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:

\[
\hat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i) + \mathcal{R}_n(\theta) \right\}.
\]
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:

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

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:

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

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

\[
y_i \sim Q(\cdot \mid \langle \theta^*, x_i \rangle),
\]

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

\[ y_i \sim Q(\cdot | \langle \theta^*, x_i \rangle), \]

where \(\theta \in \mathbb{R}^p\) has “low-dimensional structure”

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

\[
\hat{\theta} \in \arg \min_{\theta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log Q(y_i | \langle x_i, \theta \rangle) + \mathcal{R}_\lambda(\theta) \right\}.
\]
Vignette A: Regression with non-convex penalties

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

\[
\hat{\theta} \in \arg \min_\theta \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + e^{\langle x_i, \theta \rangle}) - y_i \langle x_i, \theta \rangle \right\} + R_\lambda(\theta) \right\}.
\]

Many non-convex penalties are possible:
- capped \( \ell_1 \)-penalty
- SCAD penalty \( \text{(Fan & Li, 2001)} \)
- MCP penalty \( \text{(Zhang, 2006)} \)
Convex and non-convex regularizers

Regularizers

$R_\lambda(t)$

- MCP
- $L_1$
- Capped $L_1$
Algorithm generating sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ to solve

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

Global minimizer of population risk

$$
\theta^* := \arg\min_{\theta} \mathbb{E}_{X,Y} \left[ \mathcal{L}(\theta; X, Y) \right]_{\mathcal{L}(\theta)}
$$
Statistical error versus optimization error

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

\[
\hat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(\theta; x_i, y_i) + R_\lambda(\theta) \right\}.
\]

Global minimizer of population risk

\[
\theta^* := \arg\min_{\theta} \mathbb{E}_{X,Y} \left[ L(\theta; X, Y) \right] \quad \bar{L}(\theta)
\]

Goal of statistician

Provide bounds on Statistical error: \( \|\theta^t - \theta^*\| \) or \( \|\hat{\theta} - \theta^*\| \)
Statistical error versus optimization error

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

\[
\hat{\theta} \in \arg \min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; x_i, y_i) + R_\lambda(\theta) \right\}.
\]

Global minimizer of population risk

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

**Goal of statistician**

Provide bounds on Statistical error: \( \|\theta^t - \theta^*\| \) or \( \|\hat{\theta} - \theta^*\| \)

**Goal of optimization-theorist**

Provide bounds on Optimization error: \( \|\theta^t - \hat{\theta}\| \)
Logistic regression with non-convex regularizer

log error plot for logistic regression with SCAD, a = 3.7

\[
\log(\|\beta - \beta^*\|_2)
\]

iteration count

log error plot for logistic regression with SCAD, a = 3.7

- opt err
- stat err
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 $\ell_\infty$-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 $\ell_\infty$-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:

\[ y = X \theta^* + \varepsilon \]

\[ n \times p \]

\[ n \]

\[ S \]

\[ S^c \]
**Vignette B: Error-in-variables regression**

Begin with high-dimensional sparse regression:

\[ y = X \theta^* + \varepsilon \]

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

\[ Z = F(X, W) \]

Here \( W \in \mathbb{R}^{n \times p} \) is a stochastic perturbation.
Example: Missing data

Missing data:

\[
\begin{align*}
Z & \quad = \quad X \\
\begin{array}{c}
n \times p \\
\end{array} & \quad = \quad \begin{array}{c}
n \times p
\end{array} \\
W & \quad \subset R^{n \times p}
\end{align*}
\]

Here \( W \in \mathbb{R}^{n \times p} \) is multiplicative perturbation (e.g., \( W_{ij} \sim \text{Ber}(\alpha) \)).
Example: Additive perturbations

Additive noise in covariates:

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

$$\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \theta^T \left( \frac{X^T X}{n} \right) \theta - \langle \theta, \frac{X^T y}{n} \rangle + R_\lambda(\theta) \right\}.$$
A second look at regularized least-squares

Equivalent formulation:

\[ \hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \theta^T \left( \frac{X^T X}{n} \right) \theta - \langle \theta, \frac{X^T y}{n} \rangle + \mathcal{R}_\lambda(\theta) \right\}. \]

Population view: unbiased estimators

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

Equivalent formulation:

\[ \hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \theta^T \left( \frac{X^T X}{n} \right) \theta - \langle \theta, \frac{X^T y}{n} \rangle + \mathcal{R}_\lambda(\theta) \right\}. \]

Population view: unbiased estimators

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

A general family of estimators

\[ \hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \theta^T \hat{\Gamma} \theta - \theta^T \hat{\gamma} + \lambda_n^2 \| \theta \|_1^2 \right\}, \]

where \((\hat{\Gamma}, \hat{\gamma})\) are unbiased estimators of \(\text{cov}(x_1)\) and \(\text{cov}(x_1, y_1)\).
Example: Estimator for missing data

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

\[ Z_{ij} = \begin{cases} X_{ij} & \text{with probability } 1 - \alpha \\ * & \text{with probability } \alpha. \end{cases} \]
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 \( \hat{Z} := \frac{Z}{(1 - \alpha)} \):

\[
\hat{\Gamma} = \frac{\hat{Z}^T \hat{Z}}{n} - \alpha \text{diag} \left( \frac{\hat{Z}^T \hat{Z}}{n} \right), \quad \text{and} \quad \hat{\gamma} = \frac{\hat{Z}^T y}{n},
\]
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 $\hat{Z} := \frac{Z}{(1 - \alpha)}$:

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

- solve (doubly non-convex) optimization problem: \cite{LohW2012}

$$\hat{\theta} \in \arg\min_{\theta \in \Omega} \left\{ \frac{1}{2} \theta^T \hat{\Gamma} \theta - \langle \hat{\gamma}, \theta \rangle + R_\lambda(\theta) \right\}.$$
Non-convex quadratic and non-convex regularizer

log error plot for corrected linear regression with MCP, $b = 1.5$

$\log(||\hat{\beta} - \beta^*||_2)$

iteration count
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 $\ell_1$-norm:

**Definition**

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

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

Here defined with respect to the $\ell_1$-norm:

**Definition**

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

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

holds with $\tau_1 = \tau_2 = 0$ for any function that is locally strongly convex around $\theta^*$
Restricted strong convexity

Here defined with respect to the $\ell_1$-norm:

**Definition**

The loss function $L_n$ satisfies RSC with parameters $(\alpha_j, \tau_j)$, $j = 1, 2$ if

$$\langle \nabla L_n(\theta^* + \Delta) - \nabla L_n(\theta^*), \Delta \rangle \geq \begin{cases} 
\alpha_1 \|\Delta\|^2 - \tau_1 \frac{\log p}{n} \|\Delta\|^1 & \text{if } \|\Delta\| \leq 1 \\
\alpha_2 \|\Delta\|^2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 & \text{if } \|\Delta\| > 1.
\end{cases}$$

- holds with $\tau_1 = \tau_2 = 0$ for any function that is locally strongly convex around $\theta^*$
- 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 $\lambda L$.
- For some $\mu > 0$, the function $\tilde{\mathcal{R}}_\lambda(t) = \mathcal{R}_\lambda(t) + \mu t^2$ 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 \( \lambda L \).
- For some \( \mu > 0 \), the function \( \tilde{\mathcal{R}}_\lambda(t) = \mathcal{R}_\lambda(t) + \mu t^2 \) is convex.

Includes (among others):

- rescaled \( \ell_1 \) loss: \( \mathcal{R}_\lambda(t) = \lambda |t| \).
- MCP penalty and SCAD penalties (Fan et al., 2001; Zhang, 2006)
- does not include capped \( \ell_1 \)-penalty
Main statistical guarantee

- regularized $M$-estimator
  \[
  \hat{\theta} \in \arg \min_{\|\theta\|_1 \leq M} \left\{ \mathcal{L}_n(\theta) + \mathcal{R}_\lambda(\theta) \right\}.
  \]

- loss function satisfies $(\alpha, \tau)$ RSC, and regularizer is regular (with parameters $(\mu, L)$)
Main statistical guarantee

- **regularized** $M$-estimator

\[ \hat{\theta} \in \arg \min_{\|\theta\|_1 \leq M} \left\{ \mathcal{L}_n(\theta) + \mathcal{R}_\lambda(\theta) \right\}. \]

- loss function satisfies $(\alpha, \tau)$ RSC, and regularizer is regular (with parameters $(\mu, L)$)

- **local optimum** $\hat{\theta}$ defined by conditions

\[ \langle \nabla \mathcal{L}_n(\hat{\theta}) + \nabla \mathcal{R}_\lambda(\hat{\theta}), \theta - \hat{\theta} \rangle \geq 0 \quad \text{for all feasible } \theta. \]
Main statistical guarantee

- regularized $M$-estimator

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

- loss function satisfies $(\alpha, \tau)$ RSC, and regularizer is regular (with parameters $(\mu, L)$)

- local optimum $\hat{\theta}$ defined by conditions

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

Theorem (Loh & W., 2013)

Suppose $M$ is chosen such that $\theta^*$ is feasible, and $\lambda$ satisfies the bounds

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

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

$$
\|\hat{\theta} - \theta^*\|_2 \leq \frac{6 \lambda_n \sqrt{s}}{4 (\alpha - \mu)} \quad \text{where } s = \|\beta^*\|_0.
$$
Geometry of local/global optima

Consequence:
All \{ \textit{local}, \textit{global} \} optima are within distance $\epsilon_{\text{stat}}$ of the target $\theta^*$. 
Consequence:

All \{ local, global \} optima are within distance \( \varepsilon_{\text{stat}} \) of the target \( \theta^* \).

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

\[
\varepsilon_{\text{stat}} \approx \sqrt{\frac{s \log p}{n}},
\]

which is minimax optimal.
Empirical results (rescaled)

![Graph showing empirical results with rescaled sample sizes and mean squared error for different values of p: 128, 256, 512, 1024.]
Comparisons between different penalties

comparing penalties for corrected linear regression

$I_2$ norm error vs $n/(k \log p)$ for different values of $p$: $p=64$, $p=128$, $p=256$. The graph illustrates how the $I_2$ norm error decreases as $n/(k \log p)$ increases, with different lines representing different values of $p$.
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:  

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

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:  

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

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

Simple updates:

$$
\theta^{t+1} = \arg \min_{\theta \in \Omega} \left\{ \| \theta - \alpha^t \nabla f(\theta^t) \|_2^2 + g(\theta) \right\}.
$$
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: \((\text{Nesterov, 2007})\)

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

where \(f\) is differentiable, and \(g\) is convex, sub-differentiable.

Simple updates:

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

Not directly applicable with \(f = \mathcal{L}_n\) and \(g = \mathcal{R}_\lambda\) (since \(\mathcal{R}_\lambda\) can be non-convex).
Define modified loss functions and regularizers:

\[ \tilde{\mathcal{L}}_n(\theta) := \mathcal{L}_n(\theta) - \mu\|\theta\|^2_2, \quad \text{and} \quad \tilde{\mathcal{R}}_\lambda(\theta) := \mathcal{R}_\lambda(\theta) + \mu\|\theta\|^2_2. \]
Composite gradient on a convenient splitting

- Define modified loss functions and regularizers:

\[ \tilde{L}_n(\theta) := L_n(\theta) - \mu \|\theta\|_2^2, \quad \text{and} \quad \tilde{R}_\lambda(\theta) := R_\lambda(\theta) + \mu \|\theta\|_2^2. \]

- Apply composite gradient descent to the objective

\[ \min_{\theta \in \Omega} \{ \tilde{L}_n(\theta) + \tilde{R}_\lambda(\theta) \}. \]
Composite gradient on a convenient splitting

- Define modified loss functions and regularizers:
  \[ \tilde{L}_n(\theta) := \mathcal{L}_n(\theta) - \mu \|\theta\|_2^2, \quad \text{and} \quad \tilde{R}_\lambda(\theta) := \mathcal{R}_\lambda(\theta) + \mu \|\theta\|_2^2. \]
  
  \( \tilde{L}_n(\theta) \) are non-convex, and \( \tilde{R}_\lambda(\theta) \) are convex.

- Apply composite gradient descent to the objective
  \[ \min_{\theta \in \Omega} \left\{ \tilde{L}_n(\theta) + \tilde{R}_\lambda(\theta) \right\}. \]

- Converges to local optimum \( \hat{\theta} \) (Nesterov, 2007)
  \[ \langle \nabla \tilde{L}_n(\theta) + \nabla \tilde{R}_\lambda(\hat{\theta}), \theta - \hat{\theta} \rangle \geq 0 \quad \text{for all feasible } \theta. \]
Define modified loss functions and regularizers:
\[
\tilde{L}_n(\theta) := L_n(\theta) - \mu \|\theta\|_2^2, \quad \text{and} \quad \tilde{R}_\lambda(\theta) := R_\lambda(\theta) + \mu \|\theta\|_2^2.
\]

Apply composite gradient descent to the objective
\[
\min_{\theta \in \Omega} \left\{ \tilde{L}_n(\theta) + \tilde{R}_\lambda(\theta) \right\}.
\]

converges to local optimum \(\hat{\theta}\) (Nesterov, 2007)
\[
\langle \nabla \tilde{L}_n(\theta) + \nabla \tilde{R}_\lambda(\hat{\theta}), \theta - \hat{\theta} \rangle \geq 0 \quad \text{for all feasible } \theta.
\]

will show that convergence is geometrically fast with constant stepsize
Theoretical guarantees on computational error

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

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

- population minimizer $\beta^*$ is $s$-sparse
Theoretical guarantees on computational error

- implement Nesterov’s composite method with constant stepsize to \((\tilde{L}_n, \tilde{R}_\lambda)\) split.

- fixed global optimum \(\hat{\beta}\) defines the statistical error \(\epsilon_{\text{stat}}^2 = \|\hat{\beta} - \beta^*\|_2\).

- population minimizer \(\beta^*\) is \(s\)-sparse

- loss function satisfies \((\alpha, \tau)\) RSC and smoothness conditions, and regularizer is \((\mu, L)\)-good
Theoretical guarantees on computational error

- implement Nesterov’s composite method with constant stepsize to \((\widetilde{L}_n, \widetilde{R}_\lambda)\) split.
- fixed global optimum \(\hat{\beta}\) defines the statistical error \(\epsilon_{\text{stat}}^2 = \|\hat{\beta} - \beta^*\|_2^2\).
- population minimizer \(\beta^*\) is \(s\)-sparse
- loss function satisfies \((\alpha, \tau)\) RSC and smoothness conditions, and regularizer is \((\mu, L)\)-good

**Theorem (Loh & W., 2013)**

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

\[
\|\theta^t - \hat{\theta}\|_2^2 \leq \frac{2}{\alpha - \mu} \left(\delta^2 + 128\tau \frac{s \log p}{n} \epsilon_{\text{stat}}^2\right) \quad \text{for all } t \geq T(\delta) \text{ iterations,}
\]

*where \(T(\delta) \approx \frac{\log(1/\delta)}{\log(1/\kappa)}\).*
Optimization error $\hat{\Delta}^t := \theta^t - \hat{\theta}$ decreases geometrically up to statistical tolerance:

$$\|\theta^{t+1} - \hat{\theta}\|^2 \leq \kappa^t \|\theta^0 - \hat{\theta}\|^2 + o(\|\theta^* - \hat{\theta}\|^2)$$

for all $t = 0, 1, 2, \ldots$. 

Stat. error $\epsilon_{\text{stat}}^2$
Non-convex linear regression with SCAD

log error plot for corrected linear regression with SCAD, $a = 2.5$

$\log(||\beta^t - \beta^*||_2)$

iteration count
Non-convex linear regression with SCAD

log error plot for corrected linear regression with SCAD, $a = 3.7$

$$\log(||\beta^t - \beta^*||_2)$$

iteration count

$\log$ error plot for corrected linear regression with SCAD, $a = 3.7$

- opt err
- stat err
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: