Robust high-dimensional linear regression: A statistical perspective

Po-Ling Loh

University of Wisconsin - Madison
Departments of ECE & Statistics

STOC workshop on robustness and nonconvexity
Montreal, Canada

June 23, 2017

Introduction: Robust regression

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)

Introduction: Robust regression

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations

Introduction: Robust regression

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations
- Local stability captured by *influence function*

$$IF(x; T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t}$$

Introduction: Robust regression

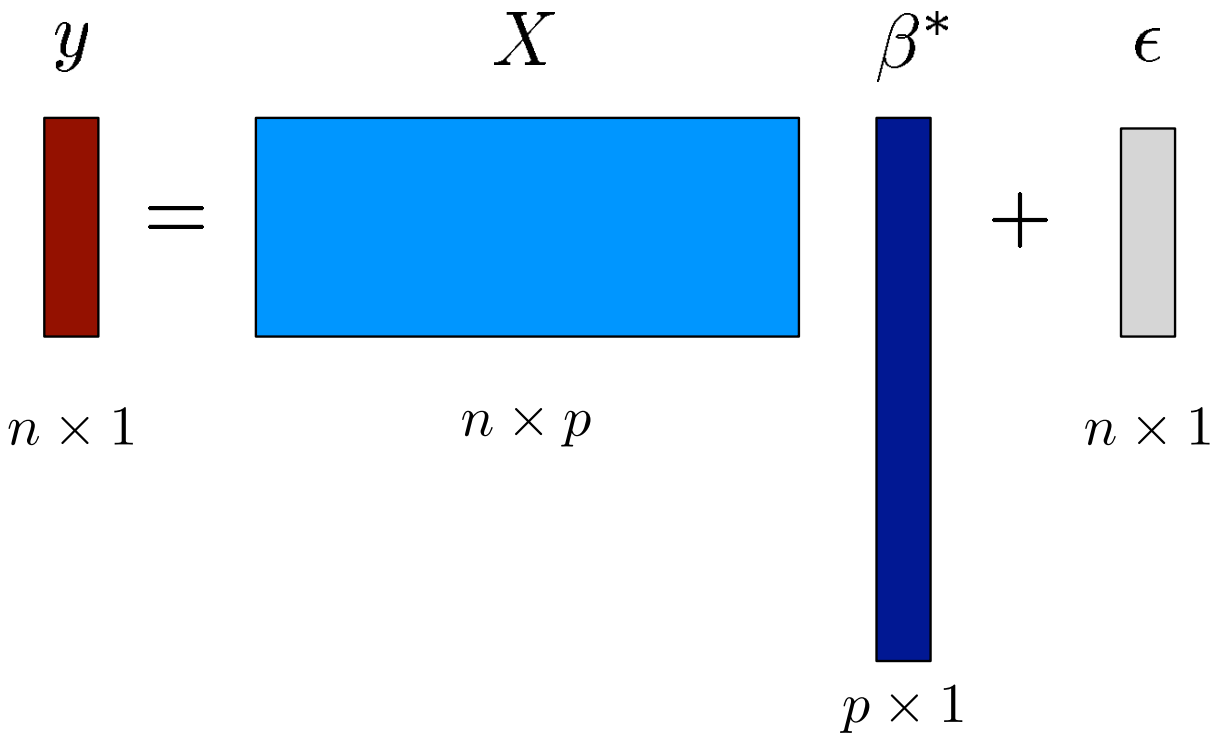
- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations
- Local stability captured by *influence function*

$$IF(x; T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t}$$

- Global stability captured by *breakdown point*

$$\epsilon^*(T; X_1, \ldots, X_n) = \min \left\{ \frac{m}{n} : \sup_{X^m} \| T(X^m) - T(X) \| = \infty \right\}$$

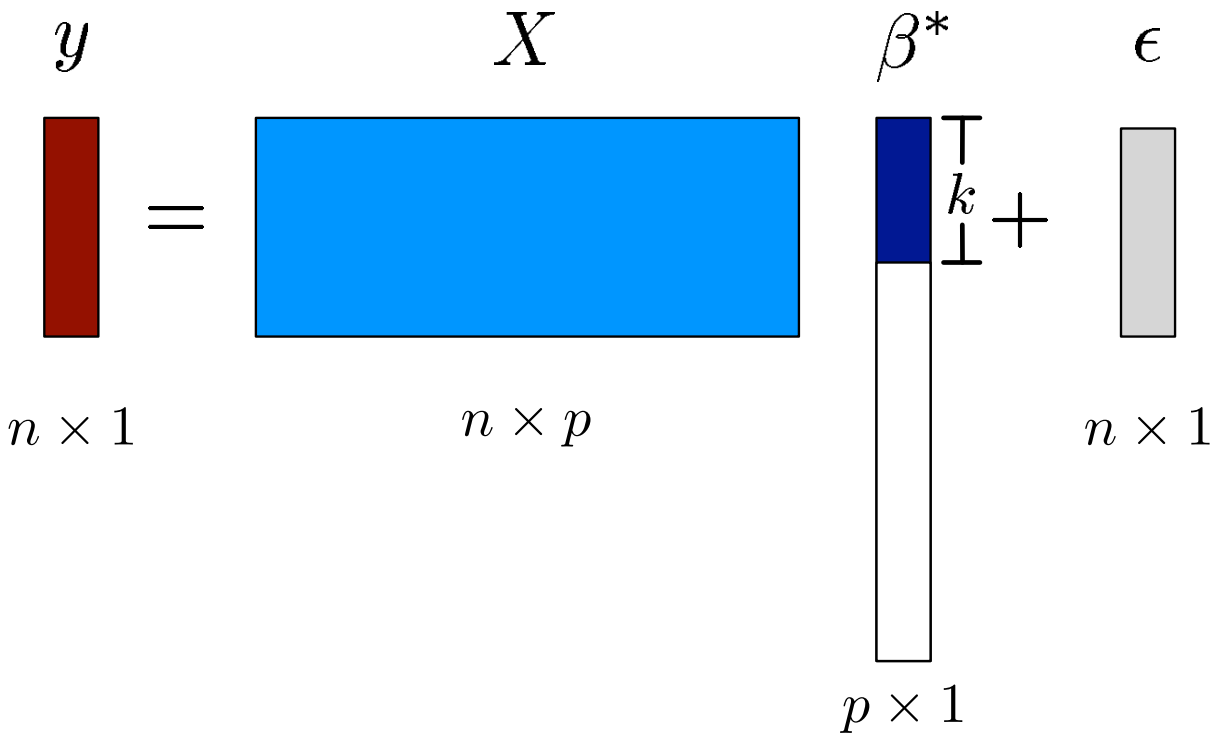
High-dimensional linear models



- Linear model:

$$y_i = x_i^T \beta^* + \epsilon_i, \qquad i = 1, \ldots, n$$

High-dimensional linear models



- Linear model:

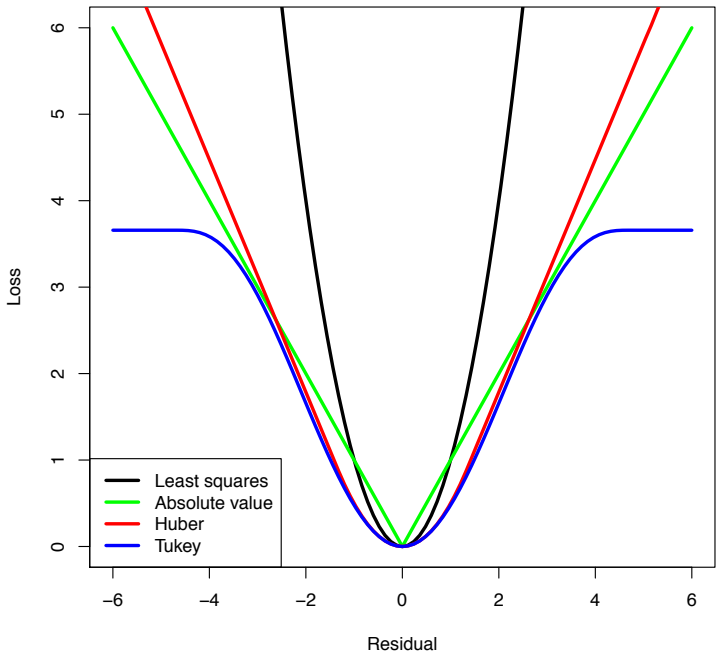
$$y_i = x_i^T \beta^* + \epsilon_i, \qquad i = 1, \ldots, n$$

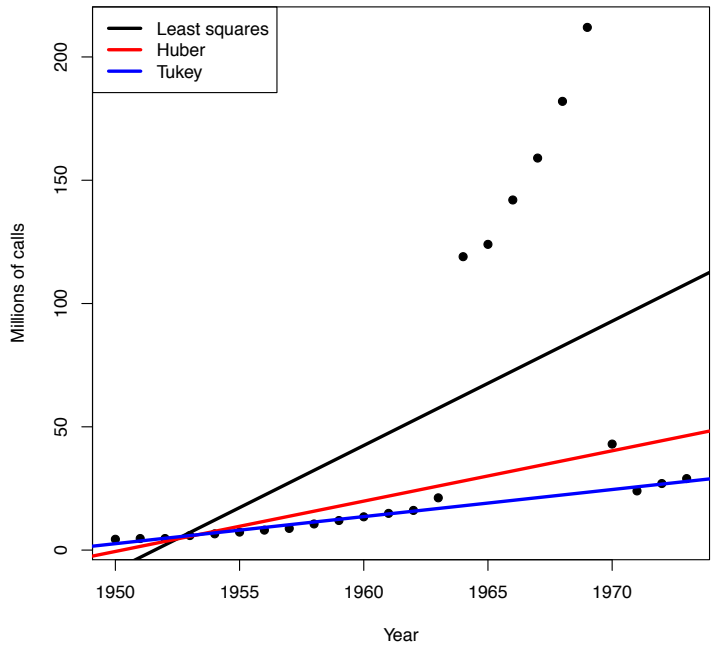
- When $p \gg n$, assume sparsity: $\|\beta^*\|_0 \leq k$

Robust M-estimators

- Generalization of OLS appropriate for robust statistics:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) \right\}$$



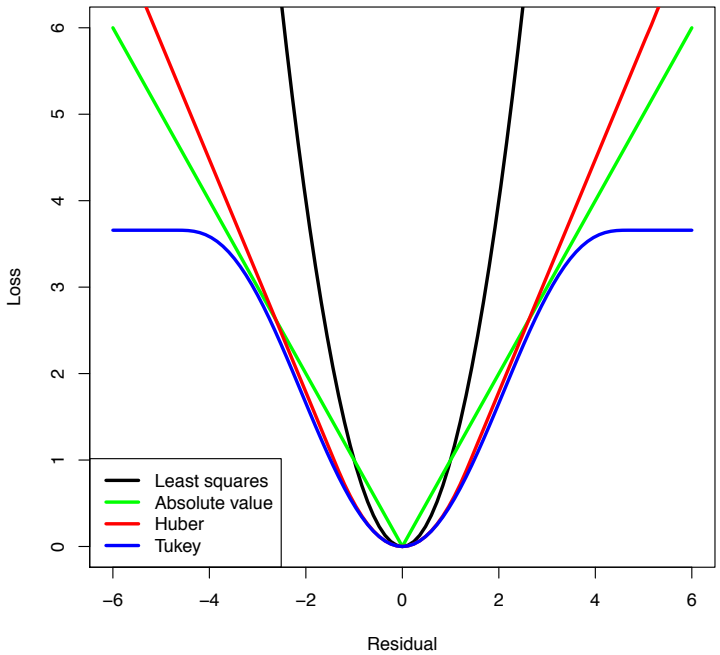


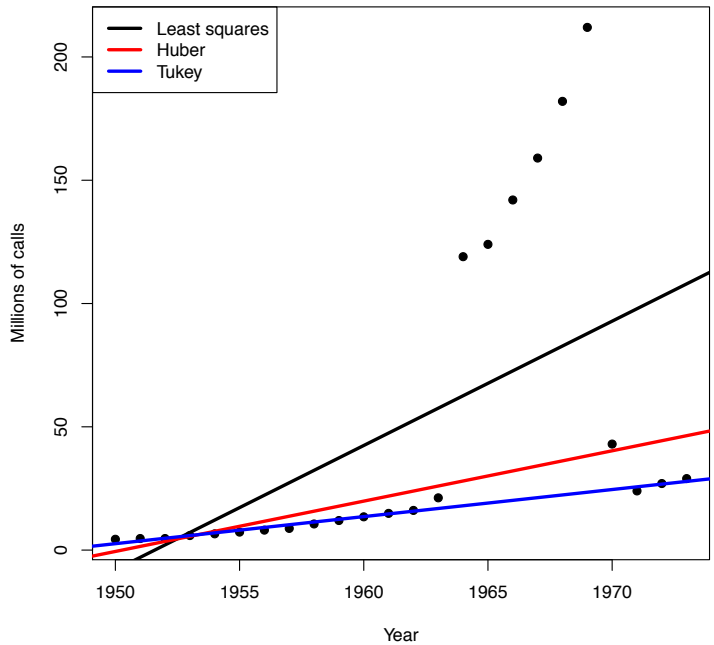
Robust M-estimators

- Generalization of OLS appropriate for robust statistics:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) \right\}$$

- Extensive theory for p fixed, $n \to \infty$





Classes of loss functions

- **Bounded** ℓ' limits influence of outliers:

$$IF((x,y);T,F) = \lim_{t\to 0^+} \frac{T((1-t)F + t\delta_{(x,y)}) - T(F)}{t}$$
$$\propto \ \ell'(x^T\beta - y)x$$

where $F \sim F_\beta$ and T minimizes M-estimator

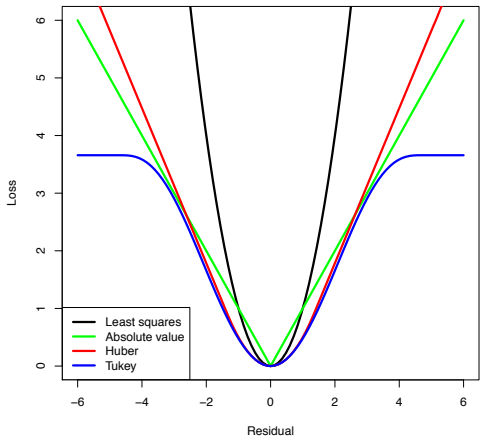
Classes of loss functions

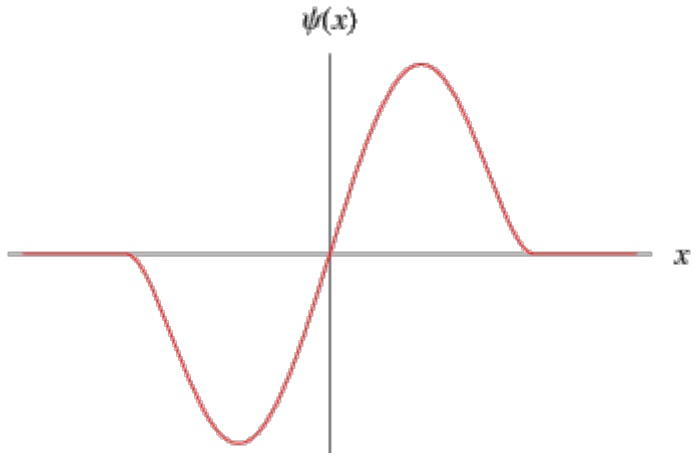
- **Bounded** ℓ' limits influence of outliers:

$$IF((x,y);T,F) = \lim_{t\to 0^+} \frac{T((1-t)F + t\delta_{(x,y)}) - T(F)}{t}$$
$$\propto \ \ell'(x^T\beta - y)x$$

  where $F \sim F_\beta$ and T minimizes M-estimator
- **Redescending** M-estimators have *finite rejection point*:

$$\ell'(u) = 0, \qquad \text{for } |u| \geq c$$





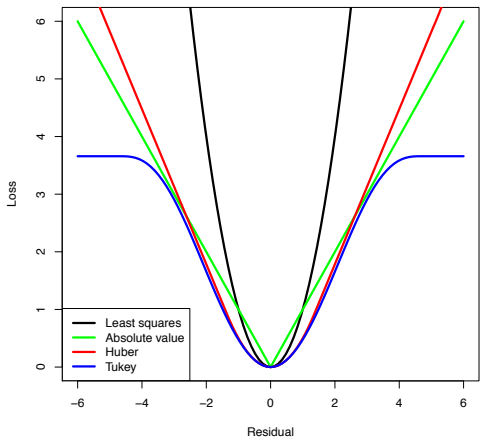
Classes of loss functions

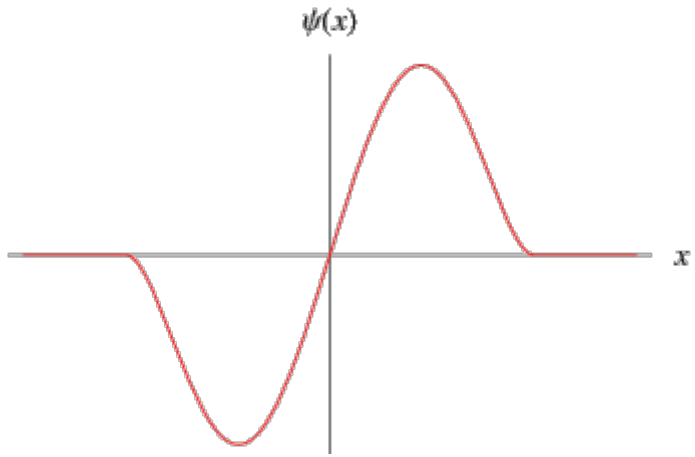
- **Bounded** ℓ' limits influence of outliers:

$$IF((x,y);T,F) = \lim_{t\to 0^+} \frac{T((1-t)F + t\delta_{(x,y)}) - T(F)}{t}$$
$$\propto \ \ell'(x^T\beta - y)x$$

  where $F \sim F_\beta$ and T minimizes M-estimator
- **Redescending** M-estimators have *finite rejection point*:

$$\ell'(u) = 0, \qquad \text{for } |u| \geq c$$





- But bad for optimization!!

High-dimensional M-estimators

- **Natural idea:** For $p > n$, use regularized version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

High-dimensional M-estimators

- **Natural idea:** For $p > n$, use regularized version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

Complications:

- Optimization for nonconvex ℓ?
- Statistical theory? Are certain losses provably better than others?

Overview of results

- When $\|\ell'\|_\infty < C$, **global optima** of high-dimensional M-estimator satisfy

$$\|\widehat{\beta} - \beta^*\|_2 \leq C\sqrt{\frac{k \log p}{n}},$$

regardless of distribution of ϵ_i

Overview of results

- When $\|\ell'\|_\infty < C$, **global optima** of high-dimensional M-estimator satisfy

$$\|\widehat{\beta} - \beta^*\|_2 \leq C\sqrt{\frac{k \log p}{n}},$$

  regardless of distribution of ϵ_i

- **Compare to Lasso theory:** Requires sub-Gaussian ϵ_i's

Overview of results

- When $\|\ell'\|_\infty < C$, **global optima** of high-dimensional M-estimator satisfy

$$\|\widehat{\beta} - \beta^*\|_2 \leq C\sqrt{\frac{k \log p}{n}},$$

  regardless of distribution of ϵ_i
- **Compare to Lasso theory:** Requires sub-Gaussian ϵ_i's
- If $\ell(u)$ is *locally* convex/smooth for $|u| \leq r$, any **local optima** within radius cr of β^* satisfy

$$\|\widetilde{\beta} - \beta^*\|_2 \leq C'\sqrt{\frac{k \log p}{n}}$$

Overview of results

- When $\|\ell'\|_\infty < C$, **global optima** of high-dimensional M-estimator satisfy

$$\|\widehat{\beta} - \beta^*\|_2 \leq C\sqrt{\frac{k \log p}{n}},$$

regardless of distribution of ϵ_i

- **Compare to Lasso theory:** Requires sub-Gaussian ϵ_i's
- If $\ell(u)$ is *locally* convex/smooth for $|u| \leq r$, any **local optima** within radius cr of β^* satisfy

$$\|\widetilde{\beta} - \beta^*\|_2 \leq C'\sqrt{\frac{k \log p}{n}}$$

* in order to verify RE condition w.h.p., need $\mathrm{Var}(\epsilon_i) < cr^2$, as well

Overview of results

- When $\|\ell'\|_\infty < C$, **global optima** of high-dimensional M-estimator satisfy

$$\|\widehat{\beta} - \beta^*\|_2 \leq C\sqrt{\frac{k \log p}{n}},$$

  regardless of distribution of ϵ_i
- **Compare to Lasso theory:** Requires sub-Gaussian ϵ_i's
- If $\ell(u)$ is *locally* convex/smooth for $|u| \leq r$, any **local optima** within radius cr of β^* satisfy

$$\|\widetilde{\beta} - \beta^*\|_2 \leq C'\sqrt{\frac{k \log p}{n}}$$

  * in order to verify RE condition w.h.p., need $\mathrm{Var}(\epsilon_i) < cr^2$, as well
- Local optima may be obtained via **two-step algorithm**

Theoretical insight

- Lasso analysis (e.g., van de Geer '07, Bickel et al. '08):

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \underbrace{\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1}_{\mathcal{L}_n(\beta)} \right\}$$

Theoretical insight

- Lasso analysis (e.g., van de Geer '07, Bickel et al. '08):

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \underbrace{\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1}_{\mathcal{L}_n(\beta)} \right\}$$

- Rearranging *basic inequality* $\mathcal{L}_n(\widehat{\beta}) \leq \mathcal{L}_n(\beta^*)$ and assuming $\lambda \geq 2 \left\| \frac{X^T \epsilon}{n} \right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

Theoretical insight

- Lasso analysis (e.g., van de Geer '07, Bickel et al. '08):

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \underbrace{\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1}_{\mathcal{L}_n(\beta)} \right\}$$

- Rearranging *basic inequality* $\mathcal{L}_n(\widehat{\beta}) \leq \mathcal{L}_n(\beta^*)$ and assuming $\lambda \geq 2\left\|\frac{X^T\epsilon}{n}\right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- Sub-Gaussian assumptions on x_i's and ϵ_i's provide $\mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$ bounds, minimax optimal

Theoretical insight

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

Theoretical insight

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- $\ell'(\epsilon)$ sub-Gaussian whenever ℓ' bounded

Theoretical insight

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- $\ell'(\epsilon)$ sub-Gaussian whenever ℓ' bounded
  $\implies$ can achieve estimation error

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\sqrt{\frac{k \log p}{n}},$$

without assuming ϵ_i is sub-Gaussian

Technical challenges

- Lasso analysis also requires verifying restricted eigenvalue (RE) condition on design matrix, more complicated for general ℓ

Technical challenges

- Lasso analysis also requires verifying restricted eigenvalue (RE) condition on design matrix, more complicated for general ℓ
- When ℓ is nonconvex, local optima $\widetilde{\beta}$ may exist that are not global optima

Technical challenges

- Lasso analysis also requires verifying restricted eigenvalue (RE) condition on design matrix, more complicated for general ℓ
- When ℓ is nonconvex, local optima $\widetilde{\beta}$ may exist that are not global optima
- Want error bounds on $\|\widetilde{\beta} - \beta^*\|_2$ as well, or algorithms to find $\widetilde{\beta}$ efficiently

Related work: Nonconvex regularized M-estimators

- Composite objective function

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \mathcal{L}_n(\beta) + \sum_{j=1}^{p} \rho_\lambda(\beta_j) \right\}$$

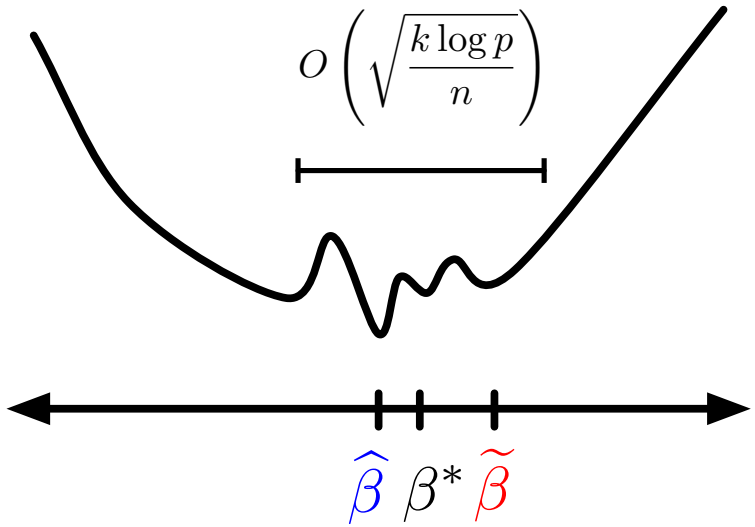
Related work: Nonconvex regularized M-estimators

- Composite objective function

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \mathcal{L}_n(\beta) + \sum_{j=1}^{p} \rho_\lambda(\beta_j) \right\}$$

- Assumptions:
  - $\mathcal{L}_n$ satisfies *restricted strong convexity* with curvature α (Negahban et al. '12)
  - ρ_λ has *bounded subgradient* at 0, and $\rho_\lambda(t) + \mu t^2$ *convex*
  - $\alpha > \mu$

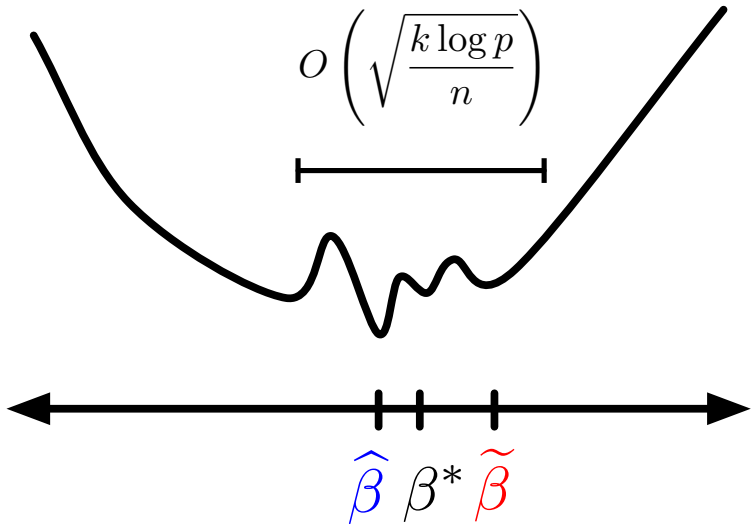
Stationary points (L. & Wainwright '15)



- Stationary points statistically indistinguishable from global optima

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}),\ \beta - \widetilde{\beta} \rangle \geq 0, \qquad \forall \beta \text{ feasible}$$

Stationary points (L. & Wainwright '15)



- Stationary points statistically indistinguishable from global optima

$$\langle \nabla \mathcal{L}_n(\widetilde{\beta}) + \nabla \rho_\lambda(\widetilde{\beta}),\ \beta - \widetilde{\beta} \rangle \geq 0, \qquad \forall \beta \text{ feasible}$$

- Under suitable distributional assumptions, for $\lambda \asymp \sqrt{\frac{\log p}{n}}$ and $R \asymp \frac{1}{\lambda}$,

$$\|\widetilde{\beta} - \beta^*\|_2 \leq c\sqrt{\frac{k \log p}{n}} \approx \text{statistical error}$$

Mathematical statement

Theorem (L. & Wainwright '15)

Suppose R is chosen s.t. β^* is feasible, and λ satisfies

$$\max\left\{\|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\ \alpha\sqrt{\frac{\log p}{n}}\right\} \precsim \lambda \precsim \frac{\alpha}{R}.$$

For $n \geq \frac{C\tau^2}{\alpha^2} R^2 \log p$, **any stationary point** $\widetilde{\beta}$ *satisfies*

$$\|\widetilde{\beta} - \beta^*\|_2 \precsim \frac{\lambda\sqrt{k}}{\alpha - \mu}, \qquad \text{where } k = \|\beta^*\|_0.$$

Mathematical statement

Theorem (L. & Wainwright '15)

Suppose R is chosen s.t. β^* is feasible, and λ satisfies

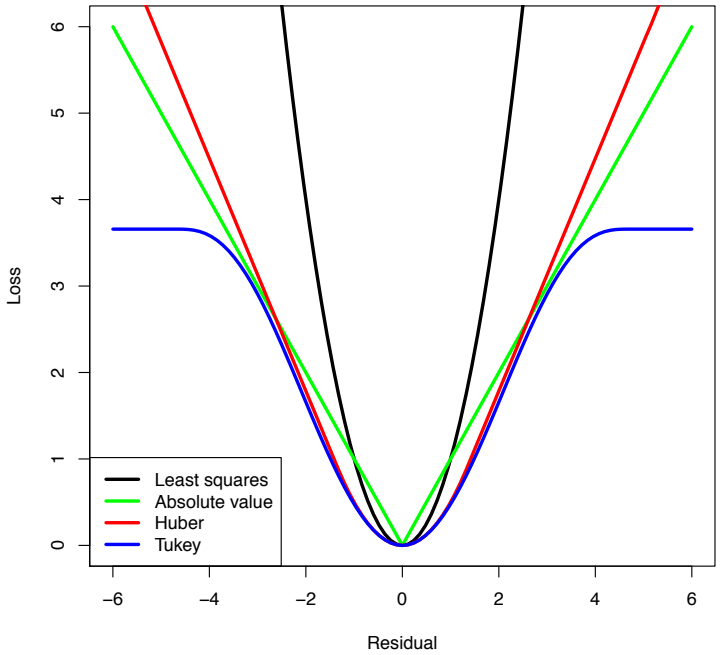
$$\max\left\{\|\nabla \mathcal{L}_n(\beta^*)\|_\infty,\ \alpha\sqrt{\frac{\log p}{n}}\right\} \precsim \lambda \precsim \frac{\alpha}{R}.$$

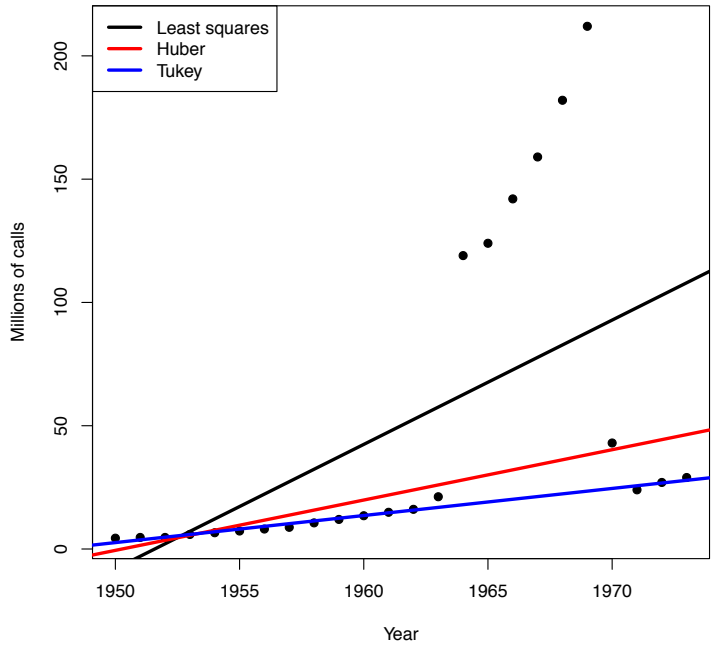
For $n \geq \frac{C\tau^2}{\alpha^2} R^2 \log p$, **any stationary point** $\widetilde{\beta}$ *satisfies*

$$\|\widetilde{\beta} - \beta^*\|_2 \precsim \frac{\lambda\sqrt{k}}{\alpha - \mu}, \qquad \text{where } k = \|\beta^*\|_0.$$

- **New ingredient for robust setting:** ℓ convex only in *local* region
  $\Longrightarrow$ need for *local* consistency results

Local statistical consistency



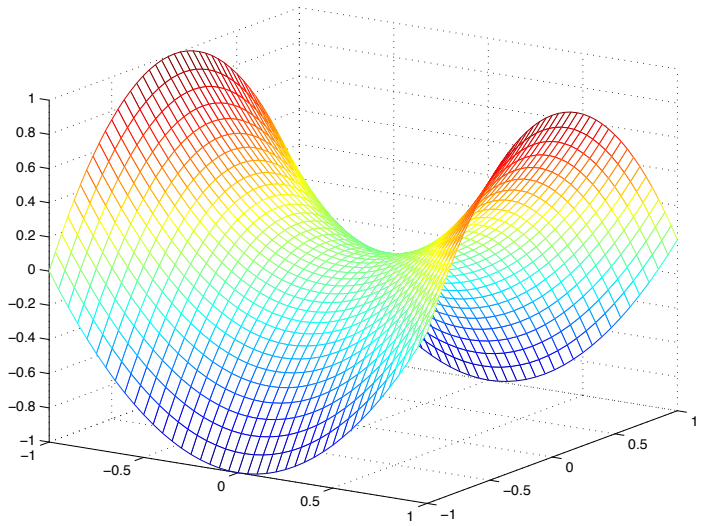


- **Challenge in robust statistics:** Population-level nonconvexity of loss
  $\Longrightarrow$ need for *local* optimization theory

Local RSC condition

- **Local** RSC condition: For $\Delta := \beta_1 - \beta_2$,

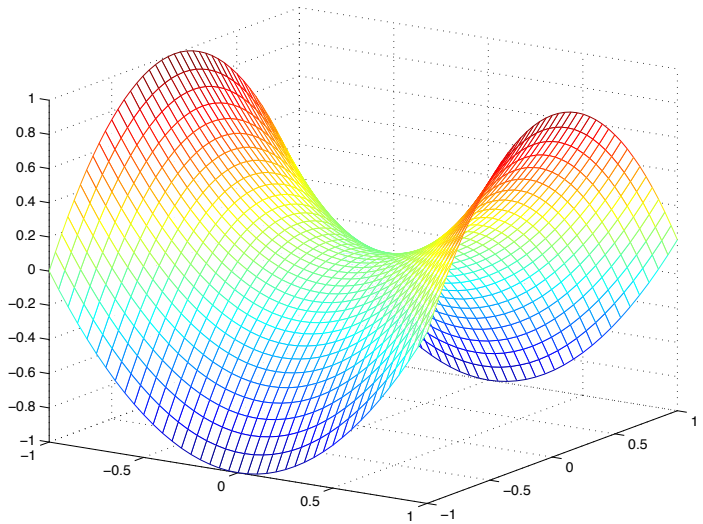
$$\langle \nabla \mathcal{L}_n(\beta_1) - \nabla \mathcal{L}_n(\beta_2),\, \Delta \rangle \geq \alpha \|\Delta\|_2^2 - \tau \frac{\log p}{n} \|\Delta\|_1^2, \quad \forall \|\beta_j - \beta^*\|_2 \leq r$$



Local RSC condition

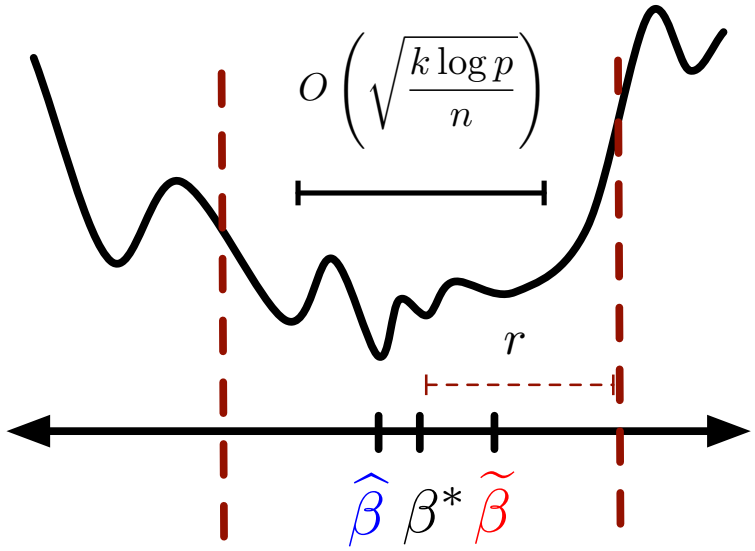
- **Local** RSC condition: For $\Delta := \beta_1 - \beta_2$,

$$\langle \nabla \mathcal{L}_n(\beta_1) - \nabla \mathcal{L}_n(\beta_2),\, \Delta \rangle \geq \alpha \|\Delta\|_2^2 - \tau \frac{\log p}{n} \|\Delta\|_1^2, \quad \forall \|\beta_j - \beta^*\|_2 \leq r$$

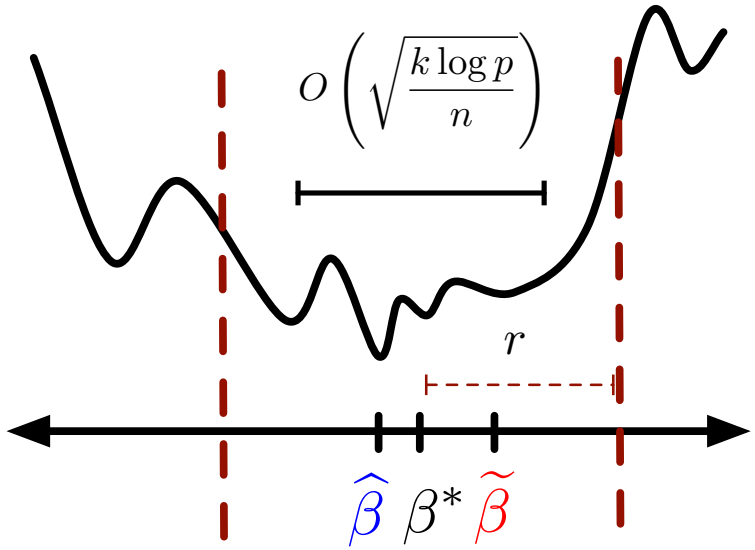


- Only requires restricted curvature within constant-radius region around β^*

Consistency of local stationary points



Consistency of local stationary points



Theorem (L. '17)

Suppose $\mathcal{L}_n$ satisfies α-local RSC and ρ_λ is μ-amenable, with $\alpha > \mu$. Suppose $\|\ell'\|_\infty \leq C$ and $\lambda \asymp \sqrt{\frac{\log p}{n}}$. For $n \gtrsim \frac{\tau}{\alpha - \mu} k \log p$, **any stationary point $\widetilde{\beta}$ s.t. $\|\widetilde{\beta} - \beta^*\|_2 \leq r$** *satisfies*

$$\|\widetilde{\beta} - \beta^*\|_2 \lesssim \frac{\lambda\sqrt{k}}{\alpha - \mu}.$$

Optimization theory

- **Question:** How to obtain sufficiently close local solutions?

Optimization theory

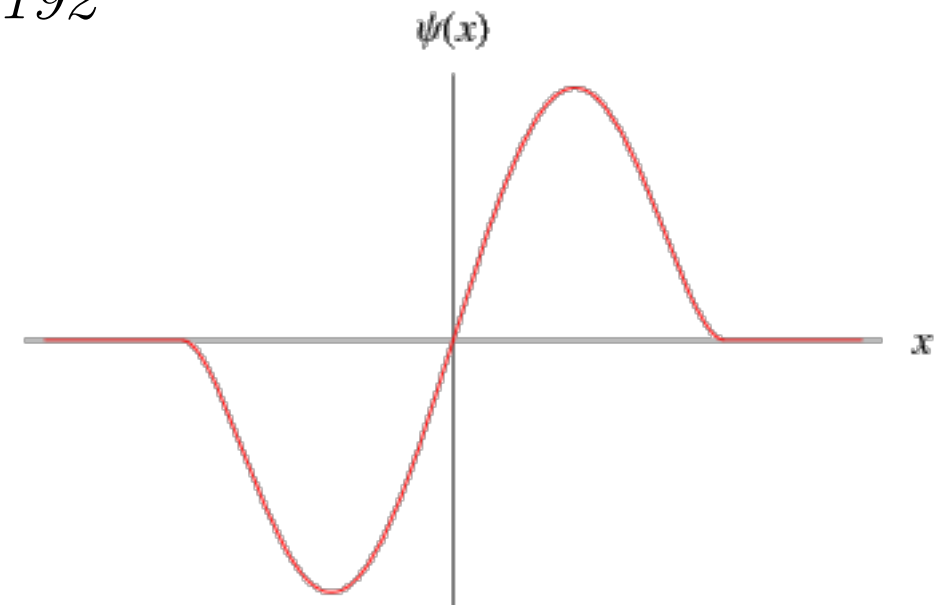
- **Question:** How to obtain sufficiently close local solutions?
- **Goal:** For regularized M-estimator

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \rho_\lambda(\beta) \right\},$$

where ℓ satisfies α-local RSC, find stationary point such that $\|\widetilde{\beta} - \beta^*\|_2 \leq r$

Wisdom from Huber

Descending ψ-functions are tricky, especially when the starting values for the iterations are non-robust. . . . It is therefore preferable to start with a monotone ψ, iterate to death, and then append a few (1 or 2) iterations with the nonmonotone ψ. — *Huber 1981, pp. 191–192*



Two-step algorithm (L. '17)

- Use *composite gradient descent* (Nesterov '07):
  - Iterative method to solve

$$\widehat{\beta} \in \arg\min_{\beta \in \Omega} \{\mathcal{L}_n(\beta) + \rho_\lambda(\beta)\},$$

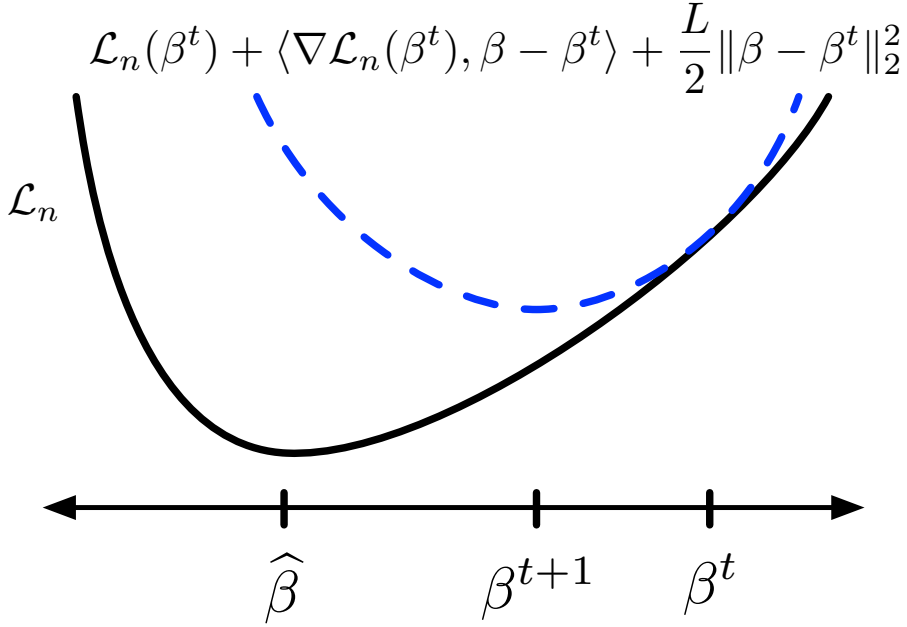
$\mathcal{L}_n$ differentiable, ρ_λ convex & subdifferentiable

Two-step algorithm (L. '17)

- Use *composite gradient descent* (Nesterov '07):
  - Iterative method to solve

$$\widehat{\beta} \in \arg\min_{\beta \in \Omega} \{\mathcal{L}_n(\beta) + \rho_\lambda(\beta)\},$$

$\mathcal{L}_n$ differentiable, ρ_λ convex & subdifferentiable

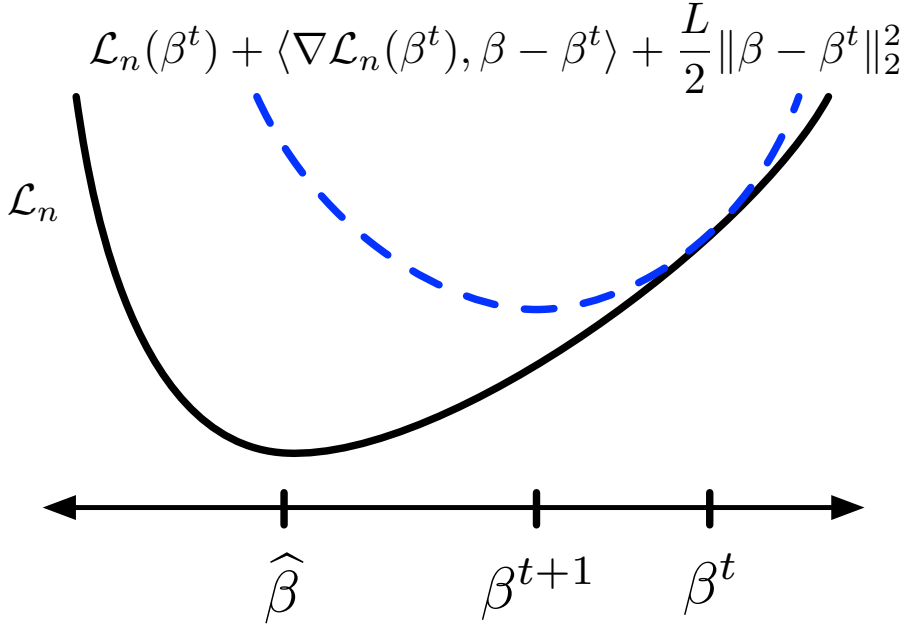


Two-step algorithm (L. '17)

- Use *composite gradient descent* (Nesterov '07):
  - Iterative method to solve

$$\widehat{\beta} \in \arg\min_{\beta \in \Omega} \{\mathcal{L}_n(\beta) + \rho_\lambda(\beta)\},$$

$\mathcal{L}_n$ differentiable, ρ_λ convex & subdifferentiable



  - Updates:

$$\beta^{t+1} \in \arg\min_{\beta \in \Omega} \left\{ \mathcal{L}_n(\beta^t) + \langle \nabla \mathcal{L}_n(\beta^t), \beta - \beta^t \rangle + \frac{L}{2}\|\beta - \beta^t\|_2^2 + \rho_\lambda(\beta) \right\}$$

Two-step algorithm (L. '17)

- **Two-step M-estimator:** Finds local stationary points of nonconvex, robust loss + μ-amenable penalty

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \rho_\lambda(\beta) \right\}$$

Two-step algorithm (L. '17)

- **Two-step M-estimator:** Finds local stationary points of nonconvex, robust loss + μ-amenable penalty

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \rho_\lambda(\beta) \right\}$$

Algorithm

1. Run composite gradient descent on **convex**, robust loss + ℓ_1-penalty until convergence, output $\widehat{\beta}_H$

Two-step algorithm (L. '17)

- **Two-step M-estimator:** Finds local stationary points of nonconvex, robust loss + μ-amenable penalty

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \rho_\lambda(\beta) \right\}$$

Algorithm

1. Run composite gradient descent on **convex**, robust loss + ℓ_1-penalty until convergence, output $\widehat{\beta}_H$
2. Run composite gradient descent on **nonconvex**, robust loss + μ-amenable penalty, input $\beta^0 = \widehat{\beta}_H$

Two-step algorithm (L. '17)

- **Two-step M-estimator:** Finds local stationary points of nonconvex, robust loss + μ-amenable penalty

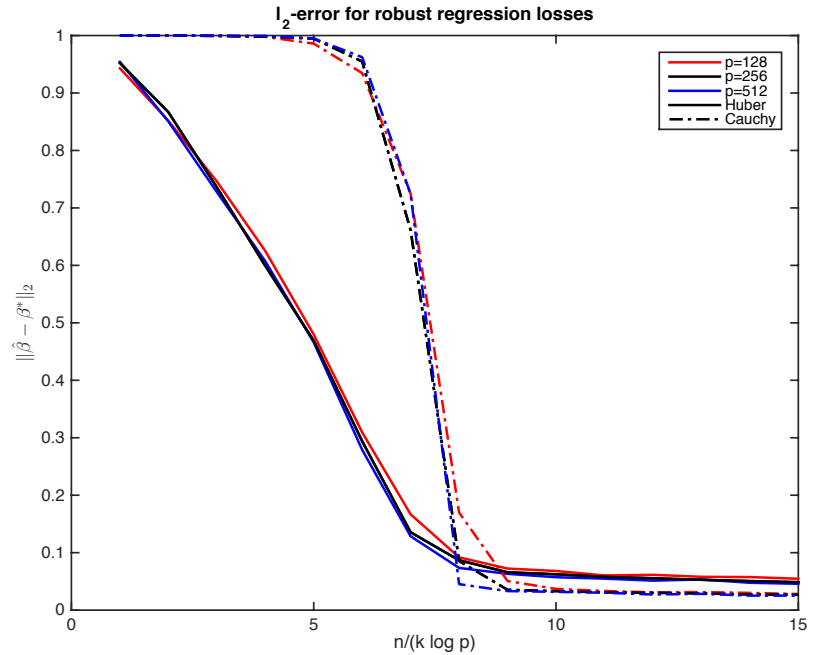
$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \rho_\lambda(\beta) \right\}$$

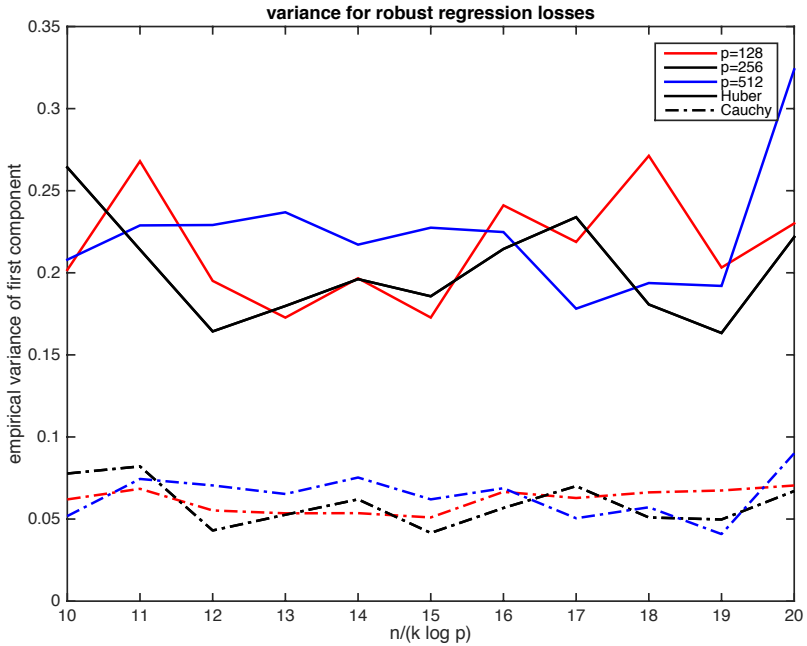
Algorithm

1. Run composite gradient descent on **convex**, robust loss + ℓ_1-penalty until convergence, output $\widehat{\beta}_H$
2. Run composite gradient descent on **nonconvex**, robust loss + μ-amenable penalty, input $\beta^0 = \widehat{\beta}_H$

- **Important:** We want to optimize original nonconvex objective, since it leads to more *efficient* (lower-variance) estimators

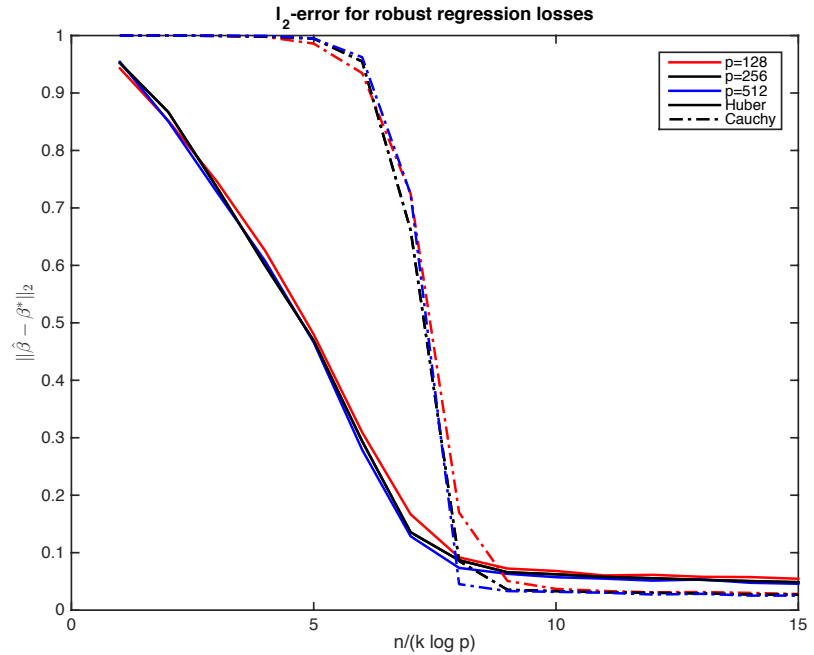
Simulation

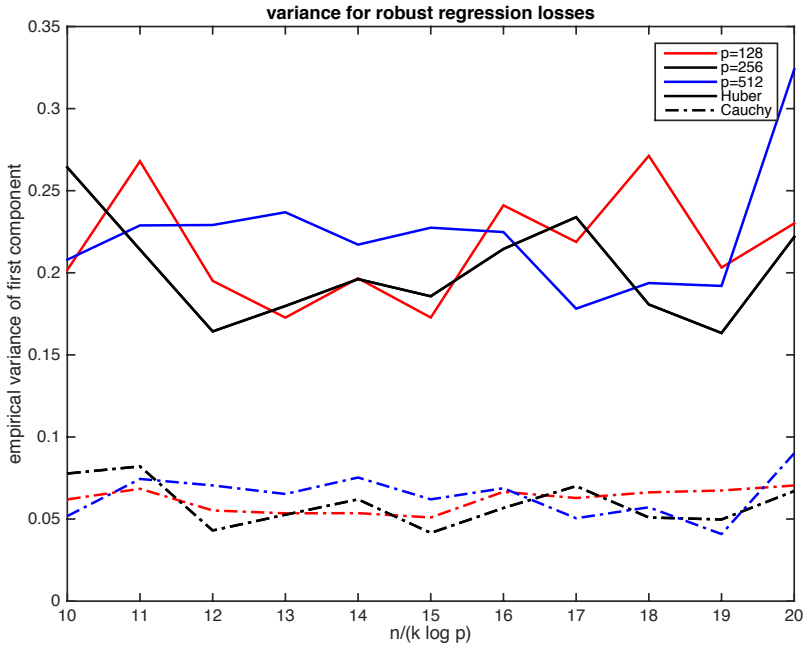




- ℓ_2-error and empirical variance of M-estimators when errors follow Cauchy distribution (SCAD regularizer)

Simulation





- ℓ_2-error and empirical variance of M-estimators when errors follow Cauchy distribution (SCAD regularizer)
- Can prove geometric convergence of two-step algorithm to desirable local optima (L. '17)

Summary

- Loss functions with desirable robustness properties in low-dimensional regression **also good for high dimensions**:

  bounded influence $\iff$ $\|\ell'\|_\infty \leq C$ $\iff$ $O\left(\sqrt{\frac{k \log p}{n}}\right)$ consistency

Summary

- Loss functions with desirable robustness properties in low-dimensional regression **also good for high dimensions**:

  bounded influence $\iff$ $\|\ell'\|_\infty \leq C$ $\iff$ $O\left(\sqrt{\frac{k \log p}{n}}\right)$ consistency

- **Two-step optimization procedure:** First step for consistency, second step for efficiency

Loh (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics.*

Trailer

- **Problem:** Loss function ℓ in some sense calibrated to scale of ϵ_i

Trailer

- **Problem:** Loss function ℓ in some sense calibrated to scale of ϵ_i
- Better objective (joint location/scale estimator):

$$(\widehat{\beta}, \widehat{\sigma}) \in \arg\min_{\beta,\sigma} \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T\beta}{\sigma}\right)\sigma + a\sigma}_{\mathcal{L}_n(\beta,\sigma)} + \lambda\|\beta\|_1 \right\}$$

Trailer

- **Problem:** Loss function ℓ in some sense calibrated to scale of ϵ_i
- Better objective (joint location/scale estimator):

$$(\widehat{\beta}, \widehat{\sigma}) \in \arg\min_{\beta,\sigma} \Bigg\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T\beta}{\sigma}\right)\sigma + a\sigma}_{\mathcal{L}_n(\beta,\sigma)} + \lambda\|\beta\|_1 \Bigg\}$$

- However, location/scale estimation notoriously difficult even in low dimensions

Trailer

- Another idea: *MM*-estimator

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T \beta}{\widehat{\sigma}_0} \right) + \lambda \|\beta\|_1 \right\},$$

  using robust estimate of scale $\widehat{\sigma}_0$ based on preliminary estimate $\widehat{\beta}_0$

Trailer

- Another idea: *MM*-estimator

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T \beta}{\widehat{\sigma}_0} \right) + \lambda \|\beta\|_1 \right\},$$

  using robust estimate of scale $\widehat{\sigma}_0$ based on preliminary estimate $\widehat{\beta}_0$

- How to obtain $(\widehat{\beta}_0, \widehat{\sigma}_0)$?

Trailer

- Another idea: *MM*-estimator

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T \beta}{\widehat{\sigma}_0} \right) + \lambda \|\beta\|_1 \right\},$$

  using robust estimate of scale $\widehat{\sigma}_0$ based on preliminary estimate $\widehat{\beta}_0$

- How to obtain $(\widehat{\beta}_0, \widehat{\sigma}_0)$?
  - *S*-estimators/LMS:

$$\widehat{\beta}_0 \in \arg\min_{\beta} \{\widehat{\sigma}(r(\beta))\},$$

  where $\widehat{\sigma}(r) = r_{(n - \lfloor n\delta \rfloor)}$

Trailer

- Another idea: *MM*-estimator

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T \beta}{\widehat{\sigma}_0} \right) + \lambda \|\beta\|_1 \right\},$$

  using robust estimate of scale $\widehat{\sigma}_0$ based on preliminary estimate $\widehat{\beta}_0$

- How to obtain $(\widehat{\beta}_0, \widehat{\sigma}_0)$?
  - *S*-estimators/LMS:

$$\widehat{\beta}_0 \in \arg\min_{\beta} \{\widehat{\sigma}(r(\beta))\},$$

    where $\widehat{\sigma}(r) = r_{(n - \lfloor n\delta \rfloor)}$
  - LTS:

$$\widehat{\beta}_0 \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n - \lfloor n\alpha \rfloor} (y_i - x_i^T \beta)^2_{(i)} + \lambda \|\beta\|_1 \right\}$$

Trailer

- Maybe an entirely different approach is necessary . . .

Loh (2017). Scale estimation for high-dimensional robust regression.
Coming soon?

Thank you!