Part 1: Optimization and Applications

Mário A. T. Figueiredo[1] and Stephen J. Wright[2]

[1]Instituto de Telecomunicações,
Instituto Superior Técnico, Lisboa, Portugal

[2]Computer Sciences Department,
University of Wisconsin,
Madison, WI, USA

HIM, January, 2016

Inference via Optimization

Many **inference** problems are formulated as **optimization** problems:

- image reconstruction
- image restoration/denoising
- supervised learning
- unsupervised learning
- statistical inference
- ...

Standard formulation:

- observed data: y
- unknown mathematical object (signal, image, vector, matrix,...): x
- inference criterion:

$$\widehat{x} \in \arg\min_{x} g(x, y)$$

Inference via Optimization

Inference criterion:

$$\widehat{x} \in \arg\min_{x} g(x, y) = \{x : \; g(x,y) \leq g(z,y), \; \forall_z\}$$

Question 1: how to build g? Where does it come from?

Answer: from the application domain (machine learning, signal processing, inverse problems, system identification, statistics, computer vision, bioinformatics,...);
... examples ahead.

Question 2: how to solve the optimization problem?

Answer: the focus of this tutorial.

Regularized Optimization

Inference criterion: $\widehat{x} \in \arg\min_{x} g(x, y)$

Typical structure of g: $g(x, y) = h(x, y) + \tau\psi(x)$

- $h(x, y) \quad \rightarrow$ how well x "fits" / "explains" the data y;
  (data term, log-likelihood, loss function, observation model,...)
- $\psi(x) \quad \rightarrow$ knowledge/constraints/structure: the **regularizer**
- $\tau \geq 0$: the **regularization parameter** (or constant).
- Since y is fixed, we often write simply $f(x) = h(x, y)$,

$$\min_{x} f(x) + \tau\psi(x)$$

Probabilistic/Bayesian Interpretations

Inference criterion: $\widehat{x} \in \arg\min_{x} g(x, y)$

Typical structure of g: $g(x, y) = h(x, y) + \tau\psi(x)$

- Likelihood (observation model): $p(y|x) = \frac{1}{Z_l} \exp\big(-h(x, y)\big)$
- Prior: $p(x) = \frac{1}{Z_p} \exp\big(-\tau\psi(x)\big)$
- Posterior: $p(x|y) = \frac{p(y|x)\; p(x)}{p(y)}$
- Log-posterior: $\log\; p(x|y) = K(y) - h(x, y) - \tau\psi(x) = K(y) - g(x, y)$
- $\widehat{x}$ is a maximum a posteriori (MAP) estimate.

Regularizers

Inference criterion: $$\min_x \; f(x) + \tau\psi(x)$$

Typically, the unknown is a vector $x \in \mathbb{R}^n$
or a matrix $x \in \mathbb{R}^{n \times m}$

Common regularizers impose/encourage one (or a combination of) the following characteristics:

- small norm (vector or matrix)
- sparsity (few nonzeros)
- specific nonzero patterns (*e.g.*, group/tree structure)
- low-rank (matrix)
- smoothness or piece-wise smoothness

Unconstrained vs Constrained Formulations

- Tikhonov regularization: $\min_x f(x) + \tau\psi(x)$

- Morozov regularization: $$\begin{array}{ll} \min_x & \psi(x) \\ \text{subject to} & f(x) \le \varepsilon \end{array}$$

- Ivanov regularization: $$\begin{array}{ll} \min_x & f(x) \\ \text{subject to} & \psi(x) \le \delta \end{array}$$

Under mild conditions, these are all *"equivalent"*.

Morozov and Ivanov can be written as Tikhonov using indicator functions (more later).

Which one is more convenient is problem-dependent.

Example: Under- and Over-Constrained Systems

A simple linear inverse problem: from $y = A\,x$, find x $\quad (A \in \mathbb{R}^{m\times n})$

- Trivial case, A is invertible: $x = A^{-1}y$
- Over-determined system $(m > n)$; least squares solution $(\text{rank}(A) = n)$:

$$\widehat{x} = \arg\min_x \sum_{i=1}^{n} (y_i - (Ax)_i)^2 = \arg\min_x \|y - A\,x\|_2^2 = (A^T A)^{-1} A^T y$$

- Under-determined system $(m < n)$; minimum norm solution $(\text{rank}(A) = m)$:

$$\widehat{x} = \left\{ \begin{array}{l} \arg\min_x \|x\|_2^2 \\ \text{s.t. } Ax = y \end{array} \right\} = A^T (AA^T)^{-1} y$$

- Non-trivial cases: resort to optimization and regularization.
- Quadratic (Euclidean) losses and regularizers have a long and rich history: Gauss, Legendre, Wiener, Moore-Penrose, Tikhonov, ...

Norms: A Quick Review

Consider some real vector space $\mathcal{V}$, for example, $\mathbb{R}^n$ or $\mathbb{R}^{n\times n}$, ...

Some function $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ is a norm if it satisfies:

- $\|\alpha x\| = |\alpha|\ \|x\|$, for any $x \in \mathcal{V}$ and $\alpha \in \mathbb{R}$ (homogeneity);
- $\|x + x'\| \leq \|x\| + \|x'\|$, for any $x, x' \in \mathcal{V}$ (triangle inequality);
- $\|x\| = 0 \ \Rightarrow\ x = 0$.

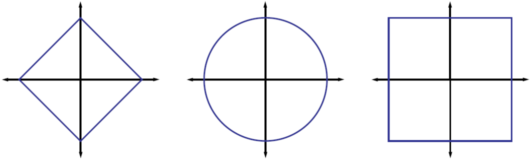
Examples:

- $\mathcal{V} = \mathbb{R}^n$, $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$ (called ℓ_p norm, for $p \geq 1$).
- $\mathcal{V} = \mathbb{R}^n$, $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p = \max\{|x_1|, ..., |x_n|\}$
- $\mathcal{V} = \mathbb{R}^{n\times n}$, $\|X\|_* = \operatorname{trace}\left(\sqrt{X^T X}\right)$ (matrix nuclear norm)

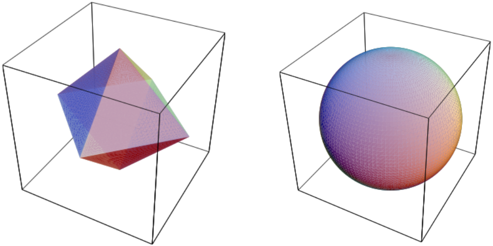
Also important (but not a norm): $\|x\|_0 = \lim_{p\to 0} \|x\|_p^p = |\{i : x_i \neq 0\}|$

Norm balls

Radius r ball in ℓ_p norm: $\qquad B_p(r) = \{x \in \mathbb{R}^n : \|x\|_p \le r\}$



$p = 1$ $\qquad p = 2$ $\qquad p = \infty$



$p = 1$ $\qquad p = 2$

Examples: Back to Under-Constrained Systems

A simple linear inverse problem: from $y = A\,x$, find x $\quad (A \in \mathbb{R}^{m\times n})$

- Under-determined system $(m < n)$; minimum norm solution:

$$\widehat{x} = \left\{ \begin{array}{l} \arg\min\limits_x \|x\|_2^2 \\ \text{s.t. } Ax = y \end{array} \right\} = A^*(AA^*)^{-1}y \neq x \text{ (in general)}$$

- Can we hope to recover x? Yes! ...if x is sparse enough $(\|x\|_0 < k)$ and A satisfies some **conditions**, using

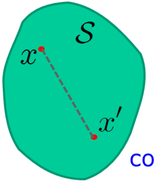
$$\widehat{x} = \begin{array}{l} \arg\min\limits_x \|x\|_0 \\ \text{s.t. } Ax = y \end{array}$$

Several proofs, under different conditions (more later).

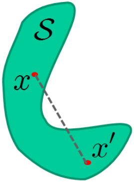
But, this is a hard problem! ℓ_0 "norm" is not **convex**.

Convex and strictly convex sets

$\mathcal{S}$ is convex if $x, x' \in \mathcal{S} \Rightarrow \forall \lambda \in [0,1], \;\; \lambda x + (1-\lambda) x' \in \mathcal{S}$

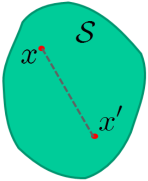


convex

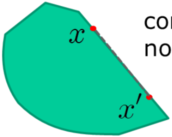


non-convex

$\mathcal{S}$ is strictly convex if $x, x' \in \mathcal{S} \Rightarrow \forall \lambda \in (0,1), \;\; \lambda x + (1-\lambda) x' \in \operatorname{int}(\mathcal{S})$



strictly convex



convex, but not strictly

Review of Basics: Convex Functions

Extended real valued function: $f : \mathbb{R}^N \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$

Domain: $\mathrm{dom}(f) = \{x : f(x) \neq +\infty\}$

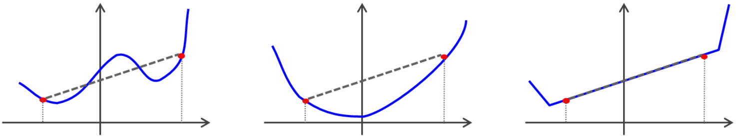
f is proper if $\mathrm{dom}(f) \neq \emptyset$

f is convex if

$$\forall \lambda \in [0,1], x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

f is strictly convex if

$$\forall \lambda \in (0,1), x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x')$$



non-convex strictly convex convex, not strictly

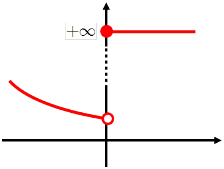
Lower Semi-Continuity: Why Is It Important?

A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is lower semi-continuous (l.s.c.) if

$$\liminf_{x \to x_0} f(x) \geq f(x_0), \quad \text{for any } x_0 \in \mathrm{dom}(f)$$

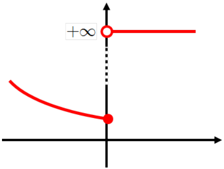
or, equivalently, $\{x : f(x) \leq \alpha\}$ is a closed set, for any $\alpha \in \mathbb{R}$

$$f(x) = \begin{cases} e^{-x}, & \text{if } x < 0 \\ +\infty, & \text{if } x \geq 0 \end{cases}$$



$\mathrm{dom}(f) =]-\infty, 0[, \quad \arg\min_x f(x) = \emptyset$

$$f(x) = \begin{cases} e^{-x}, & \text{if } x \leq 0 \\ +\infty, & \text{if } x > 0 \end{cases}$$



$\mathrm{dom}(f) =]-\infty, 0], \quad \arg\min_x f(x) = \{0\}$

Unless stated otherwise, we only consider l.s.c. functions.

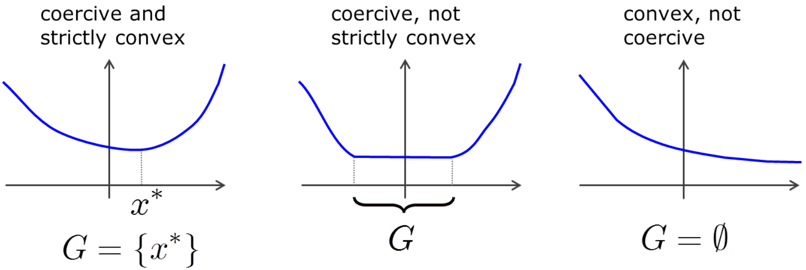
Coercivity, Convexity, and Minima

$$f : \mathbb{R}^N \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$$

f is coercive if $\displaystyle\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$

if f is coercive, then $G \equiv \arg\min_x f(x)$ is a non-empty set

if f is strictly convex, then G has at most one element



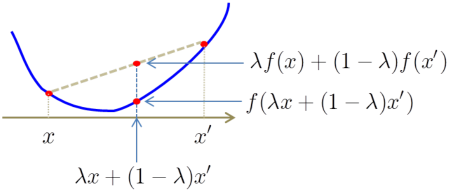
Another Important Concept: Strong Convexity

Recall the definition of convex function: $\forall \lambda \in [0,1]$,

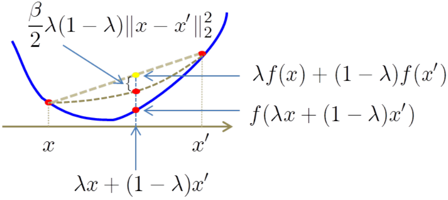
$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

A $\beta-$strongly convex function satisfies a stronger condition: $\forall \lambda \in [0,1]$

$$f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x') - \frac{\beta}{2}\lambda(1-\lambda)\|x - x'\|_2^2$$



convexity



strong convexity

Strong convexity $\underset{\not\Leftarrow}{\Rightarrow}$ strict convexity.

A Little More on Convex Functions

Let $f_1, ..., f_N : \mathbb{R}^n \to \bar{\mathbb{R}}$ be convex functions. Then

- $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, defined as $f(x) = \max\{f_1(x), ..., f_N(x)\}$, is convex.
- $g : \mathbb{R}^n \to \bar{\mathbb{R}}$, defined as $g(x) = f_1(L(x))$, where L is affine, is convex.
  Note: L is affine $\Leftrightarrow$ $L(x) - L(0)$ is linear; e.g. $L(x) = Ax + b$.
- $h : \mathbb{R}^n \to \bar{\mathbb{R}}$, defined as $h(x) = \sum_{j=1}^{N} \alpha_j f_j(x)$, for $\alpha_j > 0$, is convex.

An important function: the indicator of a set $C \subset \mathbb{R}^n$,

$$\iota_C : \mathbb{R}^n \to \bar{\mathbb{R}}, \;\; \iota_C(x) = \begin{cases} 0 & \Leftarrow \;\; x \in C \\ +\infty & \Leftarrow \;\; x \notin C \end{cases}$$

If C is a closed convex set, ι_C is a l.s.c. convex function.

The Case of Differentiable Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable and consider its Hessian matrix at x, denoted $\nabla^2 f(x)$ (or $Hf(x)$):

$$\left(\nabla^2 f(x)\right)_{ij} = \frac{\partial f}{\partial x_i \partial x_j}, \quad \text{for } i, j = 1, ..., n.$$

- f is convex $\Leftrightarrow$ its Hessian $\nabla^2 f(x)$ is positive semidefinite $\forall_x$
- f is strictly convex $\Leftarrow$ its Hessian $\nabla^2 f(x)$ is positive definite $\forall_x$
- f is β-strongly convex $\Leftrightarrow$ its Hessian $\nabla^2 f(x) \succeq \beta I$, with $\beta > 0$, $\forall_x$.

More on the Relationship Between ℓ_1 and ℓ_0

Finding the sparsest solution is NP-hard (Muthukrishnan, 2005).

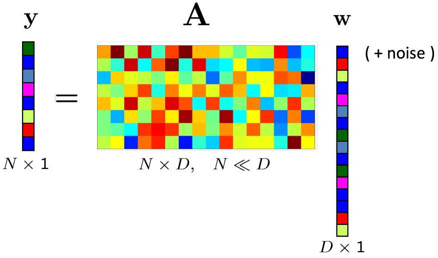
$$\begin{aligned} \widehat{w} \;&=\; \arg\min_{w} \|w\|_0 \\ &\quad \text{s.t. } \|Aw - y\|_2^2 \leq \delta. \end{aligned}$$

The related best subset selection problem is also NP-hard (Amaldi and Kann, 1998; Davis et al., 1997).

$$\begin{aligned} \widehat{w} \;&=\; \arg\min_{w} \|Aw - y\|_2^2 \\ &\quad \text{s.t. } \|w\|_0 \leq \tau. \end{aligned}$$

Under conditions, replacing ℓ_0 with ℓ_1 yields "similar" results: central issue in compressive sensing (CS) (Candès et al., 2006a; Donoho, 2006)

Compressive Sensing in a Nutshell



Even in the noiseless case, it seems impossible to recover w from y ...unless, w is sparse and A has some properties.

If w is sparse enough and A has certain properties, then w is stably recovered via (Haupt and Nowak, 2006)

$$\begin{aligned} \widehat{w} &= \arg\min_{w} \|w\|_0 \\ &\text{s. t. } \|Aw - y\| \leq \delta \end{aligned}$$

NP-hard!

Compressive Sensing in a Nutshell

Under some conditions on A (e.g., the restricted isometry property (RIP)), ℓ_0 can be replaced with ℓ_1 (Candès et al., 2006b):

$$\begin{aligned} \widehat{w} \;\; = \;\; & \arg\min_{w} \|w\|_1 \\ & \text{subject to } \|Aw - y\| \leq \delta \end{aligned}$$

convex problem

Matrix A satisfies the RIP of order k, with constant $\delta_k \in (0,1)$, if

$$\|w\|_0 \leq k \;\Rightarrow\; (1-\delta_k)\|w\|_2^2 \leq \|Aw\|_2^2 \leq (1+\delta_k)\|w\|_2^2$$

...i.e., for k-sparse vectors, A is approximately an isometry.

Other properties (spark and null space property (NSP)) can be used; caveat: checking RIP, NSP, spark is **NP-hard** (Tillmann and Pfetsch, 2012).

Examples: Back to Under-Constrained Systems

Let $\bar{x}$ be the sparsest solution of $Ax = y$, where $A \in \mathbb{R}^{m \times n}$ and $m < n$.

$$\bar{x} = \arg\min \|x\|_0 \;\text{ s.t. }\; Ax = y.$$

Consider the ℓ_1 norm version: $\quad \min_x \|x\|_1 \;\text{ s.t. }\; Ax = y$

Advantage: this is a convex problem! Fact: **all norms are convex**.

Of course, $\bar{x}$ solves this problem too, if $\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1, \;\; \forall v \in \ker(A)$.

Recall: $\ker(A) = \{x \in \mathbb{R}^n : \; Ax = 0\}$ is the kernel (a.k.a. null space) of A.

Next: elementary analysis by Yin and Zhang (2008), based on work by Kashin (1977) and Garnaev and Gluskin (1984).

Equivalence Between ℓ_1 and ℓ_0 Optimization

- Minimum ℓ_0 (sparsest) solution: $\bar{x} \in \arg\min \|x\|_0$ s.t. $Ax = y$.
- Minimum ℓ_1 solution(s): $G = \arg\min \|x\|_1$ s.t. $Ax = y$.
- $\bar{x} \in G$, if $\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1, \ \forall v \in \ker(A)$
- Let $S = \{i : \bar{x}_i \neq 0\}$ and $Z = \{1, ..., n\} \setminus S$

$$
\begin{aligned}
\|\bar{x} + v\|_1 &= \|\bar{x}_S + v_S\|_1 + \|v_Z\|_1 \\
&\geq \|\bar{x}_S\|_1 + \|v_Z\|_1 - \|v_S\|_1 && (\|a+b\| \geq \|a\| - \|b\|) \\
&= \|\bar{x}\|_1 + \|v\|_1 - 2\|v_S\|_1 \\
&\geq \|\bar{x}\|_1 + \|v\|_1 - 2\sqrt{k}\|v\|_2. && (\|a\|_1 \leq \sqrt{n}\,\|a\|_2)
\end{aligned}
$$

Hence, $\bar{x} \in G$, if $\frac{1}{2}\frac{\|v\|_1}{\|v\|_2} \geq \sqrt{k}, \ \forall v \in \ker(A)$

...but, in general, we have only: $1 \leq \frac{\|v\|_1}{\|v\|_2} \leq \sqrt{n}$

However, we may have $\frac{\|v\|_1}{\|v\|_2} \gg 1$, if v is restricted to a random subspace.

Bounding the ℓ_1/ℓ_2 Ratio in Random Matrices

If the elements of $A \in \mathbb{R}^{m\times n}$ are sampled i.i.d. from $\mathcal{N}(0,1)$ (zero mean, unit variance Gaussian), then, with high probability,

$$\frac{\|v\|_1}{\|v\|_2} \geq \frac{C\sqrt{m}}{\sqrt{\log(n/m)}}, \quad \text{for all } v \in \ker(A),$$

for some constant C (based on concentration of measure phenomena).

Thus, with high probability, $\bar{x} \in G$, if

$$m \geq \frac{4}{C^2} k \log n$$

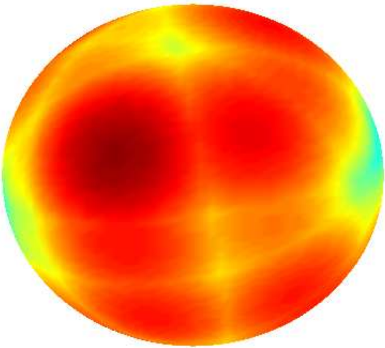
Conclusion: Can solve under-determined system, where A has i.i.d. $\mathcal{N}(0,1)$ elements, by solving

$$\min_x \|x\|_1 \;\; s.t. \;\; Ax = b,$$

(a convex problem), if the solution is sparse enough.

Ratio $\|v\|_1/\|v\|_2$ on Random Null Spaces

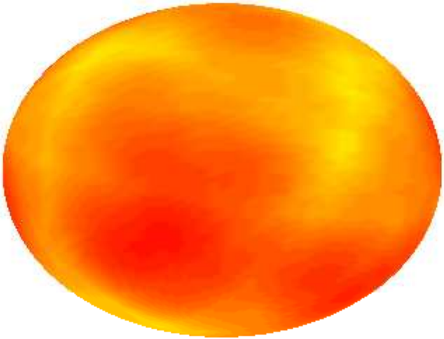
Random $A \in \mathbb{R}^{4\times 7}$, showing ratio $\|v\|_1$ for $v \in \ker(A)$ with $\|v\|_2 = 1$



Blue: $\|v\|_1 \approx 1$. Red: ratio $\approx \sqrt{7}$. Note that $\|v\|_1$ is well away from the lower bound of 1 over the whole nullspace.

Ratio $\|v\|_1/\|v\|_2$ on Random Null Spaces

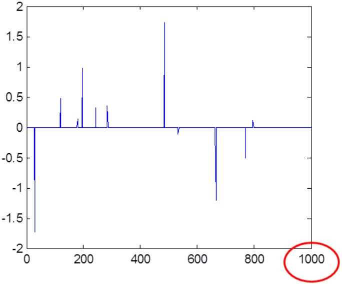
The effect grows more pronounced as m/n grows.
Random $A \in \mathbb{R}^{17\times 20}$, showing ratio $\|v\|_1$ for $v \in N(A)$ with $\|v\|_2 = 1$.



Blue: $\|v\|_1 \approx 1$. Red: $\|v\|_1 \approx \sqrt{20}$. Note that $\|v\|_1$ is closer to upper bound throughout.

When Data is Noisy

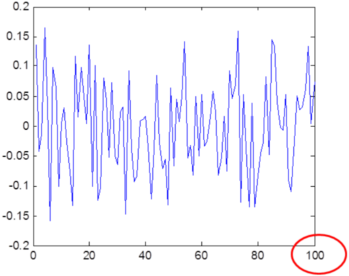
Sparse vector $\mathbf{x}$



$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$$

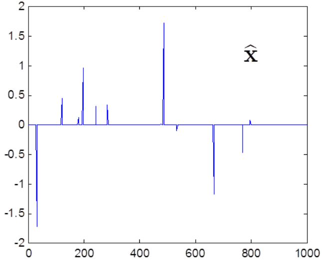
Random matrix

Observed data $\mathbf{y}$



Under certain conditions, "perfect" recovery is possible

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + 2\,\lambda\,\|\mathbf{x}\|_1 \right\}$$



[Candès, Romberg, Tao, 2004 – 2006]
[Donoho, 2006]

The Ubiquitous

- Lasso (*least ab* ...) (Tibshirani, 1996)
  a.k.a. *basis pu* ...

$$\min_x \frac{1}{2}\|A ... \quad \text{s.t.} \quad \|x\|_1 \le \delta$$

or, more gene ...

$$\min_x \; ... \; \|x\|_1 \le \delta$$

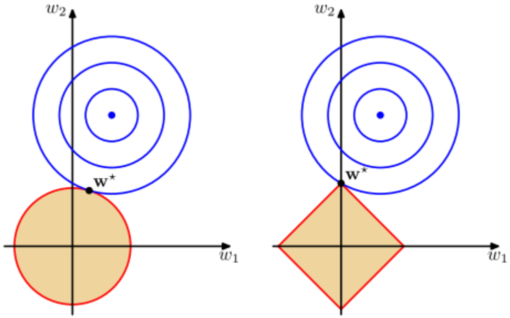
- Geology/geophysics
  - Claerbout and Muir (1973)
  - Taylor et al. (1979)
  - Levy and Fullager (1981)
  - Oldenburg et al. (1983)
  - Santosa and Symes (1988)
- Radio astronomy
  - Högbom (1974)
  - Schwarz (1978)
- Fourier transform spectroscopy
  - Kawata et al. (1983)
  - Mammone (1983)
  - Minami et al. (1985)
- NMR spectroscopy
  - Barkhuijsen (1985)
  - Newman (1988)
- Medical ultrasound
  - Papoulis and Chamzas (1979)

from (Goyal et al, 2010)

- Widely used o ... essive sensing
  (statistics, sig ...
- Many extensio ... arsity (more later).
- Why does ℓ_1 y ...
- How to solve these problems? (this tutorial)

Why ℓ_1 Yields Sparse Solution

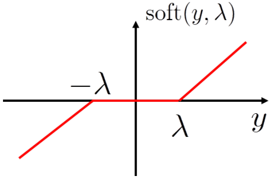
$$w^* = \arg\min_w \|Aw - y\|_2^2 \quad \text{s.t.} \quad \|w\|_2 \le \delta \qquad \text{vs} \qquad w^* = \arg\min_w \|Aw - y\|_2^2 \quad \text{s.t.} \quad \|w\|_1 \le \delta$$



Why ℓ_1 Yields Sparse Solution

The simplest problem with ℓ_1 regularization

$$\widehat{w} = \arg\min_w \frac{1}{2}(w-y)^2 + \lambda|w| = \text{soft}(y,\lambda) = \begin{cases} y-\lambda & \Leftarrow \quad y > \lambda \\ 0 & \Leftarrow \quad |y| \leq \lambda \\ y+\lambda & \Leftarrow \quad y < -\lambda \end{cases}$$



...by the way, how is this solved? (more later).

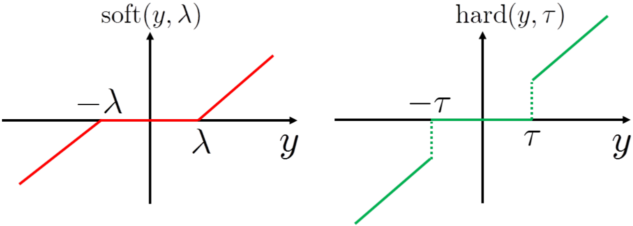
Contrast with the squared ℓ_2 (ridge) regularizer (linear scaling):

$$\widehat{w} = \arg\min_w \frac{1}{2}(w-y)^2 + \frac{\lambda}{2} w^2 = \frac{1}{1+\lambda}\, y$$

More on the Relationship Between ℓ_1 and ℓ_0

The ℓ_0 "norm" (number of non-zeros): $\|w\|_0 = |\{i : w_i \neq 0\}|$.
Not a norm, not convex, but in the simple case...

$$\widehat{w} = \arg\min_w \frac{1}{2}(w-y)^2 + \lambda |w|_0 = \text{hard}(y, \sqrt{2\lambda}) = \begin{cases} y & \Leftarrow \quad |y| > \sqrt{2\lambda} \\ 0 & \Leftarrow \quad |y| \leq \sqrt{2\lambda} \end{cases}$$



Another Application: Images

Natural images are well represented by a few coefficients in some bases.

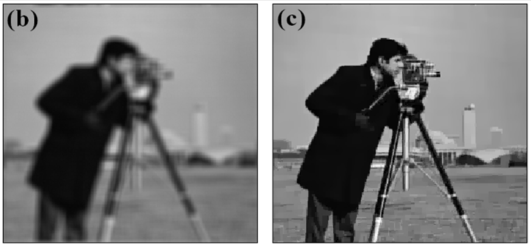
- Images ($N \times M \equiv n$ pixels) are represented by vectors $x \in \mathbb{R}^n$
- Typical images have representations $x = Ww$ that are sparse ($\|w\|_0 \ll n$) on some bases ($W^T W = W W^T = I$), such as wavelets.

Original 1000×1000 image $x \in \mathbb{R}^{10^6}$ ...only its 25000 largest coefficients.

- Also (even more) true with an over-complete tight frame; W is "fat" (more columns than rows) and $WW^T = I$, but $W^T W \neq I$.

Application to Image Deblurring/Deconvolution

blurred restored



$$\widehat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \tau\,\|\mathbf{x}\|_1$$

$$\mathbf{A} = \mathbf{B}\mathbf{W}$$

wavelet basis (or tight frame)

convolution (blur)

Application to Magnetic Resonance Imaging

$$\widehat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \tau\,\|\mathbf{x}\|_1$$

$$\mathbf{A} = \mathbf{M}\mathbf{U}\mathbf{W}$$

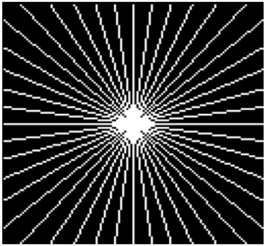
binary mask

wavelet basis (or tight frame)

discrete Fourier transform

original

acquired slices in DFT domain



reconstruction $\mathbf{W}\widehat{\mathbf{x}}$

Machine/Statistical Learning: Linear Regression

Data N pairs $(x_1, y_1), ..., (x_N, y_N)$, where $x_i \in \mathbb{R}^d$ (feature/variable vectors) and $y_i \in \mathbb{R}$ (outputs).

Goal: find "good" linear function: $\widehat{y} = \sum_{j=1}^{d} w_j x_j + w_{d+1} = [x^T \ 1]w$

Assumption: data generated i.i.d. by some underlying distribution $P_{X,Y}$

Mean squared error: $\min_w \mathbb{E}\big(Y - [X^T \ \mathbf{1}]w\big)^2$ impossible! $P_{X,Y}$ unknown

Empirical error: $\min_w \frac{1}{N} \sum_{i=1}^{N} \big(y_i - [x_i^T \ 1]w\big)^2 = \min_w \frac{1}{N} \|y - Aw\|_2^2,$

design matrix: $A_{ij} = (x_i)_j$ (j-th component of i-th sample), $A_{i(d+1)} = 1$

Regularization: $\min_w \|y - Aw\|_2^2 + \tau\psi(w)$

Machine/Statistical Learning: Linear Classification

Data N pairs $(x_1, y_1), ..., (x_N, y_N)$, where $x_i \in \mathbb{R}^d$ (feature vectors) and $y_i \in \{-1, +1\}$ (labels).

Goal: find "good" linear classifier (*i.e.*, find the optimal weights):

$$\widehat{y} = \text{sign}([x^T\ 1]w) = \text{sign}\Big(w_{d+1} + \sum_{j=1}^{d} w_j x_j\Big)$$

Assumption: data generated i.i.d. by some underlying distribution $P_{X,Y}$

Expected error: $\min_{w \in \mathbb{R}^{d+1}} \mathbb{E}\big(1_{Y([X^T\ \mathbf{1}]w)<0}\big)$ impossible! $P_{X,Y}$ unknown

Empirical error (EE): $\min_w \frac{1}{N} \sum_{i=1}^{N} h\big(\underbrace{y_i\ ([x^T\ 1]w)}_{\text{margin}}\big)$, where $h(z) = 1_{z<0}$.

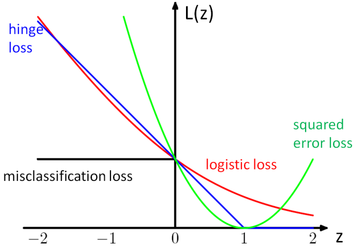
Convexification: EE neither convex nor differentiable (NP-hard problem).
Solution: replace $h : \mathbb{R} \to \{0, 1\}$ with convex loss $L : \mathbb{R} \to \mathbb{R}_+$.

Criterion: $$\underbrace{\min_w \sum_{i=1}^{N} L\big(\underbrace{y_i\,(w^T x_i + b)}_{\text{margin}}\big)}_{f(w)} + \tau\psi(w)$$

Regularizer: $\psi = \ell_1 \;\Rightarrow\;$ encourage sparseness $\;\Rightarrow$ feature selection

Convex losses: $L : \mathbb{R} \to \mathbb{R}_+$ is a (preferably convex) loss function.

- Misclassification loss: $L(z) = 1_{z<0}$
- Hinge loss: $L(z) = \max\{1 - z, 0\}$
- Logistic loss: $L(z) = \frac{\log\big(1+\exp(-z)\big)}{\log 2}$
- Squared loss: $L(z) = (z-1)^2$



Machine/Statistical Learning: General Formulation

This formulation covers a wide range of linear ML methods:

$$\min_w \underbrace{\sum_{i=1}^{N} L\big(y_i\ ([x^T\ 1]w)\big)}_{f(w)} + \tau\psi(w)$$

- Least squares regression: $L(z) = (z-1)^2,\ \ \psi(w) = 0.$
- Ridge regression: $L(z) = (z-1)^2,\ \ \psi(w) = \|w\|_2^2.$
- Lasso regression: $L(z) = (z-1)^2,\ \ \psi(w) = \|w\|_1$
- Logistic regression: $L(z) = \log(1+\exp(-z))$ (ridge, if $\psi(w) = \|w\|_2^2$
- Sparse logistic regression: $L(z) = \log(1+\exp(-z)),\ \ \psi(w) = \|w\|_1$
- Support vector machines: $L(z) = \max\{1-z, 0\},\ \ \psi(w) = \|w\|_2^2$
- Boosting: $L(z) = \exp(-z),$
- ...

Machine/Statistical Learning: Nonlinear Problems

What about non-linear functions?

Simply use $\widehat{y} = \phi(x, w) = \sum_{j=1}^{D} w_j \, \phi_j(x)$, where $\phi_j : \mathbb{R}^d \to \mathbb{R}$

Essentially, nothing changes; computationally, a lot may change!

$$\min_w \underbrace{\sum_{i=1}^{N} L\big(y_i \; \phi(x, w)\big)}_{f(w)} + \tau \psi(w)$$

Key feature: $\phi(x, w)$ is still linear with respect to w, thus f inherits the convexity of L.

Examples: polynomials, radial basis functions, wavelets, splines, kernels,...

Recover the linear case, letting $D = d + 1$, $f_j(x) = x_j$, and $f_{d+1} = 1$.

Structured Sparsity

ℓ_1 regularization promotes sparsity

A very simple sparsity pattern: prefer models with small cardinality

Can we promote less trivial sparsity patterns? How?

Group/structured regularization.

Structured Sparsity and Groups

Main goal: to promote structural patterns, not just penalize cardinality

Group sparsity: discard/keep entire *groups* of features (Bach et al., 2012)

- density inside each group
- sparsity with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*

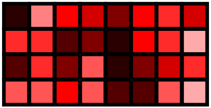
Yields statistical gains if the assumption is correct (Stojnic et al., 2009)

Many applications:

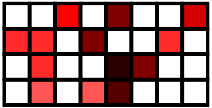
- feature template selection (Martins et al., 2011)
- multi-task learning (Caruana, 1997; Obozinski et al., 2010)
- learning the structure of graphical models (Schmidt and Murphy, 2010)

"Grid" Sparsity

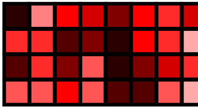
For feature spaces that can be arranged as a grid (examples next)



dense



sparse



dense

Goal: push *entire columns* to have zero weights

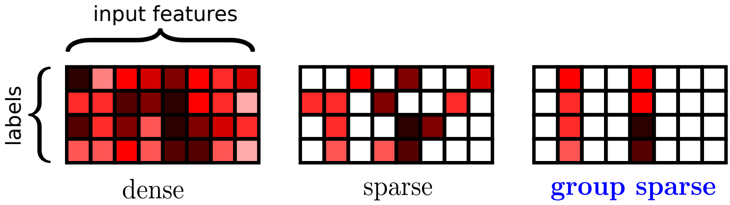
The groups are the columns of the grid

Example: Sparsity with Multiple Classes

In multi-class (more than just 2 classes) classification, a common formulation is

$$\widehat{y} = \arg \max_{y \in \{1,...,K\}} x^T w_y$$

Weight vector $w = (w_1, ..., w_K) \in \mathbb{R}^{Kd}$ has a natural group/grid organization:



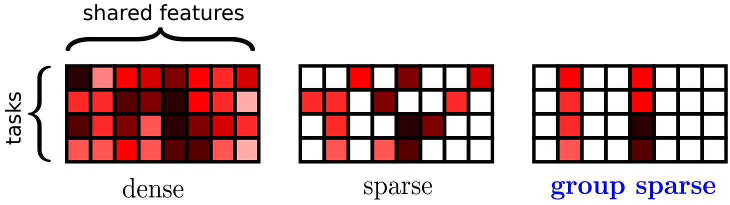
Simple sparsity is wasteful: may still need to keep all the features

Structured sparsity: discard some input features (feature selection)

Example: Multi-Task Learning

Same thing, except now rows are tasks and columns are **features**

Example: simultaneous regression (seek function into $\mathbb{R}^d \to \mathbb{R}^b$)

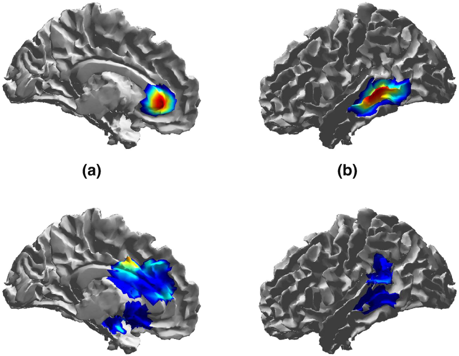


Goal: discard features that are irrelevant for *all* tasks

Approach: one group per feature (Caruana, 1997; Obozinski et al., 2010)

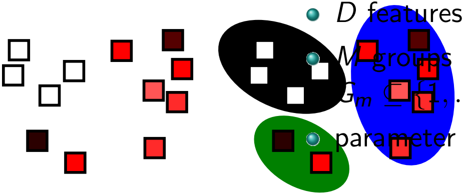
Example: Magnetoencephalograpy (MEG)

Group: localized cortex area at localized time period (Bolstad et al., 2009)



(a) (b) (c) (d)

Group Sparsity



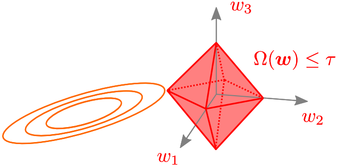
- D features
- M groups $G_1, \ldots, G_M$, each $G_m \subseteq \{1, \ldots, D\}$
- parameter subvectors $x_{G_1}, \ldots, x_{G_M}$

Group-Lasso (Bakin, 1999; Yuan and Lin, 2006):

$$\boxed{\psi(x) = \sum_{m=1}^{M} \lambda_m \|x_{G_m}\|_2}$$

- Intuitively: the ℓ_1 **norm** of the ℓ_2 **norms**
- Technically, still a norm (called a *mixed* norm, denoted $\ell_{2,1}$)
- Weighted version: λ_m are prior weights for groups (groups may have different sizes)

Lasso versus group-Lasso



$$\boxed{\Omega(\boldsymbol{w}) = |w_1| + |w_2| + |w_3|}$$

Composite Absolute Penalties (Zhao et al., 2009)

A mixed-norm regularization:

$$\boxed{\psi(x) = \left(\sum_{m=1}^{M} \|x_m\|_q^r \right)^{1/r}}$$

The r-norm of the q-norms ($r \geq 1, q \geq 1$)

Technically, this is also a norm, called a **mixed norm**, denoted $\ell_{q,r}$

- The most common choice: $\ell_{2,1}$ norm
- Another frequent choice: $\ell_{\infty,1}$ norm (Turlach et al., 2005; Quattoni et al., 2009; Graça et al., 2009; Eisenstein et al., 2011; Wright et al., 2009)

Three Scenarios

- Non-overlapping Groups
- Tree-structured Groups
- Graph-structured Groups

Non-overlapping Groups

Assume that $G_1, \ldots, G_M$ (where $G_m \subset \{1, ..., d\}$) constitute a partition:

$$\bigcup_{i=1}^{M} G_m = \{1, ..., d\} \quad \text{and} \quad i \neq j \;\Rightarrow\; G_i \cap G_j = \emptyset$$

$$\boxed{\psi(x) = \sum_{m=1}^{M} \lambda_m \|x_{G_m}\|_2}$$

Trivial choices of groups recover *unstructured* regularizers:

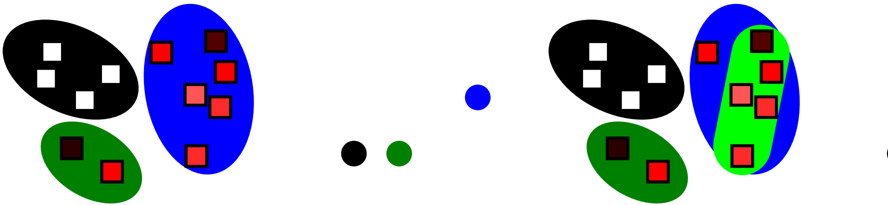
- ℓ_2-regularization: one large group $G_1 = \{1, \ldots, d\}$
- ℓ_1-regularization: d singleton groups $G_m = \{m\}$

Examples of non-trivial groups:

- label-based groups
- task-based groups

Tree-Structured Groups

Assumption: if two groups overlap, one is contained in the other
$\Rightarrow$ hierarchical structure (Kim and Xing, 2010; Mairal et al., 2010)



- What is the **sparsity pattern**?
- If a group is discarded, all its descendants are also discarded

Matrix Inference Problems

Sparsest solution:

- From $Bx = b \in \mathbb{R}^p$, find $x \in \mathbb{R}^n$ $(p < n)$.
- $\min_x \|x\|_0$ s.t. $Bx = b$
- Yields exact solution, under some conditions.

Lowest rank solution:

- From $\mathcal{B}(X) = b \in \mathbb{R}^p$, find $X \in \mathbb{R}^{m\times n}$ $(p < m\,n)$.
- $\min_X \operatorname{rank}(X)$ s.t. $\mathcal{B}(X) = b$
- Yields exact solution, under some conditions.

Both $NP-$hard (in general); the same is true of noisy versions:

$$\min_{X\in\mathbb{R}^{m\times n}} \operatorname{rank}(X) \text{ s.t. } \|\mathcal{B}(X) - b\|_2^2$$

Under some conditions, the same solution is obtained by replacing $\operatorname{rank}(X)$ by the nuclear norm $\|X\|_*$ (as any norm, it is convex) (Recht et al., 2010)

Matrix Nuclear Norm (and Other Norms)

- Also known as trace norm; the ℓ_1-type norm for matrices $X \in \mathbb{R}^{m \times n}$
- Definition: $\|X\|_* = \text{trace}\left(\sqrt{X^T X}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i,$

  the σ_i are the singular values of X.
- Particular case of Schatten q-norm: $\|X\|_q = \left(\sum_{i=1}^{\min\{m,n\}} (\sigma_i)^q \right)^{1/q}$.
- Two other notable Schatten norms:
  - Frobenius norm: $\|X\|_2 = \|X\|_F = \sqrt{\sum_{i=1}^{\min\{m,n\}} (\sigma_i)^2} = \sqrt{\sum_{i,i} X_{i,j}^2}$
  - Spectral norm: $\|X\|_\infty = \max\left\{\sigma_1, ..., \sigma_{\min\{m,n\}}\right\}$

Nuclear Norm Regularization

Tikhonov formulation: $\min_X \underbrace{\|\mathcal{B}(X) - b\|_2^2}_{f(X)} + \underbrace{\tau\|X\|_*}_{\tau\psi(X)}$

Linear observations: $\mathcal{B} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$, $\big(\mathcal{B}(X)\big)_i = \langle B_{(i)}, X\rangle$,

$$B_{(i)} \in \mathbb{R}^{m\times n}, \text{ and } \langle B, X\rangle = \sum_{ij} B_{ij}X_{ij} = \text{trace}(B^T X)$$

Particular case: matrix completion, each matrix $B_{(i)}$ has one 1 and is zero everywhere else.

Why does the nuclear norm favor low rank solutions? Let $Y = U\Lambda V^T$ be the singular value decomposition, where $\Lambda = \text{diag}\big(\sigma_1, ..., \sigma_{\min\{m,n\}}\big)$; then

$$\arg\min_X \frac{1}{2}\|Y - X\|_F^2 + \tau\|\Lambda\|_* = U \underbrace{\text{soft}(X, \tau)}_{\text{may yield zeros}} V^T$$

...singular value thresholding (Ma et al., 2011; Cai et al., 2010)

Another Matrix Inference Problem: Inverse Covariance

Consider n samples $y_1, ..., y_n \in \mathbb{R}^d$ of a Gaussian r.v. $Y \sim \mathcal{N}(\mu, C)$; the log-likelihood is

$$L(P) = \log p(y_1, ..., y_n | P) = \log\det(P) - \text{trace}(SP) + \text{constant}$$

where $S = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^T$ and $P = C^{-1}$ (inverse covariance).

Zeros in P reveal conditional independencies between components of Y:

$$P_{ij} = 0 \quad \Leftrightarrow \quad Y_i \perp\!\!\!\perp Y_j | \{Y_k,\ k \neq i, j\}$$

...exploited to infer (in)dependencies among Gaussian variables. Widely used in computational biology and neuroscience, social network analysis, ...

Sparsity (presence of zeros) in P is encouraged by solving

$$\min_{P \succ 0} \underbrace{-\log\det(P) + \text{trace}(SP)}_{f(P)} + \tau \underbrace{\|\text{vect}(P)\|_1}_{\psi(P)}$$

where $\text{vect}(P) = [P_{1,1}, ..., P_{d,d}]^T$.

Atomic-Norm Regularization

Key concept in sparse modeling: synthesize "object" using a few atoms:

$$x = \sum_{i=1}^{|\mathcal{A}|} c_i \, a_i$$

- $\mathcal{A}$ is the set of atoms (the atomic set), or building blocks.
- $c_i \geq 0$ are weights; x is simple/sparse object $\Rightarrow \quad \|c\|_0 \ll |\mathcal{A}|$
- Formally, $\mathcal{A}$ is a compact subset of $\mathbb{R}^n$

The (Minkowski) gauge of $\mathcal{A}$ is:

$$\|x\|_{\mathcal{A}} = \inf\big\{ t > 0 : \; x \in t \, \mathrm{conv}(\mathcal{A}) \big\}$$

Assuming that $\mathcal{A}$ centrally symmetry about the origin ($a \in \mathcal{A} \; \Rightarrow \; -a \in \mathcal{A}$), $\| \cdot \|_{\mathcal{A}}$ is a norm, called the atomic norm Chandrasekaran et al. (2012).

The atomic norm

$$\|x\|_{\mathcal{A}} = \inf\left\{t > 0 : \ x \in t\,\mathrm{conv}(\mathcal{A})\right\}$$
$$= \inf\left\{\sum_{i=1}^{|\mathcal{A}|} c_i : \ x = \sum_{i=1}^{|\mathcal{A}|} c_i\, a_i, \ c_i \geq 0\right\}$$

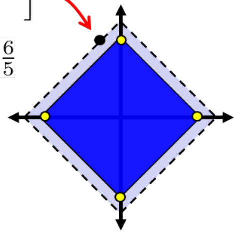
...assuming that the centroid of $\mathcal{A}$ is at the origin.

Example: the ℓ_1 norm as an atomic norm

- $\mathcal{A} = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right\}$
- $\mathrm{conv}(\mathcal{A}) = B_1(1)$ (ℓ_1 unit ball).
- $\|x\|_{\mathcal{A}} = \inf\left\{t > 0 : \ x \in t\, B_1(1)\right\}$
  $= \|x\|_1$

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

$\|x\|_{\mathcal{A}} = \frac{6}{5}$



Atomic Norms: More Examples

Examples with easy forms:

- *sparse vectors*

  $\mathcal{A} = \{\pm e_i\}_{i=1}^N$

  $\text{conv}(\mathcal{A}) = \text{cross-polytope}$

  $\|x\|_{\mathcal{A}} = \|x\|_1$

- *low-rank matrices*

  $\mathcal{A} = \{A : \text{rank}(A) = 1, \|A\|_F = 1\}$

  $\text{conv}(\mathcal{A}) = \text{nuclear norm ball}$

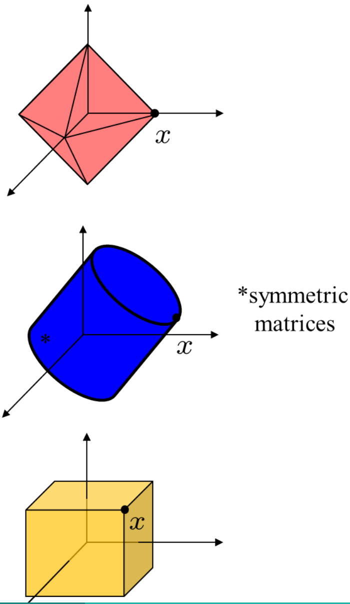
  $\|x\|_{\mathcal{A}} = \|x\|_\star$

- *binary vectors*

  $\mathcal{A} = \{\pm 1\}^N$

  $\text{conv}(\mathcal{A}) = \text{hypercube}$

  $\|x\|_{\mathcal{A}} = \|x\|_\infty$



Atomic-Norm Regularization

Given an atomic set $\mathcal{A}$, we can adopt an Ivanov formulation

$$\min f(x) \text{ s.t. } \|x\|_{\mathcal{A}} \leq \delta$$

(for some $\delta > 0$) tends to recover x with sparse atomic representation.

Can formulate algorithms for the various special cases — but is a general approach available for this formulation?

Yes! Conditional Gradient (a.k.a. Frank-Wolfe). More later!

Summary

- Many inference, learning, signal/image processing problems can be formulated as optimization problems.
- Sparsity-inducing regularizers play an important role in these problems
- There are several way to induce sparsity
- It is possible to formulate structured sparsity
- It is possible to extend the sparsity rationale to other objects, namely matrices
- Atomic norms provide a unified framework for sparsity/simplicity regularization

References I

Amaldi, E. and Kann, V. (1998). On the approximation of minimizing non zero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209:237–260.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statistical Science*, 27:450–468.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University.

Bolstad, A., Veen, B. V., and Nowak, R. (2009). Space-time event sparse penalization for magnetoelectroencephalography. *NeuroImage*, 46:1066–1081.

Cai, J.-F., Candès, E., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM JOurnal on Optimization*, 20:1956–1982.

Candès, E., Romberg, J., and Tao, T. (2006a). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509.

Candès, E., Romberg, J., and Tao, T. (2006b). Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics*, 59:1207–1223.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Chandrasekaran, V., Recht, B., Parrilo, P., and Willsky, A. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12:805–849.

References II

Chen, S., Donoho, D., and Saunders, M. (1995). Atomic decomposition by basis pursuit. Technical report, Department of Statistics, Stanford University.

Davis, G., Mallat, S., and Avellaneda, M. (1997). Greedy adaptive approximation. *Journal of Constructive Approximation*, 13:57–98.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306.

Eisenstein, J., Smith, N. A., and Xing, E. P. (2011). Discovering sociolinguistic associations with structured sparsity. In *Proc. of ACL*.

Garnaev, A. and Gluskin, E. (1984). The widths of an Euclidean ball. *Doklady Akademii Nauk*, 277:1048–1052.

Graça, J., Ganchev, K., Taskar, B., and Pereira, F. (2009). Posterior vs. parameter sparsity in latent variable models. *Advances in Neural Information Processing Systems*.

Haupt, J. and Nowak, R. (2006). Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52:4036–4048.

Kashin, B. (1977). Diameters of certain finite-dimensional sets in classes of smooth functions. *Izvestiya Akademii Nauk. SSSR: Seriya Matematicheskaya*, 41:334–351.

Kim, S. and Xing, E. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proc. of ICML*.

Ma, S., Goldfarb, D., and Chen, L. (2011). Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming (Series A)*, 128:321–353.

References III

Mairal, J., Jenatton, R., Obozinski, G., and Bach, F. (2010). Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*.

Martins, A. F. T., Smith, N. A., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2011). Structured Sparsity in Structured Prediction. In *Proc. of Empirical Methods for Natural Language Processing*.

Muthukrishnan, S. (2005). *Data Streams: Algorithms and Applications*. Now Publishers, Boston, MA.

Obozinski, G., Taskar, B., and Jordan, M. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

Quattoni, A., Carreras, X., Collins, M., and Darrell, T. (2009). An efficient projection for $l_{1,\infty}$ regularization. In *Proc. of ICML*.

Recht, B., Fazel, M., and Parrilo, P. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501.

Schmidt, M. and Murphy, K. (2010). Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proc. of AISTATS*.

Stojnic, M., Parvaresh, F., and Hassibi, B. (2009). On the reconstruction of block-sparse signals with an optimal number of measurements. *Signal Processing, IEEE Transactions on*, 57(8):3075–3085.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B.*, pages 267–288.

References IV

Tillmann, A. and Pfetsch, M. (2012). The computational complexity of RIP, NSP, and related concepts in compressed sensing. Technical report, arXiv/1205.2081.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Wright, S., Nowak, R., and Figueiredo, M. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57:2479–2493.

Yin, W. and Zhang, Y. (2008). Extracting salient features from less data via ℓ_1-minimization authors. *SIAG/OPT Views-and-News*, 19:11–19.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (B)*, 68(1):49.

Zhao, P., Rocha, G., and Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497.