High-Dimensional Data Analysis

Lecture 9 - Relaxing the Sparse Recovery Problem

Fall semester - 2024

Dr. Eng. Valentin Leplat

Innopolis University

October 31, 2024

Outline

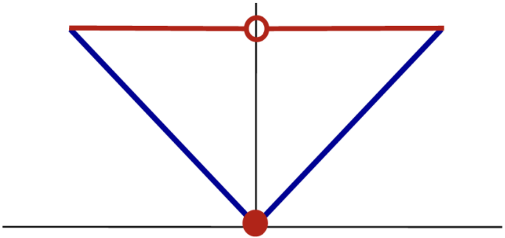
Convex Functions and Convexification

Why Convexification?

Intuitive reasons why ℓ^0 minimization:

$$\min_x \|x\|_0 \text{ subject to } Ax = y \tag{1}$$

is challenging:



Not amenable to local search methods such as gradient descent.

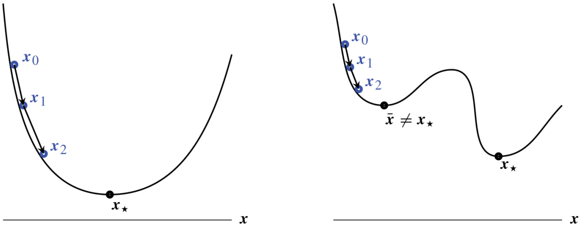
Convex versus Nonconvex Functions

For minimizing a generic function:

$$\min_x f(x), \qquad x \in \mathcal{C} \text{ (a convex set)}, \tag{2}$$

conduct **local gradient descent search:**

$$x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k) \tag{3}$$



(a) convex (b) nonconvex

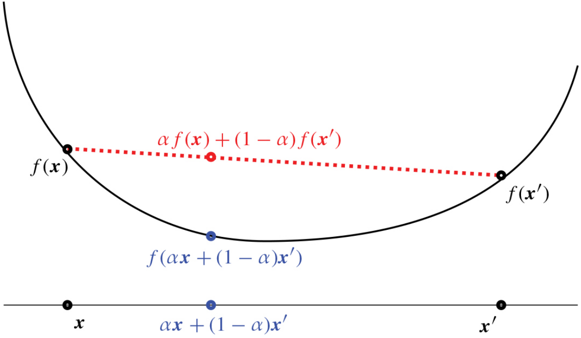
Intuitively, **convexity leads to global optimality.**

Convex Functions

Definition: Convex Function

A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for every pair of points $x, x^{'} \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ it satisfies the Jensen's inequality:

$$f(\alpha x + (1-\alpha)x^{'}) \leqslant \alpha f(x) + (1-\alpha)f(x^{'}) \tag{4}$$



Global Optimality

Proposition 1

Any local minimum of a convex function is also a global minimum.

Proof. Let $\bar{x}$ be a local minimum: $\forall x : \|x - \bar{x}\|_2 \leqslant \epsilon$, we have $f(\bar{x}) \leqslant f(x)$. Assume $x^\star$ is the global minimum and $f(\bar{x}) > f(x^\star)$. Choose λ such that $x_\lambda = \lambda\bar{x} + (1-\lambda)x^\star$ satisfies $\|x_\lambda - \bar{x}\|_2 \leqslant \epsilon$. Then

$$\begin{aligned} f(\bar{x}) &\leqslant f(x_\lambda) \\ &= f(\lambda\bar{x} + (1-\lambda)x^\star) \\ &\leqslant \lambda f(\bar{x}) + (1-\lambda)f(x^\star) \\ &< f(\bar{x}) \end{aligned} \tag{5}$$

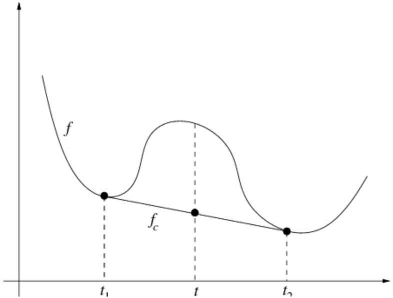
We have a contradiction, we must have $\bar{x} = x^\star$, and the result follows.

ℓ^1 Norm as Convex Surrogate for ℓ^0 Norm

Convex Envelope

Definition: Lower Convex Envelope

A function $f_c(x)$ is said to be a (lower) **convex envelope** of $f(x)$ if for all convex functions $g \leqslant f$ we have $g \leqslant f_c$.



Lower convex envelope f_c is well and uniquely defined and is equivalent to the *convex biconjugate* function f^{**} of f.

The ℓ^1 Norm as Envelope of ℓ^0 Norm

$$\forall x \in \mathbb{R}^n : \qquad \|x\|_0 = \sum_{i=1}^{n} \mathbb{1}_{x_i \neq 0}, \qquad \|x\|_1 = \sum_{i=1}^{n} |x_i| \tag{6}$$

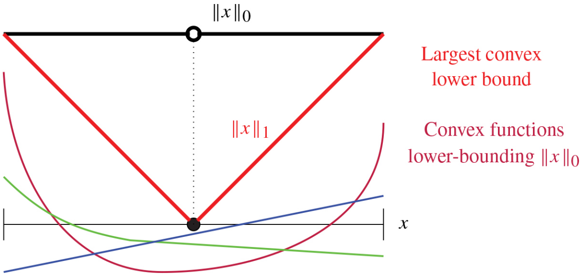


Figure: Convex surrogates for the ℓ^0 norm. $|x|$ is the convex envelope of $\|x\|_0$ on $[-1, 1]$.

The ℓ^1 Norm as Envelope of ℓ^0 Norm

Theorem

The function $\|.\|_1$ is the convex envelope of $\|.\|_0$, over the set $B_\infty = \{x | \|x\|_\infty \leqslant 1\}$ of vectors whose elements all have magnitude at most one.

Proof. Consider the cube $C = [0,1]^n$ with vertex vectors $\sigma \in \{0,1\}^n$ (each i-th component of σ is whether 0 or 1). For any convex function $g \leqslant \|.\|_0$ and since $\forall x \in C : x = \sum_i \lambda_i \sigma_i$ with $\lambda_i \geqslant 0$ $\forall i$ and $\sum_i \lambda_i = 1$,

$$\begin{aligned} g(x) &= g(\sum_i \lambda_i \sigma_i) \leqslant \sum_i \lambda_i g(\sigma_i) && \text{[Jensen's inequality]} \\ &\leqslant \sum_i \lambda_i \|\sigma_i\|_0 = \sum_i \lambda_i \|\sigma_i\|_1 && [\sigma_i \text{ are binary}] \\ &= \|x\|_1 \leqslant \|x\|_0 \end{aligned} \tag{7}$$

Repeat the argument for each orthant.

Sparsity Promoting Property of Norms

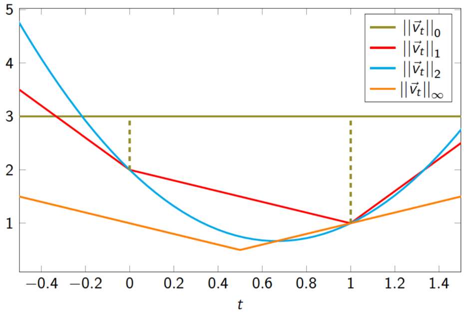
A Toy Problem: given a vector

$$\vec{v}(t) = [t, t-1, t-1] \in \mathbb{R}^3, \quad (8)$$

find t such that $\vec{v}$ is sparse.

Strategy: given a certain norm $\|.\|$,

$$\min_t f(t) = \|\vec{v}(t)\|.$$



Simple Algorithm for ℓ^1 Minimization

Minimizing the ℓ^1 Norm

Replace ℓ^0 minimization:

$$\min_x \|x\|_0 \text{ subject to } Ax = y \tag{9}$$

with the relaxed ℓ^1 minimization:

$$\min_x \|x\|_1 \text{ subject to } Ax = y \tag{10}$$

Two technical difficulties:

1. **Nontrivial constraints**: In the problem (10), the solution must satisfy $Ax = y$.
2. **Nondifferentiable objective**: ℓ^1 norm in problem (10) is not differentiable. So around points of interest the gradient $\nabla f(x)$ does not exist.

ℓ^1 Minimization via Linear Programming

Recall problem (10):

$$\min_x \|x\|_1 \text{ subject to } Ax = y$$

Let $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$, and let $z = \begin{pmatrix} x^+ \\ x^- \end{pmatrix} \in \mathbb{R}^{2n}$ and we have:

$$\|x\|_1 = e^T(x^+ + x^-) = e^T z \qquad \text{and } Ax = [A \ \ -A]z$$

with e a all-ones column vector of size $2n$. The the ℓ^1 minimization is equivalent to an LP problem:

$$\min_z e^T z \text{ subject to } [A \ \ -A]z = y, z \geqslant 0$$

This LP problem can be solved in polynomial time (with IPM).

Minimizing the ℓ^1 Norm via Local Descent

For minimizing a function with **constraints**:

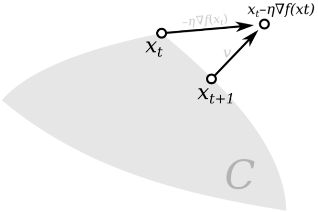
$$\min_x f(x), \qquad x \in \mathcal{C} \text{ (a convex set)}, \tag{11}$$

Basic Strategy: projected gradient descent (PGD):

$$x_{k+1} \leftarrow \mathcal{P}_{\mathcal{C}}[x_k - \eta \nabla f(x_k)] \tag{12}$$

where $\mathcal{P}_{\mathcal{C}}[.]$ projects a point, say z, to the nearest point in $\mathcal{C}$:

$$\mathcal{P}_{\mathcal{C}}[z] = \text{argmin}_{x \in \mathcal{C}} \|z - x\|_2^2 = h(x). \tag{13}$$



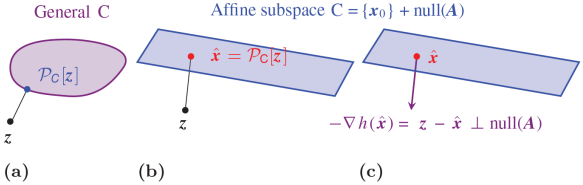
For general $\mathcal{C}$, the projection may not exist, or may not be unique (how ?).
However: for closed convex sets, the projection is well defined and satisfies a wealth of useful properties.

Project onto an affine subspace: $\mathcal{C} = \{x|Ax = y\}$

How to find the nearest point $\hat{x} = \mathcal{P}_{\mathcal{C}}[.]$ to a point z.

In this special case, and if A has full row rank, $\hat{x}$ satisfies two conditions:

1. Feasibility: $\hat{x} \in \mathcal{C}$, i.e., $A\hat{x} = y$.
2. Optimality: $z - \hat{x} \perp null(A)$



From these conditions, we have:

$$\hat{x} = \mathcal{P}_{x|Ax=y}[z] = z - A^H(AA^H)^{-1}[Az - y]$$

Directly check? Or derive alternatively? (use KKT conditions)

Minimizing ℓ^1 Norm: Nondifferentiability

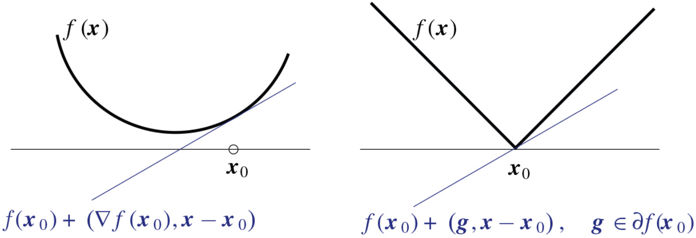
Try to solve:

$$\min_x \|x\|_1 \qquad \text{subject to } Ax = y \tag{14}$$

using projected gradient descent:

$$\min_x f(x) : \qquad x_{k+1} \leftarrow \mathcal{P}_{\mathcal{C}}[x_k - \eta \nabla f(x_k)] \tag{15}$$

But $\|x\|_1$ is non differentiable.



(a) differentiable (b) nondifferentiable

Design Strategies for All Local Descent Methods

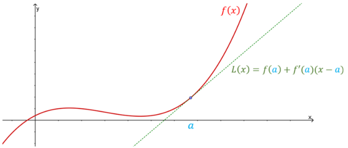
Minimization via local descent:

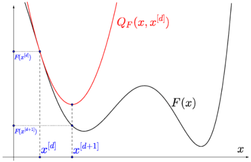
$$\begin{aligned} \min_x f(x): \qquad & x_k \rightarrow x_{k+1} \\ \text{such that} \qquad & f(x_k) \geqslant f(x_{k+1}) \end{aligned} \tag{16}$$

At current iterate x_k, find **a local surrogate** $\hat{f}(x; x_k) \approx f(x)$ such that

$$x_{k+1} := \text{argmin}_{x \in \mathcal{C}} \hat{f}(x; x_k) \text{ easy to find !} \tag{17}$$

where $\hat{f}(x; x_k)$ could be linear, quadratic, higher-order; or upper-bound (conservative) or lower-bound (accelerating).





Subgradient and Subdifferential

Generalizing the gradient $\nabla f(x)$ at x_0 with the property[1]:

$$f(x) \geqslant f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle, \qquad \forall x \in \mathbb{R}^n \tag{18}$$

Definition: Subgradient and Subdifferential

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a **convex** function. A *subgradient* of f at x_0 is any vector $u \in \mathbb{R}^n$ satisfying

$$f(x) \geqslant f(x_0) + \langle u, x - x_0 \rangle, \qquad \forall x \tag{19}$$

The *subdifferential* of f at x_0 is the set of all subgradients of f at x_0:

$$\partial f(x_0) = \{u | \forall x \in \mathbb{R}^n, f(x) \geqslant f(x_0) + \langle u, x - x_0 \rangle\}. \tag{20}$$

[1] for convex functions !

Sugradient method

With these definitions in mind:

- we might imagine that in the non-smooth case, a suitable replacement for the gradient descent algorithm might be the *subgradient method*.
- which choose (somehow) $g_k \in \partial f(x_k)$,
- and then proceeds in the direction of $-g_k$

$$x_{k+1} \leftarrow x_k - \alpha_k g_k$$

- Recall that we will need to incorporate the projection onto the feasible set $\mathcal{C}$.
- In any case, we need an expression for the subdifferential of the ℓ^1 Norm.

Subgradient and Subdifferential of ℓ^1 Norm

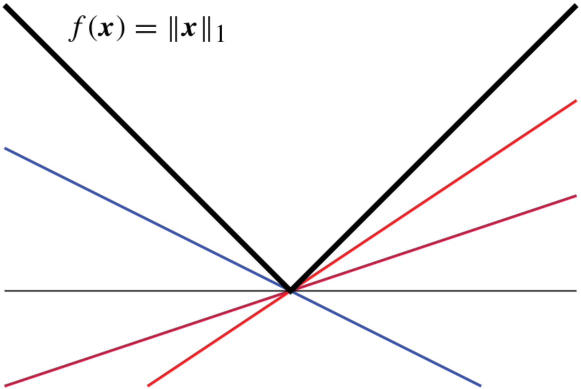


Figure: In blue, purple, and red, three linear lower bounds $g(x) := f(x_0) + \langle u, x - x_0 \rangle$, taken at $x_0 = 0$, with slope $u = -\frac{1}{2}$, $\frac{1}{3}$, and $\frac{2}{3}$, respectively. Any slope $u \in [-1, 1]$ defines a linear lower bound on $f(x)$ around $x_0 = 0$. So, $\partial|.|(0) = [-1, 1]$. For $x_0 > 0$, the only linear lower bound has slope $u = 1$; for $x_0 < 0$, the only linear lower bound has slope $u = 1$. So, $\partial|.|(x) = \{-1\}$ for $x < 0$ and $\partial|.|(x) = \{1\}$ for $x > 0$.

Subgradient and Subdifferential of ℓ^1 Norm

The following lemma extends this observation to higher-dimensional $x \in \mathbb{R}^n$

Lemma: Subdifferential of $\|.\|_1$

Let $x \in \mathbb{R}^n$, with $I = \mathrm{supp}(x)$,

$$\partial\|.\|_1(x) = \{v \in \mathbb{R}^n | P_I v = \mathrm{sign}(x), \|v\|_\infty \leqslant 1\}. \tag{21}$$

Here, $P_I \in \mathbb{R}^{n\times n}$ is the orthogonal projector onto coordinates I:

$$[P_I v]_j = \begin{cases} v_j & j \in I \\ 0 & j \notin I \end{cases} \tag{22}$$

Proof: on request.

Minimizing the ℓ^1 Norm: Projected Subgradient

To solve:

$$\min_x \|x\|_1 \qquad \text{subject to } Ax = y \tag{23}$$

using projected subgradient descent:

$$x_{k+1} \leftarrow \mathcal{P}_{\mathcal{C}}[x_k - \alpha_k g_k], \quad g_k \in \partial f(x_k). \tag{24}$$

Algorithm (ℓ^1 Minimization via Projected Subgradient Descent):

1: **Input:** a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and a vector $\boldsymbol{y} \in \mathbb{R}^m$.
2: Compute $\boldsymbol{\Gamma} \leftarrow \boldsymbol{I} - \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{A}$, and $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{A}^\dagger \boldsymbol{y} = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{y}$.
3: $\boldsymbol{x}_0 \leftarrow \boldsymbol{0}$.
4: $t \leftarrow 0$.
5: **repeat many times**
6: $\quad t \leftarrow t + 1$;
7: $\quad \boldsymbol{x}_t \leftarrow \tilde{\boldsymbol{x}} + \boldsymbol{\Gamma}\left(\boldsymbol{x}_{t-1} - \frac{1}{t}\operatorname{sign}(\boldsymbol{x}_{t-1}) \right)$;
8: **end while**

Minimizing the ℓ^1 Norm: Projected Subgradient

Remarks

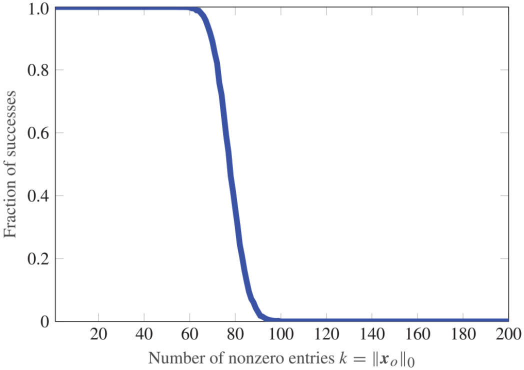
- In may aspects, this is a bad method for solving the ℓ^1 problem.
- It is correct but it converges very slowly compared to methods which exploit a certain piece of problem-specific structure, which we will describe later in the course.
- The main virtue of the Algorithm is that it is simple and intuitive,
- and also serves our exposition by introducing or reminding us of subgradients and projection operators :).

Minimizing the ℓ^1 Norm: Simulations

Solve:

$$\min_x \|x\|_1 \qquad \text{s.t. } Ax = y.$$

A is of size 200×400. Fraction of success across 50 trials.



Remark: although the method does not *always* succeed, it *does* succeed whenever the target solution x_o is *sufficiently* sparse.

Minimizing the ℓ^1 Norm: More Simulations

Solve:

$$\min_x \|x\|_1 \qquad \text{s.t. } Ax = y.$$

For the following settings:

1. Recover a sparse signal in the temporal domain using a random A matrix with independent and identically distributed (iid) Gaussian entries, with size $m \times n$ where $m \leqslant n$.
2. Recover a sparse signal in the frequency domain using $A = \Phi\Psi^{-1}$, and where Ψ is the Discrete Fourier Transform (DFT) matrix, and Φ is a matrix which select a subset of m rows of Ψ^{-1}.
3. Recover a sparse image using linear measurements with $A = \Phi\Psi^T$, and where Φ is a random matrix with iid Gaussian entries and Ψ is the Discrete Cosine Transform (DCT) matrix (see here and here for info).

Demo

Sparse Error Correction via ℓ^1 Minimization

Error Correction via ℓ^1 Minimization

- In the work of Benjamin Logan: shown that ℓ^1 minimization can be used to remove sparse errors in band limited signals
- We consider here a discretized analog of this result, in which we consider a finite-dimensional signal $y \in \mathbb{C}^n$.

Let $F \in \mathbb{C}^{n \times n}$ be the **Discrete Fourier Transform** (DFT) basis for $\mathbb{C}^n$, that is we have:

$$F_{kl} = \frac{1}{\sqrt{n}} \exp\left\{2\pi i \frac{kl}{n}\right\}, k = 0, ..., n-1, \; l = 0, ..., n-1 \tag{25}$$

Let $f_0, ..., f_{n-1}$ denote the columns of the DFT matrix[2]:

$$F = \left[f_0 | \quad \cdots \quad | f_{n-1}\right] \in \mathbb{C}^{n \times n} \tag{26}$$

[2]expression of a discrete Fourier transform (DFT) as a transformation matrix, which can be applied to a discrete signal through matrix multiplication.

Error Correction via ℓ^1 Minimization

Example of F: $n = 8$ (Eight-point)

$$F = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & \frac{1-i}{\sqrt{2}} & -i & \frac{-1-i}{\sqrt{2}} & -1 & \frac{-1+i}{\sqrt{2}} & i & \frac{1+i}{\sqrt{2}} \\ 1 & -i & -1 & i & 1 & -i & -1 & i \\ 1 & \frac{-1-i}{\sqrt{2}} & i & \frac{1-i}{\sqrt{2}} & -1 & \frac{1+i}{\sqrt{2}} & -i & \frac{-1+i}{\sqrt{2}} \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & \frac{-1+i}{\sqrt{2}} & -i & \frac{1+i}{\sqrt{2}} & -1 & \frac{1-i}{\sqrt{2}} & i & \frac{-1-i}{\sqrt{2}} \\ 1 & i & -1 & -i & 1 & i & -1 & -i \\ 1 & \frac{1+i}{\sqrt{2}} & i & \frac{-1+i}{\sqrt{2}} & -1 & \frac{-1-i}{\sqrt{2}} & -i & \frac{1-i}{\sqrt{2}} \end{bmatrix}$$

Error Correction via ℓ^1 Minimization

Let $B \in \mathbb{C}^{n\times(d+1))}$ be a submatrix of the d lowest-frequency elements of this basis and their conjugates:

$$B = \left[f_{-\frac{d-1}{2}} \middle| \quad \cdots \quad \middle| f_{\frac{d-1}{2}} \right] \quad \in \mathbb{C}^{n\times(d+1))}, \tag{27}$$

where we use f_{-i} to indicate the conjugate of f_i.
Let us imagine that $x_o = Bw_o \in \mathrm{col}(B)$, and

$$y = x_o + e_o, \qquad \text{where } \|e_o\|_0 \leqslant k \tag{28}$$

Our goal: recover x_o [3].
A discrete analog of the Logan's theorem would be to solve:

$$\min_x \|y - x\|_1 \qquad \text{s.t. } x \in \mathrm{col}(B). \tag{29}$$

[3] which is equivalent to removing e_o

Error Correction via ℓ^1 Minimization

- This problem is very much equivalent to the sparse signal recovery problem discussed so far.
- To see this: Let A be a matrix whose rows span the left null-space of B, i.e. $\text{rank}(A) = n - d$, and $AB = 0$,

Then $Ax_o = ABw_o = 0$, and

$$\bar{y} = Ay = A(x_o + e_0) = Ae_0 \tag{30}$$

To solve for e_o:

$$\min_{e} \|e\|_1 \qquad \text{s.t. } Ae = \bar{y} \tag{31}$$

Error Correction via ℓ^1 Minimization

According to Logan's Theorem, this succeeds if $d \times k \leqslant c\frac{\pi}{2}$.

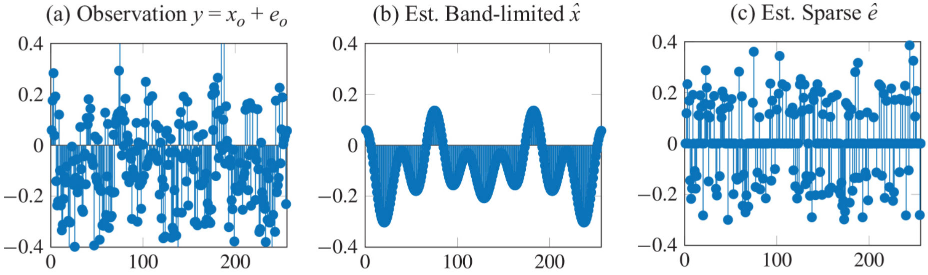


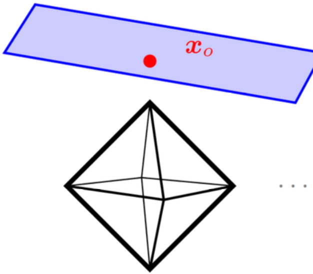
Figure: Logan's phenomenon. (a) The superposition $y = x_o + e_0$ of a band-limited signal x_o and a sparse error e_o. (b) Estimate $\hat{x}$ by ℓ^1 minimization. (c) Estimate $\hat{e}$ by ℓ^1 minimization. Both estimates are accurate to within relative error 10^{-6}.

Next: Towards a Rigorous Justification

Given $y = Ax_o$ with x_o sparse:

$$\begin{aligned} \mathbf{NP}: &\quad \min_x \|x\|_0 \qquad \text{subject to } Ax = y \\ \mathbf{P}: &\quad \min_x \|x\|_1 \qquad \text{subject to } Ax = y \end{aligned} \tag{32}$$

When and Why does ℓ^1 minimization work?



Summary

Summary

We have seen :

- Definition of *convex functions* and the crucial notion of *convex* envelope.
- The ℓ^1 norm as the convex envelope of ℓ^0 norm.
- A simple algorithm for ℓ^1 minimization, the *projected subgradient* algorithm.
- The example of *sparse error correction* via ℓ^1 minimization.

Goodbye, So Soon

THANKS FOR THE ATTENTION

- v.leplat@innopolis.ru
- sites.google.com/view/valentinleplat/