IHT dies hard: Provable accelerated Iterative Hard Thresholding

Rajiv Khanna, Anastasios Kyrillidis
rajivak@utexas.edu, anastasios.kyrillidis@ibm.com
TEXAS The University of Texas at Austin
IBM Research

Motivation

Goal: Study the momentum mechanics in the Iterative Hard Thresholding algorithm for non-convex sparsity constrained optimization.

L0 constrained minimization:

$$\min_{\|x\|_0 \leq k} f(x)$$

Main result: Linear convergence of Accelerated IHT under mild assumptions for general convex functions.

Background

Iterative Hard Thresholding (IHT): is a first order projected gradient descent optimization method that performs a non-convex projection onto the constraint set in each iteration.

$$x_{i+1} = \Pi_{k,\mathcal{A}}\left(x_i - \mu_i \nabla f(x_i)\right), \text{ where } \mu_i \in \mathbb{R}.$$

Momentum in first order methods: The acceleration in gradient descent type algorithms comes from adding a momentum term in the iterates.

$$x_{i+1} = \Pi_{k,\mathcal{A}}\left(u_i - \mu_i \nabla_{\mathcal{T}_i} f(u_i)\right)$$
$$u_{i+1} = x_{i+1} + \tau \cdot (x_{i+1} - x_i).$$

Restricted Strong Concavity (RSC)/Smoothness (RSM): A function $l()$ over subset Ω of its domain satisfies m_Ω-RSC/M_Ω-RSM if for all $\mathbf{X}, \mathbf{Y} \in \Omega$

$$-\frac{m_\Omega}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2 \geq \ell(\mathbf{Y}) - \ell(\mathbf{X}) - \langle \ell(\mathbf{X}), \mathbf{Y}-\mathbf{X}\rangle \geq -\frac{M_\Omega}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2$$

Prior Art:

- Acceleration in IHT guarantees convergence if the objective function decreases in each iteration [1], skips acceleration if that is not the case.
- Limited studies on acceleration parameter and convergence for special case of quadratic function [2,3]

Our results generalize previous studies [2,3] under milder assumptions than earlier [1] works.

Algorithm and its properties

We maintain two sequence of iterates and their respective supports, similar to classic accelerated gradient descent.

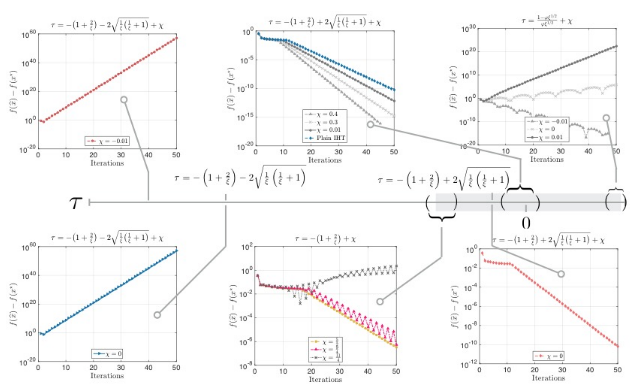
Algorithm 1 Accelerated IHT algorithm

1: **Input:** Tolerance η, T, $\alpha, \beta > 0$, model $\mathcal{A}$, $k \in \mathbb{Z}_+$.
2: **Initialize:** $x_0, u_0 \leftarrow 0$, $\mathcal{U}_0 \leftarrow \{\emptyset\}$. Set $\xi = 1 - \frac{\alpha}{\beta}$; select τ s.t. $|\tau| \leq \frac{1-\varphi\xi^{1/2}}{\varphi\xi^{1/2}}$, where $\varphi = \frac{1+\sqrt{5}}{2}$.
3: **repeat**
4: $\quad \mathcal{T}_i \leftarrow \text{supp}_{\mathcal{A}}\left(\Pi_{k,\mathcal{A}}\left(\nabla_{\mathcal{U}_i^c} f(u_i)\right)\right) \cup \mathcal{U}_i$
5: $\quad \bar{u}_i = u_i - \frac{1}{\beta}\nabla_{\mathcal{T}_i} f(u_i)$
6: $\quad x_{i+1} = \Pi_{k,\mathcal{A}}\left(\bar{u}_i\right)^{\dagger}$
7: $\quad u_{i+1} = x_{i+1} + \tau\left(x_{i+1} - x_i\right)$ where $\mathcal{U}_{i+1} \leftarrow \text{supp}_{\mathcal{A}}(u_{i+1})$
8: **until** $\|x_i - x_{i-1}\| \leq \eta\|x_i\|$ or after T iterations.
9: †*Optional*: Debias step on x_{i+1}, restricted on the support $\text{supp}_{\mathcal{A}}(x_{i+1})$.

Under certain assumptions the iterates in Algorithm 1 follow the recurrent relationship:-

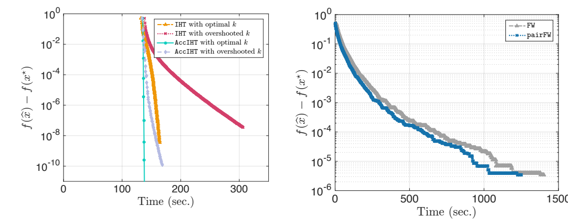
$$\begin{bmatrix}\|x_{i+1}-x^\star\|_2 \\ \|x_i - x^\star\|_2\end{bmatrix} \leq \begin{bmatrix}\left(1-\frac{\alpha}{\beta}\right)\cdot|1+\tau| & \left(1-\frac{\alpha}{\beta}\right)\cdot|\tau| \\ 1 & 0\end{bmatrix} \cdot \begin{bmatrix}\|x_i - x^\star\|_2 \\ \|x_{i-1}-x^\star\|_2\end{bmatrix}$$

From the above guarantee, it can be shown that Algorithm 1 guarantees an ε-approximate solution in $O\left(\log \frac{1-\alpha/\beta}{\varepsilon\cdot(|\lambda_1|-|\lambda_2|)}\right)$ iterations.



Experiments

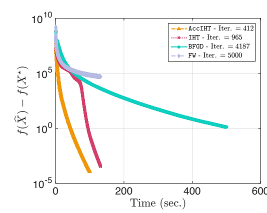
Sparse Linear Regression:



Group sparse, L2-norm regularized logistic regression

Algorithm	Test error	Time (sec)
FW [8]	0.2938	58.45
FW-Away [8]	0.2938	40.34
FW-Pair [8]	0.2938	38.22
IHT [14]	0.2825	5.24
Algorithm 1	0.2881	3.45

Low rank image completion from subset of entries



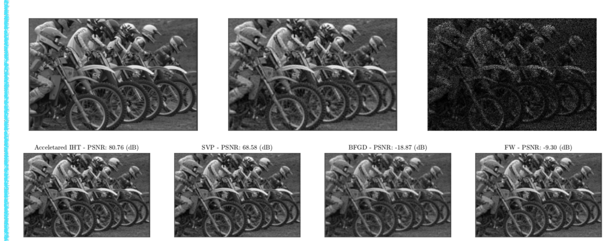


Figure: Reconstruction performance in image denoising settings.

References:
[1] T. Blumensath. Accelerated iterative hard thresh- olding. Signal Processing, 92(3):752–756, 2012
[2] A. Kyrillidis and V. Cevher. Matrix recipes for hard thresholding methods. Journal of mathemat- ical imaging and vision, 48(2):235–265, 2014.
[3] K. Wei. Fast iterative hard thresholding for com- pressed sensing. IEEE Signal Processing Letters, 22(5):593–597, 2015