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

Prateek Jain*, Ambuj Tewari†, and Purushottam Kar*

*Microsoft Research, INDIA
†University of Michigan, Ann Arbor, USA

Microsoft Research

UNIVERSITY OF MICHIGAN

The Goal

Analyze a class of effective and scalable iterative methods for high-dimensional statistical estimation problems.

High-dimensional M-estimation

Example: Sparse least squares regression
Given: n samples $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, $y_i \approx \langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i \rangle$ where $\bar{\boldsymbol{\theta}}$ is sparse
Task: Recover a sparse $\boldsymbol{\theta}^{\text{est}} \in \mathbb{R}^p$ such that $\boldsymbol{\theta}^{\text{est}} \approx \bar{\boldsymbol{\theta}}$

Points to note:
- Severely under-specified problem $n \ll p$
- Model sparsity $\|\bar{\boldsymbol{\theta}}\|_0 = s^* \ll p$

The good news:
- Consistent estimation possible with structural assumptions
  - Sparsity, low rank
- Poly-time estimation routines assuming RSC/RSS
  - Convex relaxations (LASSO), greedy methods

The not-so-good news:
- The above estimation routines do not scale at all!
  - Convex relaxations: non-smooth ⇒ slow rates
  - Greedy methods: incremental approach ⇒ slow progress

Setting the Stage

Given data samples, sparse estimation can be formulated as

$$\boldsymbol{\theta}^* = \underset{\|\boldsymbol{\theta}\|_0 \le s^*}{\arg\min}\ f(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}; \mathbf{z}_{1:n}) \qquad (1)$$

Examples: (label noise: $\xi_i \sim \mathcal{N}(0, \sigma^2)$)

1. Sparse LS regression: $y_i = \langle \bar{\boldsymbol{\theta}}, \mathbf{x}_i \rangle + \xi_i$, $\mathbf{x}_i \sim \mathcal{N}(\bar{\mathbf{x}}, \Sigma)$, $\|\bar{\boldsymbol{\theta}}\|_0 \le s^*$
$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{z}_{1:n}) = \frac{1}{n}\sum (y_i - \langle \mathbf{x}_i, \boldsymbol{\theta} \rangle)^2$$
2. Regression with feature noise: feature noise can be
   - additive: $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{w}_i$ with $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \Sigma_W)$
   - obliterative: $\tilde{\mathbf{x}}_i = \mathbf{x}_i$ w.p. $1 - \nu$ and $*$ otherwise

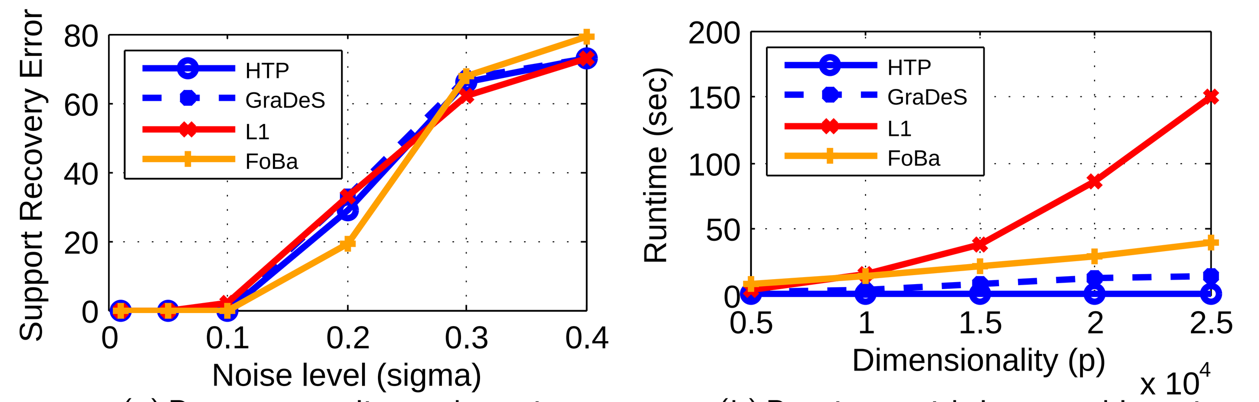
   Let $\hat{\Gamma} = \tilde{X}^\top \tilde{X}/n - \Sigma_W$ and $\hat{\gamma} = \tilde{X}^\top Y/n$
$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{z}_{1:n}) = \frac{1}{2}\boldsymbol{\theta}^\top \hat{\Gamma} \boldsymbol{\theta} - \hat{\gamma}^\top \boldsymbol{\theta}$$
   Note: the above is non-convex for $n \ll p$
3. Low-rank matrix regression: $y_i = \text{tr}(\overline{W} X_i^T) + \xi_i$, rank$(\overline{W}) = s^*$
$$\mathcal{L}(W; Z_{1:n}) = \frac{1}{n}\sum (y_i - \text{tr}(W X_i^T))^2$$

Iterative Hard Thresholding-style Methods

- Family of projected gradient descent-style methods
- Take gradient step along $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$ and project onto feasible set
  - Sparsity: $P_s(\boldsymbol{z})$: take s-largest elements of $\boldsymbol{z}$ by magnitude
  - Low rank: $PM_s(W)$: take top-s singular components of W
- Very popular, methods of choice for large-scale applications
  - IHT, GraDeS, HTP, CoSaMP, SP, OMPR(ℓ), . . .

IHT Methods in Practice

- Give comparable recovery quality as L_1 or greedy
- Much more scalable than L_1, greedy methods



(a) Recovery quality under noise (b) Runtimes with large problem sizes

Challenges ...

- Current analyses deficient in analyzing statistical models

Restricted Strong Convexity/Smoothness
A function f satisfies RSC/RSS with constants α_{2s} and L_{2s} if for all $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2$ such that $\|\boldsymbol{\theta}^1\|_0, \|\boldsymbol{\theta}^2\|_0 \le s$, we have $\frac{\alpha_{2s}}{2}\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_2^2 \le f(\boldsymbol{\theta}^1) - f(\boldsymbol{\theta}^2) - \langle \boldsymbol{\theta}^1 - \boldsymbol{\theta}^2, \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^2) \rangle \le \frac{L_{2s}}{2}\|\boldsymbol{\theta}^1 - \boldsymbol{\theta}^2\|_2^2$

- All known bounds require $\boldsymbol{\kappa} = L_{2s}/\alpha_{2s} <$ constant
  - For LS objective, this reduces to the RIP condition
  - Best known constant $\boldsymbol{\kappa} < 3$ (or $\delta_{2s} < 0.5$) due to OMPR(ℓ)
  - Completely silent otherwise
- Assumption untrue: practical settings exhibit large $\boldsymbol{\kappa}$
$$\Sigma_X = \begin{bmatrix} 1 & 1-\epsilon \\ 1-\epsilon & 1 \end{bmatrix}$$

Note: even with infinite samples, $\boldsymbol{\kappa} = \Omega(1/\epsilon)$

The Big Question

Can we show provable recovery guarantees for popular IHT-style methods under statistical settings with high condition numbers?

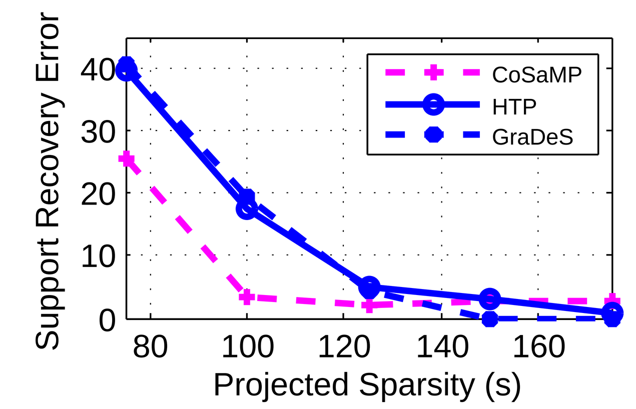
Iterative Hard-thresholding

Includes algorithms such as IHT, GraDeS

Algorithm 1 (IHT)
1. while not converged
2. $\quad \boldsymbol{\theta}^{t+1} \leftarrow P_s(\boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t))$

Theorem: **IHT** guarantees $f(\boldsymbol{\theta}^\tau) - f(\boldsymbol{\theta}^*) \le \epsilon$ for $\tau \ge \frac{L_{2s}}{\alpha_{2s}} \log \frac{1}{\epsilon}$.

Proof Idea: Key idea is to use a *relaxed* projection $s \ge \left(\frac{L}{\alpha}\right)^2 s^*$.



Note: on large $\boldsymbol{\kappa} = 50$ problems, relaxed projection really helps

Crucial: $P_s(\cdot)$ provides strong contraction if $s \gg s^*$. If $\boldsymbol{\theta} = P_s(\mathbf{z})$,
$$\|\boldsymbol{\theta} - \mathbf{z}\|_2^2 \le \frac{p-s}{p-s^*}\|\boldsymbol{\theta}^* - \mathbf{z}\|_2^2$$

Two-stage Hard-thresholding

Includes algorithms such as CoSaMP, Subspace pursuit

Algorithm 2 (TsHT)
1. while not converged
2. $\quad \mathbf{g}^t \leftarrow \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^t), S^t \leftarrow supp(\boldsymbol{\theta}^t)$
3. $\quad \boldsymbol{\beta}^t \leftarrow FC\left(f; S^t \cup \{\text{largest } \ell \text{ elements of } |\mathbf{g}^t_{\bar{S}^t}|\}\right)$
4. $\quad \mathbf{z}^t \leftarrow P_s(\boldsymbol{\beta}^t)$
5. $\quad \boldsymbol{\theta}^{t+1} \leftarrow FC(f; supp(\mathbf{z}^t))$

- Utilizes a fully corrective step
$$FC(f; S) = \arg\min_{supp(\boldsymbol{\theta}) \subseteq S} f(\boldsymbol{\theta})$$
- Similar convergence bounds as **IHT** - better constants
- Key idea 1: large distance from optima implies a large gradient
$$\|\mathbf{g}^t_{S^t \cup S^*}\| \ge 2\alpha_{2s}(f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*)) + \alpha_{2s}^2\|\boldsymbol{\theta}^t_{S^t \setminus S^*}\|$$
- Key idea 2: projection doesn't undo progress made by $FC(\cdot)$
$$f(\mathbf{z}^t) - f(\boldsymbol{\beta}^t) \le \frac{L_{2s}}{\alpha_{2s}} \cdot \frac{\ell}{s + \ell - s^*} \cdot (f(\boldsymbol{\beta}^t) - f(\boldsymbol{\theta}^*))$$
- Analyze Partial Hard-thresholding methods OMPR(ℓ) as well

Guarantees for High Dimensional Statistical Estimation

Theorem: If $\boldsymbol{\theta}^{\text{est}}$ is an ϵ_{opt}-optimal solution to (1), then
$$\|\boldsymbol{\theta}^{\text{est}} - \bar{\boldsymbol{\theta}}\|_2 \le \frac{\sqrt{s+s^*}\,\|\nabla_{\boldsymbol{\theta}} \mathcal{L}(\bar{\boldsymbol{\theta}}; \mathbf{z}_{1:n})\|_\infty}{\alpha_{s+s^*}} + \sqrt{\frac{\epsilon_{\text{opt}}}{\alpha_{s+s^*}}}$$

Proof Idea: IHT results, RSC/RSS and Hölder's inequality

- Results hold even for non-convex $\mathcal{L}(\cdot)$
  - Only RSC and RSS need to hold
  - Essential for noisy regression models

	Sparse LS regression	Regression with feature noise
RSC (α_k)	$\frac{\sigma_{\min}(\Sigma)}{2} - \frac{k \log p}{n}$	$\frac{\sigma_{\min}(\Sigma)}{2} - \frac{k\tau(p)}{n}$
RSS (L_k)	$2\sigma_{\max}(\Sigma) + \frac{k \log p}{n}$	$\frac{3\sigma_{\max}(\Sigma)}{2} + \frac{k\tau(p)}{n}$
$\|\nabla \mathcal{L}(\cdot)\|_\infty$	$\sigma\sqrt{\frac{\log p}{n}}$	$\tilde{\sigma}\|\bar{\boldsymbol{\theta}}\|_2\sqrt{\frac{\log p}{n}}$
$\|\boldsymbol{\theta}^{\text{est}} - \bar{\boldsymbol{\theta}}\|_2$	$\frac{\boldsymbol{\kappa}(\Sigma)}{\sigma_{\min}(\Sigma)}\sigma\sqrt{\frac{s^* \log p}{n}} + \sqrt{\frac{\epsilon_{\text{opt}}}{\sigma_{\min}(\Sigma)}}$	$\frac{\boldsymbol{\kappa}(\Sigma)}{\sigma_{\min}(\Sigma)}\tilde{\sigma}\|\bar{\boldsymbol{\theta}}\|_2\sqrt{\frac{s^* \log p}{n}} + \sqrt{\frac{\epsilon_{\text{opt}}}{\sigma_{\min}(\Sigma)}}$

$^*\tau(p) = \log p \cdot \frac{(\|\Sigma\|^2 + \|\Sigma_W\|^2)^2}{\sigma_{\min}(\Sigma)}$

$^{**}\tilde{\sigma} = (\|\Sigma_W\| + \sigma)\sqrt{\|\Sigma\|^2 + \|\Sigma_W\|^2}$