Non-Convex Regularized Loss Minimization (cont'd)

2022年3月17日 18:10

Example: Non-convex regularizer

Consider a 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) 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).

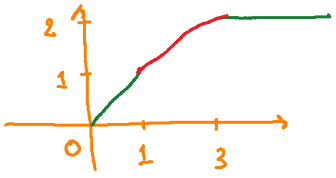
(5) R_λ is differentiable at any $t \neq 0$ and subdifferentiable at $t = 0$ with $\lim_{t \downarrow 0} R_\lambda'(t) = \lambda L$ for some $L > 0$.

e.g.:

(1) Smoothly Clipped Absolute Deviation (SCAD)

$$R_\lambda(t) = \begin{cases} \lambda |t| & , \ |t| \le \lambda, \\ -\frac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)} & , \ \lambda \le |t| \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & , \ |t| > a\lambda \end{cases}$$

$\lambda = 1$, $a = 3$



where $a > 2$ is fixed. Note that R_λ is piecewise quadratic. Conditions (1) - (3) can be easily verified. For condition (4), by taking $\mu = \frac{1}{a-1}$, we have $R_{\lambda,\mu}(t) = \frac{a\lambda}{a-1}|t| - \frac{\lambda^2}{2(a-1)}$ for $\lambda \le |t| \le a\lambda$. This can be used to complete the verification of (4) (Exercise: check all pieces) For condition

(5), we compute, for any $t \neq 0$,

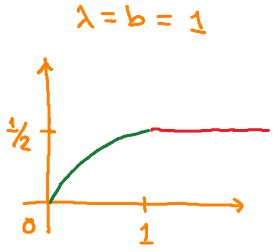
$$R_\lambda'(t) = \mathrm{sgn}(t) \cdot \left[\lambda \mathbb{1}_{\{|t| \leq \lambda\}} + \frac{(a\lambda - |t|)_+}{a-1} \mathbb{1}_{\{|t| > \lambda\}} \right]$$

and at $t=0$, we have $\partial R_\lambda(0) = [-\lambda, \lambda]$. Hence,

$$\lim_{t \downarrow 0} R_\lambda'(t) = \lambda \quad \Rightarrow \quad L = 1.$$

(2) Minimax Concave Penalty (MCP)

$$R_\lambda(t) = \lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz$$



where $b > 0$ is fixed. Conditions (1)-(3) are easily seen to be satisfied.

Now, note that for $t > 0$,

$$\int_0^t \left(1 - \frac{z}{\lambda b}\right)_+ dz = \int_0^{\min\{t, \lambda b\}} \left(1 - \frac{z}{\lambda b}\right) dz$$

$$= \min\{t, \lambda b\} - \frac{1}{2\lambda b} \left(\min\{t, \lambda b\}\right)^2$$

By taking $\mu = \frac{1}{b}$, one can verify that $R_{\lambda,\mu}$ is convex (Exercise)

So condition (4) is satisfied. Lastly, we compute

$$R_\lambda'(t) = \lambda \cdot \mathrm{sgn}(t) \cdot \left(1 - \frac{|t|}{\lambda b}\right)_+ \quad \text{for } t \neq 0$$

and $\partial R_\lambda(0) = [-\lambda, \lambda]$ and $\lim_{t \downarrow 0} R_\lambda'(t) = \lambda$, so $L = 1$.

Hence, condition (5) is satisfied. Note that MCP is also piecewise quadratic.

Preliminaries: How R_λ interacts with the ℓ_1-norm.

Proposition 1: Under the assumptions on R_λ,

(i) R_λ is λL-Lipschitz

(ii) $\forall\, \theta \in \mathbb{R}^d$, $\quad \lambda L \|\theta\|_1 \le R_\lambda(\theta) + \frac{\mu}{2}\|\theta\|_2^2$

$\|\theta\|_1 = \sum_{i=1}^d |\theta_i|$ $\quad R_\lambda(\theta) = \sum_{i=1}^d R_\lambda(\theta_i)$ $\quad \|\theta\|_2^2 = \sum_{i=1}^d |\theta_i|^2$

Proof:

(i) Suppose $0 < t_1 < t_2$. Then,

$$\frac{R_\lambda(t_2)}{t_2} \le \frac{R_\lambda(t_1)}{t_1} \quad \Rightarrow \quad t_1 R_\lambda(t_2) \le t_2 R_\lambda(t_1)$$

(Condition (3))

$$\Rightarrow \quad \frac{R_\lambda(t_2) - R_\lambda(t_1)}{t_2 - t_1} \le \frac{R_\lambda(t_1)}{t_1}$$

$$\le \lim_{t \downarrow 0} \frac{R_\lambda(t)}{t}$$

$$= \lambda L \quad \text{(Condition (5))}$$

$$\Rightarrow \quad 0 \le R_\lambda(t_2) - R_\lambda(t_1) \le \lambda L (t_2 - t_1)$$

The other cases are similar.

(ii) It suffices to prove that for $t \in \mathbb{R}$,

$$\lambda L |t| \le R_\lambda(t) + \frac{\mu}{2} |t|^2.$$

by separability. This is trivial when $t = 0$. For $t > 0$, Since RHS is convex by condition (4), for any $s > 0$,

$$\left(R_\lambda(t) + \tfrac{\mu}{2} t^2\right) \ge \left(R_\lambda(s) + \tfrac{\mu}{2} s^2\right) + \left(R_\lambda'(s) + \mu s\right)(t - s)$$

Taking $s \downarrow 0$ yields the result. The case where $t < 0$ is similar.

Proposition 2: Under the assumptions on R_λ, let $\theta \in \mathbb{R}^d$ and S be the index set of the k largest elements of θ in magnitude. Then, for any $\xi > 0$,

(1) — $\xi R_\lambda(\theta_S) - R_\lambda(\theta_{S^c}) \le \lambda L \left(\xi \|\theta_S\|_1 - \|\theta_{S^c}\|_1 \right)$

$= \sum_{i \in S} R_\lambda(\theta_i)$

Moreover, if $\theta^* \in \mathbb{R}^d$ is k-sparse, then for any $\theta \in \mathbb{R}^d$,

(2) — $R_\lambda(\theta^*) - R_\lambda(\theta) \le \lambda L \left(\|\Delta_T\|_1 - \|\Delta_{T^c}\|_1 \right)$, $\Delta = \bigcup_{T, |T|=k}$

where $\Delta = \theta - \theta^*$ and T is the index set of the k largest elements in Δ in magnitude.

Proof: For (1), define $f(t) = \frac{t}{R_\lambda(t)}$ for $t > 0$ We compute

$$\|\theta_{S^c}\|_1 = \sum_{j \in S^c} |\theta_j| = \sum_{j \in S^c} R_\lambda(\theta_j) \cdot f(|\theta_j|)$$

$$\le \sum_{j \in S^c} R_\lambda(\theta_j) \cdot f(\|\theta_{S^c}\|_\infty) \quad (\because \text{condition (3)}, f \text{ non-decreasing})$$

$$= R_\lambda(\theta_{S^c}) \cdot f(\|\theta_{S^c}\|_\infty)$$

Also,

$$R_\lambda(\theta_S) \, f(\|\theta_{S^c}\|_\infty) = \sum_{j \in S} R_\lambda(\theta_j) \cdot f(\|\theta_{S^c}\|_\infty)$$

$$\le \sum_{j \in S} R_\lambda(\theta_j) \cdot f(|\theta_j|) \quad (\because S \text{ is the index set of the } k \text{ largest elements in } \theta)$$

$$= \|\theta_S\|_1$$

Hence,

$$\xi R_\lambda(\theta_S) - R_\lambda(\theta_{S^c}) \le \frac{1}{f(\|\theta_{S^c}\|_\infty)} \left[\xi \|\theta_S\|_1 - \|\theta_{S^c}\|_1 \right]$$

$$\le \left(\lim_{t \searrow 0} \frac{R_\lambda(t)}{t} \right) \left[\xi \|\theta_S\|_1 - \|\theta_{S^c}\|_1 \right] \quad (\text{condition (3)})$$

$$\le \lambda L \left[\xi \|\theta_S\|_1 - \|\theta_{S^c}\|_1 \right]. \quad (\text{condition (5)})$$

For (2), we first make a claim:

Claim: For any t_1, t_2,

$$R_\lambda(t_1 + t_2) \le R_\lambda(t_1) + R_\lambda(t_2)$$

With the claim, let $A = \text{supp}(\theta^*)$ and for any $\theta \in \mathbb{R}^d$,

$$R_\lambda(\theta^*) - R_\lambda(\theta) = R_\lambda(\theta^*_A) + \underbrace{R_\lambda(\theta^*_{A^c})}_{=0} - R_\lambda(\theta_A) - R_\lambda(\theta_{A^c})$$

$\Delta = \theta - \theta^*$

$$\le R_\lambda(\Delta_A) - R_\lambda(\theta_{A^c}) \qquad (\because R_\lambda(\theta^*_A) = R_\lambda(\theta_A - \Delta_A) \le R_\lambda(\theta_A) + R_\lambda(\Delta_A))$$

↳ claim

$$= R_\lambda(\Delta_A) - R_\lambda(\Delta_{A^c})$$

$$\le R_\lambda(\Delta_T) - R_\lambda(\Delta_{T^c}) \qquad \Delta = \boxed{}\ T$$

$$\le \lambda L\left(\|\Delta_T\|_1 - \|\Delta_{T^c}\|_1\right) \qquad (\because \text{result of first part})$$