A Unified Framework for High-Dimensional Analysis for M-Estimators with Decomposable Regularizers

20 April, 2023

36-709 Advanced Statistical Theory I: Final Presentation

Fan Pu Zeng

Paper

- Paper: A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers (NeurIPS 2009)
- Authors: Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright and Bin Yu

Motivation

- In high-dimensional statistical inference, it is common for the number of parameters p to be comparable or greater than the sample size n
- For an estimator in this regime to be consistent. we have to assume that the model has some sort of low-dimensional structure:
    1. Sparse vectors
    2. Sparse/structured matrices (i.e band matrices)
    3. Low-rank matrices
    4. Etc..
- Lots of recent work done for these special cases, but is there a way to understand these estimators in a general sense?

Motivation

Yes!

Using two key properties of loss and regularization functions to ensure fast convergence:

- Decomposability
- Restricted Strong Convexity

Motivation

Using their unified framework, the authors were able to re-derive existing results, and also obtain new bounds on consistency and convergence rates.

Let's now try to understand this exciting result!

Problem Formulation

Goal: *Define*

- $Z_1^n := \{Z_1, \cdots, Z_n\}$ n *i.i.d observations drawn from distribution* $\mathbb{P}$ *with some paramter* θ^*
- $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$ *a convex and differentiable loss function, such that* $\mathcal{L}(\theta; Z_1^n)$ *returns the loss of* θ *on observations* Z_1^n
- $\lambda_n > 0$ *a user-defined regularization penalty*
- $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}_+$ *a norm-based regularizer*

Problem Formulation

Goal: *We solve for the convex optimization problem*

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \right\}, \tag{1}$$

and we are interested in deriving bounds on

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|. \tag{2}$$

for some error norm $\|\cdot\|$.

Decomposability of $\mathcal{R}$

The first property of our analysis is the decomposability of our norm-based regularizer $\mathcal{R}$.

Let $\mathcal{M} \subseteq \overline{\mathcal{M}} \subseteq \mathbb{R}^p$. $\mathcal{M}$ is the model subspace, which capture the constraints of the model (i.e sparse support or low-rank). $\overline{\mathcal{M}}^{\perp}$ is the orthogonal complement of the closure of $\mathcal{M}$, $\overline{\mathcal{M}}$.

Definition 1 (Decomposability): *Given a pair of subspaces* $M \subseteq \overline{\mathcal{M}}$, *a norm-based regularizer is decomposable with respect to* $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ *if*

$$\mathcal{R}(\theta + \gamma) = \mathcal{R}(\theta) + \mathcal{R}(\gamma). \tag{3}$$

Decomposability of $\mathcal{R}$

By the triangle inequality for norms, we always have

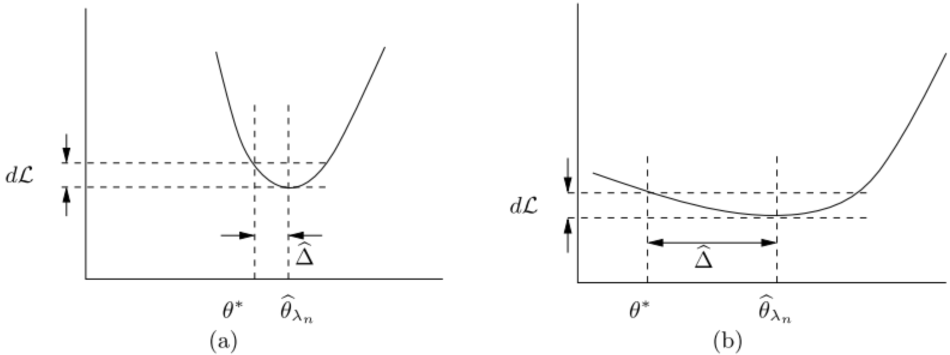
$$\mathcal{R}(\theta + \gamma) \leq \mathcal{R}(\theta) + \mathcal{R}(\gamma). \tag{4}$$

So decomposability holds when the inequality is tight.

Restricted Strong Convexity of $\mathcal{L}$

Let $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ be the difference our optimal solution and true parameter.

When is it the case that a small loss difference $|\mathcal{L}(\theta^*) - \mathcal{L}(\widehat{\theta}_{\lambda_n})|$ implies that $\widehat{\Delta}$ is small?



Restricted Strong Convexity of $\mathcal{L}$

We use the notion of strong convexity to say that a function is "not too flat". Since $\mathcal{L}$ is differentiable, define

$$\delta\mathcal{L}(\Delta, \theta^*) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla\mathcal{L}(\theta^*), \Delta \rangle. \tag{5}$$

to be the error in the first-order Taylor series expansion of $\mathcal{L}$ at θ^*.

Sufficient Condition for Strong Convexity: *$\mathcal{L}$ is strongly convex with parameter $\kappa > 0$ if $\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa \|\Delta\|^2$ for all $\Delta \in \mathbb{R}^p$ in a neighborhood of θ^*.*

Restricted Strong Convexity of $\mathcal{L}$

We now present the second condition for the analysis, the restricted strong convexity condition:

Definition 2 (Restricted Strong Convexity): *The loss function $\mathcal{L}$ satisfies a restricted strong convexity (RSC) condition with curvature $\kappa_{\mathcal{L}}$ and tolerance function $\tau_{\mathcal{L}}$ if*

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}^2(\theta^*) \tag{6}$$

for all $\Delta \in \mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*)$, where

$$\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*) := \left\{ \delta \in \mathbb{R}^p \mid \mathcal{R}(\Delta_{\overline{\mathcal{M}}^{\perp}}) \leq 3\mathcal{R}(\Delta_{\overline{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^{\perp}}) \right\} \tag{7}$$

is a star-shaped set (and possibly also a cone).

Back to Decomposability

- Why is decomposability important?
- It helps to constrain the error vector $\widehat{\Delta}$ to lie in $\mathbb{C}(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^*)$, which allows strong convexity to hold.

Bounds for General Models

We can now state the main result of the paper. Also, recall our convex optimization program:

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \right\}, \tag{8}$$

Theorem 3 (Bounds for General Models): *Suppose $\mathcal{R}$ decomposable, $\mathcal{L}$ satisfies RSC with curvature $\kappa_{\mathcal{L}}$ and tolerance $\tau_{\mathcal{L}}$, and $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$. Then any optimal solution $\widehat{\theta}_{\lambda_n}$ to Program 8 satisfies the bound*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|^2 \leq 9 \frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2} \Psi^2(\overline{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}} \left(2\tau_{\mathcal{L}}^2(\theta^*) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) \right), \tag{9}$$

where $\mathcal{R}^*$ is the dual norm of $\mathcal{R}$, and Ψ is the subspace compatibility constant which reflects the degree of compatibility between the regularizer and error norm over the subspace $\overline{\mathcal{M}}$, $\theta^*_{\mathcal{M}^\perp}$ is the projection of θ^* onto the subspace $\mathcal{M}^\perp$.

Bounds for General Models

Theorem 4 (Bounds for General Models): … *Then any optimal solution $\widehat{\theta}_{\lambda_n}$ to Program 8 satisfies the bound*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|^2 \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(\overline{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\left(2\tau_{\mathcal{L}}^2(\theta^*) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\right). \tag{10}$$

Observations

- Hard to choose $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*))$ in practice, since θ^* unknown. In practice, use concentration inequalities to get bounds that hold whp.
- This actually provides a family of bounds for each pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ of subspaces. Trade-off in contribution of error between the two terms.

Application: Sparse Linear Regression

- Recall the sparse linear regression problem, where we assume that θ^* has at most s non-zero coefficients
- Then a natural M-estimator is the Lasso:

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1. \tag{11}$$

- For any subset $S \subseteq \{1, 2, \ldots, p\}$, the ℓ_1-norm is decomposable with respect to the subspace $\mathcal{M}(S) = \{\theta \in \mathbb{R}^p \mid \theta_{S^c} = 0\}$ and its orthogonal complement.
- It turns out we can also model strong convexity via the restricted eigenvalue condition.

Application: Sparse Linear Regression

In addition to the conditions of the main theorem, if we assume $\theta^* \in \mathcal{M}$, RSC holds, and $\tau_{\mathcal{L}}(\theta^*) = 0$, then we obtain the following rates:

Corollary 5 (Sparse Linear Regression): *Consider an s-sparse instance of the linear regression model such that X satisfies the restricted eigenvalue (RE) condition and column normalization condition. Given the Lasso program with regularization parameter $\lambda_n = 4\sigma\sqrt{\frac{\log p}{n}}$, with probability at least $1 - c_1 \exp(-c_2 n \lambda_n^2)$, any optimal solution $\widehat{\theta}_{\lambda_n}$ satisfies both*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{64\sigma^2}{\kappa_{\mathcal{L}}^2} \frac{s \log p}{n}, \tag{12}$$

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_1 \leq \frac{24\sigma}{\kappa_{\mathcal{L}}} s \sqrt{\frac{\log p}{n}}. \tag{13}$$

Other Applications

The paper also shows applications to derive rates for the following settings:

- Lasso estimates with exact sparsity
- Lasso estimates with weakly sparse models
- Generalized linear models
- Group-sparse settings

Thank you for your kind attention! Any questions?