Convex Learning Problems

Peng Lingwei

July 19, 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* $\forall \alpha \in [0,1]$*, we have*

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

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* $\alpha \in [0,1]$*,*

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

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 $\forall \mathbf{w}_1, \mathbf{w}_2 \in C$, we have*

$$\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho\|\mathbf{w}_1 - \mathbf{w}_2\|.$$

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

Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, 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 $\mathbf{w}$ is $\beta - smooth$ if its gradient is $\beta - Lipschitz$; namely, $\forall \mathbf{v}, \mathbf{w}$ we have*

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

$\beta - Smoothness$ implies that

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

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

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

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

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

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

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 $f : \mathcal{H} \to \mathbb{R}, f(\mathbf{w}) = l(\mathbf{w}, z)$).*

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

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

$$(1 - \mu)^m \geq e^{-2\mu m} = 0.99.$$

If $\hat{w} < -1/(2\mu)$, then $L_{\mathcal{D}_1}(\hat{w}) = \mu(\hat{w})^2 + (1 - \mu)(\hat{w}\mu + 1)^2 \geq \mu(\hat{w})^2 \geq 1/(4\mu)$. However, $\min_w L_{\mathcal{D}_1}(w) \leq L_{\mathcal{D}_\infty(0)} = (1 - \mu)$, it follows that, $L_{\mathcal{D}_\infty}(\hat{w}) - \min_w L_{\mathcal{D}_1}(w) \geq \frac{1}{4\mu} - (1 - \mu) > \epsilon$. ($\mu < 0.0051$, which means $1/(4\mu) - (1 - \mu) > 48 \gg \epsilon$).

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

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 definition of two families of learning problems, convex-Lipschitz-bounded and convex-smooth-bounded.

12.2.2 Convex-Lipschitz/Smooth-Bounded Learning Problems

Definition 12.6. *(Convex-Lipschitz-Bounded Learning Problem). A learning problem, $(\mathcal{H}, Z, l)$, is called Convex-Lipschitz-Bounded, with parameters ρ, 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 $\rho -$ Lipschitz function.*

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

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

Definition 12.7. *(Convex-Smooth-Bounded Learning Problem). A learning problem, $(\mathcal{H}, Z, l)$, is called Convex-Smooth-Bounded, with parameters β, 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,nonnegative and $\beta -$ Smooth function.*

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

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

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 $\mathbf{w}', s.t. \epsilon \in (0,1) and \|\mathbf{w}' - \mathbf{w}\| \leq epsilon$. Then,$\langle \mathbf{w}',\mathbf{x}\rangle = \langle \mathbf{w},\mathbf{x}\rangle - \langle \mathbf{w}'-\mathbf{w}, x\rangle = -1 - \langle \mathbf{w}'-\mathbf{w}, x\rangle \leq -1+\epsilon\|\mathbf{x}\| \leq -1+\epsilon < 0$, 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.

Definition 12.8. *(hinge loss).*

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

Then, we have:

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

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

We can further rewrite the upper bound as follows:

$$L_{\mathcal{D}}^{0-1} \leq \min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w}) + \left(\min_{\mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{hinge}(\mathbf{w}) - \min_{mathbf{w}\in\mathcal{H}} L_{\mathcal{D}}^{0-1}(\mathbf{w})\right) + \epsilon.$$

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

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