A Unified Framework for High-Dim. Analysis of *M*-Estimators with Decomposable Regularizers

Daniel Cirkovic[1] Shuren He[1] Jiyoung Park[1]

[1]Department of Statistics
Texas A&M University

October 10, 2025

Introduction

Introduction I

Modern statistical problems require analysis of estimators in the regime where $p >> n$.

- Estimators are not consistent unless the model is constrained.

Before the work of [NRWY12], many different M-estimation procedures had been analyzed independently.

- Ex: sparse regression and covariance/low-rank matrix estimation.

[NRWY12] reveal the unifying principles that support the analysis of such estimators.

Section 2

Setting

Let $Z_1^n = \{Z_1, \ldots, Z_n\}$ denote n identically distributed observations with marginal distribution $\mathbb{P}$.

Further, let $\mathcal{L} : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$ be a loss function that is convex and differentiable in θ. The risk is then given by $\bar{\mathcal{L}}(\theta) = \mathbb{E}_{Z_1^n}[\mathcal{L}(\theta; Z_1^n)]$.

Define the parameter of interest and corresponding estimator by

$$\theta^\star \in \text{arg min}_{\theta \in \mathbb{R}^p} \bar{\mathcal{L}}(\theta),$$
$$\hat{\theta}_{\lambda_n} \in \text{arg min}_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \right\},$$

where $\lambda_n > 0$ and $\mathcal{R}$ is a norm.

Intuition

Recall that

$$\theta^\star \in \arg\min_{\theta\in\mathbb{R}^p} \bar{\mathcal{L}}(\theta) = \arg\min_{\theta\in\mathbb{R}^p} \mathbb{E}_{Z_1^n}[\mathcal{L}(\theta; Z_1^n)],$$
$$\hat{\theta}_{\lambda_n} \in \arg\min_{\theta\in\mathbb{R}^p} \left\{\mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta)\right\}.$$

If we have any hope of $\hat{\theta}_{\lambda_n}$ being close to $\theta^\star$, we need that $\lambda_n \to 0$.

- But if $\lambda_n \to 0$ too fast, we aren't accomplishing anything.

In addition, $\theta^\star$ should not be too penalized by $\mathcal{R}$.

- Equivalently, deviations from the model constraints should be penalized as much as possible.

The closeness of $\hat{\theta}_{\lambda_n}$ and $\theta^\star$ is measured by closeness of

$$\mathcal{L}(\hat{\theta}_{\lambda_n}) + \lambda_n \mathcal{R}(\hat{\theta}_{\lambda_n}) \text{ and } \mathcal{L}(\theta^\star) + \lambda_n \mathcal{R}(\theta^\star).$$

Decomposability

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

$$\mathcal{R}(\theta + \gamma) = \mathcal{R}(\theta) + \mathcal{R}(\gamma), \qquad \text{for all } \theta \in \mathcal{M} \text{ and } \gamma \in \overline{\mathcal{M}}^{\perp}.$$

- The model subspace $\mathcal{M}$ captures the constraints of the model.
- The perturbation space $\overline{\mathcal{M}}^{\perp}$ captures deviations away from the model.
  - Want to penalize $\gamma \in \overline{\mathcal{M}}^{\perp}$ as much as possible.

Example: Sparse vectors

For any set $S \subset \{1, 2, \ldots, p\}$ with $|S| = s$, define

$$\mathcal{M}(S) = \{\theta \in \mathbb{R}^p : \theta_j = 0 \text{ for all } j \notin S\},$$
$$\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S) = \{\theta \in \mathbb{R}^p : \theta_j = 0 \text{ for all } j \in S\}.$$

Here, $\mathcal{R}(\theta) = \|\theta\|_1$ is clearly decomposable with respect to the pair $(\mathcal{M}(S), \mathcal{M}^{\perp}(S))$.

Example: Low-rank matrices

In many applications (image compression, matrix completion, etc), one assumes a signal plus noise model

$$Y = \Theta^\star + E,$$

where $Y, \Theta^\star, E \in \mathbb{R}^{p_1 \times p_2}$ and $\operatorname{rank}(\Theta^\star) = r < p_1 \wedge p_2$.

One common estimation procedure in such a model is least squares with a nuclear norm penalization

$$\hat{\Theta}_{\lambda_n} = \arg\min_{\Theta \in \mathbb{R}^{p_1 \times p_2}} \left\{ \|Y - \Theta\|_F + \lambda_n \|\Theta\|_{\text{nuc}} \right\},$$

where

$$\|\Theta\|_{\text{nuc}} = \sum_{i=1}^{p_1 \wedge p_2} \sigma_i(\Theta).$$

$\mathcal{M} \neq \overline{\mathcal{M}}$

Let $\mathcal{U} = \text{col}(\Theta^\star)$ and $\mathcal{V} = \text{row}(\Theta^\star)$. We can define

$$\mathcal{M}(\mathcal{U},\mathcal{V}) = \left\{\Theta \in \mathbb{R}^{p_1 \times p_2} : \text{row}(\Theta) \subset \mathcal{V}, \text{col}(\Theta) \subset \mathcal{U}\right\},$$
$$\overline{\mathcal{M}}^\perp(\mathcal{U},\mathcal{V}) = \left\{\Theta \in \mathbb{R}^{p_1 \times p_2} : \text{row}(\Theta) \subset \mathcal{V}^\perp, \text{col}(\Theta) \subset \mathcal{U}^\perp\right\}.$$

Suppose $\Theta^\star = USV^T$. Note that any $A \in \mathcal{M}(\mathcal{U},\mathcal{V}), B \in \overline{\mathcal{M}}^\perp(\mathcal{U},\mathcal{V})$ can be represented as

$$A = U \begin{bmatrix} \Gamma_{11} & 0 \\ 0 & 0 \end{bmatrix} V^T, \qquad B = U \begin{bmatrix} 0 & 0 \\ 0 & \Gamma_{22} \end{bmatrix} V^T,$$

for appropriate matrices $\Gamma_{11}, \Gamma_{22} \in \mathbb{R}^{r \times r}$. Clearly $\langle A, B \rangle = \text{tr}(A'B) = 0$ so

$$\|A + B\|_{\text{nuc}} = \|A\|_{\text{nuc}} + \|B\|_{\text{nuc}}.$$

A key consequence of decomposability

For a given inner product $\langle \cdot, \cdot \rangle$ the dual norm of $\mathcal{R}$ is given by

$$\mathcal{R}^\star(v) = \sup_{u \neq 0} \frac{\langle u, v \rangle}{\mathcal{R}(u)}.$$

Lemma (Lemma 1 of [NRWY12])

Suppose the regularization parameter λ_n satisfies

$$\lambda_n \geq 2\mathcal{R}^\star \left(\nabla \mathcal{L}(\theta^\star; Z_1^n) \right).$$

Then for any pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ over which $\mathcal{R}$ is decomposable, the error $\hat{\Delta} = \hat{\theta}_{\lambda_n} - \theta^\star$ belongs to the set

$$\mathcal{C}\left(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^\star \right) \equiv \left\{ \Delta \in \mathbb{R}^p : \mathcal{R}(\Delta_{\overline{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\overline{\mathcal{M}}}) + 4\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp} \right) \right\}.$$

Restricted strong convexity (RSC)

Recall that we would like to relate closeness of $\mathcal{L}(\theta^\star + \hat{\Delta}) - \mathcal{L}(\theta^\star)$ to the smallness of $\hat{\Delta}$.

- In classical situations, this is resolved through strong convexity

$$\delta\mathcal{L}(\Delta, \theta^\star) = \mathcal{L}(\theta^\star + \Delta) - \mathcal{L}(\theta^\star) - \langle \nabla\mathcal{L}(\theta^\star), \Delta \rangle \geq \kappa \|\Delta\|^2,$$

  for some $\kappa > 0$ and all Δ in a neighborhood of $\theta^\star$.
- This is unrealistic in the high-dimensional setting.
- Luckily, Lemma 1 states that we only require convexity over $\mathcal{C}\left(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^\star\right)$.

Restricted strong convexity (RSC)

Definition

The loss function satisfies a **restricted strong convexity** *condition with curvature* $\kappa_{\mathcal{L}} > 0$ *and tolerance function* $\tau_{\mathcal{L}}$ *if*

$$\delta\mathcal{L}(\Delta, \theta^\star) \geq \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}^2(\theta^\star), \qquad \textit{for all } \Delta \in \mathcal{C}\left(\mathcal{M}, \overline{\mathcal{M}}^\perp; \theta^\star\right).$$

For many loss functions, it is possible to prove that with high probability

$$\delta\mathcal{L}(\Delta, \theta^\star) \geq \kappa_1 \|\Delta\|^2 - \kappa_2 g(n, p) \mathcal{R}^2(\Delta), \qquad \text{for all } \|\Delta\| \leq 1,$$

which implies a form of RSC as long as $\mathcal{R}(\Delta)$ is sufficiently small compared to $\|\Delta\|$.

Subspace compatibility constant

Definition

For any subspace $\mathcal{M}$ *of* $\mathbb{R}^p$*, the* **subspace compatibility constant** *with respect to the pair* $(\mathcal{R}, \|\cdot\|)$ *is given by*

$$\psi(\mathcal{M}) \equiv \sup_{u \in \mathcal{M} \backslash \{0\}} \frac{R(u)}{\|u\|}.$$

Hence if $\theta^{\star} \in \mathcal{M}$ and $\Delta \in \mathcal{C}\left(\mathcal{M}, \overline{\mathcal{M}}^{\perp}; \theta^{\star}\right)$

$$\mathcal{R}(\Delta_{\overline{\mathcal{M}}^{\perp}}) \leq 3\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right),$$

and thus by triangle inequality $\mathcal{R}(\Delta) \leq 4\mathcal{R}(\Delta_{\overline{\mathcal{M}}}) \leq 4\Psi(\overline{M})\|\Delta\|$.

Hence, the previous RSC condition becomes

$$\delta\mathcal{L}(\Delta, \theta^{\star}) \geq \left(\kappa_1 - 16\kappa_2\Psi^2(\overline{\mathcal{M}})g(n,p)\right)\|\Delta\|^2, \qquad \text{for all } \|\Delta\| \leq 1.$$

Section 3

Deviation Bound for $\|\hat{\theta}_{\lambda_n} - \theta^*\|^2$

To prove Lemma 2.1, we need to construct a function: $\mathcal{F}(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n \{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\} \leq 0$. Since $\mathcal{F}(0) = 0$, the optimal error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ must satisfy $\mathcal{F}(\widehat{\Delta}) \leq 0$. In order to control $\mathcal{F}$, we need to bound both difference of loss functions, and a difference of regularizers. They can be bounded by the following lemma:

Lemma (Lemma 3 of [NRWY12])

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) - 2\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)$$

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda_n}{2}\left[\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) + \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right)\right]$$

Let us first prove the first statement of lemma 3.1

$$\begin{aligned}
\mathcal{R}\left(\theta^* + \Delta\right) &= \mathcal{R}\left(\theta^*_{\mathcal{M}} + \theta^*_{\mathcal{M}^\perp} + \Delta_{\overline{\mathcal{M}}} + \Delta_{\overline{\mathcal{M}}^\perp}\right) \\
&\geq \mathcal{R}\left(\theta^*_{\mathcal{M}}\right) + \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right) - \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) \\
\mathcal{R}\left(\theta^* + \Delta\right) - \mathcal{R}\left(\theta^*\right) &\geq \mathcal{R}\left(\theta_{\mathcal{M}^*}\right) + \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right) \\
&- \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) - \mathcal{R}\left(\theta^*\right) \\
&\geq \mathcal{R}\left(\theta_{\mathcal{M}^*}\right) + \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right) \\
&- \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) - \left\{\mathcal{R}\left(\theta^*_{\mathcal{M}}\right) + \mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\} \\
&= \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) - 2\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)
\end{aligned}$$

Using the convexity of loss function, we have:

$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq -|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle|$$

Applying the duality and $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$, we obtain:

$$-|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \geq -\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))\mathcal{R}(\Delta) \geq -\frac{\lambda_n}{2}\left[\mathcal{R}(\Delta_{\overline{\mathcal{M}}}) + \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right)\right]$$

We can now complete the proof of Lemma 2.1. Combining the two lower bounds of (3.1) we obtain

$$\begin{aligned} 0 \geq \mathcal{F}(\widehat{\Delta}) &\geq \lambda_n \left\{\mathcal{R}\left(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\widehat{\Delta}_{\overline{\mathcal{M}}}\right) - 2\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\} \\ &\quad - \frac{\lambda_n}{2}\left[\mathcal{R}\left(\widehat{\Delta}_{\overline{\mathcal{M}}}\right) + \mathcal{R}\left(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}\right)\right] \\ &= \frac{\lambda_n}{2}\left\{\mathcal{R}\left(\widehat{\Delta}_{\overline{\mathcal{M}}^\perp}\right) - 3\mathcal{R}\left(\widehat{\Delta}_{\overline{\mathcal{M}}}\right) - 4\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\} \end{aligned}$$

Theorem (Theorem 1 of [NRWY12])

G1 *Assume the regularizer* $\mathcal{R}(\cdot)$ *is a norm, and let* $\left(\mathcal{M}, \mathcal{M}^{\perp}\right)$ *be any subspace pair over which* $\mathcal{R}(\cdot)$ *is decomposable.*

G2 *Assume the loss* $\mathcal{L}_n(\cdot)$ *is convex and differentiable, and* $\mathcal{L}_n(\theta)$ *satisfies the RSC condition w.r.t.* $\left(\mathcal{M}, \mathcal{M}^{\perp}\right)$ *at* $\theta = \theta^*$.

G3 *Assume* $\lambda_n \geq 2\mathcal{R}^* \left(\nabla \mathcal{L}\left(\theta^*\right)\right)$ *holds.*

Then any optimal solution $\widehat{\theta}_{\lambda_n}$ *to the convex program satisfies the bound:*

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

Sketch of proving Theorem 3.2:

- $\mathcal{F}(\Delta) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n \{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\}$.
- Constructing set $\mathbb{K}(\delta) := \mathbb{C} \cap \{\|\Delta\| = \delta\}$ such that $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{K}(\delta)$.
- Utilizing the property of "star shape" set to prove that if $\Delta \in \mathbb{C}$, $\{t\Delta \mid t \in (0,1)\} \subset \mathbb{C}$.
- $\mathcal{F}(\widehat{\Delta}) < 0$ since $\mathcal{F}(0) = 0$. If $\|\widehat{\Delta}\| > \delta$, there exist such $t^* \in (0,1)$ such that $\|t^*\widehat{\Delta}\| = \delta$. Considering that $\mathcal{F}$ is convex and $\mathcal{F}(0) = 0$, $\mathcal{F}\left(t^*\widehat{\Delta}\right) \leq t^*\mathcal{F}(\widehat{\Delta}) < 0$, which brings contradiction. So $\|\widehat{\Delta}\| \leq \delta$.
- Finding a lower bound for δ such that $\mathcal{F}(\Delta) > 0 \quad \forall \Delta \in \mathbb{K}(\delta)$ holds:

$$\delta^2 := 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(\overline{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\left\{2\tau_{\mathcal{L}}^2(\theta^*) + 4\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\}$$

By applying RSC condition and lemma 3.1, we have:

$$\begin{aligned}\mathcal{F}(\Delta) &= \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda_n \{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\} \\ &\geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}^2(\theta^*) + \lambda_n \{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\} \\ &\geq \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle + \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}(\theta^*)^2 \\ &+ \lambda_n \left\{\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) - 2\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\}\end{aligned}$$

Since we have $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$ (Assumption G3), we can obtain a bound for $|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle|$:

$$\begin{aligned}|\langle \nabla \mathcal{L}(\theta^*), \Delta \rangle| \leq \mathcal{R}^*(\nabla \mathcal{L}(\theta^*))\mathcal{R}(\Delta) &\leq \frac{\lambda_n}{2}\mathcal{R}(\Delta) \\ &\leq \frac{\lambda_n}{2}\left(\mathcal{R}(\Delta_{\overline{\mathcal{M}}^\perp}) + \mathcal{R}(\Delta_{\overline{\mathcal{M}}})\right)\end{aligned}$$

Plug this back, we have:

$$\begin{aligned}\mathcal{F}(\Delta) &\geq \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}^2(\theta^*) + \lambda_n \left\{\frac{1}{2}\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}^\perp}\right) - \frac{3}{2}\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) - 2\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\} \\ &\geq \kappa_{\mathcal{L}} \|\Delta\|^2 - \tau_{\mathcal{L}}^2(\theta^*) - \frac{\lambda_n}{2}\left\{3\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) + 4\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\}\end{aligned}$$

Since $\mathcal{R}\left(\Delta_{\overline{\mathcal{M}}}\right) \leq \Psi(\overline{\mathcal{M}})\left\|\Delta_{\overline{\mathcal{M}}}\right\| \leq \Psi(\overline{\mathcal{M}})\|\Delta\|$.

$$\mathcal{F}(\Delta) \geq \kappa_{\mathcal{L}}\|\Delta\|^2 - \tau_{\mathcal{L}}^2\left(\theta^*\right) - \frac{\lambda_n}{2}\left\{3\Psi(\overline{\mathcal{M}})\|\Delta\| + 4\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\} \geq 0$$

$$\rightarrow \|\Delta\|^2 \geq \delta^2 := 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(\overline{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\left\{2\tau_{\mathcal{L}}^2\left(\theta^*\right) + 4\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right)\right\}$$

Explanation for Bound of $\|\hat{\theta}_{\lambda_n} - \theta^*\|^2$ I

- This bound is deterministic and does not require strictly convex.
- This bound is a family of bounds indexed by different choices of $(\mathcal{M}, \mathcal{M}^\perp)$.
- Ignoring the $\tau_{\mathcal{L}}$. The error bound consists of two terms: estimation error $\mathcal{E}_{\text{err}}$ and approximation error $\mathcal{E}_{\text{app}}$.

$$\mathcal{E}_{\text{err}} := 9\frac{{\lambda_n}^2}{\kappa\mathcal{L}^2}\Psi^2(\overline{\mathcal{M}}) \quad \text{and} \quad \mathcal{E}_{\text{app}} := 4\frac{\lambda_n}{\kappa_{\mathcal{L}}}\mathcal{R}\left(\theta^*_{\mathcal{M}^\perp}\right).$$

- $\tau_{\mathcal{L}}$ is the tolerance term reflecting the degree of this nonidentifiability.

- As a special case of Theorem 3.2, consider the case $\theta^* \in \mathcal{M}$ and RSC condition holds over $\mathbb{C}\left(\mathcal{M}, \overline{\mathcal{M}}, \theta^*\right)$.

Corollary (Corollary 1 of [NRWY12])

Then for every $\lambda_n \geq 2\mathcal{R}^* \{\nabla \mathcal{L}_n (\boldsymbol{\theta}^*)\}$,

$$\left\| \widehat{\theta}_{\lambda_n} - \theta^* \right\| \leq 3 \frac{\lambda_n}{\kappa_{\mathcal{L}}} \Psi(\mathcal{M})$$

Since $\mathcal{R}(\Delta) \leq 4\Psi(\overline{\mathcal{M}})\|\Delta\|$, *we also have:*

$$\mathcal{R}\left(\widehat{\boldsymbol{\theta}}_{\lambda_n} - \theta^*\right) \leq 12 \frac{\lambda_n}{\kappa_{\mathcal{L}}} \Psi^2(\mathcal{M})$$

Section 4

Lasso regression I

Let us consider M-estimator for lasso regression:

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

Under this setting:

$$\begin{aligned} \delta\mathcal{L}\left(\Delta, \theta^*\right) :=& \mathcal{L}\left(\theta^* + \Delta\right) - \mathcal{L}\left(\theta^*\right) - \left(\nabla\mathcal{L}\left(\theta^*\right), \Delta\right\rangle \\ =& \left\langle \Delta, \frac{1}{n} X^T X \Delta \right\rangle = \frac{1}{n} \|X\Delta\|_2^2 \end{aligned}$$

The cone set is:

$$\mathbb{C}(S) := \left\{\Delta \in \mathbb{R}^p \mid \left\|\Delta_{S^c}\right\|_1 \leq 3 \left\|\Delta_S\right\|_1\right\}$$

Two types RE condition I

- The restricted strong convexity with respect to ℓ_2-norm:

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa_{\mathcal{L}} \|\Delta\|_2^2 \quad \text{for all } \Delta \in \mathbb{C}(S) \tag{2}$$

- The restricted strong convexity with respect to ℓ_1-norm:

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa'_{\mathcal{L}} \frac{\|\Delta\|_1^2}{|S|} \quad \text{for all } \Delta \in \mathbb{C}(S) \tag{3}$$

Two types RE condition II

- If $\Delta \in \mathbb{R}^p$, equation (2) can be rewritten as:

$$\frac{\|X\Delta\|_2^2}{n\|\Delta\|_2} \geq \kappa_{\mathcal{L}} \quad \text{for all } \Delta \in \mathbb{R}^p/\{\mathbf{0}\}$$

  Which is equivalent as:

$$\lambda_{\min}\left(X^{\mathrm{T}}X\right) \geq \kappa_{\mathcal{L}}$$

  $p \gg n$, and we only require strongly convex in $\Delta \in \mathbb{C}(S)$.

- Equation (2) is more restrictive than equation (3) since:

$$\|\Delta\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{|S|}\|\Delta_S\|_2 \leq 4\sqrt{|S|}\|\Delta\|_2 \quad \text{for all } \Delta \in \mathbb{C}(S)$$

Matrices satisfy RE conditions I

- Σ-Gaussian ensemble: $X \in \mathbb{R}^{n \times p}$ and each row $X_i \sim N(0, \Sigma)$:

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \kappa_1 \|\Delta\|_2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \quad \text{for all } \Delta \in \mathbb{R}^p \tag{4}$$

w.p. $1 - c_1 \exp(-c_2 n)$, then for $\Delta \in \mathbb{C}(S)$:

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \kappa_1 \|\Delta\|_2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \geq \left(\kappa_1 - 4\sqrt{|S| \frac{\log p}{n}} \kappa_2 \right) \|\Delta\|_2$$

To let equation (2) hold w.h.p and $\kappa_{\mathcal{L}} = \frac{\kappa_1}{2}$, $n > 64 \left(\kappa_2/\kappa_1\right)^2 |S| \log p$.

- Similar conclusion can be got if X matrix is sampled from sub-Gaussian designs.

Exact sparsity LASSO I

- We now derive the bound for LASSO with exact sparsity, under some additional assumptions.
  - G'1 Consider a matrix $X \in \mathbb{R}^{n \times p}$ whose column is normalized, *i.e.* $\|X_j\|_2/\sqrt{n} \leq 1$ for all $j = 1, \dots p$. Note that this 1 can be arbitrary constant.
  - G'2 Let $w \in \mathbb{R}^n$ be a 0-mean sub-Gaussian vector, *i.e.*

$$\sup_{v \in \mathbb{S}^{n-1}} \|\langle w, v \rangle\|_{\psi_2} < \infty.$$

    But we make a remark that this condition can be relieved using (modified) marginal sub-Gaussian definition, *i.e.*

$$\max_{i=1,\dots n} \left\| \left\langle w, \frac{X_i}{\sqrt{n}} \right\rangle \right\|_{\psi_2} < \infty.$$

    We denote the (maximum) 'variance' ($\neq$ sub-Gaussian norm) parameter of w as σ^2.

Exact sparsity LASSO II

- Consider a linear regression problem

$$Y = X\theta^* + w \tag{5}$$

where

1. $Card(\theta^*) = s$ for some fixed $s \leq p$: Exact sparsity condition.
2. w satisfies the sub-Gaussian condition (G‘2)
3. X satisfies ℓ^2-RE condition ((2)) and column normalization condition (G‘1).

Exact sparsity LASSO III

Corollary (Corollary 2 of [NRWY12])

Under the above setting, the solution of (1) *with* $\lambda_n = 4\sigma\sqrt{\log p/n}$ *satisfies the following bound:*

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

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_1 \leq \frac{24\sigma^2}{\kappa_{\mathcal{L}}} s\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - c_1 exp(-c_2 n\lambda_n^2)$ *for some constants* $c_1, c_2 > 0$.

Exact sparsity LASSO IV

- Some remarks:
  - λ_n's asymptotic order is $\lambda_n \asymp \sqrt{\log p/n}$.
  - ℓ^2-error bound is asymptotically $\sqrt{s \log p/n}$.
  - $n\lambda_n^2 \asymp \log p$, so that the convergence probability is indeed $1 - cp$.
  - In sum, larger dimension makes the convergence of the probability faster, while making the actual bound loose.
  - Larger the σ, stronger the λ_n and loosening the bound. This is natural as σ stands for the strength of the noise.

Exact sparsity LASSO V

- Sketch of the proof: Applying Cor 3.3 with appropriate quantities.
  1. Note ℓ^2-RE $\Rightarrow$ RSC w.r.t. $\mathcal{M} = \mathcal{M}(S)$.
  2. ℓ^1-norm is decomposable w.r.t. $\mathcal{M}(S)$ and its orthogonal complement, so that $\overline{\mathcal{M}}(S) = \mathcal{M}$.
  3. $\Psi(\mathcal{M}(S)) = \sup_{\theta \in \mathcal{M} \setminus \{0\}} \frac{\|\theta\|_1}{\|\theta\|_2} = \sqrt{s}$.
  4. Since $\mathcal{R}^* = \ell^\infty$ and $\nabla \mathcal{L}(\theta^*, Y, X) = X^T w/n$, we require $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*)) = 2\|X^T w/n\|_\infty$ to apply Cor 3.3.
  5. We find that $\lambda_n = 4\sigma\sqrt{\log p/n}$ satisfies the lower bound in 4 with probability at least $1 - c_1 exp(-c_2 n \lambda_n^2)$.
  6. All quantities for Cor 3.3 are now explicit, so we apply Cor 3.3 to obtain the desired bound.

Exact sparsity LASSO VI

- Details for some steps:
  - Step 5:
    - Note that normalized column condition and (modified marginal) sub-Gaussian condition imply

$$\mathbb{P}\left(\frac{1}{n}|\langle X_i, w\rangle| \geq t\right) \leq 2exp\left(-\frac{nt^2}{2\sigma^2}\right) \qquad \forall i = 1, \dots p, \ \forall t > 0$$

$$\overset{\text{Union Bound}}{\Rightarrow} \mathbb{P}\left(\left\|\frac{X^T w}{n}\right\|_\infty \geq t\right) \leq 2exp\left(-\frac{nt^2}{2\sigma^2} + \log p\right)$$

    - We choose $t = 2\sigma\sqrt{\log p/n}, \lambda_n = 2t$ so that

$$\mathbb{P}\left(2\left\|\frac{X^T w}{n}\right\|_\infty \leq \lambda_n\right) \leq 1 - 2exp(\log p) = 1 - c_1 exp(-c_2 n \lambda_n^2).$$

Weakly sparse LASSO I

- Weakly sparse LASSO is a linear regression problem (1), (5), which $\theta^* \notin \mathcal{M}(S)$, but still 'approximated well' by $\mathcal{M}(S)$
- We first clarify the meaning of 'approximated well':
  - Fix $q \in [0,1]$. We consider the case $\theta^* \in \ell^q$-ball of radius R_q:

$$\mathbb{B}(R_q) := \{\theta \in \mathbb{R}^p : \sum_{i=1}^{p} \|\theta_i\|^q \leq R_q\}.$$

  - E.g.: $q = 0$, $R_q = s$ corresponds to at most s-sparsity.
- $\mathbb{C}(\mathcal{M}(S), \overline{\mathcal{M}}(S), \theta^*) = \{\Delta \in \mathbb{R}^p : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 + 4\|\theta^*_{S^c}\|_1\}$ is no longer a cone set ('star-shaped' set).
- Since the main change of the scheme is the change in the cone set, main modification of the analysis is to reform the RSC condition appropriately.

Weakly sparse LASSO II

- Assume the following for solving the problem (1), (5):
  1. X has normalized columns (G'1).
  2. X satisfies generalized ℓ^2-RE condition (4).
  3. w satisfies the sub-Gaussian condition (G'2).
  4. $\theta^* \in \mathbb{B}_q(R_q)$ for some $R_q > 0$: Weakly sparse condition.

Weakly sparse LASSO III

Corollary (Corollary 3 of [NRWY12])

Under the above assumptions, if q and R_q satisfies the following condition:

$$\sqrt{R_q}\left(\frac{\log p}{n}\right)^{\frac{2-2q}{4}} \leq 1,$$

then, the optimal solution of (1), $\widehat{\theta}_{\lambda_n}$, *with* $\lambda_n = 4\sigma\sqrt{\log p/n}$ *satisfies*

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq c_0 R_q \left(\frac{\sigma^2}{\kappa_1^2}\frac{\log p}{n}\right)^{1-\frac{q}{2}}$$

with probability at least $1 - c_1 exp(-c_2 n\lambda_n^2)$ *for some constants* $c_0, c_1, c_2 > 0$.

Weakly sparse LASSO IV

- Some remarks:
  - If $q = 0, R_q = s$, which corresponds to at most s-sparsity, then Cor 4.2 coincides to Cor 4.1 with $c_0 = 64$.
  - The condition on q and R_q implies that θ^* needs to be 'close enough' to the sparse set. $q \in [0, 1]$ controls the relative 'sparsifiability' of θ^*. Smaller the q, more sparse the θ^*.
  - On the other hand, if q is smaller, R_q can be larger. So, if sparsifiability is strong, then we can relax the 'required closedness' to the sparse set.
  - Convergence rate gets slower as q or R_q increases, meaning θ^* is less sparse, which is very natural.
  - This rate is the optimal minimax rate for all $q \in [0, 1]$ ([RWY09]).

Weakly sparse LASSO V

- Sketch of the proof: As mentioned, key part is RSC condition part.
  1. For some η (which will be chosen to be λ_n/κ_1 later), define the thresholded subset

$$S_\eta := \{j \in \{1, \dots, p\} : |\theta_j^*| > \eta\}.$$

  2. Use [NRWY12][Lemma 2] to obtain the RSC condition.
  3. Apply Theorem 3.2 with $\Psi^2(S_\eta) = |S_\eta|$ yielding the bound w.r.t. $|S_\eta|$ and $\|\theta^*_{S_\eta^c}\|_1$.
  4. Control $|S_\eta|$ and $\|\theta^*_{S_\eta^c}\|_1$ in terms of η, q, and R_q.
     - $R_q \geq \sum_{i=1}^p |\theta_i^*|^q \geq \sum_{S_\eta} |\theta_i^*|^q \geq \eta^q |S_\eta|$.
     - $\|\theta^*_{S_\eta^c}\|_1 = \sum_{i \in S_\eta^c} |\theta_j^*| = \sum_{i \in S_\eta^c} |\theta_j^*|^q |\theta_j^*|^{1-q} \leq R^q \eta^{1-q}$ from $\theta^* \in \mathbb{B}(R_q)$.
  5. Plug-in $\eta = \lambda_n/\kappa_1$.
  6. From here, setting λ_n and obtaining the probabilistic bound works exactly same to Cor 4.1.

GLM I

- Problem setting: We consider a GLM problem, under the following setting.
  - A design matrix $X \in \mathbb{R}^{n \times p}$ is normalized by 1.
  - Conditionally on x_i, the response y_i is drawn from the following conditional distribution:

$$\mathbb{P}_{\theta^*}(y|x) \propto exp\left(\frac{y\langle x, \theta^*\rangle - g(\langle x, \theta^*\rangle)}{c(\sigma)}\right).$$

    Here $c(\sigma)$ is a known fixed scale parameter and $g : \mathbb{R} \to \mathbb{R}$ is the link function.

- We consider the following optimization problem, called GLM LASSO:

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} (g(\langle x_i, \theta\rangle - y_i\langle x_i, \theta\rangle) + \lambda_n \|\theta\|_1 \right\} \qquad (6)$$

GLM II

- We state assumptions here:
  1. GLM problem satisfies the (modified) RSC condition. *i.e.*

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_1 \|\Delta\|_2^2 - \kappa_2 \frac{\log p}{n} \|\Delta\|_1^2 \qquad \forall \|\Delta\|_2 \leq 1.$$

  2. $\theta^* \in S$.
  3. $s(\lambda_n^2 + \log p/n) < \min\{4\kappa_1^2/9, \kappa_1/64\kappa_2\}$.
  4. $\lambda_n = 4B(\sqrt{\log p/n} + \delta)$ for some $0 < \delta < 1$.
  5. The link function g's second derivative is bounded by B^2, *i.e.* $\|g''\|_\infty \leq B^2$.

GLM III

Corollary (Corollary 9.26 of [Wai19])

Under the above assumptions, with probability at least $1 - 2exp(-2n\delta^2)$,

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 \leq \frac{9}{4}\frac{s\lambda_n^2}{\kappa_1^2}$$

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_1 \leq 6\frac{s\lambda_n}{\kappa_1}$$

GLM IV

- Some remarks:
  - The RSC condition 1 is indeed valid if x_i's are 0-mean i.i.d. variables with assumptions on second and fourth moments. See appendix and [Wai19][Theorem 9.36] for more detail.
  - The third condition determine the relationship between $n, \lambda_n, \log p, s$. If $s \log p/n$ is small, then λ_n can be larger.
  - The choice of link function effects the regularizer strength by B. Note that Poisson regression does not have such B, Poisson regression fails to fall into this setting.

GLM V

- Sketch of the proof: Apply Theorem 3.2 with appropriate quantities.
  1. We set a coordinate subspace, and as a result $\Psi(\mathcal{M}) = \sqrt{s}$.
  2. Retrive the true RSC condition on the cone set from (modified) RSC condition on a unit ball. This is satisfied by Assumption 3.
  3. We require $\lambda_n \geq 2\|\nabla\mathcal{L}(\theta^*)\|_\infty$, and we show our choice $\lambda_n = 4B(\sqrt{\log p/n} + \delta)$ guarantees to be larger than RHS with high probability.
     - First, we show that each element of $\nabla\mathcal{L}(\theta^*)$ is a sub-Gaussian element.
     - With sub-Gaussian, we can do same thing in exact LASSO case, combining sub-Gaussian and union bound to obtain probabilistic upper bound of $\|\nabla\mathcal{L}(\theta^*)\|_\infty$.
     - Set λ_n to bound the term with the stated probability.
  4. Apply Theorem 3.2.

GLM VI

- Details of the proof: sub-Gaussian of $\nabla\mathcal{L}(\theta^*)$.
  - Let $V_{ij} = (g'(\langle x_i, \theta^* \rangle) - y_i)x_{ij}$. Then, $\nabla\mathcal{L}(\theta^*) = \frac{1}{n}\sum_i V_i$. Note that V_i is 0-mean vectors under the true model.
  - We check V_{ij} is sub-Gaussian by analyzing its MGF.

$$\begin{aligned}\log \mathbb{E}(exp(-tV_{ij})) &= g(tx_{ij} + \langle x_i, \theta^* \rangle) - g(\langle x_i, \theta^* \rangle) - tx_{ij}g'(\langle x_i, \theta^* \rangle) \\ &= \frac{1}{2}t^2 x_{ij}^2 g''(sx_{ij} + \langle x_i, \theta^* \rangle) \leq \frac{1}{2}t^2 x_{ij}^2 B^2\end{aligned}$$

    by Taylor expansion and bound on g''.
  - Independence and normalized column leads to

$$\begin{aligned}\log \mathbb{E}(exp(-t\frac{1}{n}\sum_i V_{ij})) &\leq \frac{1}{n}\log \mathbb{E}(exp(-t\sum_i V_{ij})) \\ &\leq \frac{1}{2}t^2 B^2(\frac{1}{n}\sum_i x_{ij}^2) \leq \frac{1}{2}t^2 B^2.\end{aligned}$$

    proving the sub-Gaussianness of jth element of $\overline{V_i}$.

Conclusion

Conclusion

- The paper explores a broad framework of regularized *M*-estimators, capturing various problems as specific instances.
- It achieves a cohesive theoretical understanding through fundamental techniques and measures.
  - Essential elements include the decomposability of $\mathcal{R}(\cdot)$, restricted strong convexity (RSC) of $\mathcal{L}_n(\cdot)$, the dual $\mathcal{R}^*(\cdot)$, and the subspace compatibility constant $\Psi(\cdot)$.
- It establishes convergence rates for diverse scenarios:
  - These include linear regression with different sparsity types, sparse generalized linear models (GLM), and low-rank matrix recovery.

References I

- Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*.
- Garvesh Raskutti, Martin J. Wainwright, and Bin Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, 2009.
- M.J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.

Appendices

Lemma 2 of [NRWY12] I

- [NRWY12][Lemma 2]

Lemma

Assume conditions in Cor 4.2 holds and $n > 9\kappa_2|S_\eta|\log p$. *Then with* $\eta = \lambda_n/\kappa_1$, *RSC condition with* $\kappa_{\mathcal{L}} = \kappa_1/2$, *and* $\tau_{\mathcal{L}} = 2\kappa_2\sqrt{\log p/n}\|\theta^*_{S^c_\eta}\|_1$ *holds over* $\mathbb{C}(\mathcal{M}(S_\eta), \mathcal{M}^\perp(S_\eta), \theta^*)$.

- Sketch of the proof:
  1. Notice for all $\Delta \in \mathbb{C}(S_\eta)$,

$$\|\Delta\|_1 \leq 4\|\Delta_{S_\eta}\|_1 + 4\|\theta^*_{S_\eta}\|_1 \leq 4\sqrt{|S_\eta|}\|\Delta\|_2 + 4R_q\eta^{1-q}$$
$$\leq 4\sqrt{R_q}\eta^{-q/2}\|\Delta\|_2 + 4R_q\eta^{1-q}.$$

  2. Plug-in the above $\|\Delta\|_1$ to generalized RE condition (4), which leads to

$$\frac{\|X^T\Delta\|_2}{\sqrt{n}} \geq \|\Delta\|_2\left(\kappa_1 - \kappa_2\sqrt{\frac{R_q \log p}{n}}\eta^{-q/2}\right) - \kappa_2\sqrt{\frac{\log p}{n}}R_q\eta^{1-q}.$$

  3. With the setting $\lambda_n = 4\sigma\sqrt{\log p/n}$, $\eta = \lambda_n/\kappa_1$, and the condition on n, the middle term is $\leq \kappa_1/2$, which implies

$$\frac{\|X^T\Delta\|_2}{\sqrt{n}} \geq \frac{\kappa_1}{2}\|\Delta\|_2 - 2\kappa_2\sqrt{\frac{\log p}{n}}\|\theta^*_{S^c_\eta}\|_1.$$

Rademacher Complexity

Definition (Rademacher Complexity)

Let $S_n = \{x_1, \ldots, x_n\}$ be a set of points in $\mathbb{R}^d$ (a data sample) and $\mathcal{F}$ a real-valued function class. We define the empirical Rademacher complexity of $\mathcal{F}$ on the data sample as

$$\widehat{Rad}(\mathcal{F}; S_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right]$$

where ϵ_i are iid random variables which take the values ± 1 with equal probability $\frac{1}{2}$. The population Rademacher complexity is defined as

$$Rad_n(\mathcal{F}) = \mathbb{E}_{S_n \sim \mathbb{P}^n} \left[\widehat{Rad}(\mathcal{F}; S_n) \right],$$

i.e. as the expected empirical Rademacher complexity over a set of n iid data points.

Theorem 9.36 of [Wai19] I

Theorem (Theorem 9.36 of [Wai19])

Assume the following:

1. x_i*'s are i.i.d. samples of 0-mean distributions.*
2. *There exists a positive constants* $\alpha, \beta > 0$ *such that* $\mathbb{E}[\langle \Delta, x \rangle^2] \geq \alpha$ *and* $\mathbb{E}[\langle \Delta, x \rangle^4] \leq \beta$ *for all* $\Delta \in \mathbb{S}^{p-1}$.

Then, in GLM setting with general $\mathcal{R}$*,*

$$\delta \mathcal{L}(\Delta, \theta^*) \geq \frac{\kappa}{2} \|\Delta\|_2^2 - c_0 \widehat{Rad}(\mathbb{B}_{\mathcal{R}^*}(1)) \mathcal{R}(\Delta)^2 \qquad \forall \Delta \in \mathbb{S}^{p-1}$$

with probability at least $1 - c_1 exp(-c_2 n)$ *for some constant* $\kappa, c_0, c_1, c_2 > 0$.

Theorem 9.36 of [Wai19] II

- Some remarks:
  - In GLM Lasso case, $\mathcal{R} = \|\cdot\|_1$. By calculating the Rademacher complexity of ℓ^∞ dual norm ball, we can retrieve the $\log p/n$ in RSC for GLM.
  - Proof idea of [Wai19][Theorem 9.36]:
    1. Start with the Taylor series of the error term on $\theta_n = \theta^* + \Delta$ up to second term.
    2. The trick is to apply truncations on $\langle \theta^*, x_i \rangle$ and $\langle \Delta, x_i \rangle$, which still yields lower bound of the error term. This is because the error is always positive (due to basic inequality). This makes the error Lipschitz.
    3. Once we have Lipschitz, it is sufficent to control the domain with high probability, instead of the error itself. This can be done by using moment conditions.