Regularized Loss Minimization (cont'd)

2022年1月25日 9:15

Recall our setup:

- Estimator

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta; \{z_i\}_{i=1}^n) + \lambda R(\theta) \right\};$$

(R: norm)

$$z_i = (x_i, y_i) \; ; \quad y = X\theta^* + w.$$

- $m \subseteq \bar{m} \subseteq \mathbb{R}^d$: model subspace,

$\bar{m}^\perp$: perturbation subspace,

Decomposability: $R(\theta + \gamma) = R(\theta) + R(\gamma) \quad \forall \theta \in m, \gamma \in \bar{m}^\perp$

Proposition: Suppose that $\mathcal{L}$ is smooth and convex,

$\lambda \geq 2R^*(\nabla \mathcal{L}(\theta^*))$ and R is decomposable wrt $(m, \bar{m}^\perp)$.

(dual norm: $R^*(\theta) = \max_{\eta: R(\eta) \leq 1} \eta^T \theta$ e.g.: $R(\theta) = \|\theta\|_2$, $R^*(\theta) = \max_{\|\eta\|_2 \leq 1} \eta^T \theta = \|\theta\|_2$)

Then,

$$\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C} \triangleq \left\{ \Delta \in \mathbb{R}^d : R(\Delta_{\bar{m}^\perp}) \leq 3R(\Delta_{\bar{m}}) + 4R(\theta^*_{m^\perp}) \right\}$$

Remarks:

1) Consider $\mathcal{L}(\theta, \{z_i\}_{i=1}^n) = \frac{1}{2}\|y - X\theta\|_2^2$, Then, $\nabla \mathcal{L}(\theta^*) = -X^T(y - X\theta^*)$

$= -X^T w$. Hence, if $R(\theta) = \|\theta\|_1$, then $R^*(\theta) = \|\theta\|_\infty = \max_i |\theta_i|$

and we require $\lambda \geq 2\|X^T w\|_\infty$. The last quantity can be estimated if we assume probabilistic models on X and w.

2) If $\theta^* \in m$, then $R(\theta^*_{m^\perp}) = 0$ and $\mathcal{C}$ is a cone: If $\Delta \in \mathcal{C}$, then $t\Delta \in \mathcal{C} \; \forall t > 0$.

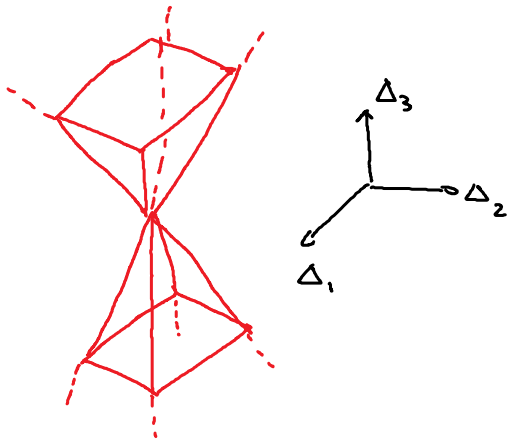
Example: $d = 3$, 1-sparse vector S = Support set = $\{3\}$

$$m(S) = \{\theta \in \mathbb{R}^3 : \theta_1 = \theta_2 = 0\} = \{\theta \in \mathbb{R}^3 : \theta_i = 0 \; \forall i \notin S\}$$

1°. $\theta^* = (0, 0, x)$ Correctly specified m

$R(\theta) = \|\theta\|_1$, $m = \bar{m}$; $m^\perp = \{\theta \in \mathbb{R}^3 : \theta_3 = 0\}$

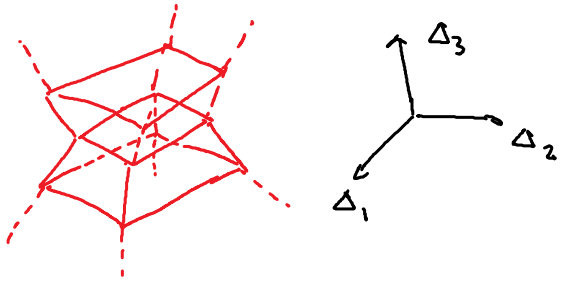
$\therefore \quad \mathcal{C} = \{ \Delta \in \mathbb{R}^3 : |\Delta_1| + |\Delta_2| \le 3|\Delta_3| \}$



2°: $\theta^* = (0, x, xx)$ incorrectly specified M

$\mathcal{C} = \{ \Delta \in \mathbb{R}^3 : |\Delta_1| + |\Delta_2| \le 3|\Delta_3| + 4|\theta^*_2| \}$

fixed



Proof of Proposition:

Define

$$\mathcal{D}(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda(R(\theta^* + \Delta) - R(\theta^*))$$

(recall: we are minimizing $\mathcal{L}(\theta) + \lambda R(\theta)$). Note that with $\hat{\Delta} = \hat{\theta} - \theta^*$, we have $\mathcal{D}(\hat{\Delta}) \le 0$ by optimality of $\hat{\theta}$

Claim 1: $R(\theta^* + \Delta) - R(\theta^*) \ge R(\Delta_{\bar{M}^\perp}) - R(\Delta_{\bar{M}}) - 2R(\theta^*_{\bar{M}^\perp})$

Claim 2: If $\lambda \ge 2R^*(\nabla \mathcal{L}(\theta^*))$ and $\mathcal{L}$ is convex, then

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

With these claims:

$$0 \ge \mathcal{D}(\hat{\Delta}) \ge \lambda\left[R(\hat{\Delta}_{\bar{M}^\perp}) - R(\hat{\Delta}_{\bar{M}}) - 2R(\theta^*_{\bar{M}^\perp})\right] \quad \leftarrow \text{Claim 1}$$

$$-\frac{\lambda}{2}\left[R(\hat{\Delta}_{\bar{M}}) + R(\hat{\Delta}_{\bar{M}^\perp})\right] \quad \leftarrow \text{claim 2}$$

$$= \frac{\lambda}{2}\left[R(\hat{\Delta}_{\bar{M}^\perp}) - 3R(\hat{\Delta}_{\bar{M}}) - 4R(\theta^*_{M^\perp})\right]$$

Proof of Claim 1: We want to prove

$$R(\theta^* + \Delta) - R(\theta^*) \geq R(\Delta_{\bar{M}^\perp}) - R(\Delta_{\bar{M}}) - 2R(\theta^*_{M^\perp})$$

We compute

$$R(\theta^* + \Delta) = R(\theta^*_M + \theta^*_{M^\perp} + \Delta_{\bar{M}} + \Delta_{\bar{M}^\perp})$$

$$\geq R(\theta^*_M + \Delta_{\bar{M}^\perp}) - R(\theta^*_{M^\perp} + \Delta_{\bar{M}}) \quad \left.\begin{matrix}\\ \\ \end{matrix}\right\} \text{triangle inequality}$$

$$\geq R(\theta^*_M + \Delta_{\bar{M}^\perp}) - R(\theta^*_{M^\perp}) - R(\Delta_{\bar{M}})$$

$$= R(\theta^*_M) + R(\Delta_{\bar{M}^\perp}) - R(\theta^*_{M^\perp}) - R(\Delta_{\bar{M}}) \quad \text{decomposability}$$

On the other hand,

$$R(\theta^*) \leq R(\theta^*_M) + R(\theta^*_{M^\perp}) \quad \text{triangle inequality}$$

Summary: We know $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C}$.

To bound the error, it suffices to show that $\mathcal{L}$ is not "too flat" on $\mathcal{C}$ This motivates the following definition:

Definition: (Restricted Strong Convexity (RSC))

We say that $\mathcal{L}$ is RSC on $\mathcal{C}$ if there exists a constant $\kappa > 0$ and a function $\tau(\cdot)$ s.t.

$$\mathcal{L}(\theta^* + \Delta) \geq \mathcal{L}(\theta^*) + \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) \quad \forall \Delta \in \mathcal{C}.$$

$\tau^2(\theta^*)$: Statistical tolerance

$\forall \Delta \in \mathcal{C}$: "restricted"

Example: Consider $\ell(\theta) = \frac{1}{2}\|y - X\theta\|_2^2$. Then,

$$\ell(\theta^* + \Delta) - \ell(\theta^*) - \nabla\ell(\theta^*)^T\Delta = \frac{1}{2}\|X\Delta\|_2^2 \qquad \text{(check)}$$

Then, the RSC property means

$$\frac{1}{2}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 - \tau^2(\theta^*)$$

Next: With the above, we have

Theorem: Under the assumptions of the Proposition and the assumption that ℓ is RSC on $\mathcal{C}$, we have

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \underbrace{\frac{9\lambda^2}{4\kappa^2}\Psi^2(\bar{m})}_{\text{tightness measure}} + \underbrace{\frac{2}{\kappa}\left[\tau^2(\theta^*) + 2\lambda R(\theta^*_{m^\perp})\right]}_{\text{model specification error}},$$

where

$$\Psi(m) = \sup_{u \in m \setminus \{0\}} \frac{R(u)}{\|u\|_2}$$

can be regarded as the Lipschitz constant of R over the subspace m.

Example: Consider $R(\theta) = \|\theta\|_1$. For $\theta \in \mathbb{R}^d$,

$$\|\theta\|_1 \leq \sqrt{d}\,\|\theta\|_2 \Rightarrow \Psi(\mathbb{R}^d) = \sqrt{d}.$$

However, consider $m = \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \notin S\}$ with $|S| = s << d$. Then, one can show

$$\Psi(m) = \sqrt{s}$$