Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 7: Motivation for Non-Smooth Optimization Problems

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

EE-556 (Fall 2015)

License Information for Mathematics of Data Slides

- This work is released under a Creative Commons License with the following terms:
- **Attribution**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.
- **Non-Commercial**
  - The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor's permission.
- **Share Alike**
  - The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor's work.
- Full Text of the License

Outline

- This lecture
  1. Deficiency of smooth models
  2. Motivation for non-smooth models
  3. Compressive sensing
  4. Subgradient descent
- Next lecture
  1. Unconstrained, non-smooth composite minimization
  2. Convergence and convergence rate characterization of various approaches

Recommended Reading

- Chapter 2 in S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.
- Section 3.2.3 in Y. Nesterov, *Introductory Lectures on Convex Optimization*. Springer Science + Business Media, 2004.

Motivation

Motivation

Nonsmooth convex optimization problems arise frequently in applications.

In some cases, nonsmooth *regularizers* are intentionally introduced to improve statistical accuracy in estimation.

This lecture gives an introduction to nonsmooth functions and optimization, including a number of specific motivating examples based on linear inverse problems.

Recap: Oracle information

Oracle Information

Algorithms are assumed to have access to *oracle information*:

- Function value $f(\mathbf{x})$
- Gradient $\nabla f(\mathbf{x})$
- Hessian $\nabla^2 f(\mathbf{x})$
- ...

Note: *How* we get such information varies between problems and applications

Recap: Oracle information

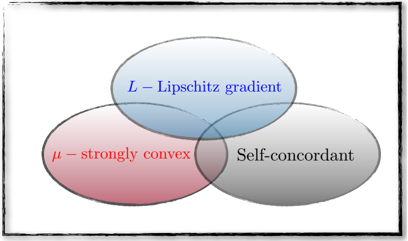
Oracle Information

Algorithms are assumed to have access to *oracle information*:

- Function value $f(\mathbf{x})$
- Gradient $\nabla f(\mathbf{x})$
- Hessian $\nabla^2 f(\mathbf{x})$
- ...

Note: *How* we get such information varies between problems and applications

For smooth objective functions, we have seen that various properties can significantly help speed up the optimization:



Differentiability in functions

Definition (Differentiability classes)

A function $f : \mathbb{R} \to \mathbb{R}$ is in the differentiability class C^k if its derivatives up to order k exist and are continuous.

- **Note:** In some fields, the word “smooth” refers specifically to the class C^∞. In optimization, it usually refers to C^1 with Lipschitz gradient.
- Some examples:

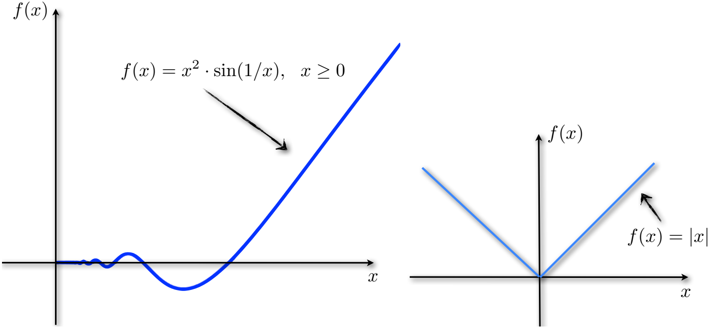


Figure: (**Left panel**) ∞-times continuously differentiable function in $\mathbb{R}$. (**Right panel**) Non-differentiable $f(x) = |x|$ in $\mathbb{R}$.

Differentiability in functions

Useful Fact 1

All convex functions are continuous (except possibly on the boundary of their domain/effective domain)

However, they need not even be differentiable: e.g. $f(x) = |x|$

Non-differentiable $\Longrightarrow$ No gradient descent, no Newton's method...

Useful fact 2:

Non-differentiable functions can still be strongly convex and/or Lipschitz continuous (but of course not Lipschitz gradient)

Non-smoothness

Many optimization problems that we would like to solve are non-smooth – how do we solve them?

This lecture: Some motivating examples, and simple techniques for solving them.

Simple examples of non-smoothness

Example 1: Simultaneously maximizing multiple objectives

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. $f(x) = \max\{x, x^2\}$

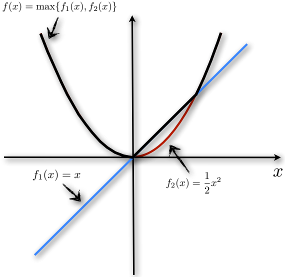


Figure: Maximum of two smooth convex functions.

Simple examples of non-smoothness

Example 2: 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).

Simple examples of non-smoothness

Example 2: 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: $\hat{\mathbf{x}}_{\mathrm{LS}} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$

- Convex, smooth, and an *explicit solution*: $\hat{\mathbf{x}}_{\mathrm{LS}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \mathbf{A}^{\dagger}\mathbf{b}$

Alternative approach: Least absolute value deviation: $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_1$

- The advantage: Improved robustness against outliers (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 $\hat{\mathbf{x}}$.

Practical performance

Denote the numerical approximation by $\mathbf{x}^\star_\epsilon$. The practical performance is determined by

$$\left\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}^\star_\epsilon - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} .$$

Sometimes *non-smooth* estimators of $\mathbf{x}^\natural$ can help *reduce the statistical error*.

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

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

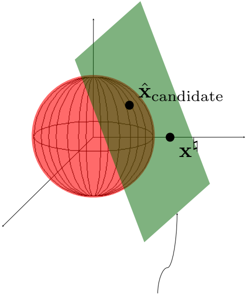
- If $\mathbf{A}$ has full column rank, $\hat{\mathbf{x}}_{\text{LS}} = \mathbf{A}^\dagger \mathbf{b}$ is uniquely defined.
- *In the case that $n < p$*, $\mathbf{A}$ cannot have full column rank, and we can only conclude that $\hat{\mathbf{x}}_{\text{LS}} \in \left\{ \mathbf{A}^\dagger \mathbf{b} + \mathbf{h} : \mathbf{h} \in \operatorname{null}(\mathbf{A}) \right\}$.

Observation: The estimation error $\left\| \hat{\mathbf{x}}_{\text{LS}} - \mathbf{x}^\natural \right\|_2$ can be *arbitrarily large*!

A candidate solution

Continuing the LS example:

- In other words, there are infinitely many solutions $\mathbf{x}$ such that $\mathbf{b} = \mathbf{A}\mathbf{x}$
- Suppose that $\mathbf{w} = 0$ (i.e. no noise). Should we just choose the one $\hat{\mathbf{x}}_{\text{candidate}}$ with the smallest norm $\|\mathbf{x}\|_2$?



$\mathbf{x}^\natural + \mathbf{h}$, $\mathbf{h} \in \text{null}(\mathbf{A})$

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 $\mathbf{x}^\natural \in \mathbb{R}^p$.

Observation: The estimation error may *not* diminish unless n is very close to p.

Intuition: The relation $n < p$ means that the dimension of the sample $\mathbf{b}$ exceeds the number of unknown variables in $\mathbf{x}^\natural$ to be solved.

Impact: It is impossible to estimate $\mathbf{x}^\natural$ accurately using $\hat{\mathbf{x}}_{\text{candidate}}$ when $n \ll p$ even if $\mathbf{w} = \mathbf{0}$.

- The statistical error $\left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\natural\right\|_2^2$ can also be arbitrarily large when $\mathbf{w} \neq \mathbf{0}$. Hence, the solution is also not robust.

Summarizing the findings so far

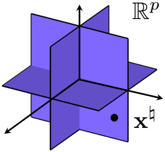
The message so far:

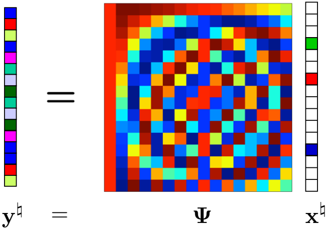
- Even in the absence of noise, we cannot recover $\mathbf{x}^\natural$ from the observations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ unless $n \geq p$
- But in applications, p might be thousands, millions, billions...
- **Can we get away with $n \ll p$ under some further assumptions on $\mathbf{x}$?**

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.

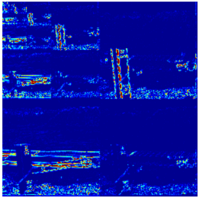




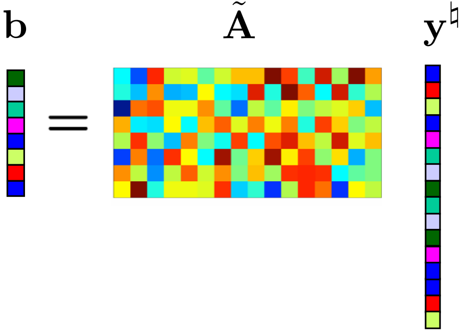
Sparse representations

$\mathbf{x}^\natural$: *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...

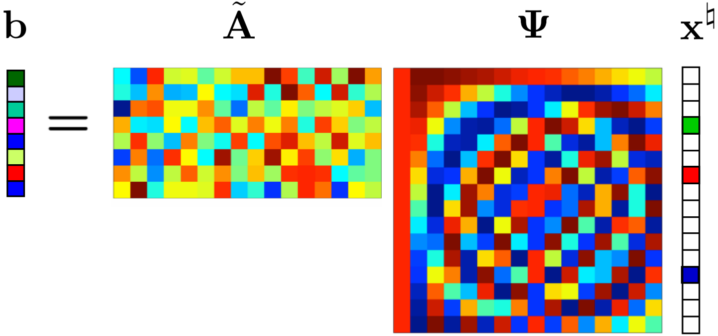


Sparse representations strike back!



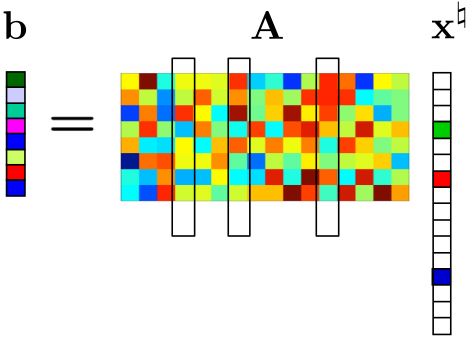
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



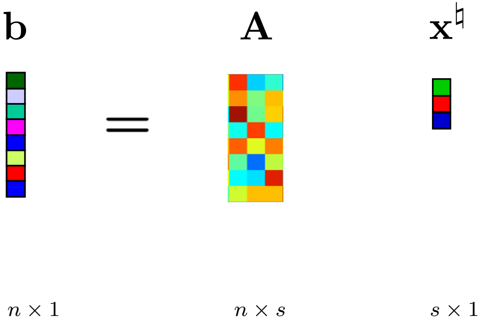
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$, $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n$

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!



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

$n \times 1$ $\quad n \times s$ $\quad s \times 1$

A fundamental impact:

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* $\mathbf{x}^\natural$.

Stability and robustness

The most basic problem is to recover $\mathbf{x}^\natural$ from noiseless measurements $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$, also given knowledge of $\mathbf{A}$. However, in practice we usually need more.

Robustness

A *robust* recovery algorithm is one that is robust to noise: If $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, then the effect of $\mathbf{w}$ on the error $\|\hat{\mathbf{x}} - \mathbf{x}^\natural\|_2^2$ is small when $\|\mathbf{w}\|_2^2$ is small.

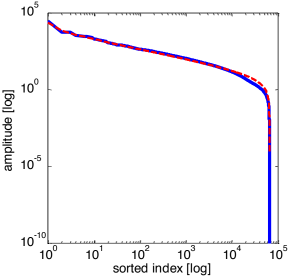
Stability

A *stable* recovery algorithm is one that is robust to signals that are not exactly sparse: If $\mathbf{x}^\natural = \mathbf{x}_s + \mathbf{x}'$ for some s-sparse signal $\mathbf{x}_s$, then the effect of $\mathbf{x}'$ on the error $\|\hat{\mathbf{x}} - \mathbf{x}^\natural\|_2^2$ is small when $\|\mathbf{x}'\|_2^2$ is small.

Compressible signals

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

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, $|\{k : |x_k| \geq t, 1 \leq k \leq p\}|$, 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 $\mathrm{b} = \mathrm{Ax} + \mathrm{w}$

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 is determined by

$$\|\mathbf{x}^\star_\epsilon - \mathbf{x}_{\mathsf{real}}\|_2 \leq \underbrace{\|\mathbf{x}^\star_\epsilon - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} + \underbrace{\left\|\mathbf{x}_{\mathsf{real}} - \mathbf{x}^\natural\right\|_2}_{\text{model error}}.$$

- A great deal of research goes into learning representations that renders the model error negligible while still keeping statistical error low.

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.

With this approach, the 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.

With this approach, the stable and robust recovery of any s-sparse signal is possible using just $n = 2s$ measurements.

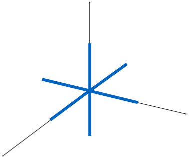
Issues

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

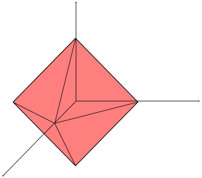
The ℓ_1-norm heuristic

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
$\left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$

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

Sparse recovery via the Lasso

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

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

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

Note: 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

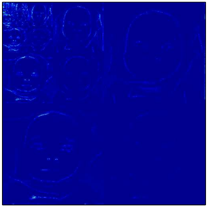
$$\left\| \hat{\mathbf{x}}_{lasso} - \mathbf{x}^\natural \right\|_2 \leq \frac{8\sigma}{\mu(\mathbf{A})} \sqrt{\frac{s \ln p}{n}},$$

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

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

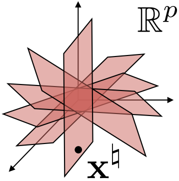
Other models with simplicity

p
pixels

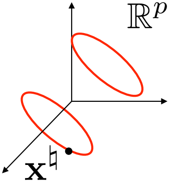


Information level:

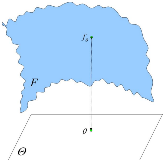
$s \ll p$
large wavelet coefficients
(blue = 0)



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 norm

Recall that we handled sparse (or compressible) vectors by solving the Lasso problem

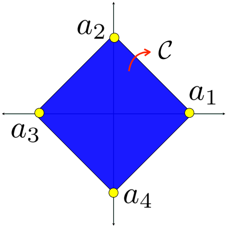
$$\hat{\mathbf{x}}_{\text{lasso}} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho\,\|\mathbf{x}\|_1$$

We observe that the ℓ_1-norm is the *atomic norm* associated with the atomic set $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$, which is indeed the convex hull of the set.

This same principle leads to effective regularizers for a wide range of atomic structures.

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

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



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

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

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 $\mathcal{A}$.

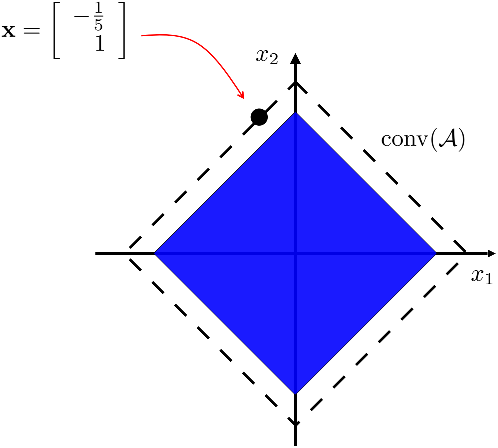
A Generalization of the Lasso

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

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

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 $\|\mathbf{x}\|_{\mathcal{A}}$?

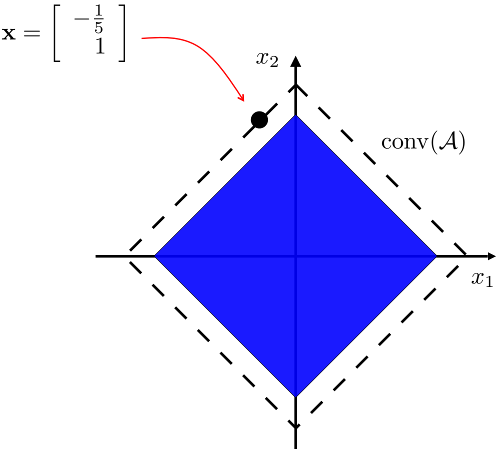


$\mathbf{x} = \begin{bmatrix} -\frac{1}{5} \\ 1 \end{bmatrix}$
x_2
$\mathrm{conv}(\mathcal{A})$
x_1

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 $\|\mathbf{x}\|_{\mathcal{A}}$?

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



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

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 $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$?

In this case, $\|\cdot\|_{\mathcal{A}}$ is the ℓ_∞-norm, and the regularized least-squares problem is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_\infty$$

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?

This is a special case of the atomic norm formulation with $\mathcal{A} = \left\{\mathbf{X} : \mathrm{rank}\ (\mathbf{X}) = 1, \|\mathbf{X}\|_F = 1, \mathbf{X} \in \mathbb{R}^{p\times p}\right\}$. It can be shown that $\|\cdot\|_{\mathcal{A}}$ is the *nuclear norm*, $\|\cdot\|_*$. The regularized least-squares problem is

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

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*.

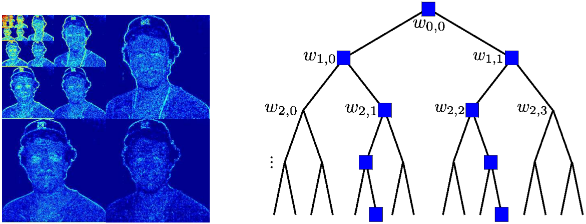


Figure: (**Left panel**) Natural image in the Wavelet domain. (**Right panel**) Rooted connected tree containing the significant coefficients.

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* [11]

How can we optimize non-smooth functions?

Recall: Gradient methods, Newton's method, etc. no longer applicable

Rest of this lecture: A simple extension of the gradient method
Next lecture: More sophisticated approaches

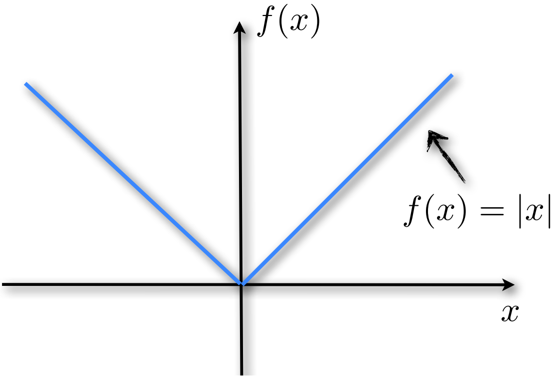


Figure: Non-differentiable at the origin

Subdifferentials and (sub)gradients in convex functions

- Subdifferential: generalizes ∇ to *nondifferentiable functions*

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 $\mathbf{x}$.

Definition

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., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

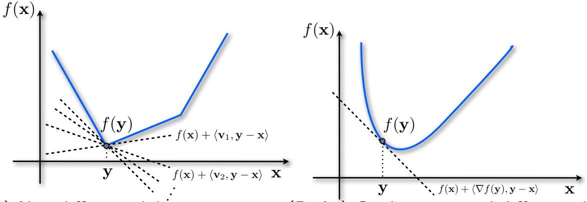

$f(\mathbf{x})$
$f(\mathbf{y})$
$f(\mathbf{x}) + \langle \mathbf{v}_1, \mathbf{y} - \mathbf{x} \rangle$
$\mathbf{y}$
$f(\mathbf{x}) + \langle \mathbf{v}_2, \mathbf{y} - \mathbf{x} \rangle$
$\mathbf{x}$
$f(\mathbf{x})$
$f(\mathbf{y})$
$\mathbf{y}$
$f(\mathbf{x}) + \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{x} \rangle$
$\mathbf{x}$

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

Subdifferentials and (sub)gradients in convex functions

Example

- $f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \longrightarrow \quad \nabla f(\mathbf{x}) = -2\mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x})$.
- $f(\mathbf{X}) = -\log\det(\mathbf{X}) \quad \longrightarrow \quad \nabla f(\mathbf{X}) = \mathbf{X}^{-1}$
- $f(x) = |x| \quad \longrightarrow \quad \partial|x| = \{\text{sgn}(x)\}$, if $x \neq 0$, but $[-1, 1]$, if $x = 0$.

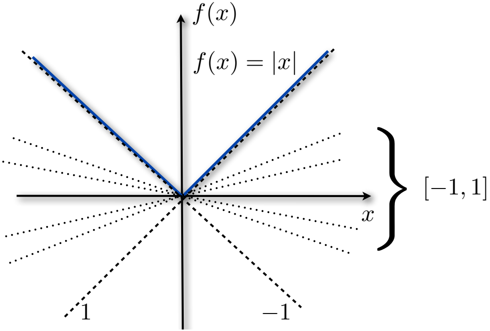


Figure: Subdifferential of $f(x) = |x|$ in $\mathbb{R}$.

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, $f \in \mathcal{F}$.
Note that (2) is unconstrained.

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:

$$\left\{ \begin{array}{ll} \mathbf{x}^{k+1} & = \mathbf{x}^k - \alpha_k \mathbf{d}^k, \end{array} \right. \tag{3}$$

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!**

Next lecture: Achieving guarantees for (many) non-smooth optimization problems that are just as good as those for smooth ones

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.

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

References II

[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.

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