Robust Regression via Hard Thresholding

Kush Bhatia† Prateek Jain† Purushottam Kar‡

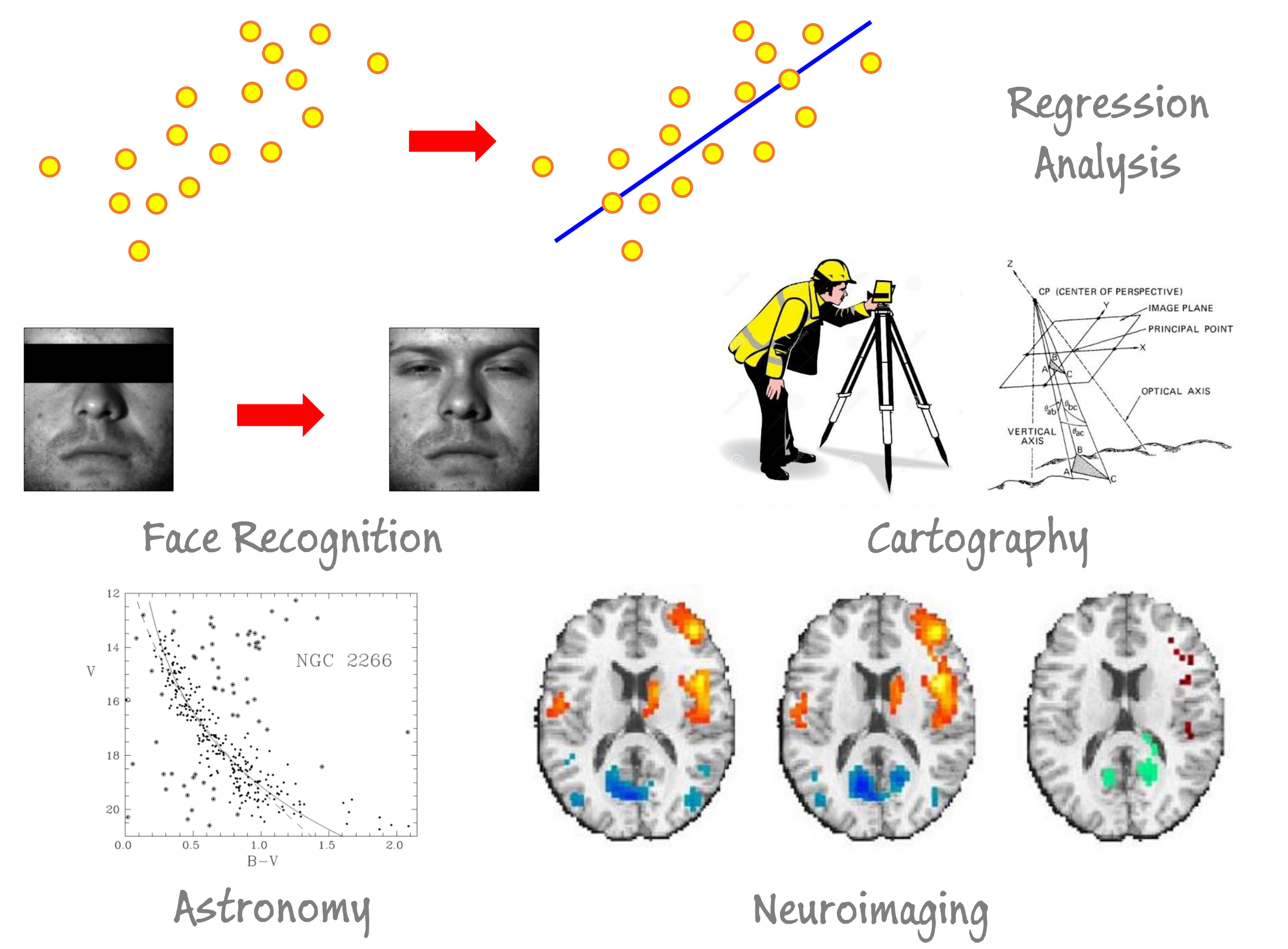
†Microsoft Research, Bengaluru, India ‡Indian Institute of Technology Kanpur, India

Microsoft Research

Full Paper: http://tinyurl.com/robustreg

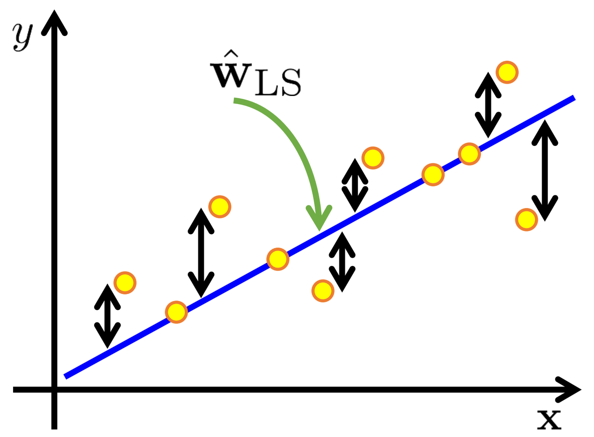
Goal: Perform statistical analysis of large-scale data in the presence of unbounded adversarial corruptions

Statistical Analysis with Corrupted Data



Ordinary Least Squares (OLS) Regression

Discovers a curve that "best fits" the entire data



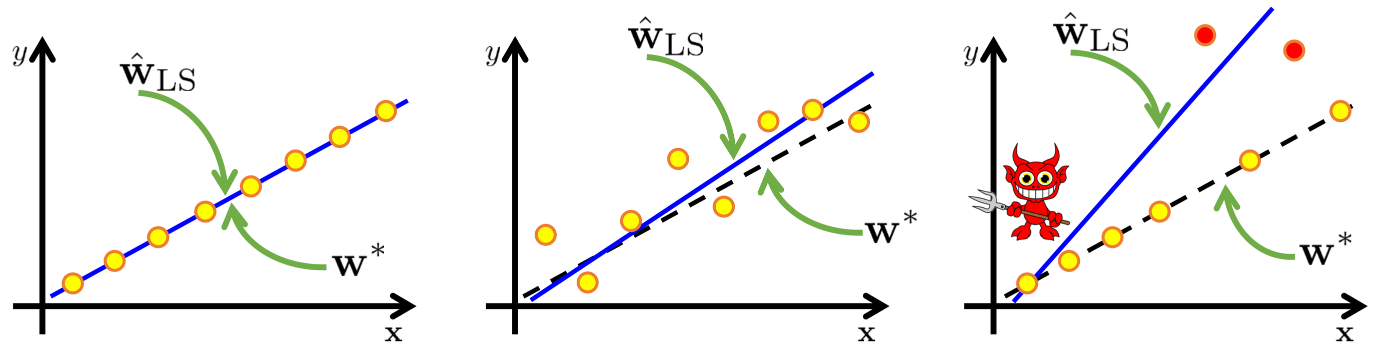
Data: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$

Response: $y_1, y_2, \ldots, y_n \in \mathbb{R}$

$$\hat{\mathbf{w}}_{\mathrm{LS}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$

Noise Model: $y_i = \langle \mathbf{w}^*, \mathbf{x}_i \rangle + b_i$

Favorable cases **Failure cases**



Exact recovery in noiseless case | Faithful recovery with white noise | Catastrophic failure under adversity

Problem Formulation: Recover the original curve in the face of a bounded number of adversarial corruptions

$$\arg\min_{\mathbf{w}, S} \sum_{i \in S} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2, \qquad |\bar{S}| \le \alpha \cdot n$$

Some existing approaches

Brute Force — *Infeasible*

Try out all possible sets S and estimate $\mathbf{w}$ using each set

Random Selection (RANSAC) — *Inefficient*

Try random sets S, estimate $\mathbf{w}$ using each and choose best

Least Median of Squares — *Expensive*

$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \operatorname{med}_{i} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$

Nguyen and Tran, 2013 — *Oblivious Adversary*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}, \mathbf{b}} \|\mathbf{w}\|_1 + \|\mathbf{b}\|_1, \text{ s.t., } X^\top \mathbf{w} + \mathbf{b} = \mathbf{y}$$

Chen et al. 2013 — *Weak Recovery Guarantees*

Trimmed inner product versions of Thresholding Regression, Lasso, and Dantzig selector

McWilliams et al., 2014

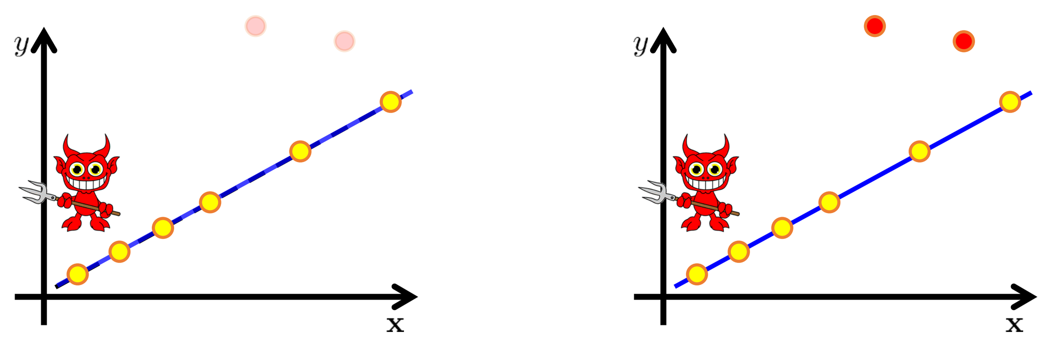
Influence and Residual weighted sub-sampling algorithms

KEY IDEA

Two unknowns: clean set of points S^*, original curve $\mathbf{w}^*$

Observation 1: given S^*, finding original curve $\mathbf{w}^*$ easy

Observation 2: given $\mathbf{w}^*$, finding clean points S^* easy



Proposal: can we alternate between estimating S^* and $\mathbf{w}^*$?

TORRENT

1. Start with any arbitrary curve $\mathbf{w}^0$ and set timer $t \leftarrow 0$
2. Repeat until convergence
   i. $r_i = |y_i - x_i^\top w^{t-1}|$ for all points
   ii. Update $S_t \leftarrow$ Points with minimum r_i
   iii. $\mathbf{w}^t \leftarrow$ UPDATE using S_t
   iv. Increment time counter $t \leftarrow t+1$

Thresholding Operator-based Robust RegrEssioN meThod

TORRENT-Variants

UPDATE TORRENT-FC

$$\mathbf{w}_{t+1} \leftarrow \arg\min_{\mathbf{w}} \sum_{i \in S_t} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$

UPDATE TORRENT-GD

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta X_{S_t} (X^\top \mathbf{w} - y_{S_t})$$

UPDATE TORRENT-HYB

If $|S_t \backslash S_{t-1}| \ge \Delta$

$\mathbf{w}_{t+1} \leftarrow$ UPDATE TORRENT-GD

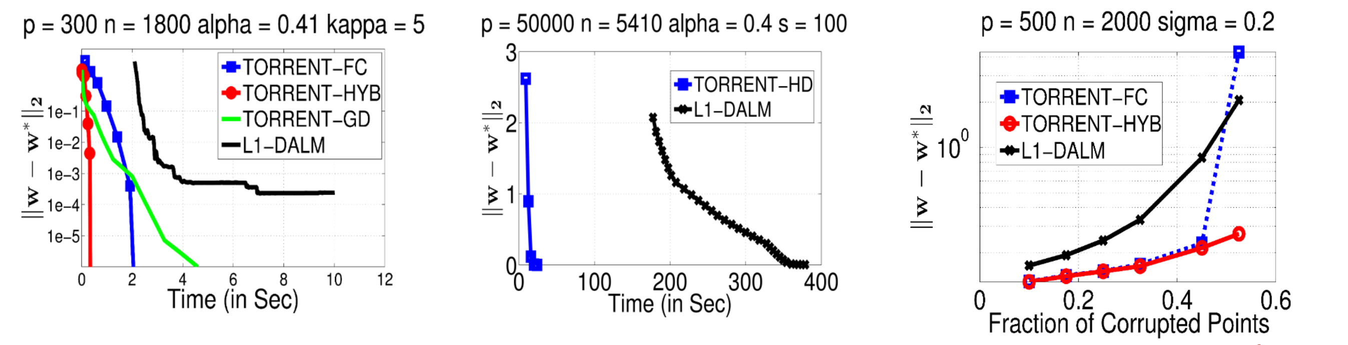
Else

$\mathbf{w}_{t+1} \leftarrow$ UPDATE TORRENT-FC

UPDATE TORRENT-HD

$$\mathbf{w}_{t+1} \leftarrow \inf_{\|w\|_0 \le s} \sum_{i \in S_t} (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2$$

Experimental Results



On regression analysis tasks, TORRENT is up to 20x faster than leading methods on low, as well as high dimensional data and can tolerate up to 40% corruption!

On face recognition tasks, TORRENT is able to recover the correct identity of the person in the presence of as much as 70% corruption!

Design Properties

Subset Strong Convexity (SSC) & Subset Strong Smoothness (SSS)

Definition: A matrix X satisfies the SSC Property (*resp.* SSS Property) at level γ with strong convexity constant λ_γ (*resp. strong smoothness constant* Λ_γ) if the following holds:

$$\lambda_\gamma \le \min_{S \in \mathcal{S}_\gamma} \lambda_{\min}(X_S X_S^\top) \le \max_{S \in \mathcal{S}_\gamma} \lambda_{\max}(X_S X_S^\top) \le \Lambda_\gamma.$$

If X has columns sampled i.i.d. from $\mathcal{N}(\mathbf{0}, I)$, w.h.p.

$$\Lambda_\gamma^{\mathrm{Gauss}} \le \gamma n \left(1 + \sqrt{\log \tfrac{1}{\gamma}}\right) + o(np)$$

$$\lambda_\gamma^{\mathrm{Gauss}} \ge n - (1-\gamma) n \left(1 + \sqrt{\log \tfrac{1}{1-\gamma}}\right) - o(np)$$

Noise Model

$$\mathbf{y} = X\mathbf{w}^* + \mathbf{b} + \boldsymbol{\eta}$$

- $\boldsymbol{\eta}$: bounded dense noise
- $\mathbf{b}$: sparse adversarial corruption, $\|\mathbf{b}\|_0 \le \alpha \cdot n$, chosen in a fully adaptive manner

Theoretical Guarantees

Low-Dimensional Setting

Theorem 1 (TORRENT-FC, $\eta = 0$): If X satisfies *SSC* and *SSS* at level γ with constants λ_γ and Λ_γ such that $\frac{(1+\sqrt{2})\Lambda_\alpha}{\lambda_{1-\alpha}} < 1$, then after $\log\frac{1}{\epsilon}$ iterations,

$$\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \epsilon$$

If each $x_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $\alpha \le \frac{1}{65}$ and $n \ge \Omega(p \log p)$, then w.h.p. $\frac{(1+\sqrt{2})\Lambda_\alpha}{\lambda_{1-\alpha}} < 0.9$

Theorem 2 (TORRENT-FC, $\eta \ne 0$): If X satisfies *SSC* and *SSS* at level γ with constants λ_γ and Λ_γ such that $\frac{4\sqrt{\Lambda_\alpha}}{\sqrt{\lambda_{1-\alpha}}} < 1$, then after $\log\frac{1}{\epsilon}$ iterations,

$$\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \epsilon + C\frac{\|\eta\|_2}{\sqrt{n}}$$

Similar convergence results proven for TORRENT-GD and TORRENT-HYB

High-Dimensional Setting

If X satisfies "restricted" equivalents of *SSC* and *SSS*, $n \ge \Omega(s\kappa \log p)$, with each $x_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ *and* $\alpha \le \frac{1}{65}$, similar convergence guarantees can be proven for TORRENT-HD