times p}$ *is a matrix of i.i.d. standard Gaussian random variables, and* $\mathbf{w} = \mathbf{0}$*. We have*

$$(1-\epsilon)\left(1 - \frac{n}{p}\right)\left\|\mathbf{x}^\natural\right\|_2^2 \leq \left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\natural\right\|_2^2 \leq (1-\epsilon)^{-1}\left(1 - \frac{n}{p}\right)\left\|\mathbf{x}^\natural\right\|_2^2$$

with probability at least $1 - 2\exp\left[-(1/4)(p-n)\epsilon^2\right] - 2\exp\left[-(1/4)p\epsilon^2\right]$*, for all* $\epsilon > 0$ *and* $\mathbf{x}^\natural \in \mathbb{R}^p$*.*

Summarizing the findings so far

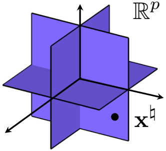
The message so far:

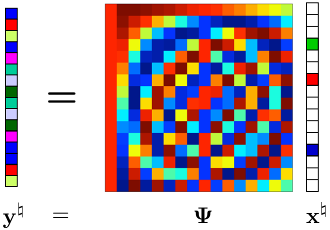
- Even in the absence of noise, we cannot recover $\mathbf{x}^\natural$ from the observations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ unless $n \geq p$
- But in applications, p might be thousands, millions, billions...
- **Can we get away with $n \ll p$ under some further assumptions on $\mathbf{x}$?**

A natural signal model

Definition (s-sparse vector)

A vector $\mathbf{x} \in \mathbb{R}^p$ is s-sparse if it has at most s non-zero entries.

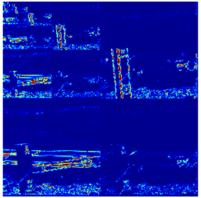




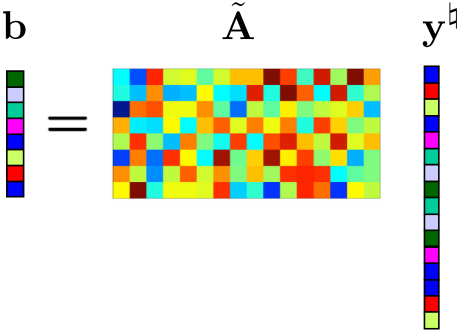
Sparse representations

$\mathbf{x}^\natural$: *sparse* transform coefficients

- Basis representations $\Psi \in \mathbb{R}^{p\times p}$
  - *Wavelets*, DCT, ...
- Frame representations $\Psi \in \mathbb{R}^{m\times p}$, $m > p$
  - Gabor, curvelets, shearlets, ...
- Other *dictionary* representations...

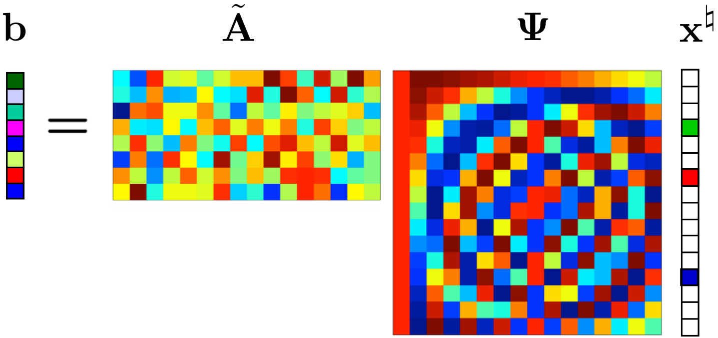


Sparse representations strike back!



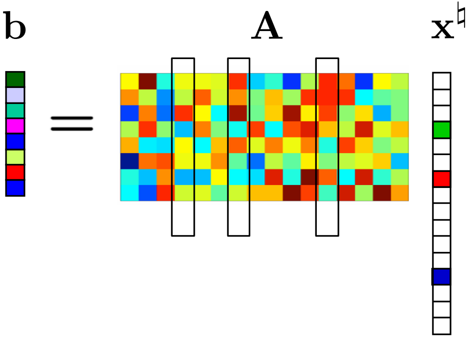
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



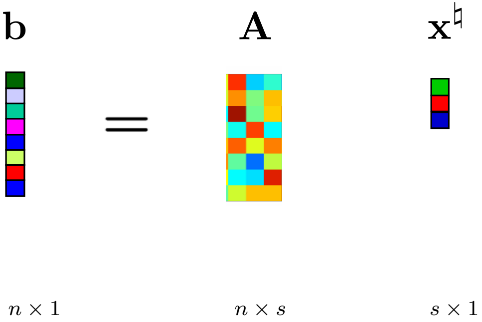
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$, $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n < p$

Sparse representations strike back!



Observations:
- The matrix $\mathbf{A}$ effectively becomes *overcomplete*.
- We could solve for $\mathbf{x}^\natural$ if we knew *the location of the non-zero entries of* $\mathbf{x}^\natural$.

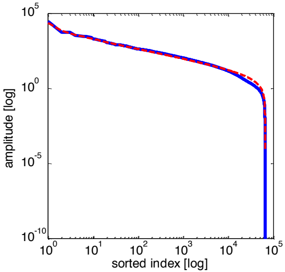
Compressible signals

◦ Real signals may not be exactly sparse, but approximately sparse, or *compressible*.

Definition (Compressible signals)

Roughly speaking, a vector $\mathbf{x} := (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is compressible if the number of its significant components (i.e., entries larger than some $\epsilon > 0$: $|\{k : |x_k| \geq \epsilon, 1 \leq k \leq p\}|$) is small.

- **Cameraman**@MIT.



- **Solid curve**: Sorted wavelet coefficients of the cameraman image.
- **Dashed curve**: Expected order statistics of generalized Pareto distribution with shape parameter 1.67.

A different tale of the linear model $\mathrm{b} = \mathrm{A}\mathrm{x} + \mathrm{w}$

A *realistic* linear model

Let $\mathbf{b} := \tilde{\mathbf{A}}\mathbf{y}^\natural + \tilde{\mathbf{w}} \in \mathbb{R}^n$.

- Let $\mathbf{y}^\natural := \Psi\mathbf{x}_{\text{real}} \in \mathbb{R}^m$ that admits a *compressible* representation $\mathbf{x}_{\text{real}}$.
- Let $\mathbf{x}_{\text{real}} \in \mathbb{R}^p$ that is *compressible* and let $\mathbf{x}^\natural$ be its *best s-term approximation*.
- Let $\tilde{\mathbf{w}} \in \mathbb{R}^n$ denote the possibly nonzero *noise* term.
- Assume that $\Psi \in \mathbb{R}^{m\times p}$ and $\tilde{\mathbf{A}} \in \mathbb{R}^{n\times m}$ are known.

Then we have

$$\begin{aligned}\mathbf{b} &= \tilde{\mathbf{A}}\Psi\left(\mathbf{x}^\natural + \mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right) + \tilde{\mathbf{w}}.\\ &:= \underbrace{\left(\tilde{\mathbf{A}}\Psi\right)}_{\mathbf{A}}\mathbf{x}^\natural + \underbrace{\left[\tilde{\mathbf{w}} + \tilde{\mathbf{A}}\Psi\left(\mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right)\right]}_{\mathbf{w}},\end{aligned}$$

equivalently, $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$.

Peeling the onion

◦ The *realistic* linear model uncovers yet another level of difficulty

Practical performance

The practical performance at time t is determined by

$$\left\|\mathbf{x}^t - \mathbf{x}_{\text{real}}\right\|_2 \le \underbrace{\left\|\mathbf{x}^t - \mathbf{x}^\star\right\|_2}_{\text{numerical error}} + \underbrace{\left\|\mathbf{x}^\star - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} + \underbrace{\left\|\mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right\|_2}_{\text{model error}}.$$

Approach 1: Sparse recovery via exhaustive search

Approach 1 for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may search over all $\binom{p}{s}$ subsets $S \subset \{1, \ldots, p\}$ of cardinality s, solve the restricted least least-squared problem $\min_{\mathbf{x}_S} \|\mathbf{b} - \mathbf{A}_S \mathbf{x}_S\|_2^2$, and return the resulting $\mathbf{x}$ corresponding to the smallest error, putting zeros in the entries of $\mathbf{x}$ outside S.

◦ Stable and robust recovery of any s-sparse signal is possible using just $n = 2s$ measurements.

Approach 1: Sparse recovery via exhaustive search

Approach 1 for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may search over all $\binom{p}{s}$ subsets $S \subset \{1, \ldots, p\}$ of cardinality s, solve the restricted least least-squared problem $\min_{\mathbf{x}_S} \|\mathbf{b} - \mathbf{A}_S\mathbf{x}_S\|_2^2$, and return the resulting $\mathbf{x}$ corresponding to the smallest error, putting zeros in the entries of $\mathbf{x}$ outside S.

◦ Stable and robust recovery of any s-sparse signal is possible using just $n = 2s$ measurements.

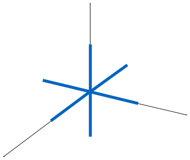
Issues

- $\binom{p}{s}$ is a huge number - too many to search!
- s is not known in practice

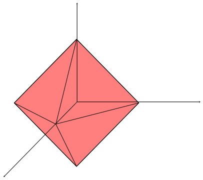
The ℓ_1-norm heuristic

Heuristic: The *ℓ_1-ball with radius c_∞* is an "approximation" of the set of sparse vectors $\hat{\mathbf{x}} \in \left\{\mathbf{x} : \|\mathbf{x}\|_0 \leq s, \|\mathbf{x}\|_\infty \leq c_\infty\right\}$ parameterized by their sparsity s and maximum amplitude c_∞.

$$\hat{\mathbf{x}} \in \left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq c_\infty\right\} \quad \text{with some } c_\infty > 0.$$



The set $\left\{\mathbf{x} : \|\mathbf{x}\|_0 \leq 1, \|\mathbf{x}\|_\infty \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$



The unit ℓ_1-norm ball
$\left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$

Remark: ◦ This heuristic leads to the so-called *Lasso* optimization problem.

Sparse recovery via the Lasso

Definition (Least absolute shrinkage and selection operator (Lasso))

$$\mathbf{x}^{\star}_{Lasso} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_1$$

with some $\rho \geq 0$.

- The second term in the objective function is called the *regularizer*.
- The parameter ρ is called the *regularization parameter*. It is used to trade off the objectives:
  - Minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$, so that the solution is consistent with the observations
  - Minimize $\|\mathbf{x}\|_1$, so that the solution has the desired sparsity structure

Remark: ◦ The Lasso has a *convex* but *non-smooth* objective function

Performance of the Lasso

Theorem (Existence of a stable solution in polynomial time [10])

This Lasso convex formulation is a second order cone program, which can be solved in polynomial time in terms of the inputs n *and* p*. Surprisingly, if the signal* $\mathbf{x}^\natural$ *is* s*-sparse and the noise* $\mathbf{w}$ *is sub-Gaussian (e.g., Gaussian or bounded) with parameter* σ*, then choosing* $\rho = \sqrt{\frac{16\sigma^2 \log p}{n}}$ *yields an error of*

$$\left\| \mathbf{x}^\star_{Lasso} - \mathbf{x}^\natural \right\|_2 \leq \frac{8\sigma}{\kappa(\mathbf{A})} \sqrt{\frac{s \ln p}{n}},$$

with probability at least $1 - c_1 \exp(-c_2 n \rho^2)$*, where* c_1 *and* c_2 *are absolute constants, and* $\kappa(\mathbf{A}) > 0$ *encodes the difficulty of the problem.*

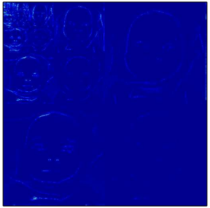
Remark:

- The number of measurements is $\mathcal{O}\left(s \ln p\right)$ – this may be *much* smaller than p!

Other models with simplicity

p

pixels

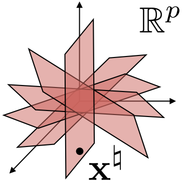


Information level:

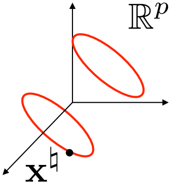
$s \ll p$

large wavelet coefficients

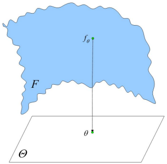
(blue = 0)



sparse signals



low-rank matrices



nonlinear models

There are many models extending far beyond sparsity, coming with other non-smooth regularizers.

Generalization via simple representations

Definition (Atomic sets & atoms [3])

An *atomic set* $\mathcal{A}$ is a set of vectors in $\mathbb{R}^p$. An *atom* is an element in an atomic set.

Terminology (Simple representation [3])

A parameter $\mathbf{x}^\natural \in \mathbb{R}^p$ admits a simple representation with respect to an atomic set $\mathcal{A} \subseteq \mathbb{R}^p$, if it can be represented as a non-negative combination of few atoms, i.e., $\mathbf{x}^\natural = \sum_{i=1}^k c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathcal{A}, \, c_i \geq 0.$

Example (Sparse parameter)

Let $\mathbf{x}^\natural$ be s-sparse. Then $\mathbf{x}^\natural$ can be represented as the non-negative combination of s elements in $\mathcal{A}$, with $\mathcal{A} := \{\pm \mathbf{e}_1, \ldots, \pm \mathbf{e}_p\}$, where $\mathbf{e}_i := (\delta_{1,i}, \delta_{2,i}, \ldots, \delta_{p,i})$ for all i.

Example (Sparse parameter with a dictionary)

Let $\Psi \in \mathbb{R}^{m \times p}$, and let $\mathbf{y}^\natural := \Psi \mathbf{x}^\natural$ for some s-sparse $\mathbf{x}^\natural$. Then $\mathbf{y}^\natural$ can be represented as the non-negative combination of s elements in $\mathcal{A}$, with $\mathcal{A} := \{\pm \psi_1, \ldots, \pm \psi_p\}$, where ψ_k denotes the kth column of Ψ.

Atomic norms

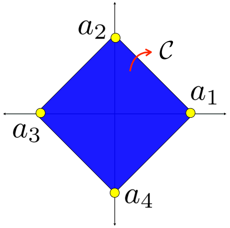
- Recall the Lasso problem

$$\mathbf{x}^\star_{\text{Lasso}} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 + \rho\,\|\mathbf{x}\|_1$$

Observations:
- ℓ_1-norm is the *atomic norm* associated with the atomic set $\mathcal{A} := \{\pm\mathbf{e}_1,\ldots,\pm\mathbf{e}_p\}$.
- The norm is closely tied with the convex hull of the set.
- We can extend the same principle for a wide range of regularizers

$$\mathcal{A} := \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}.$$

$$\mathcal{C} := \operatorname{conv}(\mathcal{A}).$$



Gauge functions and atomic norms

Definition (Gauge function)

Let $\mathcal{C}$ be a convex set in $\mathbb{R}^p$, the **gauge function** associated with $\mathcal{C}$ is given by

$$g_{\mathcal{C}}(\mathbf{x}) := \inf \{t > 0 : \mathbf{x} = t\mathbf{c} \text{ for some } \mathbf{c} \in \mathcal{C}\}.$$

Definition (Atomic norm)

Let $\mathcal{A}$ be a symmetric *atomic set* in $\mathbb{R}^p$ such that if $\mathbf{a} \in \mathcal{A}$ then $-\mathbf{a} \in \mathcal{A}$ for all $\mathbf{a} \in \mathcal{A}$. Then, the **atomic norm** associated with a symmetric atomic set $\mathcal{A}$ is given by

$$\|\mathbf{x}\|_{\mathcal{A}} := g_{\mathrm{conv}(\mathcal{A})}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p,$$

where $\mathrm{conv}(\mathcal{A})$ denotes the *convex hull* of $\mathcal{A}$.

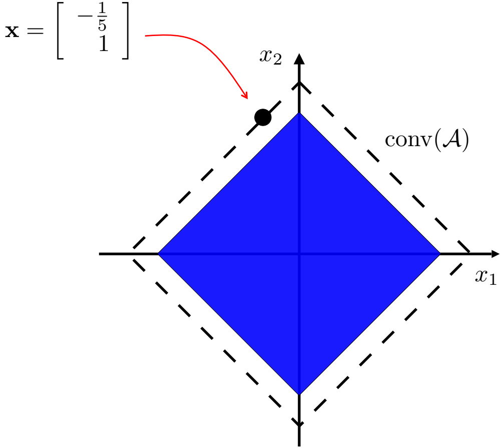
A generalization of the Lasso

Given an atomic set $\mathcal{A}$, solve the following regularized least-squares problem:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_{\mathcal{A}} \tag{1}$$

Pop quiz

Let $\mathcal{A} := \left\{(1,0)^T, (0,1)^T, (-1,0)^T, (0,-1)^T\right\}$, and let $\mathbf{x} := (-\frac{1}{5}, 1)^T$. What is $\|\mathbf{x}\|_{\mathcal{A}}$?

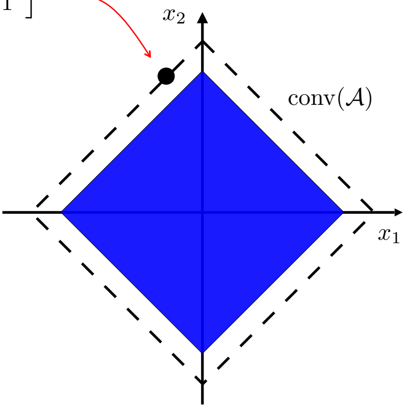


Pop quiz

Let $\mathcal{A} := \left\{(1,0)^T, (0,1)^T, (-1,0)^T, (0,-1)^T\right\}$, and let $\mathbf{x} := (-\frac{1}{5}, 1)^T$. What is $\|\mathbf{x}\|_{\mathcal{A}}$?

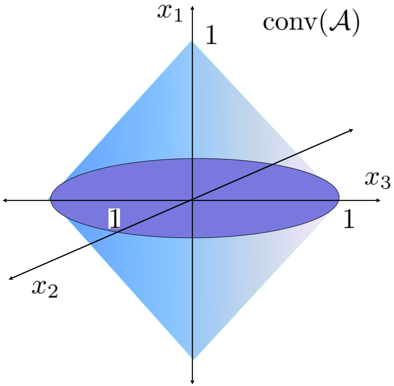
ANS: $\|\mathbf{x}\|_{\mathcal{A}} = \frac{6}{5}$.

$$\mathbf{x} = \begin{bmatrix} -\frac{1}{5} \\ 1 \end{bmatrix}$$



Pop quiz 2

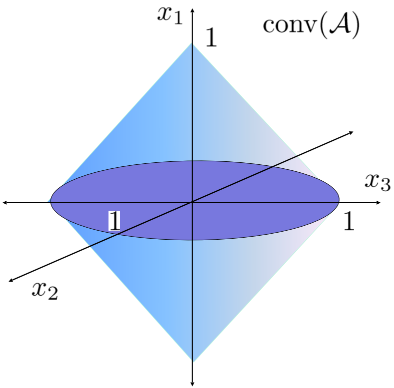
What is the expression of $\|\mathbf{x}\|_{\mathcal{A}}$ for any $\mathbf{x} := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?



Pop quiz 2

What is the expression of $\|\mathbf{x}\|_{\mathcal{A}}$ for any $\mathbf{x} := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?

ANS: $\|\mathbf{x}\|_{\mathcal{A}} = |x_1| + \left\|(x_2, x_3)^T\right\|_2$.



Application: Multi-knapsack feasibility problem

Problem formulation [9]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ which is a convex combination of k vectors in $\mathcal{A} := \{-1, +1\}^p$, and let $\mathbf{A} \in \mathbb{R}^{n\times p}$. How can we recover $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$?

The answer:

- We can use the *ℓ_∞-norm*, $\|\cdot\|_\infty$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_\infty, \rho > 0.$$

Application: Multi-knapsack feasibility problem

Problem formulation [9]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ which is a convex combination of k vectors in $\mathcal{A} := \{-1, +1\}^p$, and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. How can we recover $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$?

The answer:

- We can use the *ℓ_∞-norm*, $\|\cdot\|_\infty$ as $\|\cdot\|_\mathcal{A}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_\infty, \rho > 0.$$

The derivation:

- In this case, we have $\text{conv}(\mathcal{A}) = [-1, 1]^p$ and

$$g_{\text{conv}(\mathcal{A})}(\mathbf{x}) = \inf\left\{t > 0 : \mathbf{x} = t\mathbf{c} \text{ for some } \mathbf{c} \text{ such that } |c_i| \leq 1\ \forall i\right\}.$$

- We also have, $\forall \mathbf{x} \in \mathbb{R}^p, \mathbf{c} \in \text{conv}(\mathcal{A}), t > 0,$

$$\begin{aligned} \mathbf{x} = t\mathbf{c} &\Rightarrow \forall i, |x_i| = |tc_i| \leq t \\ &\Rightarrow g_{\text{conv}(\mathcal{A})}(\mathbf{x}) \geq \max_i |x_i|. \end{aligned}$$

- Let $\mathbf{x} \neq 0$, let $j \in \arg\max_i |x_i|$ and choose $t = \max_i |x_i|$, $c_i = x_i/t \in [-1, 1]^p$.
- Then, $\mathbf{x} = t\mathbf{c}$, and so $g_{\text{conv}(\mathcal{A})}(\mathbf{x}) \leq \max_i |x_i|$.

Application: Matrix completion

Problem formulation [2, 5]

Let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with $\mathrm{rank}(\mathbf{X}^\natural) = r$, and let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be matrices in $\mathbb{R}^{p\times p}$. How do we estimate $\mathbf{X}^\natural$ given $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $b_i = \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}^\natural\right) + w_i$, $i = 1, \ldots, n$, where $\mathbf{w} := (w_1, \ldots, w_n)^T$ denotes unknown noise?

The answer: ◦ We can use the *nuclear norm*, $\|\cdot\|_*$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{X}\in\mathbb{R}^{p\times p}} \sum_{i=1}^{n} \left(b_i - \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}\right)\right)^2 + \rho \left\|\mathbf{X}\right\|_*, \rho > 0.$$

Application: Matrix completion

Problem formulation [2, 5]

Let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with $\mathrm{rank}(\mathbf{X}^\natural) = r$, and let $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be matrices in $\mathbb{R}^{p\times p}$. How do we estimate $\mathbf{X}^\natural$ given $\mathbf{A}_1, \ldots, \mathbf{A}_n$ and $b_i = \mathrm{Tr}\left(\mathbf{A}_i \mathbf{X}^\natural\right) + w_i$, $i = 1, \ldots, n$, where $\mathbf{w} := (w_1, \ldots, w_n)^T$ denotes unknown noise?

The answer:

- We can use the *nuclear norm*, $\|\cdot\|_*$ as $\|\cdot\|_{\mathcal{A}}$. The regularized estimator is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{X}\in\mathbb{R}^{p\times p}} \sum_{i=1}^{n} \left(b_i - \mathrm{Tr}\left(\mathbf{A}_i\mathbf{X}\right)\right)^2 + \rho \|\mathbf{X}\|_*, \rho > 0.$$

The derivation:

- Let us use the following atomic set $\mathcal{A} = \left\{\mathbf{X} : \mathrm{rank}\ (\mathbf{X}) = 1, \|\mathbf{X}\|_F = 1, \mathbf{X} \in \mathbb{R}^{p\times p}\right\}$.
- Let $\forall \mathbf{X} \in \mathbb{R}^{p\times p}, \mathsf{C} = \sum_i \lambda_i \mathsf{C}_i \in \mathrm{conv}(\mathcal{A}), \sum_i \lambda_i = 1, \mathsf{C}_i \in \mathcal{A}, t > 0$. Then, we have

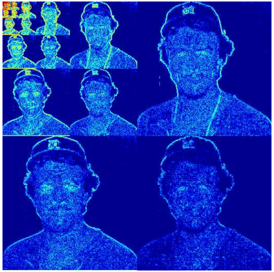
$$\mathbf{X} = t\sum_i \lambda_i \mathsf{C}_i \Rightarrow \|\mathbf{X}\|_* = t\left\|\sum_i \lambda_i \mathsf{C}_i\right\|_* \le t\sum_i \lambda_i \|\mathsf{C}_i\|_* \le t \Rightarrow g_{\mathrm{conv}(\mathcal{A})}(\mathbf{X}) \ge \|\mathbf{X}\|_*.$$

- Let $\mathbf{X} \neq 0$, let $\mathbf{X} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^t$ be its SVD decomposition, where σ_i's are its singular values.
- Let $t = \|\mathbf{X}\|_* = \sum_i |\sigma_i|$, $\mathsf{C}_i = \mathbf{u}_i\mathbf{v}_i^T \in \mathcal{A}$, $\forall i$. Then, $\mathbf{X} = t\sum_i \lambda_i \mathsf{C}_i$, $\lambda_i = \frac{|\sigma_i|}{t}$.
- Since t is feasible and $\sum_i \lambda_i = 1$, it follows that $g_{\mathrm{conv}(\mathcal{A})}(\mathbf{X}) \le \|\mathbf{X}\|_*$.

Structured Sparsity

There exist many more structures that we have not covered here, each of which is handled using different non-smooth regularizers. Some examples [1, 8]:

- **Group Sparsity:** Many signals are not only sparse, but the non-zero entries tend to cluster according to known patterns.
- **Tree Sparsity:** When natural images are transformed to the Wavelet domain, their significant entries form a *rooted connected tree*.



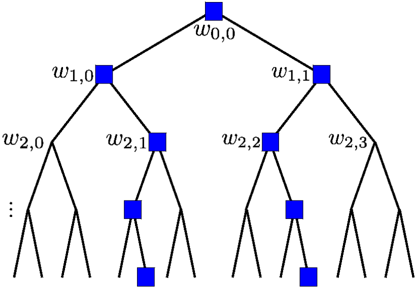


Figure: (**Left panel**) Natural image in the Wavelet domain. (**Right panel**) Rooted connected tree containing the significant coefficients.

Selection of the Parameters

In all of these problems, there remain the issues of *how to design* $\mathbf{A}$ and *how to choose* ρ.

Design of $\mathbf{A}$:

- Sometimes $\mathbf{A}$ is given "by nature", whereas sometimes it can be designed
- For the latter case, i.i.d. Gaussian designs provide good theoretical guarantees, whereas in practice we must resort to structured matrices permitting more efficient storage and computation
- See [6] for an extensive study in the context of compressive sensing

Selection of ρ:

- Theoretical bounds provide some insight, but usually the direct use of the theoretical choice does not suffice
- In practice, a common approach is *cross-validation* [4], which involves searching for a parameter that performs well on a set of known training signals
- Other approaches include *covariance penalty* [4] and *upper bound heuristic* [13]

Non-smooth unconstrained convex minimization

Problem (Mathematical formulation)

How can we find an optimal solution to the following optimization problem?

$$\boxed{F^\star := \min_{\mathbf{x} \in \mathbb{R}^p} f(\mathbf{x})} \tag{2}$$

where f is proper, closed, convex, but not everywhere differentiable.

Subdifferentials: A generalization of the gradient

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \ \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathcal{Q}\}\,.$$

Each element $\mathbf{v}$ of $\partial f(\mathbf{x})$ is called *subgradient* of f at $\mathbf{x}$.

Lemma

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function. Then, the subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ contains only the gradient, i.e., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

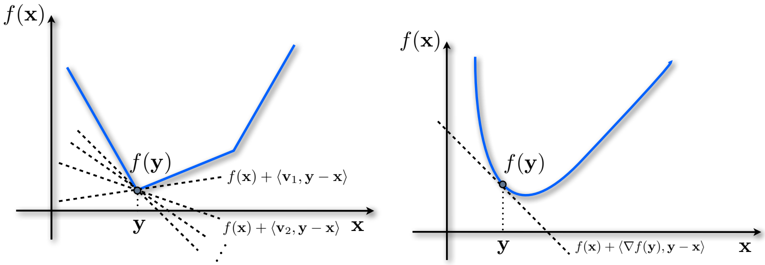


Figure: (**Left**) Non-differentiability at point $\mathbf{y}$. (**Right**) Gradient as a subdifferential with a singleton entry.

(Sub)gradients in convex functions

Example

$f(x) = |x| \qquad \longrightarrow \qquad \partial|x| = \{\text{sgn}(x)\}$, if $x \neq 0$, but $[-1, 1]$, if $x = 0$.

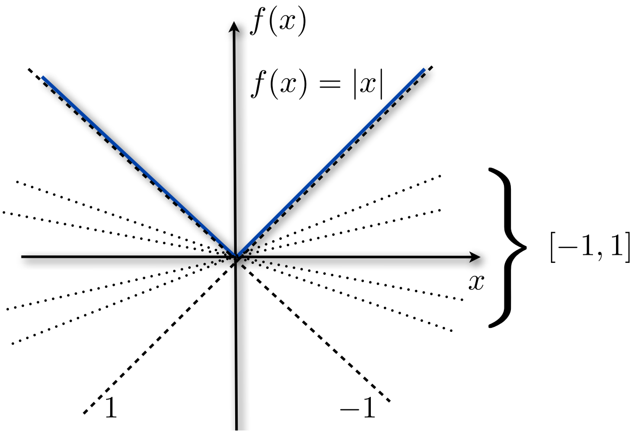


Figure: Subgradients of $f(x) = |x|$ in $\mathbb{R}$.

Subdifferentials: Two basic results

Lemma (Necessary and sufficient condition)

$$\mathbf{x}^\star \in \mathrm{dom}\,(F) \text{ *is a **globally optimal** solution to* } (2) \quad \textit{iff} \quad 0 \in \partial F(\mathbf{x}^\star).$$

Sketch of the proof.

- $\Leftarrow$: For any $\mathbf{x} \in \mathbb{R}^p$, by definition of $\partial F(\mathbf{x}^\star)$:

$$F(\mathbf{x}) - F(\mathbf{x}^\star) \geq 0^T(\mathbf{x} - \mathbf{x}^\star) = 0,$$

that is, $\mathbf{x}^\star$ is a global solution to (2).

- $\Rightarrow$: If $\mathbf{x}^\star$ is a global of (2) then for every $\mathbf{x} \in \mathrm{dom}\,(F)$, $F(\mathbf{x}) \geq F(\mathbf{x}^\star)$ and hence

$$F(\mathbf{x}) - F(\mathbf{x}^\star) \geq 0^T(\mathbf{x} - \mathbf{x}^\star), \forall \mathbf{x} \in \mathbb{R}^p,$$

which leads to $0 \in \partial F(\mathbf{x}^\star)$. □

Theorem (Moreau-Rockafellar's theorem [11])

Let ∂f and ∂g be the subdiffierential of f and g, respectively. If $f, g \in \mathcal{F}(\mathbb{R}^p)$ and $\mathrm{dom}\,(f) \cap \mathrm{dom}\,(g) \neq \emptyset$, then:

$$\boxed{\partial(f + g) = \partial f + \partial g.}$$

Non-smooth unconstrained convex minimization

Problem (**Non-smooth convex minimization**)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) \tag{3}$$

Subgradient method

The subgradient method relies on the fact that even though f is non-smooth, we can still compute its **subgradients**, informing of the local descent directions.

Subgradient method
1. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point. 2. For $k = 0, 1, \cdots,$ perform: $$\begin{cases} \mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha_k \mathbf{d}^k, \end{cases} \tag{4}$$ where $\mathbf{d}^k \in \partial f(\mathbf{x}^k)$ and $\alpha_k \in (0,1]$ is a given step size.

Convergence of the subgradient method

Theorem

Assume that the following conditions are satisfied:

1. $\|\mathbf{g}\|_2 \leq G$ *for all* $\mathbf{g} \in \partial f(\mathbf{x})$ *for any* $\mathbf{x} \in \mathbb{R}^p$.
2. $\|\mathbf{x}^0 - \mathbf{x}^\star\|_2 \leq R$

Let the stepsize be chosen as

$$\alpha_k = \frac{R}{G\sqrt{k}}$$

then the iterates generated by the subgradient method satisfy

$$\min_{0 \leq i \leq k} f(\mathbf{x}^i) - f^\star \leq \frac{RG}{\sqrt{k}}.$$

Remarks

- Condition (1) holds, for example, when f is G-Lipschitz.
- **The convergence rate of $\mathcal{O}(1/\sqrt{k})$ is the slowest we have seen so far!**

Stochastic subgradient methods

- An unbiased stochastic subgradient

$$\mathbb{E}[G(\mathbf{x})|\mathbf{x}] \in \partial f(\mathbf{x}).$$

- Stochastic gradient methods using unbiased subgradients instead of unbiased gradients work

The classic stochastic subgradient methods (SG)
1. Choose $\mathbf{x}_1 \in \mathbb{R}^p$ and $(\gamma_k)_{k\in\mathbb{N}} \in]0, +\infty[^{\mathbb{N}}$.
2. For $k = 1, \ldots$ perform: $$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k G(\mathbf{x}_k).$$

Theorem (Convergence in expectation [12])

Suppose that:

1. $\mathbb{E}[\|G(\mathbf{x}^k)\|^2] \leq M^2$,
2. $\gamma_k = \gamma_0/\sqrt{k}$.

Then,

$$\mathbb{E}[f(\mathbf{x}^k) - f(\mathbf{x}^\star)] \leq \left(\frac{D^2}{\gamma_0} + \gamma_0 M^2\right)\frac{2 + \log k}{\sqrt{k}}.$$

Remark: ◦ The rate is $\mathcal{O}(\log k/\sqrt{k})$ instead of $\mathcal{O}(1/\sqrt{k})$ for the deterministic algorithm.

Composite convex minimization

Problem (Unconstrained composite convex minimization)

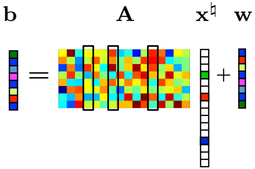
$$\boxed{F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}} \tag{5}$$

- *f and g are both proper, closed, and convex.*
- $\mathrm{dom}(F) := \mathrm{dom}(f) \cap \mathrm{dom}(g) \neq \emptyset$ *and* $-\infty < F^\star < +\infty$.
- *The solution set* $\mathcal{S}^\star := \{\mathbf{x}^\star \in \mathrm{dom}(F) : F(\mathbf{x}^\star) = F^\star\}$ *is nonempty.*

Two remarks

- Nonsmoothness: At least one of the two functions f and g is **nonsmooth**
  - General nonsmooth convex optimization methods (e.g., classical subgradient methods, level, or bundle methods) lack efficiency and numerical robustness.
    - Require $\mathcal{O}(\epsilon^{-2})$ iterations to reach a point $\mathbf{x}^\star_\epsilon$ such that $F(\mathbf{x}^\star_\epsilon) - F^\star \leq \epsilon$. Hence, to reach $\mathbf{x}^\star_{0.01}$ such that $F(\mathbf{x}^\star_{0.01}) - F^\star \leq 0.01$, we need $\mathcal{O}(10^4)$ iterations.
- Generality: it covers a wider range of problems than smooth unconstrained problems, e.g., when handling regularized M-estimation,
  - f is a loss function, a data fidelity, or negative log-likelihood function.
  - g is a regularizer, encouraging structure and/or constraints in the solution.

Example 1: Sparse regression in generalized linear models (GLMs)



Problem (Sparse regression in GLM)

Our goal is to estimate $\mathbf{x}^\natural \in \mathbb{R}^p$ given $\{b_i\}_{i=1}^n$ and $\{\mathbf{a}_i\}_{i=1}^n$, knowing that the likelihood function at y_i given $\mathbf{a}_i$ and $\mathbf{x}^\natural$ is given by $L(\langle \mathbf{a}_i, \mathbf{x}^\natural \rangle, b_i)$, and that $\mathbf{x}^\natural$ is sparse.

Optimization formulation

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \underbrace{-\sum_{i=1}^{n} \log L(\langle \mathbf{a}_i, \mathbf{x}^\natural \rangle, b_i)}_{f(\mathbf{x})} + \underbrace{\rho_n \|\mathbf{x}\|_1}_{g(\mathbf{x})} \right\}$$

where $\rho_n > 0$ is a parameter which controls the strength of sparsity regularization.

Theorem (cf. [10] for details)

Under some technical conditions, there exists $\{\rho_i\}_{i=1}^\infty$ such that with high probability,

$$\left\|\mathbf{x}^\star - \mathbf{x}^\natural\right\|_2^2 = \mathcal{O}\left(\frac{s \log p}{n}\right), \quad \operatorname{supp} \mathbf{x}^\star = \operatorname{supp} \mathbf{x}^\natural.$$

Recall ML: $\left\|\mathbf{x}_{ML} - \mathbf{x}^\natural\right\|_2^2 = \mathcal{O}(p/n)$.

Example 2: Image processing

Problem (Imaging denoising/deblurring)

Our goal is to obtain a clean image $\mathbf{x}$ *given "dirty" observations* $\mathbf{b} \in \mathbb{R}^{n\times 1}$ *via* $\mathbf{b} = \mathcal{A}(\mathbf{x}) + \mathbf{w}$, *where* $\mathcal{A}$ *is a linear operator, which, e.g., captures camera blur as well as image subsampling, and* $\mathbf{w}$ *models perturbations, such as Gaussian or Poisson noise.*

Optimization formulation

$$\text{Gaussian}: \quad \min_{\mathbf{x}\in\mathbb{R}^{n\times p}} \Big\{ \underbrace{(1/2)\|\mathcal{A}(\mathbf{x}) - \mathbf{b}\|_2^2}_{f(\mathbf{x})} + \underbrace{\rho\|\mathbf{x}\|_{\mathrm{TV}}}_{g(\mathbf{x})} \Big\}$$

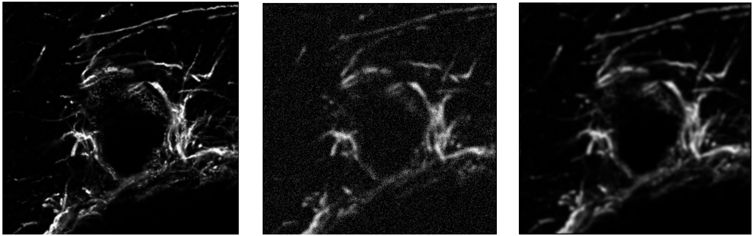
$$\text{Poisson}: \quad \min_{\mathbf{x}\in\mathbb{R}^{n\times p}} \Big\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \ln\left(\langle \mathbf{a}_i, \mathbf{x}\rangle\right)\right]}_{f(\mathbf{x})} + \underbrace{\rho\|\mathbf{x}\|_{\mathrm{TV}}}_{g(\mathbf{x})} \Big\}$$

where $\rho > 0$ is a regularization parameter and $\|\cdot\|_{\mathrm{TV}}$ is the total variation (TV) norm:

$$\|\mathbf{x}\|_{\mathrm{TV}} := \begin{cases} \sum_{i,j} |\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}| + |\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}| & \text{anisotropic case,} \\ \sum_{i,j} \sqrt{|\mathbf{x}_{i,j+1} - \mathbf{x}_{i,j}|^2 + |\mathbf{x}_{i+1,j} - \mathbf{x}_{i,j}|^2} & \text{isotropic case} \end{cases}$$

Example 3: Confocal microscopy with camera blur and Poisson observations

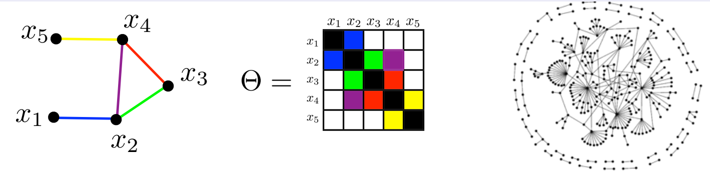
Original image $\mathbf{x}^\natural$ | Observed image $\mathbf{b}$ | Estimate $\mathbf{x}^\star$



Example 4: Sparse inverse covariance estimation

Problem (Graphical model selection)

Given a data set $\mathcal{D} := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ is a Gaussian random variable. Let Σ be the covariance matrix corresponding to the graphical model of the Gaussian Markov random field. Our goal is to learn a sparse precision matrix Θ (i.e., the inverse covariance matrix Σ^{-1}) that captures the Markov random field structure..



Optimization formulation

$$\boxed{\min_{\Theta \succ 0} \left\{ \underbrace{\mathrm{tr}(\Sigma\Theta) - \log\det(\Theta)}_{f(\mathbf{x})} + \underbrace{\lambda \|\mathrm{vec}(\Theta)\|_1}_{g(\mathbf{x})} \right\}} \tag{6}$$

where $\Theta \succ 0$ means that Θ is symmetric and positive definite and $\lambda > 0$ is a regularization parameter and vec is the vectorization operator.

Wrap up!

- Three supplementary lectures to take a look once the course is over!
  - One on compressive sensing (Math of Data Lecture 4 from 2014):
    https://archive-wp.epfl.ch/lions/wp-content/uploads/2019/01/lecture-4-2014.pdf
  - One on source separation (Math of Data Lecture 6 from 2014)
    https://archive-wp.epfl.ch/lions/wp-content/uploads/2019/01/lecture-6-2014.pdf
  - One on convexification of structured sparsity models (research presentation)
    https://www.epfl.ch/labs/lions/wp-content/uploads/2019/01/volkan-TU-view-web.pdf

References I

[1] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
Information Theory, IEEE Transactions on, 56(4):1982–2001, 2010.

[2] Emmanuel Candès and Benjamin Recht.
Exact matrix completion via convex optimization.
Found. Comput. Math., 9:717–772, 2009.

[3] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky.
The convex geometry of linear inverse problems.
Found. Comput. Math., 12:805–849, 2012.

[4] Bradley Efron.
The estimation of prediction error: Covariance penalities and cross-validation.
J. Am. Stat. Assoc., 99(467):619–632, September 2004.

[5] Steven T. Flammia, David Gross, Yi-Kai Liu, and Jens Eisert.
Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators.
New J. Phys., 14, 2012.

References II

[6] Simon Foucart and Holger Rauhut.
A mathematical introduction to compressive sensing.
Springer, 2013.

[7] Rémi Gribonval, Volkan Cevher, and Mike E. Davies.
Compressible distributions for high-dimensional statistics.
IEEE Trans. Inf. Theory, 58(8):5016–5034, 2012.

[8] Marwa El Halabi and Volkan Cevher.
A totally unimodular view of structured sparsity.
http://arxiv.org/abs/1411.1990, 2014.

[9] O. L. Mangasarian and Benjamin Recht.
Probability of unique integer solution to a system of linear equations.
Eur. J. Oper. Res., 214:27–30, 2011.

[10] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu.
A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers.
Stat. Sci., 27(4):538–557, 2012.

References III

[11] R. Tyrrell Rockafellar.
Convex analysis.
Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[12] Ohad Shamir and Tong Zhang.
Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.
In *International Conference on Machine Learning*, pages 71–79, 2013.

[13] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi.
Simple error bounds for regularized noisy linear inverse problems.
2014.
arXiv:1401.6578v1 [math.OC].