Review of "**A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring.**" by Beck, Amir, and Marc Teboulle. Publised in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009.

Direct problem are obviously easy to solve however linear inverse problems arises in a wide range of applications such as astrophysics, signal and image processing, statistical inference, and optics, to name just a few where we obtain the noisy output signals and we are interested to reconstruct the original signal or source. And objective function formulation for such problems is not only complex but non smooth due to need of regularization for a sparser solution.

Despite the presence of a nonsmoothed regularize in the objective function, the authors of this prover have proved that they can construct a faster algorithm than ISTA, called FISTA, that keeps its simplicity but shares the improved rate **$O(1/k^2)$** of the optimal gradient method devised earlier in Nesterov, Y. E. (1983) for minimizing smooth convex problems.

ISTA Formulation:

One regularization method that attracted a revived interest and considerable amount of attention in the signal processing literature is lasso (L-1) regularization in which we tend to minimize F(x):

$$\min_{\mathbf{x}}\{F(\mathbf{x}) \equiv \|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1\}$$

Here both terms in F(x) are convex making this a convex optimization problem.

However, $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$ is smooth but $\lambda\|\mathbf{x}\|_1\}$ is not however it has an easy proximal operator. Simply, we can always use sub gradient method to minimize. However, convergence will be in function values as **$O(1/\sqrt{k})$,** where k is the iteration counter which is not favorable. So, they use **proximal gradient to formulate** the problem which reduced the complexity to **$O(1/k)$** and call it as ISTA formulation.

After this paper was submitted for publication, they recently became aware of a very recent unpublished manuscript by Nesterov, Y. (2007). who has independently investigated a multistep version of an accelerated gradient-like method that solves the general problem model and, like FISTA, is proven to converge in function values as **$O(1/k^2)$.** As a consequence of the key differences between the building blocks and iterations of FISTA versus the new method of Nesterov, Y. (2007)., **the theoretical analysis and proof techniques developed here to establish the global rate convergence rate result** are completely different from that given in Nesterov, Y. (2007).

The remarkable fact is that the method developed in Nesterov, Y. (2007)., **does not require more than one gradient evaluation** at each iteration (namely, same as the gradient method), but just the computation of a **smartly chosen linear combination** of the two previous iterates. This was the main intuition behind FISTA. There was a momentum term (highlighted below) added in the equation of ISTA to control the convergence which was dependent on two of last outputs.

Algorithm-I: FISTA with constant step size

Input: $L = L(f)$ - A Lipschitz constant of ∇f.
Step 0. Take $\mathbf{y}_1 = \mathbf{x}_0 \in \mathbb{R}^n,\ t_1 = 1$.
Step k. $(k \geq 1)$ Compute

difficult to calculate each time

$$\mathbf{x}_k = \operatorname{prox}_{t_k}(g)\left(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)\right), \quad (9)$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad (10)$$

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}). \quad (11)$$

<--- smartly chosen linear combination.

Algorithm-II: FISTA with Backtracking

Step 0. Take $L_0 > 0$, some $\eta > 1$ and $\mathbf{x}_0 \in \mathbb{R}^n$.
Set $\mathbf{y}_1 = \mathbf{x}_0,\ t_1 = 1$.
Step k. $(k \geq 1)$ Find the smallest nonnegative integers i_k such that with $i = i_k$, $\bar{L} = \eta^{i_k} L_{k-1}$:

$$F(p_{\bar{L}}(\mathbf{y}_k)) \leq Q_{\bar{L}}(p_{\bar{L}}(\mathbf{y}_k), \mathbf{y}_k).$$

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

$$\mathbf{x}_k = p_{L_k}(\mathbf{y}_k),$$

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

momentum

$$\mathbf{y}_{k+1} = \mathbf{x}_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}_k - \mathbf{x}_{k-1}).$$

As an experiment of using FISTA, they have presented some preliminary numerical results for image deblurring problems, which demonstrate that FISTA can be even faster than the proven theoretical rate and can outperform ISTA by several orders of magnitude, showing the potential promise of FISTA.

Review of "Sparse reconstruction by separable approximation." By Wright, Stephen J., Robert D. Nowak, and Mário AT Figueiredo. Published in *IEEE Transactions on signal processing* 57.7 (2009)

Sparse reconstruction by separable approximation (SpaRSA) is another typical **proximity algorithm** based on sparse representation, which can be viewed as an accelerated version of ISTA (the reason why I choose this paper). SpaRSA provides a general algorithmic framework for solving the sparse representation problem and here a simple specific SpaRSA with adaptive continuation on ISTA is introduced.

One thing to note is that we had used constant regularization constant (λ) in FISTA method above. However, SpaRSA try to optimize the parameter λ by using the worm-starting technique, by choosing a more reliable approximation of $H_f(\alpha)$ (the Hessian matrix of f(α) at α^t in the second order Taylor series expansion) using Barzilai-Borwein (BB) spectral method.

Worm-starting technique to optimize λ

Hale et al. (2007) had concluded that the technique that exploits a decreasing value of λ from a warm-starting point can more efficiently solve the optimization problem than ISTA.

λ is calculated as:

$\lambda = \max\{\gamma \|X^T \boldsymbol{y}\|_\infty, \lambda\}$ where γ is a small constant.

Barzilai-Borwein (BB) spectral method to approximate H_f(α)

$H_f(\alpha)$ being complex to compute was estimated as:

- $\frac{1}{\tau} I$ in ISTA
- Lipschitz constant of ∇f(α) to replace $H_f(\alpha)$ in FISTA

SpaRSA utilizes the BB spectral method to choose the value of τ to mimic the Hessian matrix. The value of τ is required to satisfy the condition:

$$\frac{1}{\tau^{t+1}}(\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t) \approx \nabla f(\boldsymbol{\alpha}^{t+1}) - \nabla f(\boldsymbol{\alpha}^t)$$

which satisfies the minimization problem:

$$\begin{aligned}\frac{1}{\tau^{t+1}} &= \arg\min \|\frac{1}{\tau}(\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t) - (\nabla f(\boldsymbol{\alpha}^{t+1}) - \nabla f(\boldsymbol{\alpha}^t))\|_2^2 \\ &= \frac{(\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t)^T(\nabla f(\boldsymbol{\alpha}^{t+1}) - \nabla f(\boldsymbol{\alpha}^t))}{(\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t)^T(\boldsymbol{\alpha}^{t+1} - \boldsymbol{\alpha}^t)}\end{aligned}$$

SpaRSA requires that the value of λ is a decreasing sequence using **Worm-starting technique** and the value of τ should meet the condition as above.

Algorithm

We recall our initial problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min F(\boldsymbol{\alpha}) = \frac{1}{2}\|X\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

We initialize:

$$t = 0,\, i = 0,\, \boldsymbol{y}^0 = \boldsymbol{y},\, \frac{1}{\tau^0} I \approx H_f(\boldsymbol{\alpha}) = X^T X$$

Step1: calculate λ_t from **Worm-starting technique**

Step2:

$$\boldsymbol{\alpha}^{i+1} = shrink(\boldsymbol{\alpha}^i - \tau^i X^T(X^T\boldsymbol{\alpha}^t - \boldsymbol{y}), \lambda_t\tau^i).$$

Step3:
Update the value of $(1/\tau^{i+1})$ using **Barzilai-Borwein** (**BB) spectral method**

Step4:
Check for updated error bound.
If satisfied: continue to 5.
If not satisfied go to step2 updating i = i + 1

Step5:
Update y as:

$$y^{t+1} = y - X\alpha^{t+1}.$$

Step6:
If $\lambda_t = \lambda$, stop;
Otherwise, return to step 1 and t = t +1

Output: α^i

In their paper, they have reported a series of experiments which shows that SpaRSA has state of the art performance for the problems; other experiments described in that section illustrate that SpaRSA can handle a more general class of problems.

SpaRSA matches the speed of the state-of-the-art method when applied to the problem. They note that their computational experience shows relatively little difference in efficiency between monotone variants of SpaRSA and standard variants. This experience is at variance with problems in other domains, in which nonmonotone variants can have a large advantage, and they do not rule out the possibility that nonmonotonicity will play a more critical role in other instances of sparse reconstruction.