Shrinkage/Thresholding Iterative Methods

- Nonquadratic regularizers
  - Total Variation
  - $l_p$-norm
  - Wavelet orthogonal/redundant representations
  - Sparse regression
- Majorization Minimization revisited
- IST- Iterative Shrinkage Thresholding Methods
- TwIST-Two step IST
Linear inverse problems - LIPs

\[ \hat{f} = \arg\min_f L(f) \]

\[ L : \mathbb{R}^m \to \overline{\mathbb{R}} = [-\infty, +\infty] \]

\[ L(f) = \frac{1}{2} \| g - Af \|^2 + \alpha \Phi(f), \]

\( \Phi(f) \) is non-quadratic and possibly non-smooth
References


Majorization minorization (MM) framework

Let \( \hat{x} = \arg\min_x L(x) \)

Majorization Minorization algorithm: \( \hat{x}^{(t+1)} = \arg\min_x Q(x|\hat{x}^{(t)}) \)

\[ Q(x|\hat{\theta}^{(t)}) \geq L(x) \quad \text{....with equality if and only if} \quad x = \hat{x}^{(t)} \]

Easy to prove monotonicity:

\[ L(\hat{x}^{(t+1)}) \leq L(\hat{x}^{(t)}) \]

Notes: \( Q(x|\hat{x}^{(t)}) \) should be easy to maximize

EM is an algorithm of this type.
MM algorithms for LIPs

\[ L(f) = \frac{1}{2} \| g - Af \|^2 + \alpha \Phi(f) \]

IST Class: Majorize \( \| g - Af \|^2 \)

IRS Class: Majorize \( \Phi(f) \)

IST/IRS: Majorize \( \| g - Af \|^2 \) and \( \Phi(f) \)
MM algorithms: IST class

\[ L(f) = \frac{1}{2} \| g - Af \|^2 + \alpha \Phi(f) \]

\[ L_d(f) = L_d(f') + (f - f')^* \nabla L(f') + (f - f')^* A^* A(f - f') \]

Assume that \( \|A\|_2 \leq 1 \)

\[ L_d(f) \leq L_d(f') + (f - f')^* \nabla L(f') + (f - f')^* (f - f') \]

\[ = \frac{1}{2} \| f - y \| + c^{te} \]

\[ y = f + A^* (g - Af') \]
MM algorithms: IST class

\[ L(f) = \frac{1}{2} \| g - Af \|^2 + \alpha \Phi(f) \]
\[ \| A \|_2 \leq 1 \]

Majorizer:

\[ Q(f; f_k) = \frac{1}{2} \| f - y_k \|^2 + \alpha \Phi(f) \]

\[ y_k = f_k + A^*(g - Af_k) \]

Let:

\[ \psi_\alpha(y) = \arg \min_f \frac{1}{2} \| f - y \|^2 + \alpha \Phi(f) \]

**IST Algorithm**

\[ f_{k+1} = \psi_\alpha(f_k + A^*(g - Af_k)) \]
MM algorithms: IST class

Overrelaxed IST Algorithm

\[ f_{k+1} = (1 - \beta) f_k + \beta \Phi_{\alpha}(f_k + \lambda (y - Af)) \]

Convergence: [Combettes and V. Wajs, 2004]

- \( \Phi(f) \) is convex
- \( \|A\|^2 < 2 \)
- the set of minimizers, \( G \), of \( L(f) = \frac{1}{2} \|g - Af\|^2 + \alpha \Phi(f) \)
  is non-empty
- \( \beta \in ]0,1] \)
- Then \( f_k \) converges to a point in \( G \)
TwIST: Two step IST

Improve convergence rate w.r.t IST

\[ f_{k+1} = (\alpha - \beta) f_k + (1 - \alpha) f_{k-1} + \beta \Psi \alpha(f_k + A^*(g - Af_k)) \]

The two-step method has the same fixed points as the one-step counterpart.

These fixed points are minima of the objective function 
[Combettes & Wajs, 2005], [Daubechies, Defriese, De Mol, 2004].

What about convergence and rate of convergence?
Main Assumptions

(i) \[ 0 < m \leq \lambda_{\text{min}}(A^* A) \leq \lambda_{\text{max}}(A^* A) = 1 \]

(ii) \( \Phi \) convex,
    lower semi-continuous,
    positive homogeneous of degree 1
    Examples: TV, \( l_p \) and Besov \( B_{p,p}^s \) norms, for \( p \geq 1 \)

\[ \lambda_{\text{min}}(A^* A) > 0 \] and convexity of \( \Phi \) imply uniqueness of

\[ \hat{f} = \arg \min_f \left\{ \|g - Af\|^2 + \lambda \Phi(f) \right\} \]
Main Result

Let \( 0 < \alpha < 2 \) and \( 0 < \beta < 2 \alpha/(\lambda_{\text{min}} + 1) \).

Then, the two-step algorithm converges, i.e.

\[
\lim_{t \to \infty} \| f_k - \hat{f} \| = 0
\]

There is an optimal choice for \( \alpha \) and \( \beta \) for which

\[
\| f_{k+1} - \hat{f} \| \leq \frac{1 - \sqrt{m}}{1 + \sqrt{m}} \| f_k - \hat{f} \|
\]
Some Remarks

A one-step method is recovered for $\alpha = 1$

$$f_{k+1} = (1 - \beta) f_k + \beta \Psi_\lambda (f_k + A^*(g - Af_k))$$

which is an over-relaxed version of the original one-step method.

For the optimal choice of $\beta$:

$$\|f_{k+1} - \hat{f}\| \leq \frac{1 - m}{1 + m} \|f_k - \hat{f}\|$$

$$-1 / \log_{10} \frac{1 - m}{1 + m} \sim \text{number of iterations to decrease error by factor of 10.}$$

Example:

$$m = 10^{-3} \rightarrow -1 / \log \frac{1 - m}{1 + m} \sim 1150 \quad -1 / \log \frac{1 - \sqrt{m}}{1 + \sqrt{m}} \sim 35$$
Denoising with convex regularizers

\[ \Phi(\cdot) \text{ convex } \Rightarrow \left\{ \frac{1}{2} \| \cdot - y \|^2 + \alpha \Phi(\cdot) \right\} \text{ strictly convex} \]

Denoising function also known as the Moreau proximal mapping

\[ \text{prox}_{\alpha \Phi}(y) = \arg \min_f \frac{1}{2} \| f - y \|^2 + \alpha \Phi(f) \]

Moreau decomposition

\[ \text{prox}_\Phi(y) + \text{prox}_{\Phi^*}(y) = y \]

Convex conjugate

\[ \Phi^*(y) = \sup_y \langle x, y \rangle - \Phi(x) \]

Classes of convex regularizers:

1- homogeneous (TV, lp-norm (p>1))
2- p power of an lp norm
1-Homogeneous regularizers

\[ \Phi(\cdot) \text{ convex and } \Phi(tf) = t\Phi(f) \text{ for } t \geq 0 \]

Then,

\[ \psi_\alpha(y) = y - P_\alpha C(y) \]

where \( C \in \mathbb{R}^m \) is a closed convex set

\[ C = \{ f \in \mathbb{R}^m | \langle f, u \rangle \leq \Phi(u), \forall u \in \mathbb{R}^m \} \]

and \( P_A : \mathbb{R}^m \rightarrow \mathbb{R}^m \) denotes the orthogonal projection on the convex set \( A \in \mathbb{R}^m \).
Total variation regularization

Total variation [S. Osher, L. Rudin, and E. Fatemi, 1992]

\[
\Phi_{i\text{TV}}(f) = \sum_i \sqrt{(\Delta_i^h f)^2 + (\Delta_i^v f)^2}
\]

\[
\Phi_{n\text{TV}}(f) = \sum_i |\Delta_i^h f| + |\Delta_i^v f|
\]

\[
\Delta_i^h f = f_{i,j} - f_{i,j-1} \quad \Delta_i^v f = f_{i,j} - f_{i-1,j}
\]

\(\Phi\) is convex (although not strictly) and 1-homogeneous

\[
\Phi(tf) = t\Phi(f), \quad \text{for } t \geq 0
\]

Total variation is a discontinuity-preserving regularizer

\(f_1, f_2\) have the same TV
Total variation regularization

\[ \Phi(\cdot) \text{ convex and } \Phi(tf) = t\Phi(f) \text{ for } t \geq 0 \]

Then

\[ \Psi_\alpha(y) = y - P_\alpha C(y) \]

\[ C = \{ \text{div} \, p \mid p \in \mathbb{R}^m \times \mathbb{R}^m, \, |p_{i,j}| \leq 1, \, \forall i, j = 1, 2, \ldots \} \]

[Chambolle, 2004]

**Theorem 3.1.** Let \( \tau \leq 1/8 \). Then, \( \lambda \text{div} \, p^n \) converges to \( \pi_{\lambda K}(g) \) as \( n \to \infty \).

\[ p^{n+1}_{i,j} = \frac{p^n_{i,j} + \tau(\nabla(\text{div} \, p^n - g/\Omega))_{i,j}}{1 + \tau |(\nabla(\text{div} \, p^n - g/\Omega))_{i,j}|} \]
Total variation denoising
Total variation deconvolution

2000 IST iterations !!!
Weighted $l^p$-norms

$$
\Phi_{l^p_w}(f) = \|f\|_{p,w} = \left(\sum_i w_i |f_i|^p\right)^{1/p}, \quad w_i > 0
$$

$p \geq 1 \Rightarrow \Phi$ is convex (although not strictly) and 1-homogeneous

There is no closed form expression for $\Psi_\alpha$ except for some particular cases

$p = 1 \Rightarrow \Phi$, is 1-homogeneous and $\Psi_\alpha$ is decoupled

Thus

$$
\Psi_\alpha(y_i) = (\Psi_\alpha(y_1)), \ldots, \Psi_\alpha(y_N))
$$

$$
\Psi_\alpha(y_i) = y_i - P_\alpha C_i(y)
$$

$$
C_i = \{f_i \in \mathbb{R}^m | \langle f_i, u_i \rangle \leq |u_i|, \forall u \in \mathbb{R}\}
$$
Soft thresholding: p=1

\[ \Phi_{\ell^1}(f) = \|f\|_1 = \sum_i |f_i| \]

\[ p = 1 \Rightarrow \Phi, \text{ is 1-homogeneous and } \Psi_\alpha \text{ is decoupled} \]

Thus

\[ \Psi_\alpha(y) = (\Psi_\alpha(y_1)), \ldots, \Psi_\alpha(y_N)) \]

\[ \Psi_\alpha(y_i) = y_i - P_{\alpha C_i}(y) \]

\[ C_i = \{ f_i \in \mathbb{R} | \langle f_i, u_i \rangle \leq |u_i|, \forall u \in \mathbb{R} \} = [-1, 1] \]

\[ P_{\alpha C_i}(y) = \begin{cases} \alpha & y \geq \alpha \\ -\alpha & y \leq \alpha \\ y & |y| < \alpha. \end{cases} \]

\[ \Psi_\alpha(y) = y - P_{\alpha C}(y) \]

\[ \begin{cases} y - \alpha & y \geq \alpha \\ y + \alpha & y \leq \alpha \\ 0 & |y| < \alpha. \end{cases} \]
Soft thresholding: $p=1$

$$\Phi_{\ell^1}(f) = \|f\|_1 = \sum_i |f_i|$$

$$\Psi_\alpha(y) = y - P_{\alpha C}(y) = \begin{cases} y - \alpha & y \geq \alpha \\ y + \alpha & y \leq \alpha \\ 0 & |y| < \alpha. \end{cases} \equiv \text{soft}(y, \alpha)$$

![Graph showing soft thresholding function](image)
Soft thresholding: $p=1$

$$\Phi_{\ell^1}(f) = \|f\|_1 = \sum_i |f_i|$$

$$\Psi_\alpha(y) = y - P_{\alpha C}(y) = \begin{cases} 
  y - \alpha & y \geq \alpha \\
  y + \alpha & y \leq \alpha \\
  0 & |y| < \alpha.
\end{cases} \equiv \text{soft}(y, \alpha)$$

Diagram showing the soft threshold function $\text{soft}(y, \alpha)$ with $-\alpha$ and $\alpha$ intercepts.
Another way to look at it:

\[ L(f) = \frac{1}{2} \| f - y \|^2 + \alpha \Phi(f) \]

Since \( L \) is convex: the point \( \hat{f} \) is a global minimum of \( L \) if and only if

\[ 0 \in \partial L(\hat{f}) \]

where \( \partial L(f) \) is the subdifferential of \( L \) at \( f' \)

\[ \partial L(f') = \{ u \mid L(f) - L(f') \geq \langle u, f - f' \rangle \} \]
Another way to look at it:

\[ L(f) = \frac{1}{2} \| f - y \|^2 + \alpha \Phi(f) \]

\[ 0 \in \partial L(f) \iff 0 \in f - y + \alpha \partial \Phi(f) \]

\[ \iff y \in (I + \alpha \partial \Phi)(f) \]

\[ \iff f = (I + \alpha \partial \Phi)^{-1}(y) \quad \text{thanks to the uniqueness of the minimizer} \]

\[ \iff \text{prox}_{\alpha \Phi} = (I + \alpha \partial \Phi)^{-1} \quad \text{resolvent of } \partial \Phi \]
Proximity Operators

Moreau proximity operator (shrinkage/thresholding/denoising function)

\[ \text{prox}_{\tau c}(u) = \arg \min_x \left( \frac{1}{2} \|x - u\|_2^2 + \tau c(x) \right) \]

Projection onto a convex set

\[ \iota_C(x) = \begin{cases} 
0 & x \in C' \\
+\infty & x \notin C' 
\end{cases} \]

\[ \text{prox}_{\tau \iota_C}(u) = \arg \min_{x \in C} \|x - u\|_2^2 \]

Proximity operators generalize projections onto convex sets

Proximity operators have the flavor of gradient steps

Fixed points: \( u^* \) minimizes \( C \) if and only if

\[ u^* = \text{prox}_{\tau c}(u^*) \]

Moreau decomposition

\[ u = \text{prox}_c(u) + \text{prox}_{c^*}(u) \]

\( c^* \) is the convex conjugate of \( c \)

[Moreau 62], [Combettes, 01], [Combettes, Wajs, 05], [Combettes, Pesquet, 07, 11], [Parikh, Boyd, 2013]
Proximity Operators of Widely Used Convex Regularizers

**$l_1$ norm**

\[
c(z) = \|z\|_1 \Rightarrow \text{prox}_{\tau c}(u) = \text{soft}(u, \tau)
\]

\[
:= (|u| - \tau)_+ \text{sign}(u)
\]

**$l_2$ norm**

\[
c(z) = \|z\|_2 \Rightarrow \text{prox}_{\tau c}(u) = \text{vect-soft}(u, \tau)
\]

\[
:= (\|u\|_2 - \tau)_+(u/\|u\|_2)
\]

**$l_\infty$ norm**

\[
c(z) = \|z\|_\infty \Rightarrow \text{prox}_{\tau c}(u) = u - P_{B_{l_1}(\tau)}(u)
\]

**nuclear norm**

\[
c(Z) = \|Z\|_* \Rightarrow \text{prox}_{\tau c}(X) = UD_{\tau}(\Sigma)V^T
\]

\[
[U, \Sigma, V] = \text{svd}(X) \quad D_{\tau}(\Sigma) = \text{diag}(\sigma_i - \tau)_+
\]
Example: Wavelet-based restoration

\[ \theta = Wf \]

Wavelet basis

Wavelet coefficients

Approximation coefficients (g-low pass filter)

Detail coefficients (h – high pass filter)

\[ x[n] \rightarrow h[n] \rightarrow 2 \rightarrow \text{Level 1 coefficients} \]

\[ g[n] \rightarrow 2 \rightarrow h[n] \rightarrow 2 \rightarrow \text{Level 2 coefficients} \]

\[ g[n] \rightarrow 2 \rightarrow \text{Level 3 coefficients} \]

DWT, Harr, J=2
Example: Wavelet-based restoration

Histogram of coefficients - h

Histogram of coefficients – log h