Robust estimation via (non)convex M-estimation

Po-Ling Loh

University of Wisconsin - Madison
Departments of ECE & Statistics

Workshop on computational efficiency and
high-dimensional robust statistics
TTI Chicago

August 15, 2018

Outline

1. Regularized M-estimators
   - Statistical M-estimation
   - Nonconvexity
   - Consistency of local optima

2. High-dimensional robust regression
   - Statistical consistency
   - Asymptotic normality
   - Two-step M-estimators

Outline

1. Regularized M-estimators
   - Statistical M-estimation
   - Nonconvexity
   - Consistency of local optima
2. High-dimensional robust regression
   - Statistical consistency
   - Asymptotic normality
   - Two-step M-estimators

- **Prediction/regression problem:** Observe $\{(x_i, y_i)\}_{i=1}^n$, estimate

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}[\ell(\beta; x_i, y_i)], \qquad x_i \in \mathbb{R}^p, \quad y_i \in \mathbb{R}$$

Regularized M-estimators

- **Prediction/regression problem:** Observe $\{(x_i, y_i)\}_{i=1}^n$, estimate

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}[\ell(\beta; x_i, y_i)], \qquad x_i \in \mathbb{R}^p, \quad y_i \in \mathbb{R}$$

- Statistical M-estimator:

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

in high dimensions, may be ill-conditioned, large solution space

Regularized M-estimators

- **Prediction/regression problem:** Observe $\{(x_i, y_i)\}_{i=1}^n$, estimate

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}[\ell(\beta; x_i, y_i)], \qquad x_i \in \mathbb{R}^p, \quad y_i \in \mathbb{R}$$

- Regularized M-estimator:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(\beta; x_i, y_i)}_{\mathcal{L}_n(\beta)} + \rho_\lambda(\beta) \right\}$$

Example: ℓ_1-regularized OLS regression

- Linear model: $\qquad y_i = x_i^T \beta^* + \epsilon_i, \qquad \|\beta^*\|_0 \le k$

Example: ℓ_1-regularized OLS regression

- Linear model: $y_i = x_i^T \beta^* + \epsilon_i, \qquad \|\beta^*\|_0 \leq k$
- Low-dimensional M-estimator:

$$\widehat{\beta}_{\text{OLS}} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right\}$$

Example: ℓ_1-regularized OLS regression

- Linear model: $y_i = x_i^T \beta^* + \epsilon_i, \qquad \|\beta^*\|_0 \leq k$
- Low-dimensional M-estimator:

$$\widehat{\beta}_{\mathrm{OLS}} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \right\} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} y_i x_i \right)$$

Example: ℓ_1-regularized OLS regression

- Linear model: $y_i = x_i^T \beta^* + \epsilon_i, \qquad \|\beta^*\|_0 \le k$
- Low-dimensional M-estimator:

$$\widehat{\beta}_{\text{OLS}} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \beta)^2 \right\} = \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n y_i x_i \right)$$

- High-dimensional regularized M-estimator:

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

Sources of nonconvexity

- May arise in loss or regularizer

Sources of nonconvexity

- May arise in loss or regularizer
- Nonconvex loss used to correct bias, increase efficiency

Sources of nonconvexity

- May arise in loss or regularizer
- Nonconvex loss used to correct bias, increase efficiency
- Nonconvex regularizer used to reduce bias, achieve oracle result

Example: Errors-in-variables regression

- Model:

$$y_i = x_i^T \beta^* + \epsilon_i$$

observe $\{(z_i, y_i)\}_{i=1}^n$, infer β^*

Example: Errors-in-variables regression

- Model:

$$y_i = x_i^T \beta^* + \epsilon_i$$

  observe $\{(z_i, y_i)\}_{i=1}^n$, infer β^*

- OLS estimator

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

  statistically inconsistent

Corrected losses

- L. & Wainwright '12 propose natural method for correcting loss for linear regression:

$$\widehat{\beta}_{\text{OLS}} \in \arg\min_{\beta} \left\{ \frac{1}{2} \beta^T \frac{X^T X}{n} \beta - \frac{y X^T}{n} \beta + \rho_\lambda(\beta) \right\}$$

$$\widehat{\beta}_{\text{corr}} \in \arg\min_{\beta} \left\{ \frac{1}{2} \beta^T \widehat{\Gamma} \beta - \widehat{\gamma}^T \beta + \rho_\lambda(\beta) \right\}$$

$(\widehat{\Gamma}, \widehat{\gamma})$ estimators for $(\text{Cov}(x_i), \text{Cov}(x_i, y_i))$ based on $\{(z_i, y_i)\}_{i=1}^n$

Example: Additive noise

- Additive noise: $Z = X + W$, use

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \widehat{\gamma} = \frac{Z^T y}{n}$$

- However, corrected objective **nonconvex**:

$$\widehat{\beta}_{\text{corr}} \in \arg\min_{\beta} \left\{ \frac{1}{2} \beta^T \left(\frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta + \rho_\lambda(\beta) \right\}$$

Example: Additive noise

- Additive noise: $Z = X + W$, use

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \qquad \widehat{\gamma} = \frac{Z^T y}{n}$$

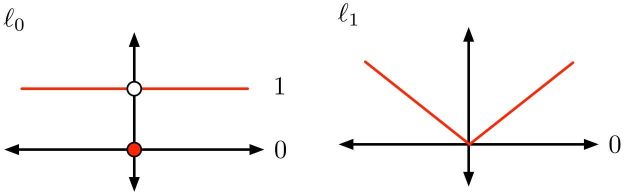
- However, corrected objective **nonconvex**:

$$\widehat{\beta}_{\text{corr}} \in \arg\min_{\beta} \left\{ \frac{1}{2} \beta^T \left(\frac{Z^T Z}{n} - \Sigma_w \right) \beta - \frac{y^T Z}{n} \beta + \rho_\lambda(\beta) \right\}$$

- Fortunately, local optima have good properties

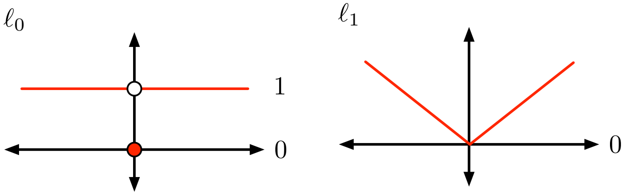
Nonconvex regularizers

- ℓ_1 is "convexified" version of ℓ_0



Nonconvex regularizers

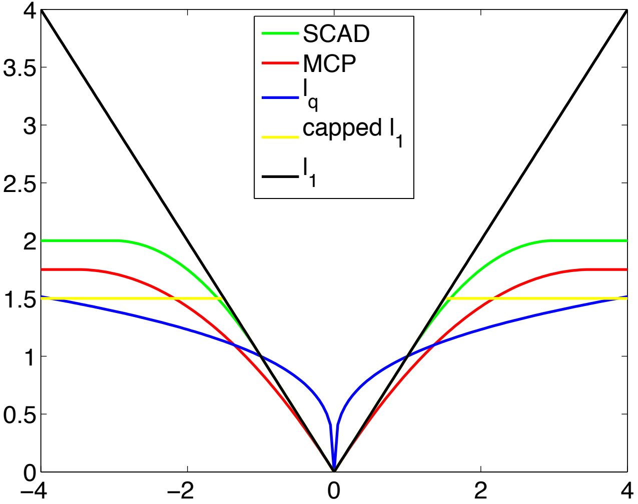
- ℓ_1 is "convexified" version of ℓ_0



- **But** ℓ_1 penalizes larger coefficients more, causes *solution bias*

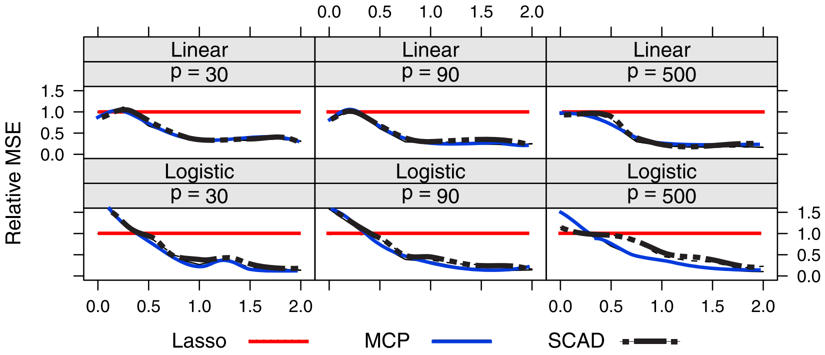
Alternative regularizers

- Various nonconvex regularizers in literature (Fan & Li '01, Zhang '10, etc.)



Empirical benefits

- Nonconvex regularizers show **significant improvement** (Breheny & Huang '11)

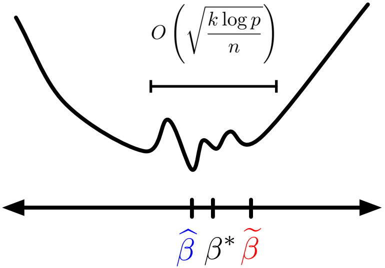


Local vs. global optima

- Optimization algorithms only guaranteed to find *local optima* (stationary points)
- Statistical theory only guarantees consistency of *global optima*

Local vs. global optima

- Optimization algorithms only guaranteed to find *local optima* (stationary points)
- Statistical theory only guarantees consistency of *global optima*



- **L. & Wainwright '13:** All stationary points of $\mathcal{L}_n(\beta) + \rho_\lambda(\beta)$ close when nonconvexity smaller than curvature

Measures of closeness

- Various measures of statistical consistency

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\beta; x_i, y_i) + \rho_\lambda(\beta) \right\}$$

Measures of closeness

- Various measures of statistical consistency

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\beta; x_i, y_i) + \rho_\lambda(\beta) \right\}$$

- **Estimation:** $\|\widehat{\beta} - \beta^*\| \to 0$
- **Prediction:** $\frac{1}{n}\sum_{i=1}^{n} \ell(\widehat{\beta}; x_i, y_i) \to 0$
- **Variable selection:** $\mathrm{supp}(\widehat{\beta}) \to \mathrm{supp}(\beta^*)$

Measures of closeness

- Various measures of statistical consistency

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(\beta; x_i, y_i) + \rho_\lambda(\beta) \right\}$$

- **Estimation:** $\|\widehat{\beta} - \beta^*\| \to 0$
- **Prediction:** $\frac{1}{n} \sum_{i=1}^{n} \ell(\widehat{\beta}; x_i, y_i) \to 0$
- **Variable selection:** $\text{supp}(\widehat{\beta}) \to \text{supp}(\beta^*)$

- Interested in cases where ℓ and ρ_λ possibly *nonconvex*

Estimation/prediction consistency

- Composite objective function

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

Estimation/prediction consistency

- Composite objective function

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

- $\mathcal{L}_n$ satisfies **restricted strong convexity** with curvature α
- ρ_λ has bounded subgradient at 0, and $\rho_\lambda(t) + \mu t^2$ convex

Estimation/prediction consistency

- Composite objective function

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

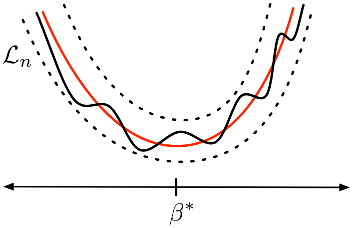
- $\mathcal{L}_n$ satisfies **restricted strong convexity** with curvature α
- ρ_λ has bounded subgradient at 0, and $\rho_\lambda(t) + \mu t^2$ convex

- **L. & Wainwright '13:** All stationary points of $\mathcal{L}_n(\beta) + \rho_\lambda(\beta)$ close when $\alpha > \mu$

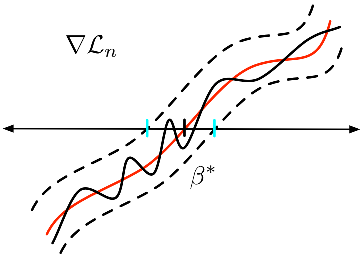
Geometric intuition

- **Population-level** convexity, **finite-sample** nonconvexity

function view

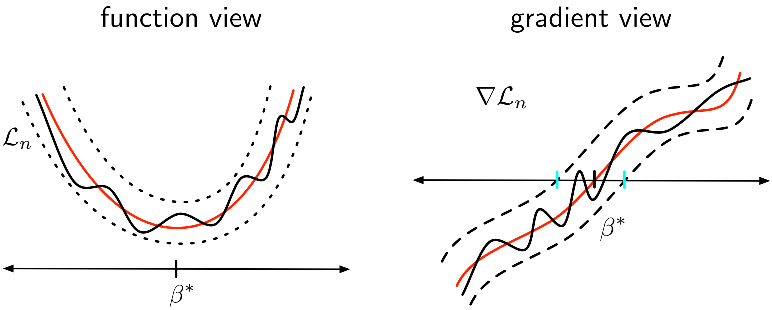


gradient view



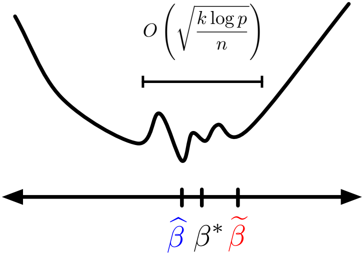
Geometric intuition

- **Population-level** convexity, **finite-sample** nonconvexity



- Population-level objective $\mathcal{L}$ strongly convex, $\alpha > \mu$
- RSC quantifies convergence rate of $\nabla\mathcal{L}_n \longrightarrow \nabla\mathcal{L}$

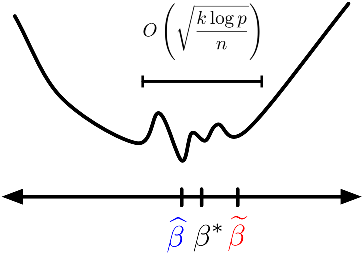
More formally



- Stationary points statistically indistinguishable from global optima

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

More formally



- Stationary points statistically indistinguishable from global optima

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

- Nonasymptotic rates: For $\lambda \asymp \sqrt{\frac{\log p}{n}}$ and $R \asymp \frac{1}{\lambda}$,

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

Technical conditions

- Requirements on loss and regularizer to ensure consistency of stationary points

Technical conditions

- Requirements on loss and regularizer to ensure consistency of stationary points
  - Restricted strong convexity of $\mathcal{L}_n$
  - Bound on nonconvexity of ρ_λ

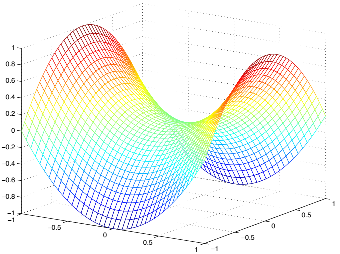
Technical conditions

- Requirements on loss and regularizer to ensure consistency of stationary points
  - Restricted strong convexity of $\mathcal{L}_n$
  - Bound on nonconvexity of ρ_λ
- "Oracle" result under additional condition on ρ_λ

Conditions on $\mathcal{L}_n$

- **Restricted strong convexity** (Negahban et al. '12):

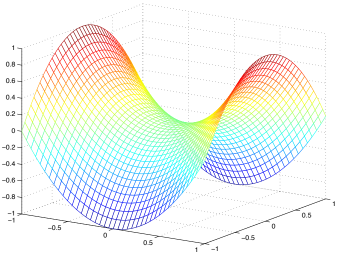
$$\langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*),\ \Delta \rangle \geq \begin{cases} \alpha \|\Delta\|_2^2 - \tau \frac{\log p}{n} \|\Delta\|_1^2, & \forall \|\Delta\|_2 \leq r \\ \alpha \|\Delta\|_2 - \tau \sqrt{\frac{\log p}{n}} \|\Delta\|_1, & \text{o.w.} \end{cases}$$



Conditions on $\mathcal{L}_n$

- **Restricted strong convexity** (Negahban et al. '12):

$$\langle \nabla \mathcal{L}_n(\beta^* + \Delta) - \nabla \mathcal{L}_n(\beta^*),\, \Delta \rangle \geq \begin{cases} \alpha \|\Delta\|_2^2 - \tau \frac{\log p}{n} \|\Delta\|_1^2, & \forall \|\Delta\|_2 \leq r \\ \alpha \|\Delta\|_2 - \tau \sqrt{\frac{\log p}{n}} \|\Delta\|_1, & \text{o.w.} \end{cases}$$



- Holds for various convex/nonconvex losses:
  - OLS & corrected OLS for linear regression, log likelihood for GLMs
  - Huber loss for robust regression

Conditions on ρ_λ

- Focus on *amenable* regularizers $\rho_\lambda(\beta) = \sum_{j=1}^p \rho_\lambda(\beta_j)$ satisfying:

Conditions on ρ_λ

- Focus on *amenable* regularizers $\rho_\lambda(\beta) = \sum_{j=1}^p \rho_\lambda(\beta_j)$ satisfying:
  - $\rho_\lambda(0) = 0$, symmetric around 0
  - Nondecreasing on $\mathbb{R}^+$
  - $t \mapsto \frac{\rho_\lambda(t)}{t}$ nonincreasing on $\mathbb{R}^+$
  - $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$ differentiable everywhere
  - $\rho_\lambda(t) + \mu t^2$ convex for some $\mu > 0$

Conditions on ρ_λ

- Focus on *amenable* regularizers $\rho_\lambda(\beta) = \sum_{j=1}^p \rho_\lambda(\beta_j)$ satisfying:
  - $\rho_\lambda(0) = 0$, symmetric around 0
  - Nondecreasing on $\mathbb{R}^+$
  - $t \mapsto \frac{\rho_\lambda(t)}{t}$ nonincreasing on $\mathbb{R}^+$
  - $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$ differentiable everywhere
  - $\rho_\lambda(t) + \mu t^2$ convex for some $\mu > 0$
  - $\rho'_\lambda(t) = 0$ for $t \geq \gamma\lambda$, for some $\gamma > 0$

Conditions on ρ_λ

- Focus on *amenable* regularizers $\rho_\lambda(\beta) = \sum_{j=1}^p \rho_\lambda(\beta_j)$ satisfying:
  - $\rho_\lambda(0) = 0$, symmetric around 0
  - Nondecreasing on $\mathbb{R}^+$
  - $t \mapsto \frac{\rho_\lambda(t)}{t}$ nonincreasing on $\mathbb{R}^+$
  - $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$ differentiable everywhere
  - $\rho_\lambda(t) + \mu t^2$ convex for some $\mu > 0$
  - $\rho'_\lambda(t) = 0$ for $t \geq \gamma\lambda$, for some $\gamma > 0$
- Examples:
  - μ-amenable: ℓ_1, SCAD, MCP, LSP

Conditions on ρ_λ

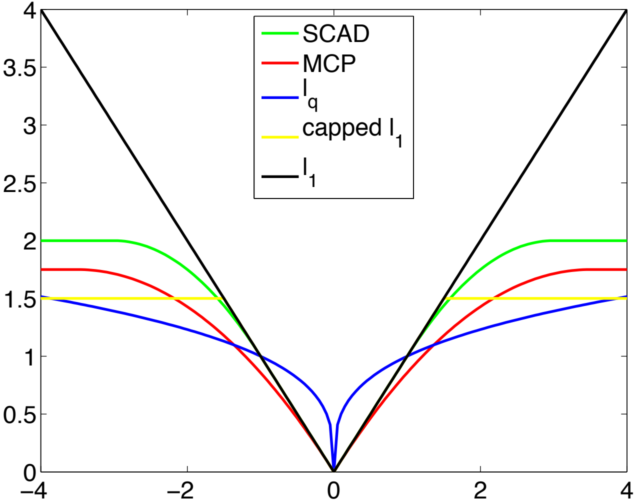
- Focus on *amenable* regularizers $\rho_\lambda(\beta) = \sum_{j=1}^p \rho_\lambda(\beta_j)$ satisfying:
  - $\rho_\lambda(0) = 0$, symmetric around 0
  - Nondecreasing on $\mathbb{R}^+$
  - $t \mapsto \frac{\rho_\lambda(t)}{t}$ nonincreasing on $\mathbb{R}^+$
  - $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$ differentiable everywhere
  - $\rho_\lambda(t) + \mu t^2$ convex for some $\mu > 0$
  - $\rho_\lambda'(t) = 0$ for $t \geq \gamma\lambda$, for some $\gamma > 0$
- Examples:
  - μ-amenable: ℓ_1, SCAD, MCP, LSP
  - (μ, γ)-amenable: SCAD, MCP

Conditions on ρ_λ

- Focus on *amenable* regularizers $\rho_\lambda(\beta) = \sum_{j=1}^p \rho_\lambda(\beta_j)$ satisfying:
    - $\rho_\lambda(0) = 0$, symmetric around 0
    - Nondecreasing on $\mathbb{R}^+$
    - $t \mapsto \frac{\rho_\lambda(t)}{t}$ nonincreasing on $\mathbb{R}^+$
    - $q_\lambda(t) := \lambda|t| - \rho_\lambda(t)$ differentiable everywhere
    - $\rho_\lambda(t) + \mu t^2$ convex for some $\mu > 0$
    - $\rho_\lambda'(t) = 0$ for $t \geq \gamma\lambda$, for some $\gamma > 0$
- Examples:
    - μ-amenable: ℓ_1, SCAD, MCP, LSP
    - (μ, γ)-amenable: SCAD, MCP
    - Neither: capped-ℓ_1, bridge penalty ($\ell_q, 0 < q < 1$)

Alternative regularizers

- Various nonconvex regularizers in literature (Fan & Li '01, Zhang '10, etc.)



- Regularized M-estimator

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \leq R} \left\{ \mathcal{L}_n(\beta) + \rho_\lambda(\beta) \right\},$$

loss function satisfies (α, τ)-RSC and regularizer is μ-amenable

Statistical consistency

- Regularized M-estimator

$$\widehat{\beta} \in \arg \min_{\|\beta\|_1 \le R} \left\{ \mathcal{L}_n(\beta) + \rho_\lambda(\beta) \right\},$$

loss function satisfies (α, τ)-RSC and regularizer is μ-amenable

Theorem (L. & Wainwright '13)

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

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

Statistical consistency

- Regularized M-estimator

$$\widehat{\beta} \in \arg\min_{\|\beta\|_1 \le R} \left\{ \mathcal{L}_n(\beta) + \rho_\lambda(\beta) \right\},$$

loss function satisfies (α, τ)-RSC and regularizer is μ-amenable

Theorem (L. & Wainwright '13)

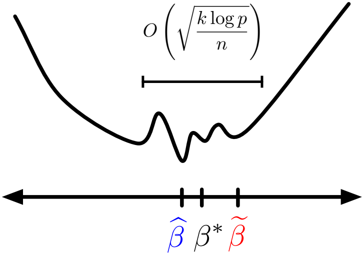
Suppose R is chosen s.t. β^* is feasible, and λ satisfies

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

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

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

More formally



- Stationary points statistically indistinguishable from global optima

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

- Nonasymptotic rates: For $\lambda \asymp \sqrt{\frac{\log p}{n}}$ and $R \asymp \frac{1}{\lambda}$,

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

Outline

1. Regularized M-estimators
   - Statistical M-estimation
   - Nonconvexity
   - Consistency of local optima

2. High-dimensional robust regression
   - Statistical consistency
   - Asymptotic normality
   - Two-step M-estimators

Robust regression functions

- Linear model:

$$y_i = x_i^T \beta^* + \epsilon_i$$

Heavy-tailed distribution on ϵ_i and/or outlier contamination

Robust regression functions

- Linear model:

$$y_i = x_i^T \beta^* + \epsilon_i$$

  Heavy-tailed distribution on ϵ_i and/or outlier contamination

- Use M-estimator

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

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

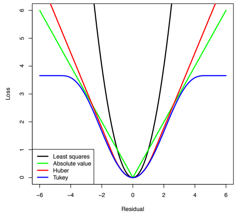
- **Bounded** ℓ' limits influence of outliers:

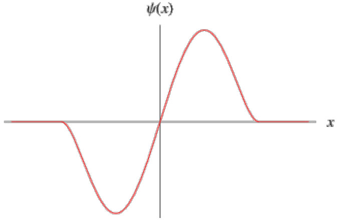
$$IF((x,y);T,F) = \lim_{t \to 0^+} \frac{T((1-t)F + t\delta_{(x,y)}) - T(F)}{t}$$
$$\propto \ell'(x^T\beta - y)x$$

  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

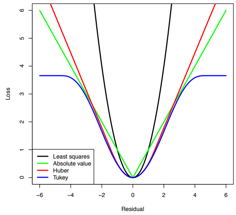
- **Bounded** ℓ' limits influence of outliers:

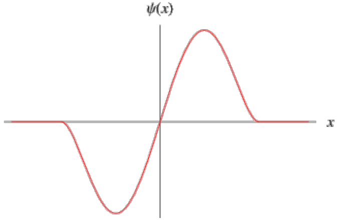
$$IF((x,y);T,F) = \lim_{t\to 0^+} \frac{T((1-t)F + t\delta_{(x,y)}) - T(F)}{t}$$
$$\propto \ell'(x^T\beta - y)x$$

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

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

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

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

- 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

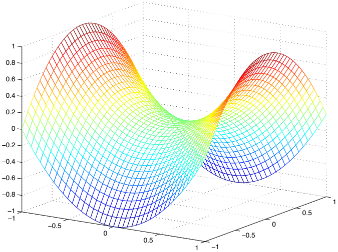
- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain
$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- $\ell'(\epsilon)$ 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

Local statistical consistency

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

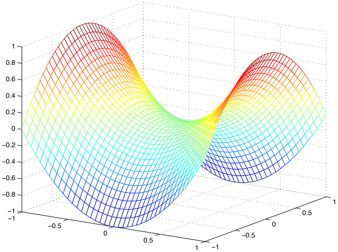
$$\langle \nabla \mathcal{L}_n(\beta_1) - \nabla \mathcal{L}_n(\beta_2),\ \Delta \rangle \geq \alpha \|\Delta\|_2^2 - \tau \frac{\log p}{n} \|\Delta\|_1^2, \quad \forall \|\beta_j - \beta^*\|_2 \leq r$$



Local statistical consistency

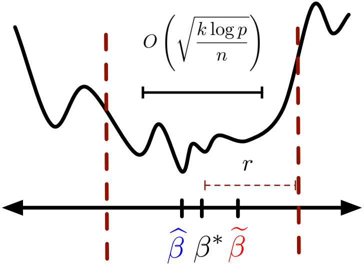
- **Local** RSC condition: For $\Delta := \beta_1 - \beta_2$,

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

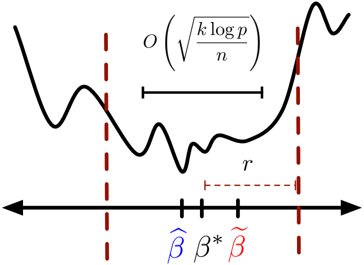


- Only requires restricted curvature within constant-radius region around β^*

Consistency of local stationary points



Consistency of local stationary points



Theorem (L. '15)

Suppose $\mathcal{L}_n$ satisfies α-local RSC and ρ_λ is μ-amenable, with $\alpha > \mu$. Suppose (λ, R) are chosen appropriately. For $n \gtrsim \frac{\tau}{\alpha-\mu} k \log p$, **any stationary point** $\widetilde{\beta}$ **s.t.** $\|\widetilde{\beta} - \beta^*\|_2 \leq r$ *satisfies*

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

Second-order considerations

- **Question:** If any bounded-derivative ℓ works for heavy-tailed distributions, why not always use Huber? LAD?

Second-order considerations

- **Question:** If any bounded-derivative ℓ works for heavy-tailed distributions, why not always use Huber? LAD?
- **Answer:** Has to do with (asymptotic) *efficiency*

Second-order considerations

- **Question:** If any bounded-derivative ℓ works for heavy-tailed distributions, why not always use Huber? LAD?
- **Answer:** Has to do with (asymptotic) *efficiency*

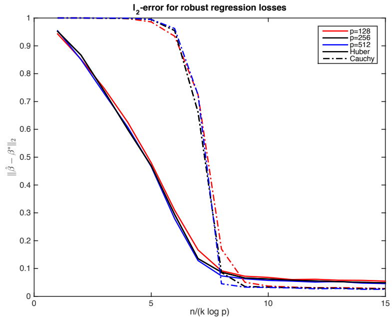
- In low-dimensional settings, MLE maximally efficient *with respect to variance*

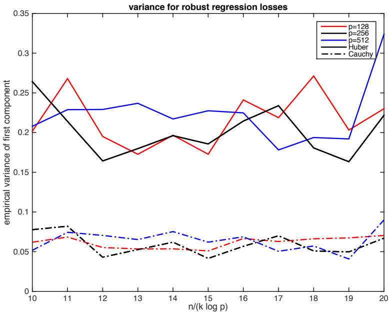
Second-order considerations

- **Question:** If any bounded-derivative ℓ works for heavy-tailed distributions, why not always use Huber? LAD?
- **Answer:** Has to do with (asymptotic) *efficiency*

- In low-dimensional settings, MLE maximally efficient *with respect to variance*
- Although MLE may behave erratically when $\frac{p}{n} \to (0, 1]$, can achieve simple asymptotic normality results under *sparsity* assumption, via oracle property

Asymptotic efficiency





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

How to obtain local efficient solutions?

- **Goal:** Nonconvex regularized M-estimator such that ℓ satisfies α-local RSC

How to obtain local efficient solutions?

- **Goal:** Nonconvex regularized M-estimator such that ℓ satisfies α-local RSC
- **Target:** Locate stationary point within radius r of β^*

How to obtain local efficient solutions?

- **Goal:** Nonconvex regularized *M*-estimator such that ℓ satisfies α-local RSC
- **Target:** Locate stationary point within radius *r* of β^*

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. 15)

- Use *composite gradient descent* starting from close initialization

Two-step algorithm (L. 15)

- Use *composite gradient descent* starting from close initialization
- **Two-step** M**-estimator:** Finds local stationary points of nonconvex, robust loss $+$ (μ, γ)-amenable penalty

Two-step algorithm (L. 15)

- Use *composite gradient descent* starting from close initialization
- **Two-step** M**-estimator:** Finds local stationary points of nonconvex, robust loss + (μ, γ)-amenable penalty

Algorithm

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

Two-step algorithm (L. 15)

- Use *composite gradient descent* starting from close initialization
- **Two-step *M*-estimator:** Finds local stationary points of nonconvex, robust loss + (μ, γ)-amenable penalty

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, γ)-amenable penalty, input $\beta^0 = \widehat{\beta}_H$

Two-step algorithm (L. 15)

- Use *composite gradient descent* starting from close initialization
- **Two-step M-estimator:** Finds local stationary points of nonconvex, robust loss + (μ, γ)-amenable penalty

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, γ)-amenable penalty, input $\beta^0 = \widehat{\beta}_H$

- Theoretical guarantees on (rate of) convergence to optimal point

Summary of two-step M-estimator

- Output is **computationally** and **statistically** efficient:

Summary of two-step *M*-estimator

- Output is **computationally** and **statistically** efficient:

- Computational guarantee on rate of convergence in each step of *M*-estimation algorithm
- Statistical guarantee on asymptotic efficiency of estimator (assuming β-min condition and (μ, γ)-amenability)

Summary

- Theory for nonconvex regularized M-estimators
  - Global RSC condition $\implies$ *all* stationary points within statistical error of β^*
  - Local RSC condition $\implies$ stationary points *in local region* within statistical error of β^*

Summary

- Theory for nonconvex regularized M-estimators
  - Global RSC condition $\Longrightarrow$ *all* stationary points within statistical error of β^*
  - Local RSC condition $\Longrightarrow$ stationary points *in local region* within statistical error of β^*
- Consequences for high-dimensional robust regression estimators
  - Consistency under relaxed distributional assumptions when ℓ' bounded

Summary

- Theory for nonconvex regularized *M*-estimators
  - Global RSC condition $\Longrightarrow$ *all* stationary points within statistical error of β^*
  - Local RSC condition $\Longrightarrow$ stationary points *in local region* within statistical error of β^*
- Consequences for high-dimensional robust regression estimators
  - Consistency under relaxed distributional assumptions when ℓ' bounded
  - Oracle estimator with (μ, γ)-amenable regularizer $\Longrightarrow$ asymptotic efficiency
  - Two-step *M*-estimator produces local oracle solutions

References

- **P. Loh** and M. J. Wainwright (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*.
- **P. Loh** (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Annals of Statistics.*

Thank you!