On Iterative Hard Thresholding Methods for High-dimensional M-Estimation

Prateek Jain[*] **Ambuj Tewari**[†] **Purushottam Kar**[*]
[*]Microsoft Research, INDIA
[†]University of Michigan, Ann Arbor, USA
{prajain,t-purkar}@microsoft.com, tewaria@umich.edu

Abstract

The use of M-estimators in generalized linear regression models in high dimensional settings requires risk minimization with hard L_0 constraints. Of the known methods, the class of projected gradient descent (also known as iterative hard thresholding IHT) methods is known to offer the fastest and most scalable solutions. However, the current state-of-the-art is only able to analyze these methods in very restrictive settings which do not hold in high dimensional statistical models. In this work we bridge this gap by providing the first analysis of IHT-style methods in the high dimensional statistical setting. Our results rely on a general analysis framework that enables us to analyze several popular hard thresholding style algorithms (such as HTP, CoSaMP, SP) in the high dimensional regression setting.

1 Introduction

Modern statistical estimation is routinely faced with real world problems where the number of parameters p handily outnumbers the number of observations n. In general, consistent estimation of parameters is not possible in such a situation. Consequently, a rich line of work has focused on models that satisfy special structural assumptions such as sparsity or low-rank structure. Under these assumptions, several works (for example, see [1, 2, 3, 4, 5]) have established that consistent estimation is information theoretically possible in the "$n \ll p$" regime as well.

The question of *efficient* estimation, however, is faced with feasibility issues since consistent estimation routines often end-up solving NP-hard problems. Examples include sparse regression which requires loss minimization with sparsity constraints and low-rank regression which requires dealing with rank constraints which are not efficiently solvable in general [6]. Interestingly, recent works have shown that these hardness results can be avoided by assuming certain natural conditions over the loss function being minimized such as restricted strong convexity/smoothness (RSC/RSS) and the use of *convex relaxations* [5] or *greedy methods* [7, 8, 9].

Despite this, certain limitations have prevented widespread use of these techniques. Relaxation-based methods typically suffer from slow rates as they solve non-smooth optimization problems. Greedy methods are slow in situations with non-negligible sparsity or relatively high rank, owing to their incremental approach of adding/removing individual support elements. Instead, the methods that do find practical applications are projected gradient (PGD) methods, also called iterative hard thresholding (IHT) methods. These methods directly project onto the underlying (non-convex) feasible sets (which can be performed efficiently for several structures such as sparsity and low rank). However, traditional analyses for convex problems viz. [10] do not apply to these techniques.

An exception to the above is the recent work [11] which analyzes PGD with non-convex penalties such as SCAD, MCP and capped L_1. However, they are unable to handle commonly used used penalties such as L_0 or low-rank constraints.

Insufficiency of RIP based Guarantees for M-estimation. PGD/IHT-style methods have been very popular in literature for sparse recovery and several algortihms including Iterative Hard Thresholding (IHT) [12], GraDeS [13], Hard Thresholding Pursuit (HTP) [14], CoSaMP [15], Subspace Pursuit (SP) [16], and OMPR(ℓ) [17] have been proposed. However, the analysis of these algorithms has traditionally assumed the Restricted Isometry property (RIP) or incoherence property. It turns out that this renders these analyses inaccessible to high-dimensional statistical estimation problems.

All existing results analyzing these methods require the condition number of the loss function, restricted to sparse vectors, to be smaller than a universal constant. The best known such constant is due to [17] that requires a bound $\delta_{2k} \leq 0.5$ on the RIP constant (or equivalently a bound $\frac{1+\delta_{2k}}{1-\delta_{2k}} \leq 3$ on the condition number). In contrast, realistic high dimensional statistical settings, wherein pairs of variables can be arbitrarily correlated, routinely offer arbitrarily large condition numbers. In particular if two variates have a covariance matrix like $\begin{bmatrix} 1 & 1-\epsilon \\ 1-\epsilon & 1 \end{bmatrix}$, then the restricted condition number (on a support set of size just 2) of the sample matrix cannot be brought down below $1/\epsilon$ even with infinitely many samples. Consequently, when $\epsilon < 1/6$, none the existing results for hard thresholding methods offer *any* guarantees. Moreover, most of these analyses consider only the least squares objective. Although recent attempts have been made to extend this to general differentiable objectives [18, 19], the results continue to require that the restricted condition number be less than a universal constant and remain unsatisfactory in a statistical setting.

Overview of Results. Our main contribution in this work is an analysis of PGD/IHT-style methods in statistical settings. Our bounds are tight, achieve known minmax lower bounds [20], and hold for arbitrary differentiable, possibly even *non-convex*, functions. Our results hold even when the underlying condition number is arbitrarily large and only require the function to satisfy RSC/RSS conditions. In particular, this reveals that these iterative methods are indeed applicable to statistical settings, a result that escaped all previous works.

Our first result shows that the PGD/IHT methods achieve global convergence if used with a relaxed projection step. More formally, if the optimal parameter is s^*-sparse and the problem satisfies RSC and RSS constraints α and L respectively, then PGD methods offer global convergence so long as they employ projection to an s-sparse set where $s \geq 4(L/\alpha)^2 s^*$. This gives convergence rates that are identical to those of convex relaxation and greedy methods for the Gaussian sparse linear model.

We are able to instantiate our methods to a variety of statistical estimation problems such as sparse linear regression, generalized linear models and low rank matrix regression. Our results effortlessly extend to the noisy setting as a corollary and give bounds similar to those of [21] that relies on solving an L_1 regularized problem. Our proofs exploit the fact that even though hard-thresholding is not the prox-operator for any convex prox function, it still provides strong contraction when projection is performed onto sets of sparsity $s \gg s^*$. This observation is crucial and allows us to provide the first unified analysis for hard thresholding based gradient descent algorithms.

2 Problem Setup and Notations

High-dimensional Sparse Estimation. Given data points $X = [X_1, \ldots, X_n]^T$, where $X_i \in \mathbb{R}^p$, and the target $Y = [Y_1, \ldots, Y_n]^T$, where $Y_i \in \mathbb{R}$, the goal is to compute an s^*-sparse $\boldsymbol{\theta}^*$ s.t.,

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta}, \|\boldsymbol{\theta}\|_0 \leq s^*} f(\boldsymbol{\theta}). \tag{1}$$

Definition 1 (RSC/RSS Properties)**.** *A differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ is said to satisfy restricted strong convexity (RSC) (resp. restricted strong smoothness (RSS)) at sparsity level $s = s_1 + s_2$ with strong convexity constraint α_s (resp. strong convexity constraint L_s) if the following holds for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ s.t. $\|\boldsymbol{\theta}_1\|_0 \leq s_1$ and $\|\boldsymbol{\theta}_2\|_0 \leq s_2$:*

$$\frac{\alpha_s}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \leq f(\boldsymbol{\theta}_1) - f(\boldsymbol{\theta}_2) - \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_2)\rangle \leq \frac{L_s}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2$$

The above definitions seamlessly extend to the *low-rank matrix regression* problem where each data point is a matrix $X_i \in \mathbb{R}^{p_1 \times p_2}$ with the L_0 norm replaced by the rank function.

Algorithm 1 Iterative Hard-thresholding

```
**Input**: Function $f$ with gradient oracle, sparsity level $s$, step-size $\eta$
$\boldsymbol{\theta}^1 = \mathbf{0}, t = 1$
**while** *not converged* **do**
    $\boldsymbol{\theta}^{t+1} = P_s(\boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t)), t = t + 1$
**end while**
**Output**: $\boldsymbol{\theta}^t$
```

3 Iterative Hard-thresholding Method

We now study the popular iterative hard thresholding method for sparse regression (see Algorithm 1 for pseudocode). The projection operator $P_s(\boldsymbol{z})$ can be implemented efficiently in this case by projecting $\boldsymbol{z}$ onto the set of s-sparse vectors by selecting the s largest elements (in magnitude) of $\boldsymbol{z}$. Most analyses of hard thresholding methods use the projection property which implies that $\|P_s(\boldsymbol{z}) - \boldsymbol{z}\|_2^2 \leq \|\boldsymbol{\theta}' - \boldsymbol{z}\|_2^2$ for all $\|\boldsymbol{\theta}'\|_0 \leq s$. However, we show (see Lemma 1) that a much stronger result holds in case $\|\boldsymbol{\theta}'\|_0 \leq s^*$ and $s^* \ll s$. This forms a crucial part of all our analyses.

Lemma 1. *For any index set I, any $\boldsymbol{z} \in \mathbb{R}^I$, let $\boldsymbol{\theta} = P_s(\boldsymbol{z})$. Then for any $\boldsymbol{\theta}^* \in \mathbb{R}^I$ such that $\|\boldsymbol{\theta}^*\|_0 \leq s^*$, we have*

$$\|\boldsymbol{\theta} - \boldsymbol{z}\|_2^2 \leq \frac{|I| - s}{|I| - s^*} \|\boldsymbol{\theta}^* - \boldsymbol{z}\|_2^2.$$

Our analysis combines the above observation with the RSC/RSS properties of f to provide geometric convergence rates for the IHT procedure below.

Theorem 1. *Let f have RSC and RSS parameters given by $L_{2s+s^*}(f) = L$ and $\alpha_{2s+s^*}(f) = \alpha$ respectively. Let Algorithm 1 be invoked with f, $s \geq 32 \left(\frac{L}{\alpha}\right)^2 s^*$ and $\eta = \frac{2}{3L}$. Also let $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}, \|\boldsymbol{\theta}\|_0 \leq s^*} f(\boldsymbol{\theta})$. Then, the τ-th iterate of Algorithm 1, for $\tau = O(\frac{L}{\alpha} \cdot \log(\frac{f(\boldsymbol{\theta}^0)}{\epsilon}))$ satisfies:*

$$f(\boldsymbol{\theta}^\tau) - f(\boldsymbol{\theta}^*) \leq \epsilon.$$

We note that a result similar to Lemma 1 holds for matrix regression as well which can be used to show similar guarantees for the IHT method for low-rank matrix regression.

4 High Dimensional Statistical Estimation

We now show how results of the previous section can be instantiated in a variety of statistical estimation problems. Suppose we have a sample of data points $Z_{1:n}$ and a loss function $\mathcal{L}(\boldsymbol{\theta}; Z_{1:n})$ that depends on a parameter $\boldsymbol{\theta}$ and the sample. Then we can show the following result.

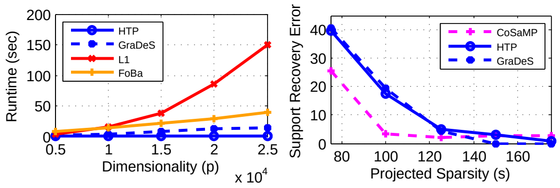
Theorem 2. *Let $\bar{\boldsymbol{\theta}}$ be any s^*-sparse vector. Suppose $\mathcal{L}(\boldsymbol{\theta}; Z_{1:n})$ is differentiable and satisfies RSC and RSS at sparsity level $s + s^*$ with parameters α_{s+s^*} and L_{s+s^*} respectively, for $s \geq 4 \left(\frac{L_{2s+s^*}}{\alpha_{2s+s^*}}\right)^2 s^*$. Let $\boldsymbol{\theta}^\tau$ be the τ-th iterate of Algorithm 1 for τ chosen as in Theorem 1 and ε be the function value error incurred by Algorithm 1. Then we have*

$$\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^\tau\|_2 \leq \frac{2\sqrt{s + s^*} \|\nabla \mathcal{L}(\bar{\boldsymbol{\theta}}; Z_{1:n})\|_\infty}{\alpha_{s+s^*}} + \sqrt{\frac{2\epsilon}{\alpha_{s+s^*}}}.$$

Note that the result does *not* require the loss function to be convex. This fact will be crucially used later. We now apply the above result to several statistical estimation scenarios.

Sparse Linear Regression. Here $Z_i = (X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ and $Y_i = \langle \bar{\boldsymbol{\theta}}, X_i \rangle + \xi_i$ where $\xi_i \sim \mathcal{N}(0, \sigma^2)$ is label noise with least squares as the loss function $\mathcal{L}(\boldsymbol{\theta}; Z_{1:n}) = \frac{1}{n}\|Y - X\boldsymbol{\theta}\|_2^2$. Suppose $X_{1:n}$ are drawn i.i.d. from a sub-Gaussian distribution with covariance Σ with $\Sigma_{jj} \leq 1\ \forall j$. [22, Lemma 6] can be used to obtain high probability RSC and RSS guarantees. Then we can show that with $n > 8c_1 s \log p / \sigma_{\min}(\Sigma)$ samples and $s = 324\kappa(\Sigma)^2 s^*$, we have, with high probability,

$$\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^\tau\|_2 \leq 145 \frac{\kappa(\Sigma)}{\sigma_{\min}(\Sigma)} \sigma \sqrt{\frac{s^* \log p}{n}} + 2\sqrt{\frac{\epsilon}{\sigma_{\min}(\Sigma)}}$$



(a) (b)

Figure 1: A comparison of IHT methods with L1 and greedy methods (FoBa) on sparse noisy linear regression tasks. 1(a) shows the variation in running times with increasing dimensionality p. 1(b) shows the recovery properties of different IHT methods under large condition number ($\kappa = 50$) setting as the size of projected set is increased to $s \gg s^*$.

Noisy and Missing Data. We now look at cases with feature noise as well. More specifically, assume that we only have access to $\tilde{X}_i$'s that are corrupted versions of X_i's. Two models of noise are popular in literature [21] a) (*additive noise*) $\tilde{X}_i = X_i + W_i$ where $W_i \sim \mathcal{N}(\mathbf{0}, \Sigma_W)$, and b) (*missing data*) $\tilde{X}$ is a $\mathbb{R} \cup \{\star\}$-valued matrix obtained by independently, with probability $\nu \in [0, 1)$, replacing each entry in X with $\star$. For the case of additive noise (missing data can be handled similarly), $Z_i = (\tilde{X}_i, Y_i)$ and $\mathcal{L}(\boldsymbol{\theta}; Z_{1:n}) = \frac{1}{2}\boldsymbol{\theta}^T\hat{\Gamma}\boldsymbol{\theta} - \hat{\gamma}^T\boldsymbol{\theta}$ where $\hat{\Gamma} = \tilde{X}^T\tilde{X}/n - \Sigma_W$ and $\hat{\gamma} = \tilde{X}^T Y/n$ are unbiased estimators of Σ and $\Sigma^T\bar{\boldsymbol{\theta}}$ respectively. [21, Lemma 1] can be used to obtain high probability RSC and RSS guarantees.

Note that $\mathcal{L}(\cdot; Z_{1:n})$ is *non-convex* but we can still apply Theorem 2 because RSC, RSS hold. Then, it is possible to show that if $n \geq c_1 s\tau(p)/\sigma_{\min}(\Sigma)$ and $s = 196\kappa(\Sigma)^2 s^*$, then with high probability,

$$\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^\tau\|_2 \leq c_2 \frac{\kappa(\Sigma)}{\sigma_{\min}(\Sigma)} \tilde{\sigma} \|\bar{\boldsymbol{\theta}}\|_2 \sqrt{\frac{s^* \log p}{n}} + 2\sqrt{\frac{\epsilon}{\sigma_{\min}(\Sigma)}},$$

where c_0, c_1, c_2 are universal constants, $\tilde{\sigma} = \sqrt{\|\Sigma_W\|_{\text{op}}^2 + \|\Sigma\|_{\text{op}}^2}(\|\Sigma_W\|_{\text{op}} + \sigma)$ and $\tau(p) = c_0\sigma_{\min}(\Sigma)\max(\frac{(\|\Sigma\|_{\text{op}}^2 + \|\Sigma_W\|_{\text{op}}^2)^2}{\sigma_{\min}^2(\Sigma)}, 1)\log p$.

5 Discussion

We showed that iterative hard thresholding algorithms can provably minimize arbitrary, possibly non-convex, differentiable objective functions assuming Restricted Strong Convexity/Smoothness (RSC/RSM) conditions. Our basic insight was to relax the stringent RIP requirement popular in literature by running these iterative algorithms with an enlarged support size. Our theoretical results put hard thresholding methods on par with those based on convex relaxation or greedy algorithms.

The results in this paper, along with full proofs, appear in [23]. In particular, [23] shows that our results can also be extended to a family of "fully corrective" methods that includes CoSaMP [15], Subspace Pursuit (SP) [16], and OMPR(ℓ) [17]. These methods keep the optimization objective fully minimized over the support of the current iterate. These algorithms have also thus far been analyzed only under RIP guarantees for the least squares objective. Using our analysis framework developed in the previous sections, we can arrive at a generic RSC-based analysis for general two-stage methods for arbitrary loss functions. Our analysis hinges upon a key observation specific to these methods that bound the error incurred by the hard thresholding step in these methods.

[23] also presents experimental evidence supporting theoretical claims made here (see Figure 1). In particular, using experiments on ill-conditioned problems, with the condition number as large as 50, we demonstrate that various IHT-style algorithms such as CoSaMP, HTP and GraDeS offer remarkably improved performance as the the projected sparsity levels s were increased beyond s^*. Experimental results also demonstrated that IHT-style methods such as HTP and GraDeS were much more scalable as compared to L_1 and greedy methods such as FoBa [24].

In future work, it would be interesting to generalize our algorithms and their analyses to more general structures. A unified analysis for general structures will probably create interesting connections with existing unified frameworks such as those based on decomposability [5] and atomic norms [25].

References

[1] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

[2] Sahand Negahban, Martin J Wainwright, et al. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

[3] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *Information Theory, IEEE Transactions on*, 57(10):6976–6994, 2011.

[4] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[5] Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, Bin Yu, et al. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[6] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[7] Ji Liu, Ryohei Fujimaki, and Jieping Ye. Forward-backward greedy algorithms for general convex smooth functions over a cardinality constraint. In *Proceedings of The 31st International Conference on Machine Learning*, pages 503–511, 2014.

[8] Ali Jalali, Christopher C Johnson, and Pradeep D Ravikumar. On learning discrete graphical models using greedy methods. In *NIPS*, pages 1935–1943, 2011.

[9] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[10] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Springer, 2004.

[11] P. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima, 2013. arXiv:1305.2436 [math.ST].

[12] Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265 – 274, 2009.

[13] Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[14] Simon Foucart. Hard thresholding pursuit: an algorithm for compressive sensing. *SIAM J. on Num. Anal.*, 49(6):2543–2563, 2011.

[15] Deanna Needell and Joel A. Tropp. CoSaMP: Iterative Signal Recovery from Incomplete and Inaccurate Samples. *Appl. Comput. Harmon. Anal.*, 26:301–321, 2008.

[16] Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Inf. Theory*, 55(5):22302249, 2009.

[17] Prateek Jain, Ambuj Tewari, and Inderjit S. Dhillon. Orthogonal matching pursuit with replacement. In *Annual Conference on Neural Information Processing Systems*, 2011.

[18] Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. Greedy sparsity-constrained optimization. *The Journal of Machine Learning Research*, 14(1):807–841, 2013.

[19] Xiaotong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *Proceedings of The 31st International Conference on Machine Learning*, 2014.

[20] Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. arXiv:1402.1918, 2014.

[21] P. Loh and M. J. Wainwright. High-dimension regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

[22] Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452—2482, 2012.

[23] A. Anonymous. On Iterative Hard Thresholding Methods for High-dimensional M-Estimation, 2014.

[24] Tong Zhang. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Trans. Inf. Theory*, 57:4689–4708, 2011.

[25] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.