# First-Order Methods for Regularized Objectives

Stephen J. Wright[1]

[1] Computer Sciences Department,
University of Wisconsin-Madison.

IMA, August 2016

# Thanks

Mario Figueiredo.

# Statistical Inference via Optimization

Many problems in **statistical inference** can be formulated as **optimization** problems:

- image reconstruction
- image restoration / denoising
- supervised learning (regression / classification)
- unsupervised learning
- ...

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:

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

**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,...) together with statistical principles.

... examples ahead.

**Question 2:** how to solve the optimization problem?

**Answer:** We'll discuss in these sessions (and see also earlier sessions: Mahoney, Duchi, ...)

# Inference and 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, often drop it for convenience and write $f(x) = h(x, y)$,

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

# 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: $p(x) = \frac{1}{Z_p} \exp\big(-\tau\psi(x)\big)$
  - Gaussian: $\psi(x) = \|x\|^2$
  - Laplacian: $\psi(x) = \|x\|_1$.
- 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: $$\min_x \; \psi(x) \quad \text{subject to} \quad f(x) \le \varepsilon$$
- Ivanov regularization: $$\min_x \; f(x) \quad \text{subject to} \quad \psi(x) \le \delta$$

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

Morozov and Ivanov can be written as Tikhonov using indicator functions.

Which one is most convenient depends on the application and context.

# Relationship Between $\ell_1$ and $\ell_0$

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

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

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

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

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

# Under-Constrained Systems: Relating $\ell_0$ and $\ell_1$

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

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

Suppose that $\bar{x}$ has $k$ nonzero elements, with $k \ll n$.

Consider the $\ell_1$ norm version: $\min_x \|x\|_1 \ \text{ s.t. } \ Ax = y$

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

$\bar{x}$ will solve this problem too, provided that
$\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1, \ \forall v \in \ker(A)$.

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 $\ell_1$ and $\ell_0$

- 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 $S = \{i : \bar{x}_i \neq 0\}$ (support of $\bar{x}$ with cardinality $k \ll n$); and $S^c = \{1, ..., n\} \setminus S$

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

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 $\ell_1/\ell_2$ Ratio in Random Kernels

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

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

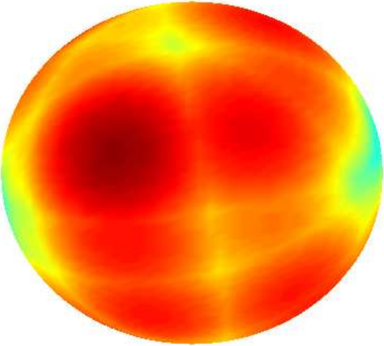
Conclusion: Can solve under-determined system, where $A$ has i.i.d. $\mathcal{N}(0,1)$ elements, by solving

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

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

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

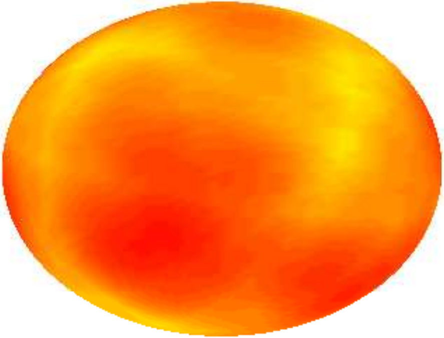
Random $A \in \mathbb{R}^{4\times 7}$, showing ratio $\|v\|_1$ for $v \in \ker(A)$ with $\|v\|_2 = 1$



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.

# Regularized Optimization

How to change these methods to handle regularized optimization?

$$\min_x \; f(x) + \lambda\psi(x),$$

where $f$ is convex and smooth, while $\psi$ is convex but usually **nonsmooth**.

Often, all that is needed is to change the update step to

$$x_{k+1} = \arg\min_x \frac{1}{2}\|x - \Phi(x_k)\|_2^2 + \alpha_k\lambda\psi(x). \tag{1}$$

where $\Phi(x_k)$ could be a steepest descent step

$$\Phi(x_k) = x_k - \alpha_k\nabla f(x_k),$$

or something more complicated (such as heavy ball, or some other accelerated method). When $\lambda = 0$, we have simply $x_{k+1} = \Phi(x_k)$, so this reverts to the standard first-order methods described above.

(1) is the shrinkage/tresholding step; how to solve it with a nonsmooth $\psi$? That's the topic of the following slides.

# Another Motivation

We can view shrinking / thresholding alternatively as a first-order subproblem with a quadratic prox term.

$$\begin{aligned} x_{k+1} &= \arg\min_x \frac{1}{2} \|x - (x_k - \alpha_k \nabla f(x_k))\|_2^2 + \alpha_k \lambda \psi(x) \\ &= \arg\min_x -\nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 + \lambda \psi(x), \end{aligned}$$

where we divided by $\alpha_k$ in the second expression and dropped the term that's independent of $x$.

This subproblem:

- makes a linear approximation to $f$ at $x_k$;
- incorporates a quadratic prox term with weight $1/\alpha_k$;
- incorporates the regularization term $\lambda\psi(x)$ explicitly, without modification.

This approach makes sense when the subproblem is easy to solve. This is true in a number of interesting cases.

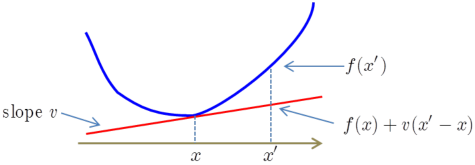
# Reminder: Subgradients

Subgradients generalize gradients for general convex functions:

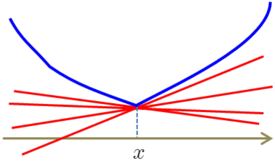
$$\boxed{v \text{ is a subgradient of } f \text{ at } x \text{ if } \; f(x') \geq f(x) + v^T(x' - x)}$$

Subdifferential: $\partial f(x) = \{\text{all subgradients of } f \text{ at } x\}$

If $f$ is differentiable, $\partial f(x) = \{\nabla f(x)\}$



linear lower bound



nondifferentiable case

Subgradients satisfy a monotonicity property: If $a \in \partial f(x)$ and $b \in \partial f(y)$, then $(a - b)^T(x - y) \geq 0$.

# A Key Tool: Moreau's Proximity Operators

Moreau (1962) proximity operator

$$\boxed{\widehat{x} \in \arg\min_{x} \frac{1}{2}\|x-y\|_2^2 + \psi(x) =: \mathrm{prox}_{\psi}(y)}$$

...well defined for convex $\psi$, since $\|\cdot - y\|_2^2$ is coercive and strictly convex.

Example: $\mathrm{prox}_{\tau|\cdot|}(y) = \mathrm{soft}(y,\tau) = \mathrm{sign}(y)\max\{|y|-\tau, 0\}$

Block separability: $x = (x_1, ..., x_N)$ (a partition of the components of $x$)

$$\psi(x) = \sum_i \psi_i(x_i) \;\Rightarrow\; (\mathrm{prox}_{\psi}(y))_i = \mathrm{prox}_{\psi_i}(y_i)$$

Relationship with subdifferential: $z = \mathrm{prox}_{\psi}(y) \;\Leftrightarrow\; z - y \in \partial\psi(z)$

Resolvent: $z = \mathrm{prox}_{\psi}(y) \;\Leftrightarrow\; 0 \in \partial\psi(z) + (z-y) \;\Leftrightarrow\; y \in (\partial\psi + I)z$

$$\mathrm{prox}_{\psi}(y) = (\partial\psi + I)^{-1} y$$

# Prox operators and the Moreau envelope

Moreau envelope:

$$M_{\lambda,\psi}(y) := \frac{1}{\lambda} \inf_x \left\{ \frac{1}{2}\|x-y\|_2^2 + \lambda\psi(x) \right\}$$

The minimizer in $M_{\lambda,\psi}(y)$ is achieved at $\text{prox}_{\lambda\psi}(y)$.

By optimality properties, we have

$$y - \text{prox}_{\lambda\psi}(y) \in \lambda\partial\psi(\text{prox}_{\lambda\psi}(y)).$$

$M_{\lambda,\psi}(y)$ can be viewed as a smoothing of $\psi$, differentiable everywhere:

$$\nabla M_{\lambda,\psi}(y) = \frac{1}{\lambda}(y - \text{prox}_{\lambda\psi}(y)).$$

By monotonicity of $\partial$, together with optimality condition above, can show that prox is a contraction, that is,

$$\|\text{prox}_{\lambda\psi}(y) - \text{prox}_{\lambda\psi}(z)\| \le \|y-z\|.$$

# Important Proximity Operators

- Soft-thresholding is the proximity operator of the $\ell_1$ norm.
- Consider the indicator $\iota_{\mathcal{S}}$ of a convex set $\mathcal{S}$;

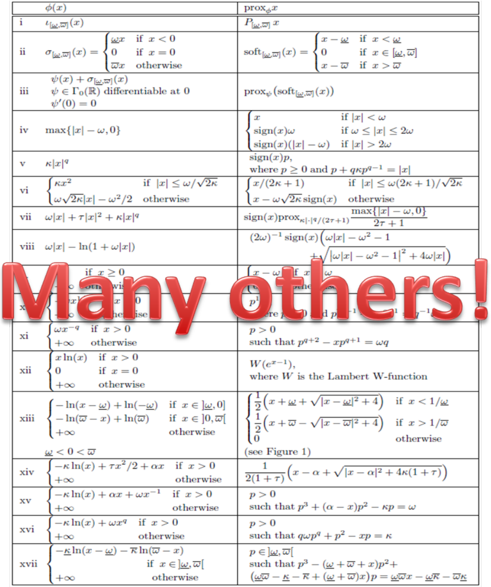
$$\text{prox}_{\iota_{\mathcal{S}}}(u) = \arg\min_{x} \frac{1}{2}\|x-u\|_2^2 + \iota_{\mathcal{S}}(x) = \arg\min_{x\in\mathcal{S}} \frac{1}{2}\|x-y\|_2^2 = P_{\mathcal{S}}(u)$$

...the Euclidean projection on $\mathcal{S}$.

- Squared Euclidean norm (separable, smooth): Exercise!
- Euclidean norm (not separable, nonsmooth):

$$\text{prox}_{\tau\|\cdot\|_2}(y) = \begin{cases} \frac{y}{\|y\|_2}(\|y\|_2 - \tau), & \text{if } \|y\|_2 > \tau \\ 0 & \text{if } \|y\|_2 \le \tau \end{cases}$$

# More Proximity Operators



| | $\phi(x)$ | $\mathrm{prox}_\phi x$ |
|---|---|---|
| i | $\iota_{[\underline{\omega},\varpi]}(x)$ | $P_{[\underline{\omega},\varpi]}\, x$ |
| ii | $\sigma_{[\underline{\omega},\varpi]}(x) = \begin{cases} \underline{\omega}x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \varpi x & \text{otherwise} \end{cases}$ | $\mathrm{soft}_{[\underline{\omega},\varpi]}(x) = \begin{cases} x - \underline{\omega} & \text{if } x < \underline{\omega} \\ 0 & \text{if } x \in [\underline{\omega},\varpi] \\ x - \varpi & \text{if } x > \varpi \end{cases}$ |
| iii | $\psi(x) + \sigma_{[\underline{\omega},\varpi]}(x)$<br>$\psi \in \Gamma_0(\mathbb{R})$ differentiable at $0$<br>$\psi'(0) = 0$ | $\mathrm{prox}_\psi\left(\mathrm{soft}_{[\underline{\omega},\varpi]}(x)\right)$ |
| iv | $\max\{\lvert x\rvert - \omega, 0\}$ | $\begin{cases} x & \text{if } \lvert x\rvert < \omega \\ \mathrm{sign}(x)\omega & \text{if } \omega \le \lvert x\rvert \le 2\omega \\ \mathrm{sign}(x)(\lvert x\rvert - \omega) & \text{if } \lvert x\rvert > 2\omega \end{cases}$ |
| v | $\kappa\lvert x\rvert^q$ | $\mathrm{sign}(x)p$,<br>where $p \ge 0$ and $p + q\kappa p^{q-1} = \lvert x\rvert$ |
| vi | $\begin{cases} \kappa x^2 & \text{if } \lvert x\rvert \le \omega/\sqrt{2\kappa} \\ \omega\sqrt{2\kappa}\lvert x\rvert - \omega^2/2 & \text{otherwise} \end{cases}$ | $\begin{cases} x/(2\kappa+1) & \text{if } \lvert x\rvert \le \omega(2\kappa+1)/\sqrt{2\kappa} \\ x - \omega\sqrt{2\kappa}\,\mathrm{sign}(x) & \text{otherwise} \end{cases}$ |
| vii | $\omega\lvert x\rvert + \tau\lvert x\rvert^2 + \kappa\lvert x\rvert^q$ | $\mathrm{sign}(x)\mathrm{prox}_{\kappa\lvert\cdot\rvert^q/(2\tau+1)} \dfrac{\max\{\lvert x\rvert - \omega, 0\}}{2\tau+1}$ |
| viii | $\omega\lvert x\rvert - \ln(1 + \omega\lvert x\rvert)$ | $(2\omega)^{-1}\,\mathrm{sign}(x)\Big(\omega\lvert x\rvert - \omega^2 - 1 + \sqrt{\left\lvert\omega\lvert x\rvert - \omega^2 - 1\right\rvert^2 + 4\omega\lvert x\rvert}\Big)$ |
| [illegible] | [illegible] if $x \ge 0$<br>$+\infty$ otherwise | $x - \omega$ if $x$ [illegible] $\omega$<br>[illegible] otherwise |
| x | [illegible] $x$ [illegible] $0$<br>$+\infty$ otherwise | $p^1$ [illegible]<br>where $p$ [illegible] $0$ and $p$ [illegible] |
| xi | $\begin{cases} \omega x^{-q} & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$ | $p > 0$<br>such that $p^{q+2} - xp^{q+1} = \omega q$ |
| xii | $\begin{cases} x\ln(x) & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ +\infty & \text{otherwise} \end{cases}$ | $W(e^{x-1})$,<br>where $W$ is the Lambert W-function |
| xiii | $\begin{cases} -\ln(x - \underline{\omega}) + \ln(-\underline{\omega}) & \text{if } x \in \,]\underline{\omega}, 0] \\ -\ln(\varpi - x) + \ln(\varpi) & \text{if } x \in \,]0, \varpi[ \\ +\infty & \text{otherwise} \end{cases}$<br>$\underline{\omega} < 0 < \varpi$ | $\begin{cases} \frac{1}{2}\left(x + \underline{\omega} + \sqrt{\lvert x - \underline{\omega}\rvert^2 + 4}\right) & \text{if } x < 1/\underline{\omega} \\ \frac{1}{2}\left(x + \varpi - \sqrt{\lvert x - \varpi\rvert^2 + 4}\right) & \text{if } x > 1/\varpi \\ 0 & \text{otherwise} \end{cases}$<br>(see Figure 1) |
| xiv | $\begin{cases} -\kappa\ln(x) + \tau x^2/2 + \alpha x & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$ | $\dfrac{1}{2(1+\tau)}\left(x - \alpha + \sqrt{\lvert x - \alpha\rvert^2 + 4\kappa(1+\tau)}\right)$ |
| xv | $\begin{cases} -\kappa\ln(x) + \alpha x + \omega x^{-1} & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$ | $p > 0$<br>such that $p^3 + (\alpha - x)p^2 - \kappa p = \omega$ |
| xvi | $\begin{cases} -\kappa\ln(x) + \omega x^q & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}$ | $p > 0$<br>such that $q\omega p^q + p^2 - xp = \kappa$ |
| xvii | $\begin{cases} -\underline{\kappa}\ln(x - \underline{\omega}) - \overline{\kappa}\ln(\varpi - x) & \text{if } x \in \,]\underline{\omega}, \varpi[ \\ +\infty & \text{otherwise} \end{cases}$ | $p \in \,]\underline{\omega}, \varpi[$<br>such that $p^3 - (\underline{\omega} + \varpi + x)p^2 +$<br>$(\underline{\omega}\varpi - \underline{\kappa} - \overline{\kappa} + (\underline{\omega} + \varpi)x)p = \underline{\omega}\varpi x - \underline{\omega}\overline{\kappa} - \varpi\underline{\kappa}$ |

Many others!

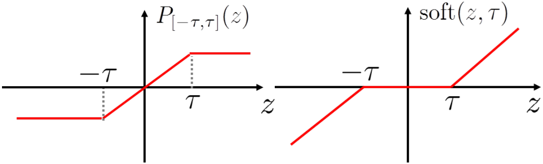
(Combettes and Pesquet, 2011)

# From Conjugates to Proximity Operators

Notice that $|u| = \sup_{x\in[-1,1]} x^T u = \sigma_{[-1,1]}(u)$, thus $|\cdot|^* = \iota_{[-1,1]}$.

Using Moreau's decomposition, we easily derive the soft-threshold:

$$\text{prox}_{\tau|\cdot|} = 1 - \text{prox}_{\iota_{[-\tau,\tau]}} = 1 - P_{[-\tau,\tau]} = \text{soft}(\cdot,\tau)$$



Conjugate of a norm: if $f(x) = \tau\|x\|_p$ then $f^* = \iota_{\{x:\|x\|_q\le\tau\}}$,

where $\frac{1}{q}+\frac{1}{p}=1$ (a Hölder pair, or Hölder conjugates).

That is, $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms:

$$\|z\|_q = \sup\{x^T z : \|x\|_p \le 1\} = \sup_{x\in B_p(1)} x^T z = \sigma_{B_p(1)}(z)$$

# From Conjugates to Proximity Operators

- Proximity of norm:

$$\boxed{\text{prox}_{\tau\|\cdot\|_p} = I - P_{B_q(\tau)}}$$

where $B_q(\tau) = \{x : \ \|x\|_q \leq \tau\}$ and $\frac{1}{q} + \frac{1}{p} = 1$.

- Example: computing $\text{prox}_{\|\cdot\|_\infty}$ (notice $\ell_\infty$ is not separable):

Since $\frac{1}{\infty} + \frac{1}{1} = 1$,

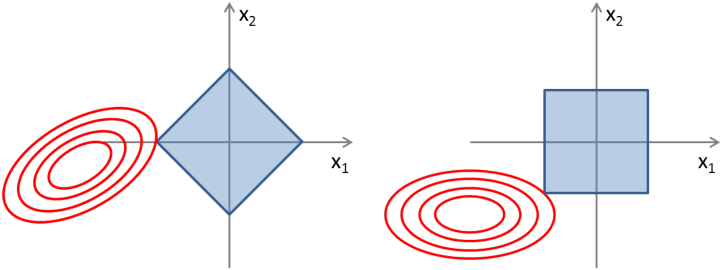
$$\text{prox}_{\tau\|\cdot\|_\infty} = I - P_{B_1(\tau)}$$

... the proximity operator of $\ell_\infty$ norm is the residual of the projection on an $\ell_1$ ball.

- Projection on $\ell_1$ ball has no closed form, but there are efficient (linear cost) algorithms (Brucker, 1984), (Maculan and de Paula, 1989).

# Geometry and Effect of $\mathrm{prox}_{\ell_\infty}$

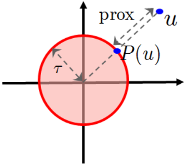
Whereas $\ell_1$ promotes sparsity, $\ell_\infty$ promotes equality (in absolute value).



# From Conjugates to Proximity Operators

The dual of the $\ell_2$ norm is the $\ell_2$ norm.

$$\mathrm{prox}_{\tau\|\cdot\|_2}(u) = u - P_{\{x:\|x\|_2\le\tau\}}(u)$$



$$= u - \begin{cases} u & \Leftarrow \quad \|u\|_2 \le \tau \\ \tau\, u/\|u\|_2 & \Leftarrow \quad \|u\|_2 > \tau \end{cases}$$

$$= \frac{u}{\|u\|_2}\max\{0, \|u\|_2 - \tau\}$$

vector soft thresholding

# Matrix Nuclear Norm and its Prox Operator

- Recall the trace/nuclear norm: $\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i.$
- The dual of a Schatten $p$-norm is a Schatten $q$-norm, with $\frac{1}{q} + \frac{1}{p} = 1$. Thus, the dual of the nuclear norm is the spectral norm:

$$\|X\|_\infty = \max\left\{\sigma_1, ..., \sigma_{\min\{m,n\}}\right\}.$$

- If $Y = U\Lambda V^T$ is the SVD of $Y$, we have

$$\begin{aligned}\mathrm{prox}_{\tau\|\cdot\|_*}(Y) &= U\Lambda V^T - P_{\{X:\max\{\sigma_1,...,\sigma_{\min\{m,n\}}\}\leq\tau\}}(U\Lambda V^T)\\ &= U\,\mathrm{soft}\big(\Lambda,\tau\big)\,V^T.\end{aligned}$$

# Another Use of Fenchel-Legendre Conjugates

- The original problem: $\min_x f(x) + \psi(x)$
- Often this has the form: $\min_x g(Ax) + \psi(x)$
- Using the definition of conjugate $g(Ax) = \sup_u u^T Ax - g^*(u)$

$$\begin{aligned}
\min_x g(Ax) + \psi(x) &= \inf_x \sup_u u^T Ax - g^*(u) + \psi(x) \\
&= \sup_u (-g^*(u)) + \inf_x \; u^T Ax + \psi(x) \\
&= \sup_u (-g^*(u)) - \underbrace{\sup_x \; -x^T A^T u - \psi(x)}_{\psi^*(-A^T u)} \\
&= -\inf_u g^*(u) + \psi^*(-A^T u)
\end{aligned}$$

- The dual $\inf_u g^*(u) + \psi^*(-A^T u)$ is sometimes easier to handle.

# Basic Proximal-Gradient Algorithm

Use basic structure:

$$x_k = \arg\min_x \|x - \Phi(x_k)\|_2^2 + \psi(x).$$

with $\Phi(x_k)$ a simple gradient descent step, thus

$$\boxed{x_{k+1} = \mathrm{prox}_{\alpha_k \psi}\big(x_k - \alpha_k \nabla f(x_k)\big)}$$

This approach goes by many names, such as

- "proximal gradient algorithm" (PGA),
- "iterative shrinkage/thresholding" (IST),
- "forward-backward splitting" (FBS)

It it has been reinvented several times in different communities: optimization, partial differential equations, convex analysis, signal processing, machine learning.

# Convergence of Prox-Gradient

$$x_{k+1} = \text{prox}_{\alpha_k \psi}(x_k - \alpha_k \nabla f(x_k)).$$

Proof makes use of "gradient map" defined by

$$G_\alpha(x) := \frac{1}{\alpha} \left( x - \text{prox}_{\alpha\psi}(x - \alpha \nabla f(x)) \right). \tag{2}$$

Can rewrite the step taken at iteration $k$:

$$x^{k+1} = x^k - \alpha_k G_{\alpha_k}(x^k) \quad \Leftrightarrow \quad G_{\alpha_k} = \frac{1}{\alpha_k}(x^k - x^{k+1}). \tag{3}$$

## Lemma

*Suppose that $\psi$ is closed convex function, $\nabla f$ has Lipschitz constant $L$.*

(a) $G_\alpha(x) \in \nabla f(x) + \partial\psi(x - \alpha G_\alpha(x))$.

(b) *For any $z$, and any $\alpha \in (0, 1/L]$, we have that*

$$\phi(x - \alpha G_\alpha(x)) \le \phi(z) + G_\alpha(x)^T(x - z) - \frac{\alpha}{2}\|G_\alpha(x)\|^2.$$

# Proof of (a)

From optimality property of the prox-operator, which is

$$y - \text{prox}_{\lambda\psi}(y) \in \lambda\partial\psi(\text{prox}_{\lambda\psi}(y)),$$

we have

$$(x - \alpha\nabla f(x)) - \text{prox}_{\alpha\psi}(x - \alpha\nabla f(x)) \in \alpha\partial\psi(\text{prox}_{\alpha\psi}(x - \alpha\nabla f(x))).$$

Now substitute from $\text{prox}_{\alpha\psi}(x - \alpha\nabla f(x)) = x - \alpha G_\alpha(x)$, to obtain

$$0 \in \alpha\partial\psi(x - \alpha G_\alpha(x)) - \alpha(G_\alpha(x) - \nabla f(x)),$$

from which (a) follows when we divide by $\alpha$.

# Proof of (b)

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2.$$

Set $y = x - \alpha G_\alpha(x)$, for any $\alpha \in (0, 1/L]$, to get

$$\begin{aligned} f(x - \alpha G_\alpha(x)) &\le f(x) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{L\alpha^2}{2}\|G_\alpha(x)\|^2 \\ &\le f(x) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2. \qquad (4) \end{aligned}$$

(Second inequality uses $\alpha \in (0, 1/L]$.) By convexity of $f$ and $\psi$, for any $z$ and any $v \in \partial\psi(x - \alpha G_\alpha(x))$ we have

$$f(z) \ge f(x) + \nabla f(x)^T (z - x) \qquad (5a)$$

$$\psi(z) \ge \psi(x - \alpha G_\alpha(x)) + v^T (z - (x - \alpha G_\alpha(x))). \qquad (5b)$$

From (a) we have $v = (G_\alpha(x) - \nabla f(x)) \in \partial\psi(x - \alpha G_\alpha(x))$, so by substituting in (5) and also using (4) we have the following...

# Proof of (b), continued

$$
\begin{aligned}
&\phi(x-\alpha G_\alpha(x)) \\
&= f(x-\alpha G_\alpha(x)) + \psi(x-\alpha G_\alpha(x)) \\
&\le f(x) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2 + \psi(x-\alpha G_\alpha(x)) \quad \text{(from (4))} \\
&\le f(z) + \nabla f(x)^T(x-z) - \alpha G_\alpha(x)^T \nabla f(x) + \frac{\alpha}{2}\|G_\alpha(x)\|^2 \\
&\qquad + \psi(z) + (G_\alpha(x) - \nabla f(x))^T(x - \alpha G_\alpha(x) - z) \quad \text{(from (5))} \\
&= f(z) + \psi(z) + G_\alpha(x)^T(x-z) - \frac{\alpha}{2}\|G_\alpha(x)\|^2,
\end{aligned}
$$

for any $\alpha \in (0, 1/L]$, where the last equality follows from cancellation of several terms in the previous line.

# Sublinear Convergence

Denote $\phi(x) = f(x) + \psi(x)$ with minimizer $x^*$ (not necessarily unique). Main convergence result:

### Theorem

*If* $\alpha_k \equiv 1/L$, *have*

$$\phi(x^T) - \phi^* \le \frac{L\|x^0 - x^*\|^2}{2T}, \quad T = 1, 2, \dots .$$

Use Lemma 1 (b) to show decrease of $\{\phi(x^k)\}$ and $\|x^k - x^*\|$. Set $x = z = x^k$ and $\alpha = \alpha_k$ and use (3) to obtain

$$\phi(x^{k+1}) = \phi(x^k - \alpha_k G_{\alpha_k}(x^k)) \le \phi(x^k) - \frac{\alpha_k}{2}\|G_{\alpha_k}(x^k)\|^2,$$

showing decrease in $\phi$.

# Proof, continued

For decrease in $\|x - x^*\|$, set $x = x^k$, $\alpha = \alpha_k$, and $z = x^*$ in Lemma 1:

$$\begin{aligned} 0 &\le \phi(x^{k+1}) - \phi^* \\ &= \phi(x^k - \alpha_k G_{\alpha_k}(x^k)) - \phi^* \\ &\le G_{\alpha_k}^T(x^k - x^*) - \frac{\alpha_k}{2}\|G_{\alpha_k}(x^k)\|^2 \\ &= \frac{1}{2\alpha_k}\left(\|x^k - x^*\|^2 - \|x^k - x^* - \alpha_k G_{\alpha_k}(x^k)\|^2\right) \\ &= \frac{1}{2\alpha_k}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right), \end{aligned} \tag{6}$$

from which $\|x^{k+1} - x^*\| \le \|x^k - x^*\|$ follows.

# Proof, continued

Set $\alpha_k = 1/L$ in (6), and sum over $k = 0, 1, 2, \dots, T-1$, obtain

$$\sum_{k=0}^{T-1} (\phi(x^{k+1}) - \phi^*) \le \frac{L}{2} \left( \|x^0 - x^*\|^2 - \|x^K - x^*\|^2 \right) \le \frac{L}{2} \|x^0 - x^*\|^2.$$

By monotonicity of $\{\phi(x^k)\}$, we have

$$T(\phi(x^T) - \phi^*) \le \sum_{k=0}^{T-1} (\phi(x^{k+1}) - \phi^*).$$

Result follows by combining these last two expressions.

# Proximal-Gradient Algorithm: Quadratic Case

- Consider the quadratic case (of great interest): $f(x) = \frac{1}{2}\|Bx - b\|_2^2$.
- Here, $\nabla f(x) = B^T(Bx - b)$ and the IST/PGA/FBS algorithm is

  $$x_{k+1} = \text{prox}_{\alpha_k \psi}\big(x_k - \alpha_k B^T(Bx - b)\big)$$

  can be implemented with only matrix-vector multiplications with $B$ and $B^T$.

  This is a very important feature in large-scale applications, such as image processing, where fast algorithms exist for computing these products (e.g. fast Fourier transforms or wavelet transforms), but these matrices cannot be formed and stored explicitly.
- In this case, some more refined convergence results are available.
- Even more refined results are available if $\psi(x) = \|x\|_1$

# More on IST/FBS/PGA for the $\ell_2$-$\ell_1$ Case

- Problem: $\widehat{x} \in G = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Bx - b\|_2^2 + \tau\|x\|_1$ (recall $B^T B \preceq LI$)
- IST/FBS/PGA becomes $\boxed{x_{k+1} = \text{soft}\big(x_k - \alpha B^T(Bx - b), \alpha\tau\big)}$ with $\alpha < 2/L$.
- The zero set: $\mathcal{Z} \subseteq \{1, ..., n\}$ : $\widehat{x} \in G \Rightarrow \widehat{x}_{\mathcal{Z}} = 0$
- Zeros are found in a finite number of iterations (Hale et al., 2008): after a finite number of iterations, we have $(x_k)_{\mathcal{Z}} = 0$.
- After that, if $B_{\mathcal{Z}}^T B_{\mathcal{Z}} \succeq mI$, with $m > 0$ (thus $\kappa(B_{\mathcal{Z}}^T B_{\mathcal{Z}}) = L/m$):

$$\|x_{k+1} - \widehat{x}\|_2 \leq \frac{1 - \kappa}{1 + \kappa}\|x_k - \widehat{x}\|_2 \quad \text{(linear convergence)}$$

  for the optimal choice $\alpha = 2/(L + m)$. (Weaker condition suffices for lienar convergence of $\{f(x_k)\}$; see above.)

# FISTA with prox operations

- Recall that FISTA — *fast iterative shrinkage-thresholding algorithm* — ((Beck and Teboulle, 2009), based on (Nesterov, 1983)) is a heavy-ball-type acceleration of IST:

Initialize: Choose $\alpha \le 1/L$, $x_0$; set $y_1 = x_0$, $t_1 = 1$;

Iterate:

$$x_k \leftarrow \text{prox}_{\tau\alpha\psi}\big(y_k - \alpha\nabla f(y_k)\big);$$

$$t_{k+1} \leftarrow \tfrac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right);$$

$$y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}).$$

- Acceleration:

$$\text{FISTA: } f(x_k) - f(\widehat{x}) \sim O\left(\frac{1}{k^2}\right) \qquad \text{IST: } f(x_k) - f(\widehat{x}) \sim O\left(\frac{1}{k}\right).$$

- When $L$ is not known, increase an estimate of $L$ until it's big enough.

# Heavy Ball Acceleration: TwIST

- TwIST (*two-step iterative shrinkage-thresholding* (Bioucas-Dias and Figueiredo, 2007)) is a heavy-ball-type acceleration of IST, for

$$\min_x \tfrac{1}{2}\|Bx - b\|_2^2 + \tau\psi(x)$$

- Iterations (with $\alpha < 2/L$)

$$x_{k+1} = (\gamma - \beta)\, x_k + (1-\gamma) x_{k-1} + \beta\, \mathrm{prox}_{\alpha\tau\psi}\big(x_k - \alpha\, B^T(Bx - b)\big)$$

- Analysis in the strongly convex case: $mI \preceq B^T B \preceq LI$, with $m > 0$. Conditioning (as above) $\kappa = L/m < \infty$.
- Optimal parameters: $\gamma = \rho^2 + 1$, $\beta = \frac{2\alpha}{m+L}$, where $\rho = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$, yield linear convergence

$$\|x_{k+1} - \widehat{x}\|_2 \le \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\|x_k - \widehat{x}\|_2 \quad \left(\text{versus } \tfrac{1-\kappa}{1+\kappa} \text{ for IST}\right)$$

# Acceleration via Larger Steps: SpaRSA

- The standard step-size $\alpha_k \leq 2/L$ in IST is too timid
- The SpARSA (sparse reconstruction by separable approximation) framework proposes bolder choices of $\alpha_k$ (Wright et al., 2009):
  - ✓ Barzilai-Borwein (see above), to mimic Newton steps — or at least get the scaling right.
  - ✓ keep increasing $\alpha_k$ until monotonicity is violated: backtrack.
- Convergence to critical points (minima in the convex case) is guaranteed for a safeguarded version: ensure sufficient decrease w.r.t. the worst value in previous $M$ iterations.

# Acceleration by Continuation

- IST/FBS/PGA can be very slow if $\tau$ is very small and/or $f$ is poorly conditioned.

- A very simple acceleration strategy: continuation/homotopy

```
Initialization: Set $\tau_0 \gg \tau$, starting point $\bar{x}$, factor $\sigma \in (0,1)$, and $k = 0$.
Iterations: Find approx solution $x(\tau_k)$ of $\min_x f(x) + \tau_k \psi(x)$, starting from $\bar{x}$;
    if $\tau_k = \tau_f$ STOP;
    Set $\tau_{k+1} \leftarrow \max(\tau_f, \sigma\tau_k)$ and $\bar{x} \leftarrow x(\tau_k)$;
```

- Often the solution path $x(\tau)$, for a range of values of $\tau$ is desired, anyway (e.g., within an outer method to choose an optimal $\tau$)

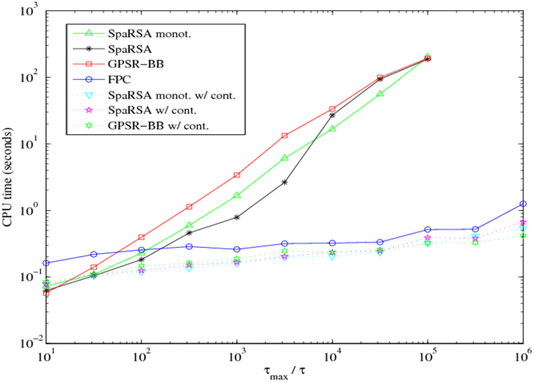
- Shown to be very effective in practice (Hale et al., 2008; Wright et al., 2009). Recently analyzed by Xiao and Zhang (2012).

# Acceleration by Continuation: An Example

Classical sparse reconstruction problem (Wright et al., 2009)

$$\widehat{x} \in \arg\min_{x} \tfrac{1}{2}\|B\,x - b\|_2^2 + \tau\|x\|_1$$

with $B \in \mathbb{R}^{1024\times 4096}$ (thus $x \in \mathbb{R}^{4096}$ and $b \in \mathbb{R}^{1024}$).



# A Final Touch: Debiasing

Consider problems of the form $\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Bx - b\|_2^2 + \tau\|x\|_1$

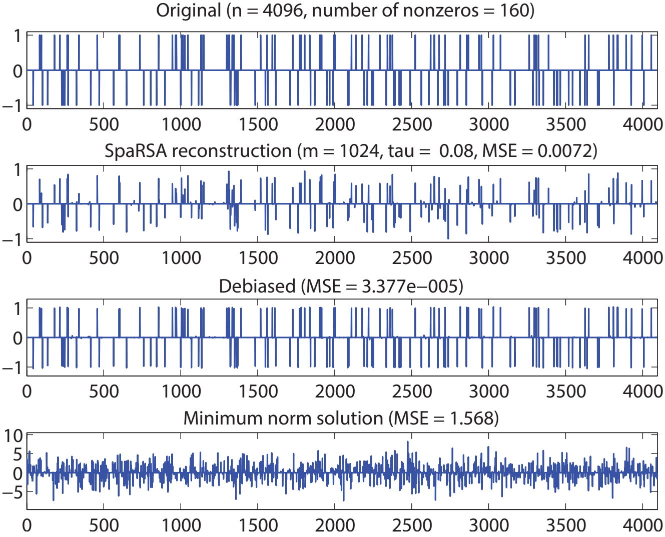
Often, the original goal was to minimize the quadratic term, after the support of $x$ had been found. But the $\ell_1$ term can cause the nonzero values of $x_i$ to be "suppressed."

Debiasing:

- ✓ find the zero set (complement of the support of $\widehat{x}$): $\mathcal{Z}(\widehat{x}) = \{1, \dots, n\} \setminus \text{supp}(\widehat{x})$.
- ✓ solve $\min_x \|Bx - b\|_2^2$ s.t. $x_{\mathcal{Z}(\widehat{x})} = 0$. (Fix the zeros and solve an unconstrained problem over the support.)

  Often, this problem has to be solved using an algorithm that only involves products by $B$ and $B^T$, since this matrix cannot be partitioned.

# Effect of Debiasing


Original (n = 4096, number of nonzeros = 160)

SpaRSA reconstruction (m = 1024, tau =  0.08, MSE = 0.0072)

Debiased (MSE = 3.377e−005)

Minimum norm solution (MSE = 1.568)

# Example: Matrix Recovery (Toh and Yun, 2010)

$$\widehat{M} \in \arg \min_{M \in \mathbb{R}^{n\times n}} \frac{1}{2}\|\Phi(M) - U\|_F^2 + \mu\|M\|_*$$

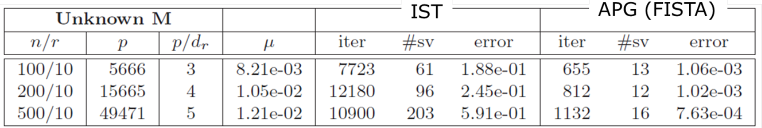
linear operator

The proximal algorithm (IST) is as before:

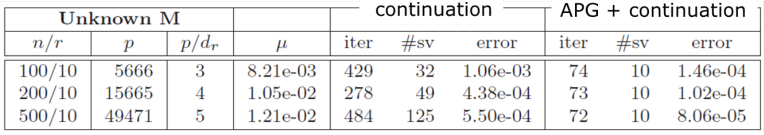
...its adjoint

$$X_{k+1} = \text{svt}_{\mu\,\beta_k}\Big(X_k - \beta_k\,\Phi^*(\Phi(X_k) - U)\Big)$$

Matrix completion: $\Phi(X) = X_\Omega$ (subset of entries) $|\Omega| = p$



| Unknown M | | | | IST | | | APG (FISTA) | | |
|---|---|---|---|---|---|---|---|---|---|
| $n/r$ | $p$ | $p/d_r$ | $\mu$ | iter | #sv | error | iter | #sv | error |
| 100/10 | 5666 | 3 | 8.21e-03 | 7723 | 61 | 1.88e-01 | 655 | 13 | 1.06e-03 |
| 200/10 | 15665 | 4 | 1.05e-02 | 12180 | 96 | 2.45e-01 | 812 | 12 | 1.02e-03 |
| 500/10 | 49471 | 5 | 1.21e-02 | 10900 | 203 | 5.91e-01 | 1132 | 16 | 7.63e-04 |



| Unknown M | | | | continuation | | | APG + continuation | | |
|---|---|---|---|---|---|---|---|---|---|
| $n/r$ | $p$ | $p/d_r$ | $\mu$ | iter | #sv | error | iter | #sv | error |
| 100/10 | 5666 | 3 | 8.21e-03 | 429 | 32 | 1.06e-03 | 74 | 10 | 1.46e-04 |
| 200/10 | 15665 | 4 | 1.05e-02 | 278 | 49 | 4.38e-04 | 73 | 10 | 1.02e-04 |
| 500/10 | 49471 | 5 | 1.21e-02 | 484 | 125 | 5.50e-04 | 72 | 10 | 8.06e-05 |

...the importance of acceleration!

# Identifying Optimal Manifolds

Identification of the manifold of the regularizer $\psi$ on which $x^*$ lies can improve algorithm performance, by focusing attention on a reduced space. We can thus evaluate *partial* gradients and Hessians, restricted to just this space.

For nonsmooth regularizer $\psi$, the optimal manifold is a smooth surface passing through $x^*$ along which the restriction of $\psi$ is smooth.

**Example:** for $\psi(x) = \|x\|_1$, have manifold consisting of $z$ with

$$z_i \begin{cases} \geq 0 & \text{if } x_i^* > 0 \\ \leq 0 & \text{if } x_i^* < 0 \\ = 0 & \text{if } x_i^* = 0. \end{cases}$$

If we know the optimal nonzero components, we know the manifold. We could restrict the search to just this set of nonzeros.

# Identification Properties of Shrink Algorithms

When the optimal manifold is partly smooth (that is, parametrizable by smooth functions and otherwise well behaved) and prox-regular, and the minimizer is nondegenerate, then the shrink approach can identify it from any sufficiently close $x$. That is,

$$S_\tau(x - \alpha \nabla f(x), \alpha)$$

lies on the optimal manifold, for $\alpha$ bounded away from 0 and $x$ in a neighborhood of $x^*$. (Consequence of Lewis and Wright (2008).)

For $\psi(x) = \|x\|_1$, shrink algorithms identify the correct nonzero set, provided there are no "borderline" components (that is, the optimal nonzero set would not change with an arbitrarily small perturbation to the data).

Can use a heuristic to identify when the nonzero set settles down, then switch to second phase to conduct a search on the reduced space of "possible nonzeros."

# 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).

# Atomic-Norm Regularization

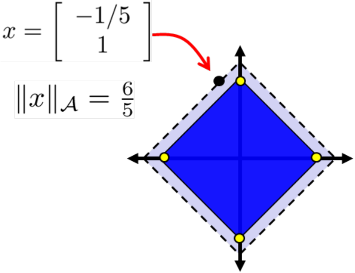
The atomic norm

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

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

Example: the $\ell_1$ 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).
- $\|x\|_{\mathcal{A}} = \inf\left\{t > 0 : \; x \in t\; B_1(1)\right\}$
  $= \|x\|_1$



$x = \begin{bmatrix} -1/5 \\ 1 \end{bmatrix}$
$\|x\|_{\mathcal{A}} = \frac{6}{5}$

# Atomic Norms: More Examples

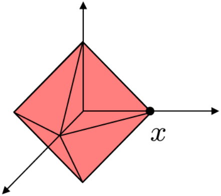
Examples with easy forms:

- *sparse vectors*

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

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

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

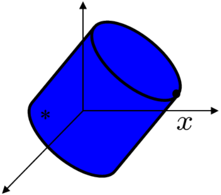


- *low-rank matrices*

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

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

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



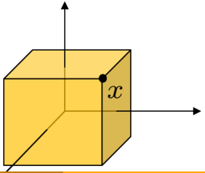
*symmetric matrices

- *binary vectors*

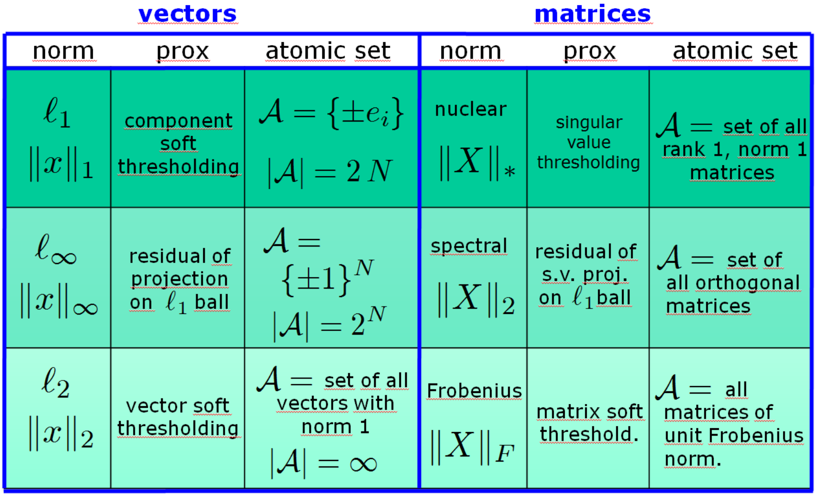
  $\mathcal{A} = \{\pm 1\}^N$

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

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



# Atomic Norms: A Unified View



| vectors | | | matrices | | |
|---|---|---|---|---|---|
| norm | prox | atomic set | norm | prox | atomic set |
| $\ell_1$<br>$\|x\|_1$ | component soft thresholding | $\mathcal{A} = \{\pm e_i\}$<br>$\|\mathcal{A}\| = 2\,N$ | nuclear<br>$\|X\|_*$ | singular value thresholding | $\mathcal{A} =$ set of all rank 1, norm 1 matrices |
| $\ell_\infty$<br>$\|x\|_\infty$ | residual of projection on $\ell_1$ ball | $\mathcal{A} = \{\pm 1\}^N$<br>$\|\mathcal{A}\| = 2^N$ | spectral<br>$\|X\|_2$ | residual of s.v. proj. on $\ell_1$ ball | $\mathcal{A} =$ set of all orthogonal matrices |
| $\ell_2$<br>$\|x\|_2$ | vector soft thresholding | $\mathcal{A} =$ set of all vectors with norm 1<br>$\|\mathcal{A}\| = \infty$ | Frobenius<br>$\|X\|_F$ | matrix soft threshold. | $\mathcal{A} =$ all matrices of unit Frobenius norm. |

# Atomic-Norm Regularization

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

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

(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).

# Conditional Gradient for Atomic-Norm Constraints

Conditional Gradient is particularly useful for optimization over atomic-norm constraints.

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

Reminder: Given the set of atoms $\mathcal{A}$ (possibly infinite) we have

$$\|x\|_{\mathcal{A}} := \inf \left\{ \sum_{a \in \mathcal{A}} c_a \, : \, x = \sum_{a \in \mathcal{A}} c_a a, \; c_a \ge 0 \right\}.$$

The search direction $v_k$ is $\tau \bar{a}_k$, where

$$\bar{a}_k := \arg\min_{a \in \mathcal{A}} \langle a, \nabla f(x_k) \rangle.$$

That is, we seek the atom that lines up best with the negative gradient direction $-\nabla f(x_k)$.

# Generating Atoms

We can think of each step as the "addition of a new atom to the basis." Note that $x_k$ is expressed in terms of $\{\bar{a}_0, \bar{a}_1, \dots, \bar{a}_k\}$.

If few iterations are needed to find a solution of acceptable accuracy, then we have an approximate solution that's represented in terms of few atoms, that is, sparse or compactly represented.

For many atomic sets $\mathcal{A}$ of interest, the new atom can be found cheaply.

Example: For the constraint $\|x\|_1 \le \tau$, the atoms are $\{\pm e_i \, : \, i = 1, 2, \dots, n\}$. if $i_k$ is the index at which $|[\nabla f(x_k)]_i|$ attains its maximum, we have

$$\bar{a}_k = -\text{sign}([\nabla f(x_k)]_{i_k}) \, e_{i_k}$$

Example: For the constraint $\|x\|_\infty \le \tau$, the atoms are the $2^n$ vectors with entries ±1. We have

$$[\bar{a}_k]_i = -\text{sign}[\nabla f(x_k)]_i, \quad i = 1, 2, \dots, n.$$

# More Examples

**Example: Nuclear Norm.** For the constraint $\|X\|_* \le \tau$, for which the atoms are the rank-one matrices, we have $\bar{A}_k = u_k v_k^T$, where $u_k$ and $v_k$ are the first columns of the matrices $U_k$ and $V_k$ obtained from the SVD $\nabla f(X_k) = U_k \Sigma_k V_k^T$.

**Example: sum-of-$\ell_2$.** For the constraint

$$\sum_{i=1}^{m} \|x_{[i]}\|_2 \le \tau,$$

the atoms are the vectors $a$ that contain all zeros except for a vector $u_{[i]}$ with unit 2-norm in the $[i]$ block position. (Infinitely many.) The atom $\bar{a}_k$ contains nonzero components in the block $i_k$ for which $\|[\nabla f(x_k)]_{[i]}\|$ is maximized, and the nonzero part is

$$u_{[i]} = -[\nabla f(x_k)]_{[i_k]} / \|[\nabla f(x_k)]_{[i_k]}\|.$$

# Other Enhancements

**Reoptimizing.** Instead of fixing the contribution $\alpha_k$ from each atom at the time it joins the basis, we can periodically and approximately reoptimize over the current basis.

- This is a finite dimension optimization problem over the (nonnegative) coefficients of the basis atoms.
- It need only be solved approximately.
- If any coefficient is reduced to zero, it can be dropped from the basis.

**Dropping Atoms.** Sparsity of the solution can be improved by dropping atoms from the basis, if doing so does not degrade the value of $f$ too much (see (Rao et al., 2013)).

In the important least-squares case, the effect of dropping can be evaluated efficiently.

# 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.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Bioucas-Dias, J. and Figueiredo, M. (2007). A new twist: two-step iterativeshrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16:2992–3004.

Brucker, P. (1984). An O(n) algorithm for quadratic knapsack problems. *Operations Research Letters*, 3:163–166.

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

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

Combettes, P. and Pesquet, J.-C. (2011). Signal recovery by proximal forward-backward splitting. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer.

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

# References II

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

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

Hale, E., Yin, W., and Zhang, Y. (2008). Fixed-point continuation for l1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19:1107–1130.

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

Maculan, N. and de Paula, G. G. (1989). A linear-time median-finding algorithm for projecting a vector on the simplex of $\mathbb{R}^n$. *Operations Research Letters*, 8:219–222.

Moreau, J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899.

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

Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Math. Doklady*, 27:372–376.

Rao, N., Shah, P., Wright, S. J., and Nowak, R. (2013). A greedy forward-backward algorithm for atomic norm constrained minimization. In *Proceedings of ICASSP*.

Toh, K.-C. and Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6:615–640.

# References III

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

Xiao, L. and Zhang, T. (2012). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*. (to appear; available at http://arxiv.org/abs/1203.3002).

Yin, W. and Zhang, Y. (2008). Extracting salient features from less data via $\ell_1$-minimization authors. *SIAG/OPT Views-and-News*, 19:11–19.