Regularized Loss Minimization (cont'd)

2022年1月28日 9:21

Recall the Setup:

- $y = X\theta^* + w, \quad X \in \mathbb{R}^{n\times d}, \quad n \ll d \;; \quad \mathcal{M} \subseteq \bar{\mathcal{M}} \subseteq \mathbb{R}^d$

model $\qquad \bar{\mathcal{M}}^\perp$: perturbation

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{ \mathcal{L}(\theta) + \lambda R(\theta) \}$$

- Localization of error vector

$$\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C} = \{ \Delta \in \mathbb{R}^d : R(\Delta_{\bar{\mathcal{M}}^\perp}) \le 3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\mathcal{M}^\perp}) \}$$

- RSC of $\mathcal{L}$ over $\mathcal{C}$

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

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

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

where

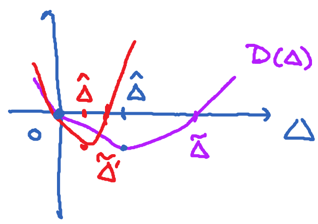
$$\psi(\mathcal{M}) = \sup_{u \in \mathcal{M}\setminus\{0\}} \frac{R(u)}{\|u\|_2}$$

Proof: Recall $\hat{\Delta} = \hat{\theta} - \theta^*$

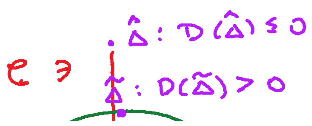
$$D(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda [R(\theta^* + \Delta) - R(\theta^*)]$$

and $D(\hat{\Delta}) \le 0$ and $D(\cdot)$ is convex

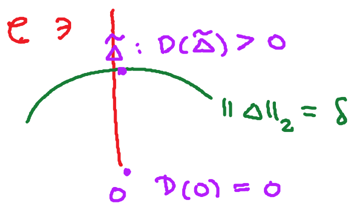
Idea: For $\Delta \in \mathcal{C}$, if $|D(\Delta)|$ is small, then $\|\Delta\|_2$ should be small



Claim: For any $\delta > 0$, if $D(\Delta) > 0$ for all $\Delta \in K(\delta) \triangleq \mathcal{C} \cap \{ \Delta \in \mathbb{R}^d : \|\Delta\|_2 = \delta \}$, then $\|\hat{\Delta}\|_2 \le \delta$.



then $\|\Delta\|_2 \leq \delta$.



To prove the theorem, it suffices to find a suitable $\delta > 0$ such that $D(\Delta) > 0$ for all $\Delta \in K(\delta)$. To do this, we compute, for $\Delta \in K(\delta)$,

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

$$\geq \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) \qquad \text{(RSC)}$$

$$+ \lambda\left[R(\Delta_{\bar{m}^\perp}) - R(\Delta_{\bar{m}}) - 2R(\theta^*_{m^\perp})\right] \qquad \text{(Claim 1 from last lecture)}$$

Idea: Lower bound the RHS by a quadratic function in $\|\Delta\|_2$

$$\geq \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) + \lambda\left[R(\Delta_{\bar{m}^\perp}) - R(\Delta_{\bar{m}}) - 2R(\theta^*_{m^\perp})\right]$$

$$- R^*(\nabla\mathcal{L}(\theta^*)) \cdot R(\Delta) \qquad \text{(generalized Cauchy-Schwarz)}$$

$\lambda/2 \geq R^*(\nabla\mathcal{L}(\theta^*))$ $\qquad$ $\leq R(\Delta_{\bar{m}^\perp}) + R(\Delta_{\bar{m}})$ (triangle inequality)

$$\geq \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) + \frac{\lambda}{2}\left[\underbrace{R(\Delta_{\bar{m}^\perp})}_{\geq 0} - 3R(\Delta_{\bar{m}}) - 4R(\theta^*_{m^\perp})\right]$$

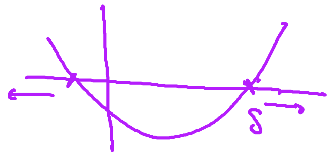
$$\geq \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) - \frac{\lambda}{2}\left[3R(\Delta_{\bar{m}}) + 4R(\theta^*_{m^\perp})\right]$$

Note: $R(\Delta_{\bar{m}}) \leq \Psi(\bar{m}) \cdot \|\Delta_{\bar{m}}\|_2$ $\qquad$ $\left[\Psi(\bar{m}) = \sup_{u \in \bar{m}\setminus\{0\}} \frac{R(u)}{\|u\|_2}\right]$

$\leq \Psi(\bar{m}) \cdot \|\Delta\|_2$

$$\geq \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) - \frac{\lambda}{2}\left[3\Psi(\bar{m}) \cdot \|\Delta\|_2 + 4R(\theta^*_{m^\perp})\right]$$

The RHS is a quadratic function in $\|\Delta\|_2$ and will be positive if $\|\Delta\|_2 > \delta$, where δ is the positive root of the quadratic function.

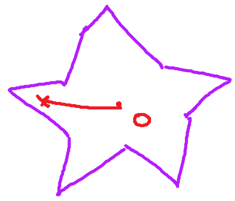


Back to the claim: What is the idea of the proof?

What we need is the following fact:

Fact: $\mathcal{C}$ is star-shaped: For any $\Delta \in \mathcal{C}$,

$$\{ t\Delta : t \in (0,1) \} \subseteq \mathcal{C}$$



Example: LASSO with exactly sparse model

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \Big\{ \underbrace{\frac{1}{2n} \|y - X\theta\|_2^2}_{\mathcal{L}(\theta)} + \lambda \underbrace{\|\theta\|_1}_{R(\theta)} \Big\} \qquad \text{— (L)}$$

By taking S to be the support of θ^* with $|S| = s$, $\mathcal{C}$ takes the form

$$\mathcal{C} = \{ \Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq 3 \|\Delta_S\|_1 \}$$

Want:

① $\mathcal{L}$ is RSC over $\mathcal{C}$:

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa \|\Delta\|_2^2 \quad \forall \Delta \in \mathcal{C}. \qquad \text{— (*)}$$

② Estimate $R^*(\nabla \mathcal{L}(\theta^*))$: $\quad R^*(\theta) = \|\theta\|_\infty$

$$\therefore \quad R^*(\nabla \mathcal{L}(\theta^*)) = \|X^T w\|_\infty$$

Proposition: Suppose that

(i) X satisfies $(*)$;

(ii) $\frac{\|X_j\|_2}{\sqrt{n}} \leq 1 \quad \forall j$, where X_j is the j^{th} col of X,

(iii) w is sub-Gaussian with parameter $\sigma > 0$: w has zero mean and for any fixed $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$,

$$\Pr\left[\, |v^T w| \geq t \,\right] \leq 2 \exp\left(-t^2/2\sigma^2\right) \quad \forall t > 0$$

(iv) θ^* is s-sparse.

Then, by taking $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$ in (L), then

$$\|\hat{\theta} - \theta^*\|_2^2 \leq 36 \frac{\sigma^2}{\kappa^2} \cdot \frac{s \log d}{n}$$

with probability $\geq 1 - 2/d$.