Part 1: Optimization and Applications

Mário A. T. Figueiredo\textsuperscript{1} and Stephen J. Wright\textsuperscript{2}

\textsuperscript{1}Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal

\textsuperscript{2}Computer Sciences Department, University of Wisconsin, Madison, WI, USA

HIM, January, 2016
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:

$$\hat{x} \in \arg \min_x g(x, y)$$
Inference via Optimization

Inference criterion:

\[ \hat{x} \in \arg \min_{x} g(x, y) = \{ x : g(x, y) \leq g(z, y), \forall z \} \]

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: \( \hat{x} \in \arg \min_x g(x, y) \)

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

- \( h(x, y) \rightarrow \) how well \( x \) “fits” / “explains” the data \( y \); 
  (data term, log-likelihood, loss function, observation model,...)

- \( \psi(x) \rightarrow \) 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: \( \hat{x} \in \arg \min_x g(x, y) \)

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

- Likelihood (observation model): \( p(y|x) = \frac{1}{Z_l} \exp(-h(x, y)) \)
- Prior: \( p(x) = \frac{1}{Z_p} \exp(-\tau \psi(x)) \)
- Posterior: \( p(x|y) = \frac{p(y|x)p(x)}{p(y)} \)
- Log-posterior: \( \log p(x|y) = K(y) - h(x, y) - \tau \psi(x) = K(y) - g(x, y) \)

\( \hat{x} \) is a maximum a posteriori (MAP) estimate.
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:**
  \[ \min_x \psi(x) \]
  subject to \( f(x) \leq \varepsilon \)

- **Ivanov regularization:**
  \[ \min_x f(x) \]
  subject to \( \psi(x) \leq \delta \)

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 $y = Ax$, find $x$ $(A \in \mathbb{R}^{m \times n})$

- Trivial case, $A$ is invertible: $x = A^{-1}y$

- Over-determined system ($m > n$); least squares solution (rank($A$) = $n$):
  \[
  \hat{x} = \arg\min_x \sum_{i=1}^{n} (y_i - (Ax)_i)^2 = \arg\min_x \|y - Ax\|_2^2 = (A^T A)^{-1} A^T y
  \]

- Under-determined system ($m < n$); minimum norm solution (rank($A$) = $m$):
  \[
  \hat{x} = \begin{cases} 
  \arg\min_x \|x\|_2^2 \\
  \text{s.t. } Ax = y 
  \end{cases} = A^T (AA^T)^{-1} y
  \]

- 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 \( \ell_p \) norm, for \( p \geq 1 \)).
- \( \mathcal{V} = \mathbb{R}^n \), \( \| x \|_{\infty} = \lim_{p \to \infty} \| x \|_p = \max \{ |x_1|, \ldots, |x_n| \} \)
- \( \mathcal{V} = \mathbb{R}^{n \times n} \), \( \| X \|_* = \text{trace} \left( \sqrt{X^TX} \right) \) (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 $\ell_p$ norm:

$$B_p(r) = \{ x \in \mathbb{R}^n : \|x\|_p \leq r \}$$
Examples: Back to Under-Constrained Systems

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

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

\[
\hat{x} = \left\{ \begin{array}{c}
\arg \min_x \|x\|_2^2 \\
\text{s.t. } Ax = y
\end{array} \right\} = A^*(AA^*)^{-1}y \neq x \text{ (in general)}
\]

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

\[
\hat{x} = \arg \min_x \|x\|_0 \\
\text{s.t. } Ax = y
\]

Several proofs, under different conditions (more later).

But, this is a hard problem! \( \ell_0 \) “norm” is not convex.
**Convex and strictly convex sets**

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

\[ S \text{ is strictly convex if } \forall x, x' \in S \Rightarrow \forall \lambda \in (0, 1), \ \lambda x + (1 - \lambda)x' \in \text{int}(S) \]
Review of Basics: Convex Functions

Extended real valued function:  \( f : \mathbb{R}^N \rightarrow \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \)

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

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

\( f \) is convex if
\[
\forall \lambda \in [0, 1], x, x' \in \text{dom}(f) \quad f(\lambda x + (1 - \lambda) x') \leq \lambda f(x) + (1 - \lambda) f(x')
\]

\( f \) is strictly convex if
\[
\forall \lambda \in (0, 1), x, x' \in \text{dom}(f) \quad f(\lambda x + (1 - \lambda) x') < \lambda f(x) + (1 - \lambda) f(x')
\]

non-convex

strictly convex

convex, not strictly
A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is lower semi-continuous (l.s.c.) if

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

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

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

$\text{dom}(f) = ]-\infty, 0[$, $\text{arg min}_x f(x) = \emptyset$

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

$\text{dom}(f) = ]-\infty, 0]$, $\text{arg min}_x f(x) = \{0\}$

Unless stated otherwise, we only consider l.s.c. functions.
Coercivity, Convexity, and Minima

\[ f : \mathbb{R}^N \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \]

\( f \) is coercive if \( \lim_{\|x\| \to +\infty} f(x) = +\infty \)

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

\[ G = \{ x^* \} \]

\[ G \]

\[ G = \emptyset \]
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: \( \forall \lambda \in [0, 1] \)

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

\[
\boxed{\text{convexity}} \quad \Rightarrow \quad \boxed{\text{strong convexity}} \quad \not\Rightarrow \quad \boxed{\text{strict convexity}}.
\]
A Little More on Convex Functions

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

- \( f : \mathbb{R}^n \to \bar{\mathbb{R}}, \) defined as \( f(x) = \max\{f_1(x), \ldots, f_N(x)\}, \) 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 \( \iff \) \( L(x) - L(0) \) 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 \( C \subset \mathbb{R}^n, \)

\[ \iota_C : \mathbb{R}^n \to \bar{\mathbb{R}}, \quad \iota_C(x) = \begin{cases} 0 & \iff x \in C \\ +\infty & \iff x \notin C \end{cases} \]

If \( C \) is a closed convex set, \( \iota_C \) is a l.s.c. convex function.
Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice differentiable and consider its Hessian matrix at $x$, denoted $\nabla^2 f(x)$ (or $Hf(x)$):

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

- $f$ is **convex** $\iff$ its Hessian $\nabla^2 f(x)$ is positive semidefinite $\forall x$.
- $f$ is **strictly convex** $\iff$ its Hessian $\nabla^2 f(x)$ is positive definite $\forall x$.
- $f$ is $\beta$-**strongly convex** $\iff$ its Hessian $\nabla^2 f(x) \succeq \beta I$, with $\beta > 0$, $\forall x$. 
More on the Relationship Between $\ell_1$ and $\ell_0$

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

$$\hat{w} = \arg \min_w \|w\|_0$$

s.t. $\|Aw - y\|_2^2 \leq \delta$.

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

$$\hat{w} = \arg \min_w \|Aw - y\|_2^2$$

s.t. $\|w\|_0 \leq \tau$.

Under conditions, replacing $\ell_0$ with $\ell_1$ yields “similar” results: central issue in compressive sensing (CS) (Candès et al., 2006a; Donoho, 2006)
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)

$$\hat{w} = \arg \min_w \|w\|_0$$

s. t. $\|Aw - y\| \leq \delta$  \hspace{1cm} \text{NP-hard!}$
Compressive Sensing in a Nutshell

Under some conditions on $A$ (e.g., the restricted isometry property (RIP)), $\ell_0$ can be replaced with $\ell_1$ (Candès et al., 2006b):

$$\hat{w} = \arg \min_w \|w\|_1$$
subject to $\|Aw - y\| \leq \delta$  

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 \textbf{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 $\ell_1$ norm version:

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

Advantage: this is a convex problem! Fact: \textbf{all norms are convex}.

Of course, $\bar{x}$ solves this problem too, if

$$\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1, \quad \forall v \in \ker(A).$$

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

Equivalence Between $\ell_1$ and $\ell_0$ Optimization

- Minimum $\ell_0$ (sparsest) solution: $\bar{x} \in \arg\min \|x\|_0$ s.t. $Ax = y$.
- Minimum $\ell_1$ 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 $Z = \{1, ..., n\} \setminus S$

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

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 $\ell_1/\ell_2$ 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

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

(a convex problem), if the solution is sparse enough.
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.
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 \( \mathbf{x} \)

\[
y = \mathbf{A} \mathbf{x} + \mathbf{n}
\]

Observed data \( \mathbf{y} \)

Random matrix

Under certain conditions, “perfect” recovery is possible

\[
\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \left\{ \| \mathbf{y} - \mathbf{A} \mathbf{x} \|^2 + 2 \lambda \| \mathbf{x} \|_1 \right\}
\]

[Candès, Romberg, Tao, 2004 – 2006]
[Donoho, 2006]
The Ubiquitous $\ell_1$ Norm

- Lasso (least absolute shrinkage and selection operator, a.k.a. basis pursuit denoising) (Tibshirani, 1996):
  \[
  \min_{x} \frac{1}{2} \|Ax - y\|_2^2 + \tau \|x\|_1 \\
  \text{or, more generally,}
  \min_{x} f(x) + \lambda \|x\|_1
  \text{s.t. } \|x\|_1 \leq \delta
  \]

Widely used outside and much earlier than compressive sensing (statistics, signal processing, ...).

Many extensions: namely to express structured sparsity (more later).

Why does $\ell_1$ yield sparse solutions? (next slides)

How to solve these problems? (this tutorial)

- Geological/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)
Why $\ell_1$ Yields Sparse Solution

\[ w^* = \arg \min_w \|Aw - y\|_2^2 \quad \text{s.t.} \quad \|w\|_2 \leq \delta \]

vs

\[ w^* = \arg \min_w \|Aw - y\|_2^2 \quad \text{s.t.} \quad \|w\|_1 \leq \delta \]
Why $\ell_1$ Yields Sparse Solution

The simplest problem with $\ell_1$ regularization

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

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

Contrast with the squared $\ell_2$ (ridge) regularizer (linear scaling):

$$\hat{w} = \arg \min_w \frac{1}{2}(w - y)^2 + \frac{\lambda}{2}w^2 = \frac{1}{1 + \lambda}y$$
More on the Relationship Between $\ell_1$ and $\ell_0$

The $\ell_0$ “norm” (number of non-zeros): $\|w\|_0 = |\{i : w_i \neq 0\}|$.

Not a norm, not convex, but in the simple case...

$$\hat{w} = \arg\min_w \frac{1}{2}(w - y)^2 + \lambda |w|_0 = \text{hard}(y, \sqrt{2\lambda}) = \begin{cases} y & \iff |y| > \sqrt{2\lambda} \\ 0 & \iff |y| \leq \sqrt{2\lambda} \end{cases}$$
Another Application: Images

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

- Images \((N \times M \equiv n \text{ pixels})\) 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 \((W^TW = WW^T = I)\), such as wavelets.

Original 1000 \(\times\) 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 \(W^TW \neq I\).
Application to Image Deblurring/Deconvolution

\[
\hat{x} \in \arg\min_x \frac{1}{2} \|Ax - y\|^2_2 + \tau \|x\|_1
\]

\[A = BW\]

convolution (blur)  wavelet basis (or tight frame)
\[ \hat{x} \in \arg\min_x \frac{1}{2} \|Ax - y\|_2^2 + \tau \|x\|_1 \]

\[ A = MUW \]

- binary mask
- wavelet basis (or tight frame)
- discrete Fourier transform

original
acquired slices in DFT domain
reconstruction \( W\hat{x} \)
Data \( N \) pairs \((x_1, y_1), \ldots, (x_N, y_N)\), where \( x_i \in \mathbb{R}^d \) (feature/variable vectors) and \( y_i \in \mathbb{R} \) (outputs).

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

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

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

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

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

Regularization: \( \min_w \|y - Aw\|_2^2 + \tau \psi(w) \)
Data \( N \) pairs \((x_1, y_1), \ldots, (x_N, y_N)\), 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):

\[
\hat{y} = \text{sign}( [x^T 1] w ) = \text{sign} \left( w_{d+1} + \sum_{j=1}^{d} w_j x_j \right)
\]

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

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

Empirical error (EE): \[
\min_{w} \frac{1}{N} \sum_{i=1}^{N} h\left( y_i \left( [x^T 1] w \right) \right), \text{ where } h(z) = 1_{z < 0}.
\]

Convexification: EE neither convex nor differentiable (NP-hard problem).

Solution: replace \( h : \mathbb{R} \to \{0, 1\} \) with convex loss \( L : \mathbb{R} \to \mathbb{R}_+ \).
Machine/Statistical Learning: Linear Classification

Criterion: \[
\min_w \sum_{i=1}^N L\left( y_i \left( w^T x_i + b \right) \right) + \tau \psi(w)
\]

Regularizer: \( \psi = \ell_1 \Rightarrow \) encourage sparseness \( \Rightarrow \) feature selection

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

- Misclassification loss: \( L(z) = 1_{z<0} \)
- Hinge loss: \( L(z) = \max\{1 - z, 0\} \)
- Logistic loss: \( L(z) = \frac{\log(1+\exp(-z))}{\log 2} \)
- Squared loss: \( L(z) = (z - 1)^2 \)
This formulation covers a wide range of linear ML methods:

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

- 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: \( L(z) = \exp(-z), \)
- ...
What about non-linear functions?

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

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

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

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

\( \ell_1 \) 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)

Goal: push *entire columns* to have zero weights

**The groups are the columns of the grid**
In multi-class (more than just 2 classes) classification, a common formulation is

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

Weight vector $w = (w_1, \ldots, w_K) \in \mathbb{R}^{Kd}$ 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 $\mathbb{R}^d \rightarrow \mathbb{R}^b$)

Goal: discard features that are irrelevant for all tasks

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

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

\begin{itemize}
  \item $D$ features
  \item $M$ groups $G_1, \ldots, G_M$, each $G_m \subseteq \{1, \ldots, D\}$
  \item Parameter subvectors $x_{G_1}, \ldots, x_{G_M}$
\end{itemize}

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

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

- Intuitively: the $\ell_1$ norm of the $\ell_2$ norms
- Technically, still a norm (called a *mixed* norm, denoted $\ell_{2,1}$)
- Weighted version: $\lambda_m$ are prior weights for groups (groups may have different sizes)
Lasso versus group-Lasso

\[ \Omega(w) = |w_1| + |w_2| + |w_3| \]

\[ \Omega(w) \leq \tau \]
A mixed-norm regularization:

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

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 $G_m \subset \{1, \ldots, d\}$) constitute a partition:

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

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

Trivial choices of groups recover unstructured regularizers:
- $\ell_2$-regularization: one large group $G_1 = \{1, \ldots, d\}$
- $\ell_1$-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
⇒ 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 $B(X) = b \in \mathbb{R}^p$, find $X \in \mathbb{R}^{m \times n}$ ($p < mn$).
- $\min_X \text{rank}(X)$ s.t. $B(X) = b$
- Yields exact solution, under some conditions.

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

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

Under some conditions, the same solution is obtained by replacing $\text{rank}(X)$ 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 $\ell_1$-type norm for matrices $X \in \mathbb{R}^{m \times n}$

- Definition: $\|X\|_* = \text{trace}(\sqrt{X^T X}) = \sum_{i=1}^{\min\{m,n\}} \sigma_i$, where the $\sigma_i$ 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: $\|X\|_\infty = \max \{ \sigma_1, \ldots, \sigma_{\min\{m,n\}} \}$
Tikhonov formulation: \[
\min_X \left\{ \left\| \mathcal{B}(X) - b \right\|_2^2 + \tau \| X \|_* \right\} \\
\underbrace{f(X)}_{\tau \mathcal{\psi}(X)}
\]

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

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

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 \( \Lambda = \text{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}}) \); then

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

...**singular value thresholding** (Ma et al., 2011; Cai et al., 2010)
Another Matrix Inference Problem: Inverse Covariance

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

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

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 \( Y \):

\[
P_{ij} = 0 \iff Y_i \perp \perp Y_j | \{Y_k, k \neq i, j\}
\]

...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 \( \text{vect}(P) = [P_{1,1}, \ldots, P_{d,d}]^T \).
Atomic-Norm Regularization

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

\[ x = \sum_{i=1}^{\vert A \vert} c_i a_i \]

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

The (Minkowski) gauge of \( A \) is:

\[ \|x\|_A = \inf \{ t > 0 : x \in t \text{ conv}(A) \} \]

Assuming that \( A \) centrally symmetry about the origin
(\( a \in A \Rightarrow -a \in A \)), \( \| \cdot \|_A \) is a norm, called the atomic norm
Chandrasekaran et al. (2012).
Atomic-Norm Regularization

The atomic norm

\[ \|x\|_A = \inf \left\{ t > 0 : \ x \in t \ \text{conv}(A) \right\} \]

\[ = \inf \left\{ \sum_{i=1}^{\|A\|} c_i : \ x = \sum_{i=1}^{\|A\|} c_i a_i, \ c_i \geq 0 \right\} \]

...assuming that the centroid of \( A \) is at the origin.

Example: the \( \ell_1 \) norm as an atomic norm

\( \mathcal{A} = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right\} \)

\( \text{conv}(\mathcal{A}) = B_1(1) \) (\( \ell_1 \) unit ball).

\[ \|x\|_A = \inf \left\{ t > 0 : \ x \in t \ B_1(1) \right\} \]

\[ = \|x\|_1 \]
Atomic Norms: More Examples

Examples with easy forms:

- **sparse vectors**
  \[ \mathcal{A} = \{ \pm e_i \}_{i=1}^N \]
  \[ \text{conv}(\mathcal{A}) = \text{cross-polytope} \]
  \[ \| x \|_A = \| x \|_1 \]

- **low-rank matrices**
  \[ \mathcal{A} = \{ A : \text{rank}(A) = 1, \| A \|_F = 1 \} \]
  \[ \text{conv}(\mathcal{A}) = \text{nuclear norm ball} \]
  \[ \| x \|_A = \| x \|_* \]

- **binary vectors**
  \[ \mathcal{A} = \{ \pm 1 \}^N \]
  \[ \text{conv}(\mathcal{A}) = \text{hypercube} \]
  \[ \| x \|_A = \| x \|_\infty \]
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!
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 ways 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 IV


