Estimators based on non-convex programs: Statistical and computational guarantees

Martin Wainwright

UC Berkeley
Statistics and EECS

Based on joint work with:

Po-Ling Loh (UC Berkeley)

Introduction

Prediction/regression problems arise throughout statistics:

- vector of predictors/covariates $x \in \mathbb{R}^p$
- response variable $y \in \mathbb{R}$
- In "big data" setting:
  - both sample size n and ambient dimension p are large
  - many problems have $p \gg n$

Introduction

Prediction/regression problems arise throughout statistics:

- vector of predictors/covariates $x \in \mathbb{R}^p$
- response variable $y \in \mathbb{R}$
- In "big data" setting:
  - both sample size n and ambient dimension p are large
  - many problems have $p \gg n$

Regularization is essential:

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i)}_{\mathcal{L}_n(\theta; x_1^n, y_1^n)} + \underbrace{\mathcal{R}_\lambda(\theta)}_{\text{Regularizer}} \Big\}.$$

Introduction

Prediction/regression problems arise throughout statistics:

- In "big data" setting:
  - both sample size n and ambient dimension p are large
  - many problems have $p \gg n$

Regularization is essential:

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i)}_{\mathcal{L}_n(\theta; x_1^n, y_1^n)} + \underbrace{\mathcal{R}_\lambda(\theta)}_{\text{Regularizer}} \Big\}.$$

For non-convex problems: "Mind the gap!"

- any global optimum is "statistically good"....
- but efficient algorithms only find local optima

Introduction

Prediction/regression problems arise throughout statistics:

- In "big data" setting:
  - both sample size n and ambient dimension p are large
  - many problems have $p \gg n$

Regularization is essential:

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i)}_{\mathcal{L}_n(\theta; x_1^n, y_1^n)} + \underbrace{\mathcal{R}_\lambda(\theta)}_{\text{Regularizer}} \Big\}.$$

For non-convex problems: "Mind the gap!"

- any global optimum is "statistically good"....
- but efficient algorithms only find local optima

Question

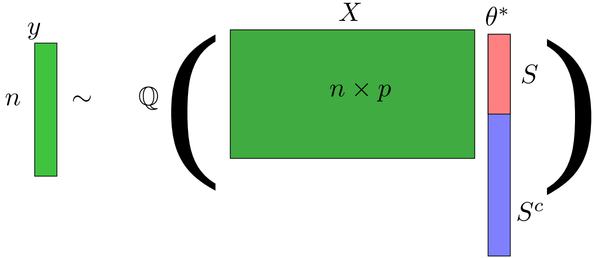
How to close this undesirable gap between statistics and computation?

Vignette A: Regression with non-convex penalties

Set-up: **Observe** (y_i, x_i) pairs for $i = 1, 2, \ldots, n$, where

$$y_i \sim \mathbb{Q}\big(\cdot \mid \langle \theta^*,\, x_i \rangle\big),$$

where $\theta \in \mathbb{R}^p$ has "low-dimensional structure"

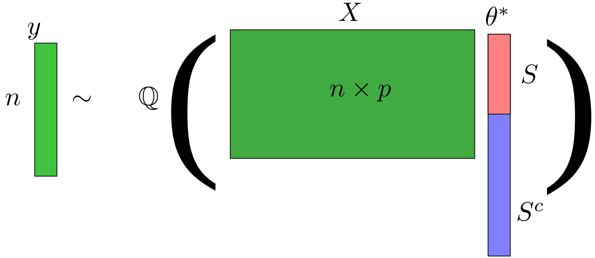


Vignette A: Regression with non-convex penalties

Set-up: **Observe** (y_i, x_i) pairs for $i = 1, 2, \ldots, n$, where

$$y_i \sim \mathbb{Q}\big(\cdot \mid \langle \theta^*,\, x_i \rangle\big),$$

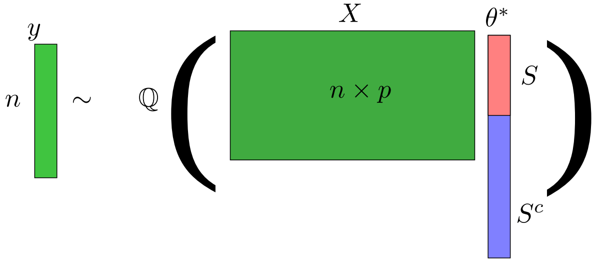
where $\theta \in \mathbb{R}^p$ has "low-dimensional structure"



Estimator: $\mathcal{R}_\lambda$-regularized likelihood

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{Q}\big(y_i \mid \langle x_i,\, \theta \rangle\big) + \mathcal{R}_\lambda(\theta) \Big\}.$$

Vignette A: Regression with non-convex penalties



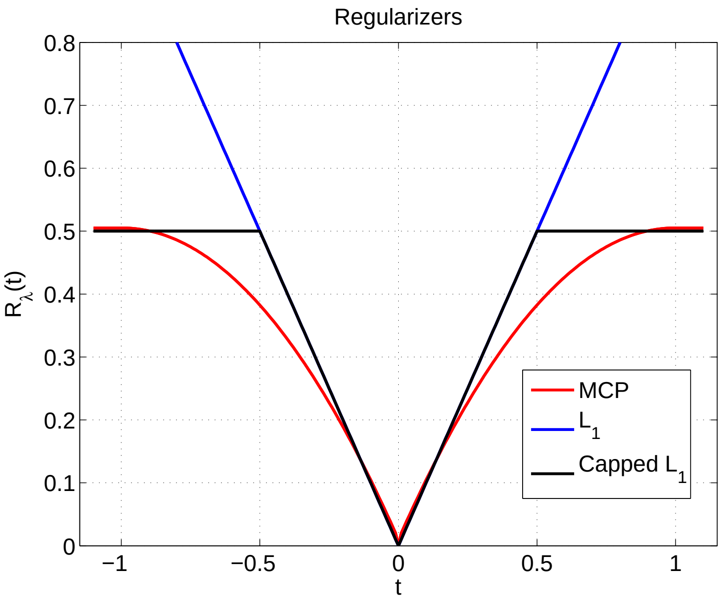
Example: Logistic regression for binary responses $y_i \in \{0, 1\}$:

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big\{ \log(1 + e^{\langle x_i,\, \theta \rangle}) - y_i \langle x_i,\, \theta \rangle \big\} + \mathcal{R}_\lambda(\theta) \Big\}.$$

Many non-convex penalties are possible:

- capped ℓ_1-penalty
- SCAD penalty (Fan & Li, 2001)
- MCP penalty (Zhang, 2006)

Convex and non-convex regularizers



Statistical error versus optimization error

Algorithm generating sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ to solve

$$\widehat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i) + \mathcal{R}_\lambda(\theta) \right\}.$$

Global minimizer of population risk

$$\theta^* := \arg\min_{\theta} \underbrace{\mathbb{E}_{X,Y}\Big[\mathcal{L}(\theta; X, Y)\Big]}_{\bar{\mathcal{L}}(\theta)}$$

Statistical error versus optimization error

Algorithm generating sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ to solve

$$\widehat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i) + \mathcal{R}_\lambda(\theta) \right\}.$$

Global minimizer of population risk

$$\theta^* := \arg\min_{\theta} \underbrace{\mathbb{E}_{X,Y}\Big[\mathcal{L}(\theta; X, Y)\Big]}_{\bar{\mathcal{L}}(\theta)}$$

Goal of statistician

Provide bounds on Statistical error: $\|\theta^t - \theta^*\|$ or $\|\widehat{\theta} - \theta^*\|$

Statistical error versus optimization error

Algorithm generating sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ to solve

$$\widehat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i) + \mathcal{R}_\lambda(\theta) \right\}.$$

Global minimizer of population risk

$$\theta^* := \arg\min_{\theta} \underbrace{\mathbb{E}_{X,Y}\Big[\mathcal{L}(\theta; X, Y)\Big]}_{\bar{\mathcal{L}}(\theta)}$$

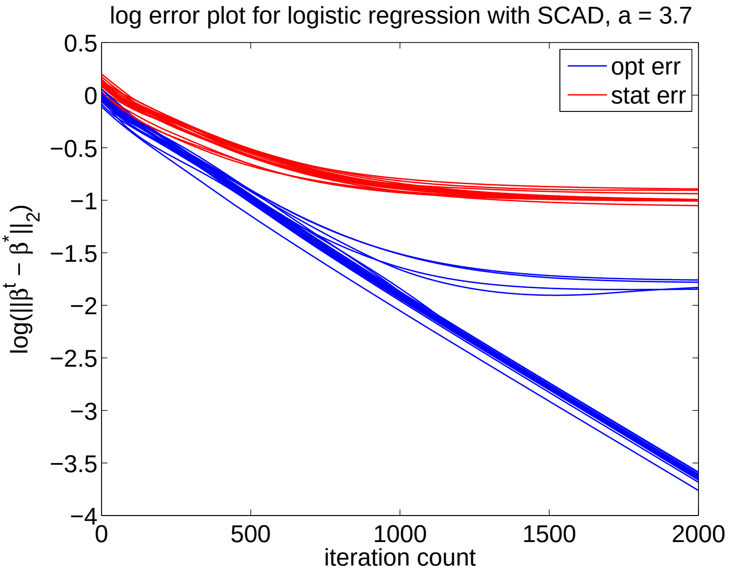
Goal of statistician

Provide bounds on Statistical error: $\|\theta^t - \theta^*\|$ or $\|\widehat{\theta} - \theta^*\|$

Goal of optimization-theorist

Provide bounds on Optimization error: $\|\theta^t - \widehat{\theta}\|$

Logistic regression with non-convex regularizer



What phenomena need to be explained?

Empirical observation #1:

From a statistical perspective, all local optima are essentially as good as a global optimum.

What phenomena need to be explained?

Empirical observation #1:

From a statistical perspective, all local optima are essentially as good as a global optimum.

Some past work:

- for least-squares loss, certain local optima are good (Zhang & Zhang, 2012)
- if initialized at Lasso solution with ℓ_∞-guarantees, local algorithm has good behavior (Fan et al., 2012)

What phenomena need to be explained?

Empirical observation #1:

From a statistical perspective, all local optima are essentially as good as a global optimum.

Some past work:

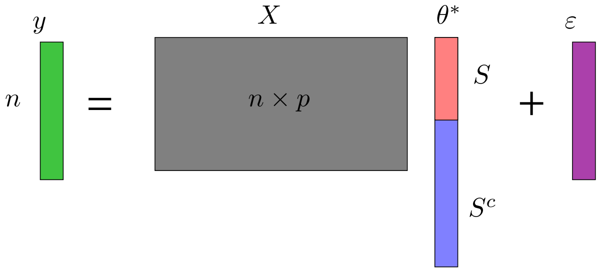
- for least-squares loss, certain local optima are good (Zhang & Zhang, 2012)
- if initialized at Lasso solution with ℓ_∞-guarantees, local algorithm has good behavior (Fan et al., 2012)

Empirical observation #2:

First-order methods converge as fast as possible up to statistical precision.

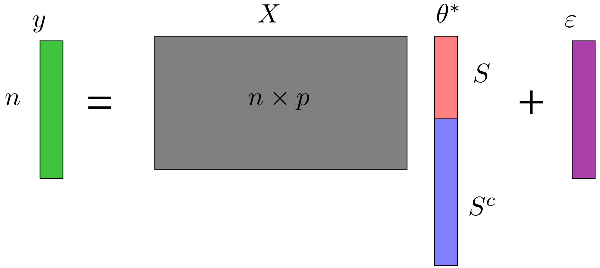
Vignette B: Error-in-variables regression

Begin with high-dimensional sparse regression:

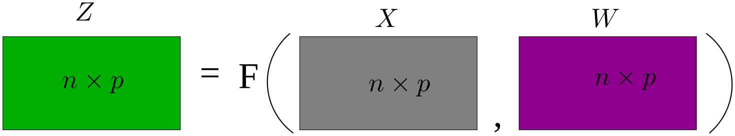


Vignette B: Error-in-variables regression

Begin with high-dimensional sparse regression:



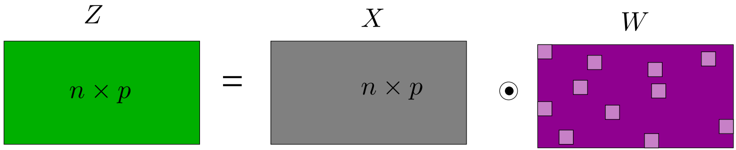
Observe $y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{n \times p}$



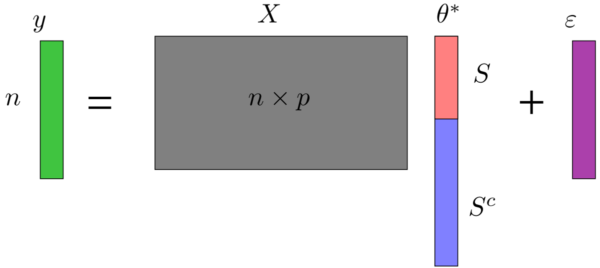
Here $W \in \mathbb{R}^{n \times p}$ is a stochastic perturbation.

Example: Missing data

Missing data:

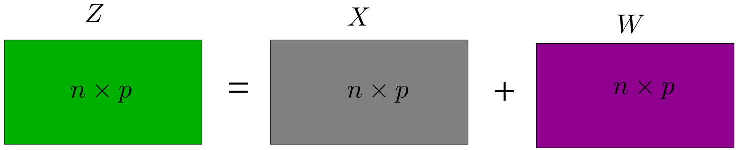


Here $W \in \mathbb{R}^{n \times p}$ is multiplicative perturbation (e.g., $W_{ij} \sim \mathrm{Ber}(\alpha)$.)

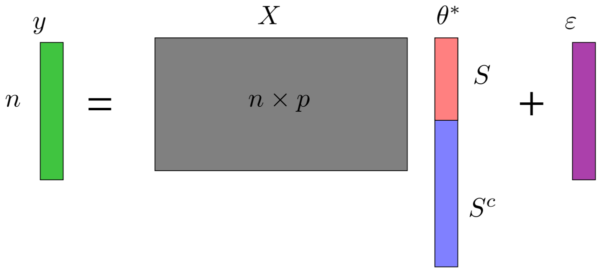


Example: Additive perturbations

Additive noise in covariates:



Here $W \in \mathbb{R}^{n \times p}$ is an additive perturbation (e.g., $W_{ij} \sim N(0, \sigma^2)$).



A second look at regularized least-squares

Equivalent formulation:

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{ \frac{1}{2} \theta^T \big(\frac{X^T X}{n}\big) \theta \;-\; \langle \theta, \, \frac{X^T y}{n} \rangle \;+\; \mathcal{R}_\lambda(\theta) \Big\}.$$

A second look at regularized least-squares

Equivalent formulation:

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{ \frac{1}{2} \theta^T \big(\frac{X^T X}{n}\big) \theta \; - \; \langle \theta, \, \frac{X^T y}{n} \rangle \; + \; \mathcal{R}_\lambda(\theta) \Big\}.$$

Population view: unbiased estimators

$$\text{cov}(x_1) = \mathbb{E}\Big[\frac{X^T X}{n}\Big], \quad \text{and} \quad \text{cov}(x_1, y_1) = \mathbb{E}\Big[\frac{X^T y}{n}\Big].$$

Corrected estimators

Equivalent formulation:

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{ \frac{1}{2} \theta^T \big(\frac{X^T X}{n}\big) \theta \; - \; \langle \theta, \; \frac{X^T y}{n} \rangle \; + \; \mathcal{R}_\lambda(\theta) \Big\}.$$

Population view: unbiased estimators

$$\operatorname{cov}(x_1) = \mathbb{E}\Big[\frac{X^T X}{n}\Big], \quad \text{and} \quad \operatorname{cov}(x_1, y_1) = \mathbb{E}\Big[\frac{X^T y}{n}\Big].$$

A general family of estimators

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta \; - \; \theta^T \widehat{\gamma} \; + \; \lambda_n^2 \|\theta\|_1^2 \Big\},$$

where $(\widehat{\Gamma}, \widehat{\gamma})$ are unbiased estimators of $\operatorname{cov}(x_1)$ and $\operatorname{cov}(x_1, y_1)$.

Example: Estimator for missing data

- observe corrupted version $Z \in \mathbb{R}^{n \times p}$

$$Z_{ij} = \begin{cases} X_{ij} & \text{with probability } 1 - \alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

Example: Estimator for missing data

- observe corrupted version $Z \in \mathbb{R}^{n \times p}$

$$Z_{ij} = \begin{cases} X_{ij} & \text{with probability } 1-\alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

- Natural unbiased estimates: set $\star \equiv 0$ and $\widehat{Z} := \frac{Z}{(1-\alpha)}$:

$$\widehat{\Gamma} \quad = \quad \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \operatorname{diag}\big(\frac{\widehat{Z}^T \widehat{Z}}{n}\big), \quad \text{and} \quad \widehat{\gamma} \quad = \quad \frac{\widehat{Z}^T y}{n},$$

Example: Estimator for missing data

- observe corrupted version $Z \in \mathbb{R}^{n \times p}$

$$Z_{ij} = \begin{cases} X_{ij} & \text{with probability } 1-\alpha \\ \star & \text{with probability } \alpha. \end{cases}$$

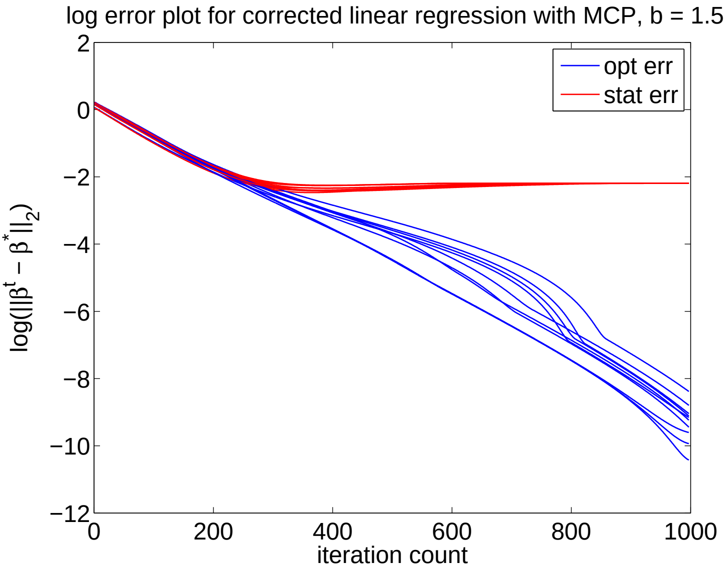
- Natural unbiased estimates: set $\star \equiv 0$ and $\widehat{Z} := \frac{Z}{(1-\alpha)}$:

$$\widehat{\Gamma} = \frac{\widehat{Z}^T \widehat{Z}}{n} - \alpha \,\text{diag}\left(\frac{\widehat{Z}^T \widehat{Z}}{n}\right), \quad \text{and} \quad \widehat{\gamma} = \frac{\widehat{Z}^T y}{n},$$

- solve (doubly non-convex) optimization problem: (Loh & W., 2012)

$$\widehat{\theta} \in \arg\min_{\theta \in \Omega} \left\{ \frac{1}{2} \theta^T \widehat{\Gamma} \theta - \langle \widehat{\gamma}, \theta \rangle + \mathcal{R}_\lambda(\theta) \right\}.$$

Non-convex quadratic and non-convex regularizer

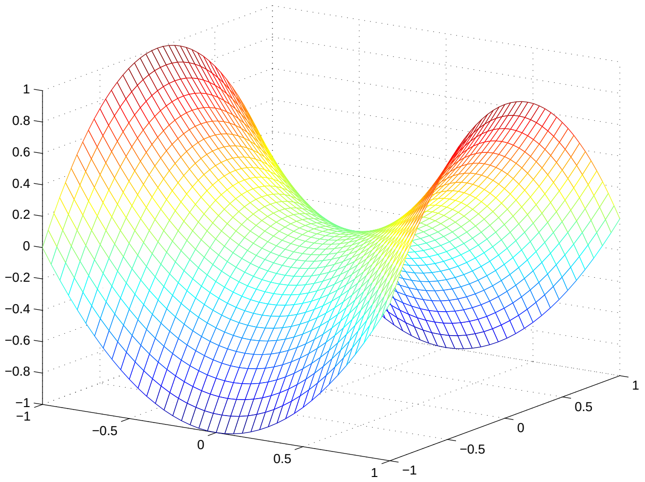


Remainder of talk

1. Why are all local optima "statistically good"?
   - Restricted strong convexity
   - A general theorem
   - Various examples

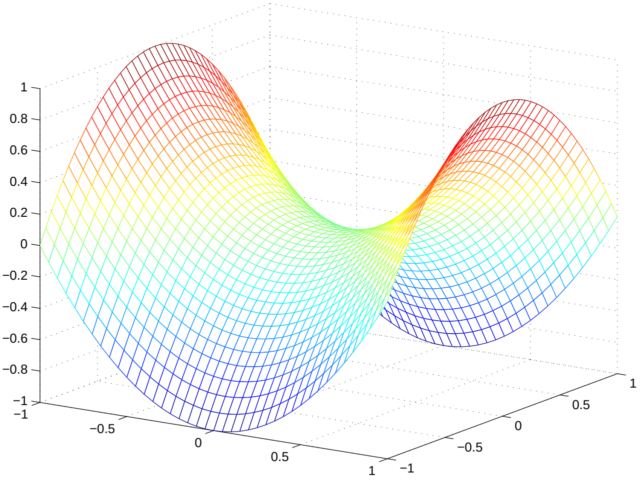
2. Why do first-order gradient methods converge quickly?
   - Composite gradient methods
   - Statistical versus optimization error
   - Fast convergence for non-convex problems

Geometry of a non-convex quadratic loss



- Loss function has directions of both positive and negative curvature.

Geometry of a non-convex quadratic loss



- Loss function has directions of both positive and negative curvature.
- Negative directions must be forbidden by regularizer.

Restricted strong convexity

Here defined with respect to the ℓ_1-norm:

Definition

The loss function $\mathcal{L}_n$ satisfies RSC with parameters (α_j, τ_j), $j = 1, 2$ if

$$\underbrace{\langle \nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*), \, \Delta \rangle}_{\text{Measure of curvature}} \geq \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2 & \text{if } \|\Delta\|_2 \leq 1 \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 & \text{if } \|\Delta\|_2 > 1. \end{cases}$$

Restricted strong convexity

Here defined with respect to the ℓ_1-norm:

Definition

The loss function $\mathcal{L}_n$ satisfies RSC with parameters (α_j, τ_j), $j = 1, 2$ if

$$\underbrace{\langle \nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*),\, \Delta \rangle}_{\text{Measure of curvature}} \geq \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2 & \text{if } \|\Delta\|_2 \leq 1 \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 & \text{if } \|\Delta\|_2 > 1. \end{cases}$$

- holds with $\tau_1 = \tau_2 = 0$ for any function that is locally strongly convex around θ^*

Restricted strong convexity

Here defined with respect to the ℓ_1-norm:

Definition

The loss function $\mathcal{L}_n$ satisfies RSC with parameters (α_j, τ_j), $j = 1, 2$ if

$$\underbrace{\langle \nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*),\, \Delta \rangle}_{\text{Measure of curvature}} \geq \begin{cases} \alpha_1 \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2 & \text{if } \|\Delta\|_2 \leq 1 \\ \alpha_2 \|\Delta\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 & \text{if } \|\Delta\|_2 > 1. \end{cases}$$

- holds with $\tau_1 = \tau_2 = 0$ for any function that is locally strongly convex around θ^*
- holds for a variety of loss functions (convex and non-convex):
  - ordinary least-squares (Raskutti, W. & Yu, 2010)
  - likelihoods for generalized linear models (Negahban et al., 2012)
  - certain non-convex quadratic functions (Loh & W, 2012)

Well-behaved regularizers

Properties defined at the univariate level $\mathcal{R}_\lambda : \mathbb{R} \to [0, \infty]$.

- Satisfies $\mathcal{R}_\lambda(0) = 0$, and is symmetric around zero ($\mathcal{R}_\lambda(t) = \mathcal{R}_\lambda(-t)$.)
- Non-decreasing and subadditive $\mathcal{R}_\lambda(s + t) \leq \mathcal{R}_\lambda(s) + \mathcal{R}_\lambda(t)$.
- Function $t \mapsto \frac{\mathcal{R}_\lambda(t)}{t}$ is nonincreasing for $t > 0$
- Differentiable for all $t \neq 0$, subdifferentiable at $t = 0$ with subgradients bounded in absolute value by λL.
- For some $\mu > 0$, the function $\widetilde{\mathcal{R}}_\lambda(t) = \mathcal{R}_\lambda(t) + \mu t^2$ is convex.

Well-behaved regularizers

Properties defined at the univariate level $\mathcal{R}_\lambda : \mathbb{R} \rightarrow [0, \infty]$.

- Satisfies $\mathcal{R}_\lambda(0) = 0$, and is symmetric around zero ($\mathcal{R}_\lambda(t) = \mathcal{R}_\lambda(-t)$.)
- Non-decreasing and subadditive $\mathcal{R}_\lambda(s+t) \leq \mathcal{R}_\lambda(s) + \mathcal{R}_\lambda(t)$.
- Function $t \mapsto \frac{\mathcal{R}_\lambda(t)}{t}$ is nonincreasing for $t > 0$
- Differentiable for all $t \neq 0$, subdifferentiable at $t = 0$ with subgradients bounded in absolute value by λL.
- For some $\mu > 0$, the function $\widetilde{\mathcal{R}}_\lambda(t) = \mathcal{R}_\lambda(t) + \mu t^2$ is convex.

Includes (among others):

- rescaled ℓ_1 loss: $\mathcal{R}_\lambda(t) = \lambda|t|$.
- MCP penalty and SCAD penalties (Fan et al., 2001; Zhang, 2006)
- does not include capped ℓ_1-penalty

Main statistical guarantee

- regularized M-estimator
$$\widehat{\theta} \in \arg \min_{\|\theta\|_1 \leq M} \Big\{ \mathcal{L}_n(\theta) + \mathcal{R}_\lambda(\theta) \Big\}.$$
- loss function satisfies (α, τ) RSC, and regularizer is regular (with parameters (μ, L))

Main statistical guarantee

- regularized M-estimator

$$\widehat{\theta} \in \arg \min_{\|\theta\|_1 \leq M} \Big\{ \mathcal{L}_n(\theta) + \mathcal{R}_\lambda(\theta) \Big\}.$$

- loss function satisfies (α, τ) RSC, and regularizer is regular (with parameters (μ, L))
- local optimum $\widehat{\theta}$ defined by conditions

$$\langle \nabla \mathcal{L}_n(\widehat{\theta}) + \nabla \mathcal{R}_\lambda(\widehat{\theta}),\, \theta - \widehat{\theta} \rangle \geq 0 \qquad \text{for all feasible } \theta.$$

Main statistical guarantee

- regularized M-estimator

$$\widehat{\theta} \in \arg \min_{\|\theta\|_1 \leq M} \Big\{ \mathcal{L}_n(\theta) + \mathcal{R}_\lambda(\theta) \Big\}.$$

- loss function satisfies (α, τ) RSC, and regularizer is regular (with parameters (μ, L))
- local optimum $\widehat{\theta}$ defined by conditions

$$\langle \nabla \mathcal{L}_n(\widehat{\theta}) + \nabla \mathcal{R}_\lambda(\widehat{\theta}), \, \theta - \widehat{\theta} \rangle \geq 0 \qquad \text{for all feasible } \theta.$$

Theorem (Loh & W., 2013)

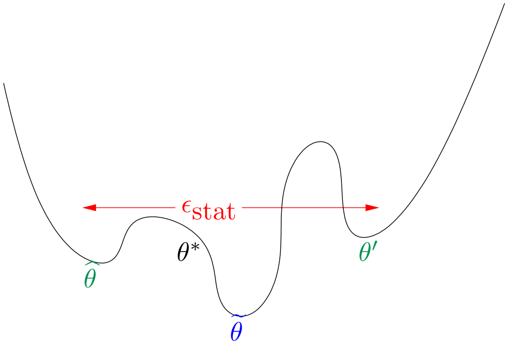
Suppose M is chosen such that θ^* is feasible, and λ satisfies the bounds

$$\max \Big\{ \|\nabla \mathcal{L}_n(\theta^*)\|_\infty, \alpha_2 \sqrt{\frac{\log p}{n}} \Big\} \;\leq\; \lambda \;\leq\; \frac{\alpha_2}{6LM}$$

Then any local optimum $\widehat{\theta}$ satisfies the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{6 \, \lambda_n \sqrt{s}}{4 \, (\alpha - \mu)} \qquad \textit{where } s = \|\beta^*\|_0.$$

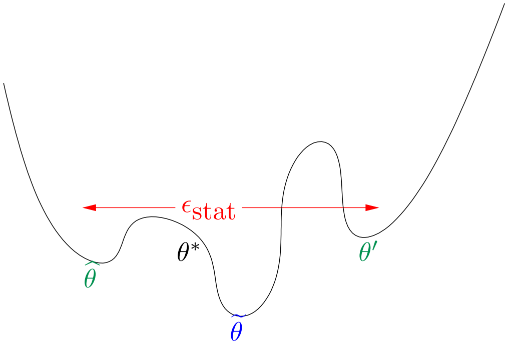
Geometry of local/global optima



Consequence:

All { local, global } optima are within distance ϵ_{stat} of the target θ^*.

Geometry of local/global optima



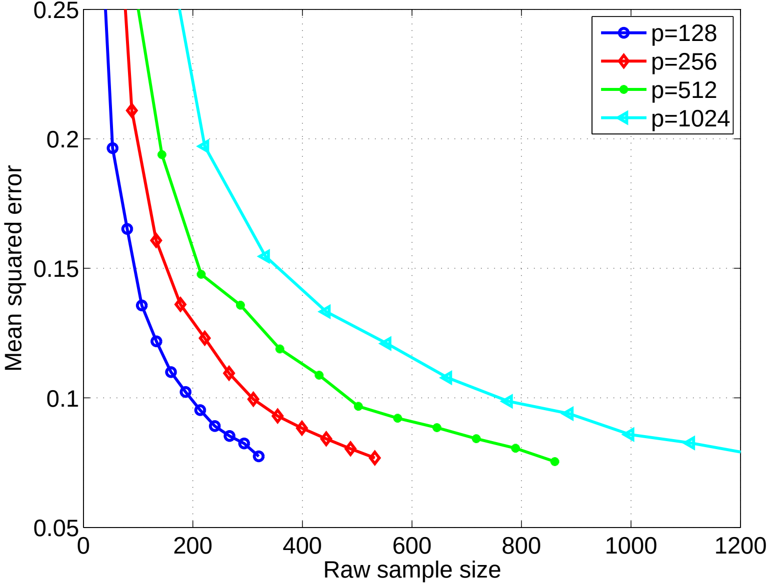
Consequence:

All { local, global } optima are within distance ϵ_{stat} of the target θ^*.

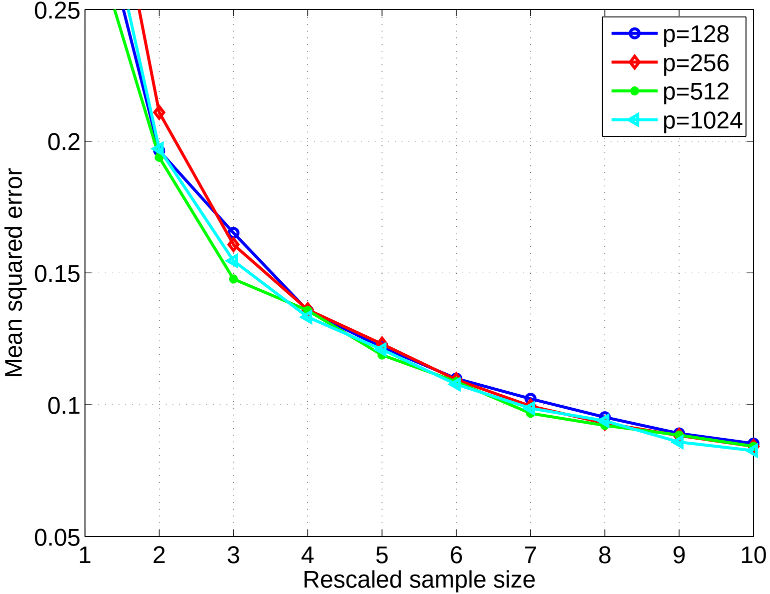
With $\lambda = c\sqrt{\frac{\log p}{n}}$, statistical error scales as

$$\epsilon_{\text{stat}} \asymp \sqrt{\frac{s \log p}{n}}, \qquad \text{which is minimax optimal.}$$

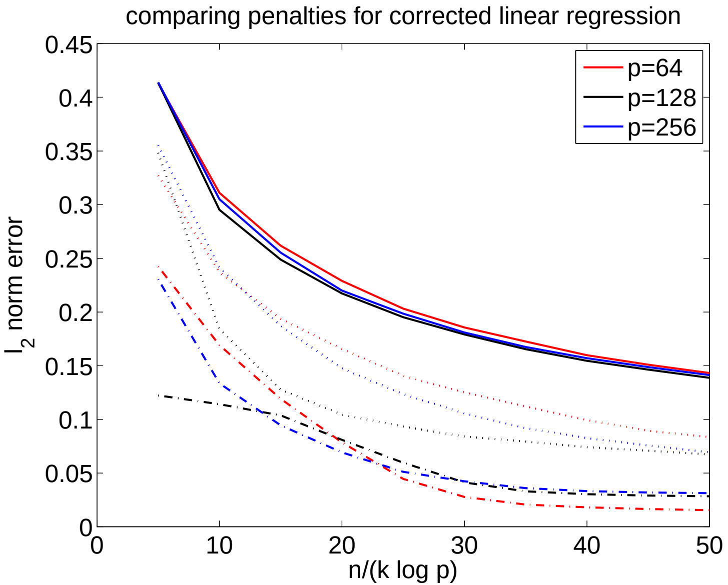
Empirical results (unrescaled)



Empirical results (rescaled)



Comparisons between different penalties



First-order algorithms and fast convergence

Thus far....

- have shown that all local optima are "statistically good"
- how to obtain a local optimum quickly?

First-order algorithms and fast convergence

Thus far....

- have shown that all local optima are "statistically good"
- how to obtain a local optimum quickly?

Composite gradient descent for regularized objectives: (Nesterov, 2007)

$$\min_{\theta \in \Omega} \left\{ f(\theta) + g(\theta) \right\}$$

where f is differentiable, and g is convex, sub-differentiable.

First-order algorithms and fast convergence

Thus far....

- have shown that all local optima are “statistically good”
- how to obtain a local optimum quickly?

Composite gradient descent for regularized objectives: (Nesterov, 2007)

$$\min_{\theta \in \Omega} \Big\{ f(\theta) + g(\theta) \Big\}$$

where f is differentiable, and g is convex, sub-differentiable.

Simple updates:

$$\theta^{t+1} = \arg\min_{\theta \in \Omega} \Big\{ \|\theta - \alpha^t \nabla f(\theta^t)\|_2^2 + g(\theta) \Big\}.$$

First-order algorithms and fast convergence

Thus far....

- have shown that all local optima are "statistically good"
- how to obtain a local optimum quickly?

Composite gradient descent for regularized objectives: (Nesterov, 2007)

$$\min_{\theta \in \Omega} \Big\{ f(\theta) + g(\theta) \Big\}$$

where f is differentiable, and g is convex, sub-differentiable.

Simple updates:

$$\theta^{t+1} = \arg\min_{\theta \in \Omega} \Big\{ \|\theta - \alpha^t \nabla f(\theta^t)\|_2^2 + g(\theta) \Big\}.$$

Not directly applicable with $f = \mathcal{L}_n$ and $g = \mathcal{R}_\lambda$ (since $\mathcal{R}_\lambda$ can be non-convex).

Composite gradient on a convenient splitting

- Define modified loss functions and regularizers:

$$\widetilde{\mathcal{L}}_n(\theta) := \underbrace{\mathcal{L}_n(\theta) - \mu\|\theta\|_2^2}_{\text{non-convex}}, \quad \text{and} \quad \widetilde{\mathcal{R}}_\lambda(\theta) := \underbrace{\mathcal{R}_\lambda(\theta) + \mu\|\theta\|_2^2}_{\text{convex}}.$$

Composite gradient on a convenient splitting

- Define modified loss functions and regularizers:

$$\widetilde{\mathcal{L}}_n(\theta) := \underbrace{\mathcal{L}_n(\theta) - \mu\|\theta\|_2^2}_{\text{non-convex}}, \quad \text{and} \quad \widetilde{\mathcal{R}}_\lambda(\theta) := \underbrace{\mathcal{R}_\lambda(\theta) + \mu\|\theta\|_2^2}_{\text{convex}}.$$

- Apply composite gradient descent to the objective

$$\min_{\theta \in \Omega} \left\{ \widetilde{\mathcal{L}}_n(\theta) + \widetilde{\mathcal{R}}_\lambda(\theta) \right\}.$$

Composite gradient on a convenient splitting

- Define modified loss functions and regularizers:

$$\widetilde{\mathcal{L}}_n(\theta) := \underbrace{\mathcal{L}_n(\theta) - \mu\|\theta\|_2^2}_{\text{non-convex}}, \quad \text{and} \quad \widetilde{\mathcal{R}}_\lambda(\theta) := \underbrace{\mathcal{R}_\lambda(\theta) + \mu\|\theta\|_2^2}_{\text{convex}}.$$

- Apply composite gradient descent to the objective

$$\min_{\theta \in \Omega} \left\{ \widetilde{\mathcal{L}}_n(\theta) + \widetilde{\mathcal{R}}_\lambda(\theta) \right\}.$$

- converges to local optimum $\widehat{\theta}$ (Nesterov, 2007)

$$\langle \nabla \widetilde{\mathcal{L}}_n(\theta) + \nabla \widetilde{\mathcal{R}}_\lambda(\widehat{\theta}), \, \theta - \widehat{\theta} \rangle \geq 0 \qquad \text{for all feasible } \theta.$$

Composite gradient on a convenient splitting

- Define modified loss functions and regularizers:

$$\widetilde{\mathcal{L}}_n(\theta) := \underbrace{\mathcal{L}_n(\theta) - \mu\|\theta\|_2^2}_{\text{non-convex}}, \quad \text{and} \quad \widetilde{\mathcal{R}}_\lambda(\theta) := \underbrace{\mathcal{R}_\lambda(\theta) + \mu\|\theta\|_2^2}_{\text{convex}}.$$

- Apply composite gradient descent to the objective

$$\min_{\theta \in \Omega} \Big\{ \widetilde{\mathcal{L}}_n(\theta) + \widetilde{\mathcal{R}}_\lambda(\theta) \Big\}.$$

- converges to local optimum $\widehat{\theta}$ (Nesterov, 2007)

$$\langle \nabla \widetilde{\mathcal{L}}_n(\theta) + \nabla \widetilde{\mathcal{R}}_\lambda(\widehat{\theta}), \, \theta - \widehat{\theta} \rangle \geq 0 \qquad \text{for all feasible } \theta.$$

- will show that convergence is geometrically fast with constant stepsize

Theoretical guarantees on computational error

- implement Nesterov's composite method with constant stepsize to $(\widetilde{\mathcal{L}}_n, \widetilde{\mathcal{R}}_\lambda)$ split.
- fixed global optimum $\widehat{\beta}$ defines the statistical error $\epsilon^2_{\text{stat}} = \|\widehat{\beta} - \beta^*\|_2$.
- population minimizer β^* is s-sparse

Theoretical guarantees on computational error

- implement Nesterov's composite method with constant stepsize to $(\widetilde{\mathcal{L}}_n, \widetilde{\mathcal{R}}_\lambda)$ split.

- fixed global optimum $\widehat{\beta}$ defines the statistical error $\epsilon^2_{\text{stat}} = \|\widehat{\beta} - \beta^*\|_2$.

- population minimizer β^* is s-sparse

- loss function satisfies (α, τ) RSC and smoothness conditions, and regularizer is (μ, L)-good

Theoretical guarantees on computational error

- implement Nesterov's composite method with constant stepsize to $(\widetilde{\mathcal{L}}_n, \widetilde{\mathcal{R}}_\lambda)$ split.
- fixed global optimum $\widehat{\beta}$ defines the statistical error $\epsilon^2_{\text{stat}} = \|\widehat{\beta} - \beta^*\|_2$.
- population minimizer β^* is s-sparse
- loss function satisfies (α, τ) RSC and smoothness conditions, and regularizer is (μ, L)-good

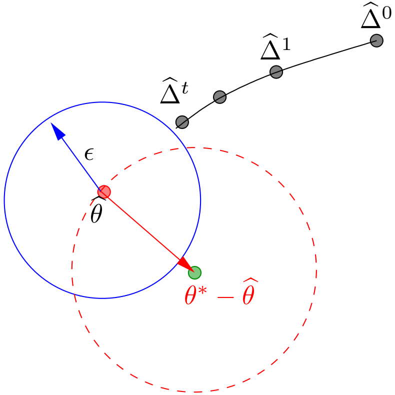
Theorem (Loh & W., 2013)

If $n \gtrsim s \log p$, there is a contraction factor $\kappa \in (0,1)$ such that for any $\delta \geq \epsilon_{stat}$, we have

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \frac{2}{\alpha - \mu}\Big(\delta^2 + 128\tau \frac{s \log p}{n} \epsilon^2_{stat}\Big) \qquad \textit{for all } t \geq T(\delta) \textit{ iterations,}$$

where $T(\delta) \asymp \frac{\log(1/\delta)}{\log(1/\kappa)}$.

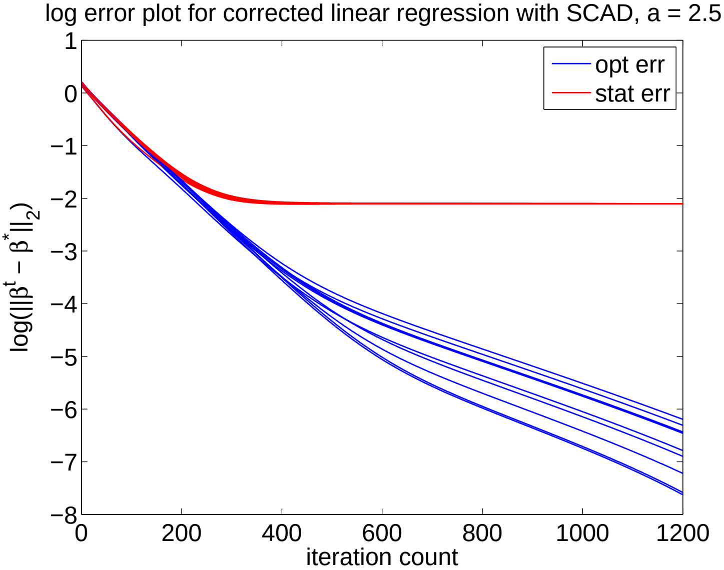
Geometry of result



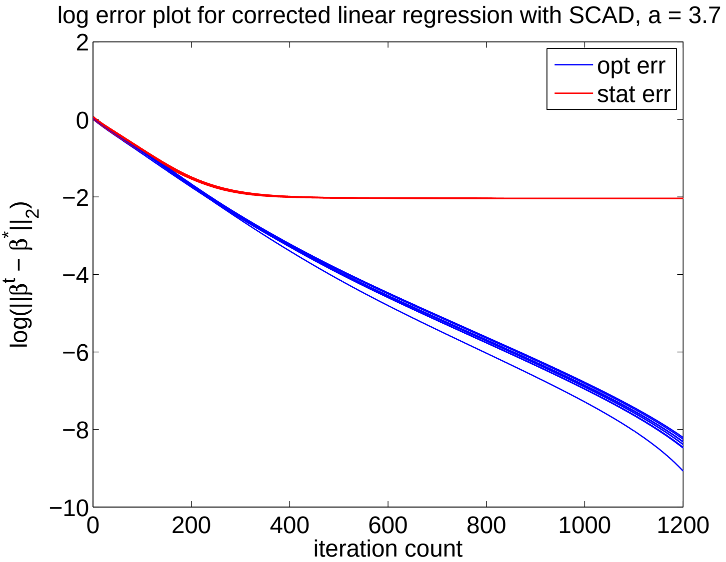
Optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ decreases geometrically up to statistical tolerance:

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + o\Big(\underbrace{\|\theta^* - \widehat{\theta}\|^2}_{\text{Stat. error } \epsilon^2_{\text{stat}}} \Big) \qquad \text{for all } t = 0, 1, 2, \ldots.$$

Non-convex linear regression with SCAD



Non-convex linear regression with SCAD



Summary

- M-estimators based on non-convex programs arise frequently
- under suitable regularity conditions, we showed that:
  - all local optima are "well-behaved" from the statistical point of view
  - simple first-order methods converge as fast as possible

Summary

- M-estimators based on non-convex programs arise frequently
- under suitable regularity conditions, we showed that:
  - all local optima are "well-behaved" from the statistical point of view
  - simple first-order methods converge as fast as possible
- many open questions
  - similar guarantees for more general problems?
  - geometry of non-convex problems in statistics?

Summary

- M-estimators based on non-convex programs arise frequently
- under suitable regularity conditions, we showed that:
  - all local optima are "well-behaved" from the statistical point of view
  - simple first-order methods converge as fast as possible
- many open questions
  - similar guarantees for more general problems?
  - geometry of non-convex problems in statistics?

Papers and pre-prints:

- Loh & W. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Pre-print arXiv:1305.2436*
- Loh & W. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40:1637–1664.
- Negahban et al. (2012). A unified framework for high-dimensional analysis of M-estimators. *Statistical Science*, 27(4): 538–557.