Convex Learning Problems

Peng Lingwei

April 24, 2019

Contents

12 Convex Learning Problem

Convex learning problems can be learn efficiently. $0-1$ loss function is nonconvex function, and is computationally hard to learn in the unrealizable case.

12.1 CONVEXITY,LIPSCHITZNESS,AND SMOOTHNESS

12.1.1 Convexity

Definition 12.1. *(Convex Set). $\forall \mathbf{u}; \mathbf{v}$, then $\forall \lambda \in [0; 1]$, we have*

$$\lambda \mathbf{u} + (1 - \lambda)\mathbf{v} \in C.$$

Definition 12.2. *(Convex Function). Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if $\forall \mathbf{u}; \mathbf{v} \in C$ and $\lambda \in [0; 1]$,*

$$f(\lambda \mathbf{u} + (1 - \lambda)\mathbf{v}) \leq \lambda f(\mathbf{u}) + (1 - \lambda) f(\mathbf{v}).$$

For convex differentiable functions,

$$\forall f(\mathbf{u}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}); \mathbf{u} - \mathbf{w} \rangle.$$

Keep Convexity:

- $g(x) = \max_{i \in [r]} f_i(x)$
- $g(x) = \sum_{i=1}^{r} w_i f_i(x)$, where for all $i; w_i \geq 0$

12.1.2 Lipschitzness

Definition 12.3. *(Lipschitzness).Let $C \subset \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is $\rho - Lipschitz$ over C if $\forall w_1, w_2 \in C$, we have*

$$\|f(w_1) - f(w_2)\| \leq \rho \|w_1 - w_2\|.$$

Intuitively, Lipschitzness constrains $f'(u)$.

Let $f(x) = g_1(g_2(x))$, where g_1 is $\rho_1 - Lipschitz$ and g_2 is $\rho_2 - Lipschitz$. Then, f is $(\rho_1 \rho_2) - Lipschitz$.

12.1.3 Smoothness

Definition 12.4. *(Smoothness).A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ at w is $\beta - smooth$ if its gradient is $\beta - Lipschitz$; namely, $\forall v, w$ we have*

$$\|\nabla f(v) - \nabla f(w)\| \leq \beta \|v - w\|.$$

$\beta - Smoothness$ implies that

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2}\|v - w\|^2.$$

Setting $v = w - \frac{1}{\beta}\nabla f(w)$, we have

$$\frac{1}{2\beta}\|\nabla f(w)\|^2 \leq f(w) - f(v).$$

If we assume that $\forall v, f(v) \geq 0$, we conclude that smoothness implies *self-bounded*:

$$\|\nabla f(w)\|^2 \leq 2\beta f(w).$$

Let $f(w) = g(\langle w, x \rangle + b)$, where $g \to \mathbb{R} \to \mathbb{R}$ is a $\beta - smooth$ function, then f is $(\beta \|x\|^2) - smooth$.

12.2 CONVEX LEARNING PROBLEMS

Symbols: a hypothesis classes set $\mathcal{H}$, a set of examples Z, and a loss function $l : \mathcal{H} \times Z \to \mathbb{R}_+$

$\mathcal{H}$ can be an arbitrary set. In this chapter, we consider hypothesis classes set $\mathcal{H} = \mathbb{R}^d$.

Definition 12.5. *(Convex Learning Problem). A learning problem, $(\mathcal{H}, Z, l)$, is called convex if the hypothesis class $\mathcal{H}$ is convex set and $\forall z \in Z$, the loss function, $l(\cdot, z)$, is a convex function (which means $f : \mathcal{H} \to \mathbb{R}, f(w) = l(w, z)$).*

Lemma 12.1. *If l is a convex loss function and the class $\mathcal{H}$ is convex, then the $ERM_{\mathcal{H}}$ problem, of minimizing the empirical loss over $\mathcal{H}$, is a convex optimization problem.*

12.2.1 Learnability of Convex Learning Problems

Is convexity a su cient condition for the learnability of a problem? The answer is NO.

Example 12.1. *(Nonlearnability of Linear Regression Even If $d = 1$). Let $\mathcal{H} = \mathbb{R}$, and the loss be the squared loss: $l(w; (x; y)) = (wx - y)^2$. We assume A is a successful PAC learner for this problem.*

Choose $= 1{=}100$, $= 1{=}2$, *let* $m \geq m(\; ; \;)$ *and set* $= \frac{\ln(100{=}99)}{2m}$. *We get two points* $z_1 = (1; 0)$ *and* $z_2 = (\; ; -1)$, *then we construct two distributions:* $\mathcal{D}_1 = \{(z_1; \;); (z_2; 1 - \;)\}$, *and* $\mathcal{D}_2 = \{(z_2; 1)\}$

The probability that all examples of the training set will be z_2 *is at least 99%.*

$$(1 - \;)^m \geq e^{-2 \; m} = 0{:}99{:}$$

If $\hat{w} < -1{=}(2 \;)$, *then* $L_{\mathcal{D}_1}(\hat{w}) = \; (\hat{w})^2 + (1 - \;)(\hat{w} \; + 1)^2 \geq \; (\hat{w})^2 \geq 1{=}(4 \;)$. *However,* $\min_w L_{\mathcal{D}_1}(w) \leq L_{\mathcal{D}_\infty(0)} = (1 - \;)$, *it follows that,* $L_{\mathcal{D}_\infty}(\hat{w}) - \min_w L_{\mathcal{D}_1}(w) \geq \frac{1}{4} - (1 - \;) > \;$. *(* $< 0{:}0051$, *which means* $1{=}(4 \;) - (1 - \;) > 48 \gg \;$*).*

If $\hat{w} \geq -1{=}(2 \;)$, *then* $L_{\mathcal{D}_2} = (\hat{w} \; + 1)^2 \geq 1{=}4 > \;$.

All in all, the problem is not PAC learnable.

In addition to the convexity requirement, we also need $\mathcal{H}$ will be bounded. But the above example is still not PAC learnble. This motivate a de nition of two families of learning problems, convex-Lipschitz-bounded and convex-smooth-bounded.

12.2.2 Convex-Lipschitz/Smooth-Bounded Learning Problems

De nition 12.6. *(Convex-Lipschitz-Bounded Learning Problem). A learning problem,* $(\mathcal{H}; Z; l)$, *is called Convex-Lipschitz-Bounded, with parameters* $\; ; B$ *if:*

- *The hypothesis class $\mathcal{H}$ is a convex set and bounded (parameter is B).*
- *For all $z \in Z$, the loss function, $l(\cdot; z)$, is a convex and* $-$ Lipschitz *function.*

Example 12.2. *Let* $\mathcal{X} = \{x \in \mathbb{R}^d : \|\mathrm{x}\| \leq \; \}$ *and* $\mathcal{Y} = \mathbb{R}$. *Let* $\mathcal{H} = \{\mathrm{w} \in \mathbb{R}^d : \|\mathrm{w}\| \leq B\}$ *and let the loss function be* $l(\mathrm{w}; (\mathrm{x}; y)) = |\langle \mathrm{w}; \mathrm{x} \rangle - y|$.

Proof. $|l(\mathrm{w}_1; (\mathrm{x}; y)) - l(\mathrm{w}_2; (\mathrm{x}; y))| \leq |\langle \mathrm{w}_1 - \mathrm{w}_2; \mathrm{x} \rangle| \leq \|\mathrm{x}\| \cdot \|\mathrm{w}_1 - \mathrm{w}_2\|$ □

De nition 12.7. *(Convex-Smooth-Bounded Learning Problem). A learning problem,* $(\mathcal{H}; Z; l)$, *is called Convex-Smooth-Bounded, with parameters* $\; ; B$ *if:*

- *The hypothesis class $\mathcal{H}$ is a convex set and bounded (parameter is B).*
- *For all $z \in Z$, the loss function, $l(\cdot; z)$, is a convex and* $-$ Smooth *function.*

Example 12.3. *Let* $\mathcal{X} = \{x \in \mathbb{R}^d : \|\mathrm{x}\|^2 \leq \; {=}2\}$ *and* $\mathcal{Y} = \mathbb{R}$. *Let* $\mathcal{H} = \{\mathrm{w} \in \mathbb{R}^d : \|\mathrm{w}\| \leq B\}$ *and let the loss function be* $l(\mathrm{w}; (\mathrm{x}; y)) = (\langle \mathrm{w}; \mathrm{x} \rangle - y)^2$.

Proof. $\|\nabla l(\mathrm{w}_1; (\mathrm{x}; y)) - \nabla l(\mathrm{w}_2; (\mathrm{x}; y))\| = 2\|\mathrm{x}\langle \mathrm{w}_1 - \mathrm{w}_2; \mathrm{x} \rangle\| = 2\|\mathrm{x}\|^2\|\mathrm{w}_1 - \mathrm{w}_2\|$ □

12.3 SURROGATE LOSS FUNCTIONS

The $0-1$ loss function is not convex.

$$l^{0-1}(\mathbf{w}; (\mathbf{x}; y)) = \mathbf{1}\{y\langle \mathbf{w}; \mathbf{x}\rangle \leq 0\}:$$

Proof. Let $\mathcal{H}$ be the class of homogeneous halfspaces in $\mathbb{R}^d$. Let $\mathbf{x} = \mathbf{e}_1; y = 1$, and consider the sample $S = \{\mathbf{x}; y\}$. Let $\mathbf{w} = -\mathbf{e}_1$. Then, $\langle \mathbf{w}; \mathbf{x}\rangle = -1$ and $L_S(h_{\mathbf{w}}) = 1$. Let $\mathbf{w}'; s:t: \in (0; 1) and \|\mathbf{w}' - \mathbf{w}\| \leq epsilon$. Then,$\langle \mathbf{w}'; \mathbf{x}\rangle = \langle \mathbf{w}; \mathbf{x}\rangle - \langle \mathbf{w}' - \mathbf{w}; x\rangle = -1 - \langle \mathbf{w}' - \mathbf{w}; x\rangle \leq -1 + \ \|\mathbf{x}\| \leq -1 + \ < 0$, which means $L_S(\mathbf{w}') = 1$. □

The requirements from a convex surrogate loss are as follows:

- It should be convex.
- It should be upper bound th original loss.

De nition 12.8. *(hinge loss).*

$$l^{hinge}(\mathbf{w}; (\mathbf{x}; y)) = \max\{0; 1 - y\langle \mathbf{w}; \mathbf{x}\rangle\}:$$

Then, we have:

$$L_{\mathcal{D}}^{hinge}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{hinge}(\mathbf{w}) + \ :$$

$$L_{\mathcal{D}}^{0-1}(A(S)) \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{hinge}(\mathbf{w}) + \ :$$

We can further rewrite the upper bound as follows:

$$L_{\mathcal{D}}^{0-1} \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) + \quad \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{hinge}(\mathbf{w}) - \min_{mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) \quad + \ :$$

The $0-1$ error of the learned predictor is upper bounded by three terms:

- Approximation error: the rst term.
- Optimization error: the second term.
- Estimation error: the third term.