Non-Convex Regularized Loss Minimization (cont'd)

2022年3月15日 9:01

Example: Linear model with additively corrupted covariates

Consider

$$\underset{\substack{\text{response}\\ \text{variable}}}{y_i} = \underset{\substack{\text{covariate}\\ \text{vector}}}{x_i^T}\overset{\text{ground truth}}{\theta^*} + \underset{\substack{\text{noise}\\ \mathbb{E}[\varepsilon_i]=0}}{\varepsilon_i}, \quad y_i \in \mathbb{R}, \quad x_i \in \mathbb{R}^d, \quad \theta^* \in \mathbb{R}^d, \quad \varepsilon_i \in \mathbb{R} \qquad -(L)$$

More compactly,

$$y = X\theta^* + \varepsilon$$

The estimation problem

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \| y - X\theta \|_2^2 + R(\theta) \right\} \qquad (LS)$$

can be viewed as a sample average version of the following idealized problem:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2} \theta^T \overset{\frac{1}{n}X^TX}{\boxed{\Sigma_x}} \theta - \theta^T \overset{\frac{1}{n}X^Ty}{\boxed{\Sigma_x \theta^*}} + R(\theta) \right\}, \qquad (I)$$

where $\Sigma_x \succ 0$ is the covariance of $\{x_i\}$, which are assumed to be iid mean-zero. This follows from

$$\Sigma_x = \mathbb{E}[xx^T], \qquad \mathbb{E}_{\varepsilon_i}[y_i x_i] = x_i x_i^T \theta^*,$$

$$X^TX = \sum_{i=1}^n x_i x_i^T, \qquad X^Ty = X^T(X\theta^* + \varepsilon)$$

Now, suppose that x_i is not observed directly, but rather we observe a $z_i \in \mathbb{R}^d$ that is related to x_i via

$$z_i = x_i + w_i, \qquad x_i \in \mathbb{R}^d, \ z_i \in \mathbb{R}^d, \ i = 1, \cdots, n \quad (n \ll d)$$

where $w_i \in \mathbb{R}^d$ is the noise vector, assumed to be zero mean and have covariance Σ_w (known), and w_i is independent of x_i and ε_i. Following our previous argument, we compute

$$\frac{1}{n}\mathbb{E}_w[Z^TZ] = \frac{1}{n}X^TX + \Sigma_w, \qquad \frac{1}{n}\mathbb{E}_w[Z^Ty] = \frac{1}{n}X^Ty$$

Then, we see that in this setting, we can use

$$\hat{\Gamma} = \frac{1}{n}Z^TZ - \Sigma_w, \qquad \hat{\gamma} = \frac{1}{n}Z^Ty$$

$$\hat{\Gamma} = \frac{1}{n}\underbrace{Z^T Z}_{d\times d} - \Sigma_w \quad , \quad \hat{\gamma} = \frac{1}{n} Z^T y$$

and this gives rise to

$$\mathbb{R}^{n\times d} \ni Z = \begin{bmatrix} - z_1^T - \\ \vdots \\ - z_n^T - \end{bmatrix}$$

$$\hat{\theta} \in \underset{\boxed{\|\theta\|_1 \leq R}}{\arg\min} \left\{ \frac{1}{2}\theta^T \hat{\Gamma} \theta - \hat{\gamma}^T \theta \right\}$$

make sure $\hat{\theta}$ exists

$$\frac{1}{n}\sum_{i=1}^{n} z_i z_i^T$$

$//$

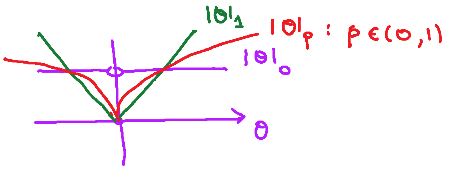
Note that $\hat{\Gamma}$ need not be psd. Indeed, $\frac{1}{n}Z^T Z$ has rank at most n but $\Sigma_w \in S_+^d$ could have rank d (e.g., $\Sigma_w = I_d$). Hence, the above formulation is non-convex in general. We can also consider the regularized version

$$\hat{\theta}' \in \underset{\boxed{\|\theta\|_1 \leq R'}}{\arg\min} \left\{ \frac{1}{2}\theta^T \hat{\Gamma} \theta - \hat{\gamma}^T \theta + \lambda \|\theta\|_1 \right\}$$

make sure $\hat{\theta}$ exists

Example: Non-convex regularizer

When dealing with sparse model, we want to use ℓ_0-quasi-norm to measure sparsity. This function is challenging, so we replace it by a convex surrogate: ℓ_1-norm. To get a better approximation of $\|\cdot\|_0$, one can consider non-convex (but continuous) surrogates



e.g. $\|\theta\|_p^p = \sum_{i=1}^{d} |\theta_i|^p$ $\quad$ ℓ_p-(quasi)-norm if $p \in (0,1)$

(Bridge penalty)

However, this is not quite desirable, because

$$\lim_{\theta \downarrow 0} (|\theta|^p)' = +\infty$$

Consider the following family of regularizers:

$$\mathbb{R}^d \ni \theta \mapsto R_\lambda(\theta) \underset{\substack{\uparrow \\ \text{Separable}}}{=} \sum_{i=1}^{d} R_\lambda(\theta_i)$$

and the function $R_\lambda : \mathbb{R} \to \mathbb{R}_+$ satisfies the following properties:

(1) $R_\lambda(0) = 0$ and $R_\lambda(t) = R_\lambda(-t)$ $\forall t \in \mathbb{R}$,

(2) R_λ is non-decreasing on $\mathbb{R}_+$

(3) For $t > 0$, the function $t \mapsto \frac{R_\lambda(t)}{t}$ is non-increasing in t.

(4) R_λ is differentiable at any $t \neq 0$ and subdifferentiable at $t = 0$ with $\lim_{t \downarrow 0} R_\lambda'(t) = \lambda L$ for some $L > 0$.

needs definition b/c R_λ may not be convex ✓ (see below)

(5) There exists a $\mu > 0$ s.t. $t \mapsto R_{\lambda,\mu}(t) \triangleq R_\lambda(t) + \frac{\mu}{2} t^2$ is convex (in other words, R_λ is μ-weakly convex).

With (5), we can write

$$R_\lambda(t) = R_{\lambda,\mu}(t) - \frac{\mu}{2} t^2 .$$

Hence, formally, we can write

$$\begin{aligned} \partial R_\lambda(t) &= \partial\left(R_{\lambda,\mu}(t) - \frac{\mu}{2} t^2 \right) \\ &= \underbrace{\partial R_{\lambda,\mu}(t)}_{\substack{\text{well-defined b/c} \\ R_{\lambda,\mu} \text{ is convex}}} - \mu t \end{aligned}$$

Hence, we can define

$$\partial R_\lambda(t) \triangleq \partial R_{\lambda,\mu}(t) - \mu t .$$