Proof (Peeling Argument)

Since $g(r) \geq \mu$ for all $r \geq 0$, define, for $m = 1, 2, \cdots$

$$A_m = \{ v \in A : 2^{m-1}\mu \leq g(h(v)) \leq 2^m \mu \}$$

For a given X, if $v \in A$ is s.t. $f(v, X) \geq 2\, g(h(v))$, then $v \in A_m$ for some m. By the union bound,

$$\Pr[\mathcal{E}] \leq \sum_{m \geq 1} \Pr[\exists v \in A_m : f(v, X) \geq 2\, g(h(v))]$$

By definition, if $v \in A_m$ and $f(v, X) \geq 2\, g(h(v))$, then

$$f(v, X) \geq 2(2^{m-1}\mu) = 2^m \mu$$

↑ g is strictly increasing

Also, $g(h(v)) \leq 2^m \mu \iff h(v) \leq g^{-1}(2^m \mu)$.

Hence,

$$\Pr[\mathcal{E}] \leq \sum_{m \geq 1} \Pr\left[\sup_{\substack{h(v) \leq g^{-1}(2^m\mu) \\ v \in A}} f(v, X) \geq \underbrace{2^m \mu}_{g(g^{-1}(2^m\mu)) = g(r)} \right]$$

($\sup_{h(v) \leq g^{-1}(2^m\mu), v \in A}$: r)

$$\leq 2 \sum_{m \geq 1} \exp\left[-c \cdot a \cdot (g(g^{-1}(2^m \mu)))^2 \right]$$

$$= 2 \sum_{m \geq 1} \exp\left[-c \cdot a \cdot 2^{2m} \cdot \mu^2 \right]$$

$$\leq 2 \sum_{m \geq 1} \exp(-4c \cdot a \cdot \mu^2 \, m) \qquad \to \sum_{m \geq 1} \alpha^m$$

geometric sum

$$= \frac{2 \exp(-4c \cdot a \cdot \mu^2)}{1 - \exp(-4c \cdot a \cdot \mu^2)}$$

$S = \alpha + \alpha^2 + \alpha^3 + \cdots$

$-) \alpha S = \quad \alpha^2 + \alpha^3 + \cdots$

$(1-\alpha)S = \alpha$

$S = \frac{\alpha}{1-\alpha}$

Recall the regularized loss minimization problem:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta) + \lambda R(\theta) \right\} \qquad \text{— (*)}$$

Assumptions

(A1) $\mathcal{L}$ is convex and continuously differentiable

(A2) R is convex, $\lambda = 1$

Under (A1) and (A2), (*) is convex. The first-order optimality condition

generalized equation

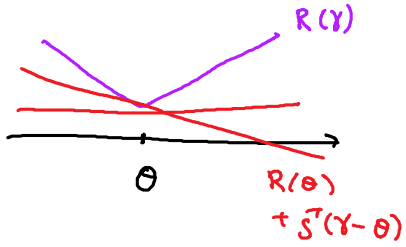
$$0 \in \nabla \mathcal{L}(\theta) + \partial R(\theta) \qquad \text{— (**)}$$

is necessary and sufficient for optimality. Here, recall that

$$\partial R(\theta) = \left\{ s \in \mathbb{R}^d : R(\gamma) \geq R(\theta) + s^T(\gamma - \theta) \ \ \forall \gamma \right\}$$

Subdifferential of R at θ

subgradient of R at θ



(Exercise) Suppose that R is a norm on $\mathbb{R}^d$.

Then, $\partial R(\theta) = \left\{ s \in \mathbb{R}^d : R^*(s) \leq 1, \ s^T\theta = R(\theta) \right\}$,

where $R^*(s) = \sup_{\theta : R(\theta) \leq 1} s^T\theta$ is the dual norm of R.

Representation of (**)

Define the proximal map associated with R by

$$\text{prox}_R(\theta) = \underset{\gamma \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\theta - \gamma\|_2^2 + R(\gamma) \right\} \qquad \text{— (†)}$$

$0 \in \gamma - \theta + \partial R(\gamma)$

Observe: $\text{prox}_R : \mathbb{R}^d \to \mathbb{R}^d$ and the output is uniquely

defined (fill in the details)

Claim: $\hat{\theta}$ is optimal for (*) iff

$$\hat{\theta} = \text{prox}_R\left(\underbrace{\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})}_{\text{one gradient step wrt } \mathcal{L}}\right) \qquad \text{— (Δ)}$$

(fixed-point equation)

This motivates the following method for solving (*):

$$\theta^{k+1} \leftarrow \text{prox}_{\alpha_k R}\left(\theta^k - \alpha_k \nabla\mathcal{L}(\theta^k)\right) \qquad \text{(proximal gradient method)}$$

where $\alpha_k > 0$ is the step size.

Proof of Claim: By considering the first-order optimality condition of (†), we see that for any $\theta \in \mathbb{R}^d$,

$$0 \in \text{prox}_R(\theta) - \theta + \partial R(\text{prox}_R(\theta))$$

Hence,

$\hat{\theta}$ is optimal for (*)

$$\iff 0 \in \nabla\mathcal{L}(\hat{\theta}) + \partial R(\hat{\theta}) \qquad \text{(from (**))}$$

$$= -\underbrace{(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta}))}_{\theta} + \underbrace{\hat{\theta}}_{\text{prox}_R(\theta)} + \underbrace{\partial R(\hat{\theta})}_{\text{prox}_R(\theta)}$$

$$\iff \hat{\theta} = \text{prox}_R(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta}))$$