A Fast Iterative Shrinkage Algorithm for Convex Regularized Linear Inverse Problems

Marc Teboulle

School of Mathematical Sciences
Tel-Aviv University, Ramat-Aviv, Israel

Joint Work with Amir Beck, Technion, Haifa

International Conference on Nonlinear Programming and Applications
NPA 2008 – April 6–9, 2008, Beijing, China
Outline

- Linear Inverse Problems with Nonsmooth Regularization
  - Formulation and Application Areas

- Current Class of Iterative Methods (ISTA):
  - Iterative Shrinkage-Threshold Algorithms

- FISTA: A Fast Iterative Shrinkage-Threshold Algorithm
  - A global rate of convergence/complexity estimate

- Numerical Examples for Image Deblurring Problems

- Conclusions
Linear Inverse Problem

**Problem:** Estimate the unknown signal $x$ from a noisy observation

$$Ax = b + w.$$ 

- $x \in \mathbb{R}^n$ - input signal – (Unknown True Image)
- $b \in \mathbb{R}^m$ - observable output – (Blurred Image)
- $w \in \mathbb{R}^m$ - unknown noise vector.
- $A \in \mathbb{R}^{m \times n}$ model – (Blurring matrix (2-dim convolution)).

**An Example:** The problem of estimating $x$ from the observed blurred and noisy image is an *Image Deblurring Problem*. 
Regularization Approaches

Classical Least Squares (LS) estimator

(LS) : \( \hat{x}_{LS} = \arg \min_x \|Ax - b\|^2. \)

A ill-conditioned – meaningless solution
Regularization Approaches

Classical Least Squares (LS) estimator

\[(LS): \hat{x}_{LS} = \arg\min_x \|Ax - b\|^2.\]

A ill-conditioned – meaningless solution

Tikhonov regularization – quadratic penalty

\[(T): \hat{x}_{TIK} = \arg\min_x \{\|Ax - b\|^2 + \lambda \|Lx\|^2\}, \ \lambda > 0.\]
Regularization Approaches

Classical Least Squares (LS) estimator

\[ (\text{LS}) : \quad \hat{x}_{\text{LS}} = \arg\min_x \|Ax - b\|^2. \]

A ill-conditioned – meaningless solution

Tikhonov regularization – quadratic penalty

\[ (T) : \quad \hat{x}_{\text{TIK}} = \arg\min_x \{\|Ax - b\|^2 + \lambda\|Lx\|^2\}, \quad \lambda > 0. \]

$l_1$-norm regularization

\[ (L_1) \quad \min_x \{F(x) \equiv \|Ax - b\|^2 + \lambda\|x\|_1\} \]

Less sensitive to outliers (as opposed to $l_2$ regularization). Has attracted a revived interest and considerable amount of attention in Signal Processing Research.
The $l_1$-Regularization Model: Old and New Applications

- LASSO in Statistics (Tibshirani (96))
- Basis pursuit denoising (Chen et al. (98))
The $l_1$-Regularization Model: Old and New Applications

- LASSO in Statistics (Tibshirani (96))
- Basis pursuit denoising (Chen et al. (98))
- Wavelet based image/signal restoration (Donoho (95), Chambolle (04)...) 
- Sparse Approximation of signals (Elad (06), Daubechies et al. (07),...)
- Compressed sensing: few measurements are enough to produce good reconstruction (Candes-Tao (06), Donoho (06)...)

♠ The term $\|x\|_1$ promotes sparsity in the optimal solution.
The $l_1$-Regularization Model: Old and New Applications

- LASSO in Statistics (Tibshirani (96))
- Basis pursuit denoising (Chen et al. (98))
- Wavelet based image/signal restoration (Donoho (95), Chambolle (04)…)
- Sparse Approximation of signals (Elad (06), Daubechies et al. (07),…)
- Compressed sensing: few measurements are enough to produce good reconstruction (Candes-Tao (06), Donoho(06)…)

♠ The term $\|x\|_1$ promotes sparsity in the optimal solution.

◊ In image deblurring/wavelet based restoration: most images have a sparse representation in wavelet domain.
♠ State of the art regularization for Image Restoration involves nonsmooth regularizers.
General Formulation with Nonsmooth Regularizers

A nonsmooth convex minimization model which covers quite a lot of interesting and disparate applications.

\[
(P) \quad \min \{ F(x) \equiv f(x) + g(x) : x \in \mathbb{R}^n \}.
\]

- \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is a smooth convex function of the type \( C^{1,1} \), i.e., continuously differentiable with Lipschitz continuous gradient \( L(f) \):

\[
\| \nabla f(x) - \nabla f(y) \| \leq L(f) \| x - y \| \quad \text{for every } x, y \in \mathbb{R}^n,
\]

where \( \| \cdot \| \) denotes the standard Euclidean norm and \( L(f) > 0 \) is the Lipschitz constant of \( \nabla f \).

- \( g : \mathbb{R}^n \rightarrow \mathbb{R} \) is a convex function which is nonsmooth.

- Problem (P) is solvable, i.e., \( X_* := \text{argmin } f \neq \emptyset \), and for \( x^* \in X_* \) we set \( F_* := F(x^*) \).
Challenges: How do we solve (P)?

- The problem in non-smooth.
Challenges: How do we solve (P)?

- The problem is non-smooth.
- In most applications, can be very large scale, e.g., in image deblurring, the dimension varies from $d = 65,536$ to $1,048,576$. 
Challenges: How do we solve (P)?

- The problem is **nonsmooth**.
- In most applications, can be **very large scale**, e.g., in image deblurring, the dimension varies from $d = 65,536$ to $1,048,576$.
- Involves **dense matrix data**, precluding the use and potential advantages of well-known methods (storage/factorization impractical), even for the $L_1$ problem (which can be reformulated as a QP or SOCP).
Challenges: How do we solve (P)?

- The problem in **nonsmooth**.
- In most applications, can be **very large scale**, e.g., in image deblurring, the dimension varies from \( d = 65,536 \) to \( 1,048,576 \).
- Involves **dense matrix data**, precluding the use and potential advantages of well-known methods (storage/factorization impractical), even for the \( L_1 \) problem (which can be reformulated as a QP or SOCP).
- This motivates the search for simple and efficient algorithms where the dominant computational effort is a relatively cheap **matrix-vector** multiplications involving \( A \) and \( A^T \).
Challenges: How do we solve (P)?

- The problem is **nonsmooth**.

- In most applications, can be **very large scale**, e.g., in image deblurring, the dimension varies from \( d = 65,536 \) to \( 1,048,576 \).

- Involves **dense matrix data**, precluding the use and potential advantages of well-known methods (storage/factorization impractical), even for the \( L_1 \) problem (which can be reformulated as a QP or SOCP).

- This motivates the search for simple and efficient algorithms where the dominant computational effort is a relatively cheap **matrix-vector** multiplications involving \( A \) and \( A^T \).

- Simple algorithms exist...But...
A Current Very Popular Algorithm

Class of *Iterative Shrinkage-Threshold Algorithms* (ISTA) for $L_1$:

$$x_{k+1} = T_{\lambda t}(x_k - tA^T(Ax_k - b)),$$ \(t > 0\) a step size

and $T_{\alpha} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the shrinkage operator defined by

$$T_{\alpha}(x)_i = (|x_i| - \alpha)_+ \text{sgn}(x_i).$$

Each iteration involves matrix-vector multiplication involving $A$ and $A^T$ followed by a shrinkage/soft-threshold step.
A Current Very Popular Algorithm

Class of *Iterative Shrinkage-Threshold Algoritms* (ISTA) for $L_1$:

$$x_{k+1} = \mathcal{T}_{\lambda t} \left( x_k - t A^T (Ax_k - b) \right), \quad t > 0 \text{ a step size}$$

and $\mathcal{T}_\alpha : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the shrinkage operator defined by

$$\mathcal{T}_\alpha(x)_i = (|x_i| - \alpha)_+ \text{sgn} (x_i).$$

Each iteration involves matrix-vector multiplication involving $A$ and $A^T$ followed by a shrinkage/soft-threshold step.

**In SP literature:** appeared under various names: Iterative denoising, Shrinkage-Thresholded, Landweber, EM wavelet based etc....: Chambolle (98); Figueiredo-Nowak (03, 05); Daubechies et al. (04),...

**In Optimization:** it is a well known algorithm....
For any $L > 0$, and a given $z$:

$$Q_L(x, z) := f(z) + \langle x - z, \nabla f(z) \rangle + \frac{L}{2} \|x - z\|^2 + g(x)$$ 

left untouched

$$\min_x F(x) \leftrightarrow \min_x Q_L(x, z)$$ which admits a unique minimizer

$$p_L(z) := \arg\min_x Q_L(x, z) = \arg\min_x \{g(x) + \frac{L}{2} \|x - (z - \frac{1}{L} \nabla f(z))\|^2\}.$$
A Basic Approximation Model: Following the well-known gradient scheme

- For any $L > 0$, and a given $z$:

$$Q_L(x, z) := f(z) + \langle x-z, \nabla f(z) \rangle + \frac{L}{2} \|x-z\|^2 + g(x) \left\downarrow \text{left untouched} \right.$$ 

$$\min_x F(x) \leftrightarrow \min_x Q_L(x, z) \text{ which admits a unique minimizer}$$

$$p_L(z) := \arg\min_x Q_L(x, z) = \arg\min_x \{g(x) + \frac{L}{2} \|x-(z-\frac{1}{L} \nabla f(z))\|^2 \}.$$ 

- **Algorithm:** $x_0 \in \mathbb{R}^n$, $x_{k+1} = p_L(x_k)$.  


A Basic Approximation Model: Following the well-known gradient scheme

For any $L > 0$, and a given $z$:

$$Q_L(x, z) := f(z) + \langle x - z, \nabla f(z) \rangle + \frac{L}{2} \|x - z\|^2 + g(x) \left/ \text{left untouched} \right.$$  

$$\min_x F(x) \leftrightarrow \min_x Q_L(x, z)$$ which admits a unique minimizer

$$p_L(z) := \arg\min_x Q_L(x, z) = \arg\min_x \left\{ g(x) + \frac{L}{2} \|x - (z - \frac{1}{L} \nabla f(z))\|^2 \right\}.$$  

Algorithm: $x_0 \in \mathbb{R}^n$, $x_{k+1} = p_L(x_k)$.

Special Case-ISTA

$g(x) := \lambda \|x\|_1$, $f(x) := \|Ax - b\|^2$, $L := t^{-1}$
For any $L > 0$, and a given $z$:

$$Q_L(x, z) := f(z) + \langle x - z, \nabla f(z) \rangle + \frac{L}{2} \|x - z\|^2 + g(x)$$

\[\text{min}_x F(x) \iff \text{min}_x Q_L(x, z)\]

which admits a unique minimizer

$$p_L(z) := \arg\min_x Q_L(x, z) = \arg\min_x \{g(x) + \frac{L}{2} \|x - (z - \frac{1}{L} \nabla f(z))\|^2\}.$$

**Algorithm:** $x_0 \in \mathbb{R}^n$, $x_{k+1} = p_L(x_k)$.

**Special Case-ISTA**

$g(x) := \lambda \|x\|_1$, $f(x) := \|Ax - b\|^2$, $L := t^{-1}$

Can be viewed as the Proximal-FB Splitting Method (Passty (79)):

$$0 \in \nabla f(x) + \partial g(x) \iff x = (I + s \partial g)^{-1}(I - s \nabla f)(x), \quad (s > 0)$$
Advantage and Drawback of ISTA

- **Advantage:** Simplicity. Useful when $p_L(\cdot)$ can be computed analytically, e.g. when $g(\cdot)$ is separable, reduces to a one dimensional minimization problem, ( $g(x) := \|x\|_p$, $p \geq 1$).

- **Drawback:** ISTA appears to be a (very) slow method.
Advantage and Drawback of ISTA

- **Advantage:** Simplicity. Useful when $p_L(\cdot)$ can be computed analytically, e.g. when $g(\cdot)$ is separable, reduces to a one dimensional minimization problem, ($g(x) := \|x\|_p$, $p \geq 1$).

- **Drawback:** ISTA appears to be a (very) slow method.

◊ Convergence analysis of methods like ISTA has been well studied in past/recent literature under various contexts and frameworks, (Facchinei-Pang, Vol II, Chap. 12, 2003).
◊ The focus is on pointwise convergence of $\{x_k\}$ and *asymptotic* rate of convergence.
Advantage and Drawback of ISTA

- **Advantage**: Simplicity. Useful when $p_L(\cdot)$ can be computed analytically, e.g. when $g(\cdot)$ is separable, reduces to a one dimensional minimization problem, ($g(x) := \|x\|_p$, $p \geq 1$).

- **Drawback**: ISTA appears to be a (very) slow method.

- Convergence analysis of methods like ISTA has been well studied in past/recent literature under various contexts and frameworks, (Facchinei-Pang, Vol II, Chap. 12, 2003).
- The focus is on pointwise convergence of $\{x_k\}$ and asymptotic rate of convergence.

Here, we focus on the **nonasymptotic** global rate of convergence and efficiency measured through functions values.

A by-product of our analysis theoretically confirms the slow convergence rate:

$$F(x_k) - F(x^*) \simeq O(1/k),$$

namely ISTA, shares a **sublinear** global rate of convergence.
Can we devise a faster method than ISTA such that:
♠ The computational effort of the new method will keep the simplicity of ISTA
♠ Its global rate of convergence will be significantly better, theoretically and practically.
Can We Do Better to Solve the NSO $\min_x \{f(x) + g(x)\}$?

- Can we devise a faster method than ISTA such that:
  - ♠ The computational effort of the new method will keep the simplicity of ISTA
  - ♠ Its global rate of convergence will be significantly better, theoretically and practically.

- **Answer:** Yes, through an equally simple scheme

\[
\bullet \quad x_{k+1} = \arg\min_x Q_L(x, y_k), \quad \leftrightarrow \quad y_k \text{ instead of } x_k
\]

The new point $y_k$ will be smartly chosen and *easy* to compute.
Can We Do Better to Solve the NSO $\min_x \{ f(x) + g(x) \}$?

- Can we devise a faster method than ISTA such that:
  - ♠ The computational effort of the new method will keep the simplicity of ISTA
  - ♠ Its global rate of convergence will be significantly better, theoretically and practically.

- **Answer:** Yes, through an equally simple scheme

  $\spadesuit \quad x_{k+1} = \arg\min_{x} Q_{L}(x, y_k), \quad \leftarrow y_k$ instead of $x_k$

  The new point $y_k$ will be smartly chosen and easy to compute.

- **Idea:** From an algorithm (Nesterov 1983), designed for minimizing a smooth convex function, and proven to be an "optimal" first order method (Yudin-Nemirovsky (80).)
Can We Do Better to Solve the NSO $\min_x \{f(x) + g(x)\}$?

- Can we devise a faster method than ISTA such that:
  - ♠ The computational effort of the new method will keep the simplicity of ISTA
  - ♠ Its global rate of convergence will be significantly better, theoretically and practically.

- **Answer:** Yes, through an equally simple scheme

  $$\blacklozenge \quad x_{k+1} = \arg\min_x Q_L(x, y_k), \quad \leftrightarrow \quad y_k \text{ instead of } x_k$$

  The new point $y_k$ will be smartly chosen and **easy** to compute.

- **Idea:** From an algorithm (Nesterov 1983), designed for minimizing a **smooth** convex function, and proven to be an "**optimal**" first order method (Yudin-Nemirovsky (80)).

- But, here our problem (P) is **nonsmooth** !.. Yet, we derive a faster algorithm than ISTA for the general NSO problem (P), proven optimal. We call it **FISTA**..
An equally simple algorithm as ISTA. Here $L(f)$ is known.

**FISTA with constant stepsize**

**Input:** $L = L(f)$ - A Lipschitz constant of $\nabla f$.

**Step 0.** Take $y_1 = x_0 \in \mathbb{R}^n$, $t_1 = 1$.

**Step k.** ($k \geq 1$) Compute

\[
x_k = p_L(y_k), \quad \leftrightarrow \quad \text{main computation as ISTA}
\]

\[
• \quad t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2},
\]

\[
•• \quad y_{k+1} = x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1}).
\]

The requested additional computation for FISTA in ($•$) and ($••$) is clearly marginal.
Knowledge of $L(f)$ is not Necessary:

FISTA With Backtracking

**FISTA with backtracking**

**Step 0.** Take $L_0 > 0$, some $\eta > 1$ and $x_0 \in \mathbb{R}^n$. Set $y_1 = x_0$, $t_1 = 1$.

**Step k.** ($k \geq 1$) Find the smallest nonnegative integers $i_k$ such that with $i = i_k$, $\bar{L} = \eta^{i_k} L_{k-1}$:

$$F(p_{\bar{L}}(y_k)) \leq Q_{\bar{L}}(p_{\bar{L}}(y_k), y_k).$$

Set $L_k = \eta^{i_k} L_{k-1}$ and compute

$$x_k = p_{L_k}(y_k),$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2},$$

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

**Note:** FISTA can be easily extended to constrained convex NSO.
Lemma 1 (Well-Known) Let $f \in C^{1,1}_{L(f)}(\mathbb{R}^n)$. Then, for any $L \geq L(f)$,

$$f(x) \leq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2} \|x - y\|^2,$$

for every $x, y \in \mathbb{R}^n$. 
**Analysis: The 3 Pillars**

**Lemma 1** (Well-Known) Let $f \in C_{L(f)}^{1,1}(\mathbb{R}^n)$. Then, for any $L \geq L(f)$,

$$f(x) \leq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2} \|x - y\|^2,$$

for every $x, y \in \mathbb{R}^n$.

**Lemma 2** (A Key Inequality) Let $x, y \in \mathbb{R}^n$ and $L > 0$ such that $F(p_L(y)) \leq Q(p_L(y), y)$. Then

$$F(x) - F(p_L(y)) \geq \frac{L}{2} \|p_L(y) - y\|^2 + L \langle y - x, p_L(y) - y \rangle.$$
**Analysis: The 3 Pillars**

**Lemma 1** (Well-Known) Let $f \in C^{1,1}_{L(f)}(\mathbb{R}^n)$. Then, for any $L \geq L(f)$,

$$f(x) \leq f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2} \|x - y\|^2, \text{ for every } x, y \in \mathbb{R}^n.$$  

**Lemma 2** (A Key Inequality) Let $x, y \in \mathbb{R}^n$ and $L > 0$ such that $F(p_L(y)) \leq Q(p_L(y), y)$. Then

$$F(x) - F(p_L(y)) \geq \frac{L}{2} \|p_L(y) - y\|^2 + L \langle y - x, p_L(y) - y \rangle.$$  

**Lemma 3** (A Recursive Relation for Function Values) The sequences $\{x_k, y_k\}$ generated via FISTA satisfy for every $k \geq 1$

$$L_k^{-1} t_k^2 v_k - L_{k+1}^{-1} t_{k+1}^2 v_{k+1} \geq (\|u_{k+1}\|^2 - \|u_k\|^2)/2,$$

where $v_k := F(x_k) - F(x^*)$, $u_k := t_k x_k - (t_k - 1) x_{k-1} - x^*$. 

Marc Teboulle – p. 14
Let \( \{x_k\}, \{y_k\} \) be generated by FISTA. Then for any \( k \geq 1 \)

\[
F(x_k) - F(x^*) \leq \frac{2\alpha L(f)\|x_0 - x^*\|^2}{(k + 1)^2},
\]

where \( \alpha = 1 \) for the constant stepsize setting and \( \alpha = \eta \) for the backtracking stepsize setting.
Theorem – Global Rate of Convergence for FISTA

Let \( \{x_k\}, \{y_k\} \) be generated by FISTA. Then for any \( k \geq 1 \)

\[
F(x_k) - F(x^*) \leq \frac{2\alpha L(f)\|x_0 - x^*\|^2}{(k + 1)^2},
\]

where \( \alpha = 1 \) for the constant stepsize setting and \( \alpha = \eta \) for the backtracking stepsize setting.

The number of iterations of FISTA required to obtain an \( \varepsilon \)-optimal solution, that is an \( \tilde{x} \) such that:

\[
F(\tilde{x}) - F_* \leq \varepsilon,
\]

is at most \( \sim O(1/\sqrt{\varepsilon}) \). This clearly improves ISTA by a square root factor.
Theorem – Global Rate of Convergence for FISTA

Let \( \{x_k\}, \{y_k\} \) be generated by FISTA. Then for any \( k \geq 1 \)

\[
F(x_k) - F(x^*) \leq \frac{2\alpha L(f)\|x_0 - x^*\|^2}{(k + 1)^2},
\]

where \( \alpha = 1 \) for the constant stepsize setting and \( \alpha = \eta \) for the backtracking stepsize setting.

The number of iterations of FISTA required to obtain an \( \varepsilon \)-optimal solution, that is an \( \tilde{x} \) such that:

\[
F(\tilde{x}) - F_* \leq \varepsilon,
\]

is at most \( \sim O(1/\sqrt{\varepsilon}) \). This clearly improves ISTA by a square root factor.

Do we practically achieve this theoretical rate?
Numerical Examples: Image Deblurring

\[
\min_x \{ \|Ax - b\|^2 + \lambda \|x\|_1 \}
\]

Compare ISTA versus FISTA on

- A Simple Test Image from Regularization Tool (Hansen, (97))
- The Cameraman Test Image
- More Simulations

- Problems are in dimension \(d\) like \(d = 256 \times 256 = 65,536\), or/and \(512 \times 512 = 262,144\).
- The \(d \times d\) matrix \(A\) is dense.
- All problems solved with fixed \(\lambda\) and Gaussian noise.
Deblurring of A Simple Test Image

original

blurred and noisy
Output of 200 Iterations of ISTA versus 50 of FISTA

ISTA: \( F_{200} = 0.42 \)

FISTA: \( F_{50} = 0.23 \)

After tens of thousands of iterations, ISTA get stuck at \( F = 0.32 \)!
Deblurring of the Cameraman

original

blurred and noisy
1000 Iterations of ISTA versus 100 of FISTA

ISTA: 1000 Iterations

FISTA: 100 Iterations
Original Versus Deblurring via FISTA

Original

FISTA: 1000 Iterations
More Simulations

- Previous simulations indicate that practically FISTA seems to be able to reach accuracies that are beyond the capabilities of ISTA.

- We further tested this hypothesis on an example with known optimal solution.

- This simulation shows that the results of FISTA are better by several order of magnitudes. After 10000 iterations our method reaches accuracy of approximately $10^{-7}$ while ISTA reaches an accuracy of $10^{-3}$.

- Moreover, the value obtained by ISTA at iteration 10000 was already obtained by FISTA at iteration 254.

- The next figure describing function values of both methods for 10000 iterations speaks for itself!
Function Values errors $F(x_k) - F(x^*)$
Conclusions

- FISTA is a very simple and promising iterative scheme. Covers a broad class of problems arising in several recent diverse/key applications.

- Appears even faster than the proven predicted theoretical rate!

- Work in progress: potential for analyzing and designing faster algorithms in other areas, and with other types of nonsmooth regularizers.
Conclusions

- FISTA is a very simple and promising iterative scheme. Covers a broad class of problems arising in several recent diverse/key applications.

- Appears **even faster than the proven predicted theoretical rate!**

- Work in progress: potential for analyzing and designing faster algorithms in other areas, and with other types of nonsmooth regularizers.

Thank you for listening!