Learning Algorithms in Depth: Stability

Ch. 13 (read Ch. 3 on optimization)

Recap: data $z^n \overset{iid}{\sim} P$ hypotheses $f \in \mathcal{F}$

e.g. ERM $\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} f(z_i)$

$$L(\hat{f}_n) = \int_{\mathcal{Z}} \hat{f}_n(z) P(dz)$$

$$L^* = \min_{f \in \mathcal{F}} L(f)$$

(universal) "consistency":

$$\sup_P \{ L(\hat{f}_n) - L^* \} \to 0 \quad \text{as } n \to \infty$$

Sufficient condition for consistency: UCEM

$$P_n(f) := \frac{1}{n} \sum_{i=1}^{n} f(z_i) = E_{P_n}[f] \quad \text{(empirical risk)}$$

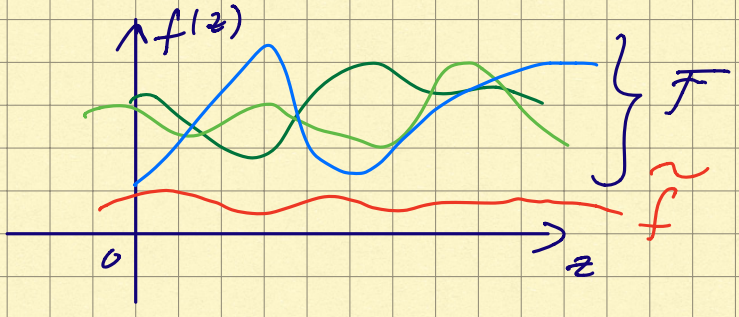
$$P(f) = E_P[f]$$

$$\sup_P E_P \sup_{f \in \mathcal{F}} | P_n(f) - P(f) | \xrightarrow{n \to \infty} 0$$

UCEM (property of $\mathcal{F}$) $\Longrightarrow$ ERM is consistent

But: learnability is possible w/o UCEM!

- $\mathcal{F}$ given no assumptions
- assume $\exists\ \tilde{f}$ s.t. $\tilde{f}(z) < \inf_{f \in \mathcal{F}} f(z) \quad \forall z$



$\tilde{\mathcal{F}} := \mathcal{F} \cup \{\tilde{f}\}$

ERM on $\tilde{\mathcal{F}}$ will always return $\tilde{f}$!

Closer look at learning algos

Framework (Vapnik, 1995)

- $\mathcal{Z}$: instance space
- $\mathcal{P}$: class of prob. dist. on $\mathcal{Z}$
- $\mathcal{F}$: closed, convex subset of a Hilbert space $\mathcal{H}$
- $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$: loss fcn
- $L_P(f) := \mathbb{E}_P[\ell(f, Z)]$

$$= \int_{\mathcal{Z}} \ell(f, z) P(dz)$$

Examples: 1) $\mathcal{H}$: H.S. of fcns $\mathcal{Z} \to \mathbb{R}$

$\ell(f, z) := f(z)$

2) $\mathcal{F} = \mathcal{H}_K$ (RKHS for some K)

$z = (x, y) \in \mathcal{X} \times \{-1, 1\}$

$\ell(f, z) = \ell(f, (x, y)) = \varphi(-y f(x))$

↳ pen. fcn.

3) $\ell(f, (x, y)) = (y - f(x))^2$

...

Main idea: want to bring out the dependence of $\ell(\cdot, \cdot)$ on both f and z

- Learning algos:

$A_n : \mathcal{Z}^n \to \mathcal{F}$ (one for each n)

$\mathcal{Z}^* := \bigcup_{n \geq 1} \mathcal{Z}^n$: all tuples over $\mathcal{Z}$

$A : \mathcal{Z}^* \to \mathcal{F}$

$A(z^n)$: random element of $\mathcal{F}$

- Risks:

$$L_P(A(z^n)) = \int_{\mathcal{Z}} \ell(A(z^n), z) \, P(dz)$$

$$L_n(A(z^n)) = \frac{1}{n} \sum_{i=1}^{n} \ell(A(z^n), z_i)$$

Motivating Example: ERM w/ Strongly Convex Losses (Ch. 3)

$(f, z) \mapsto \ell(f, z)$

$f \in \mathcal{F}$: closed, convex subset of $(\mathcal{H}, \|\cdot\|)$

Assume:

1) $\ell(f, z)$ is L-Lipschitz in $f \in \mathcal{F}$, unif. in z:

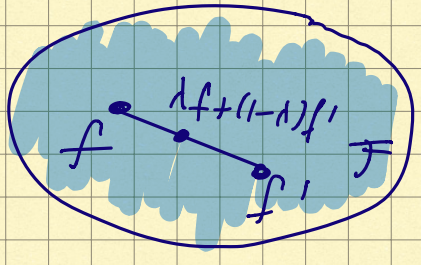
$$\sup_{z \in \mathcal{Z}} |\ell(f, z) - \ell(f', z)| \le L \|f - f'\|$$

2) $\ell(f, z)$ is m-strongly convex ($m > 0$) in $f \in \mathcal{F}$, unif. in z

Review: $\varphi : \mathcal{F} \to \mathbb{R}$

$\mathcal{F}$ - convex set $\quad f, f' \in \mathcal{F}$

$\Rightarrow \lambda f + (1-\lambda) f' \in \mathcal{F} \quad \forall \lambda \in [0,1]$



$\mathcal{F}$: closed $\quad f_n \in \mathcal{F}$, $f_n \to f \Rightarrow f \in \mathcal{F}$

$\varphi : \mathcal{F} \to \mathbb{R}$ is convex if

$$\varphi(\lambda f + (1-\lambda) f') \le \lambda \varphi(f) + (1-\lambda) \varphi(f')$$

for all $f, f' \in \mathcal{F}$, $\lambda \in [0,1]$

$\varphi : \mathcal{F} \to \mathbb{R}$ is m-strongly convex ($m \ge 0$)

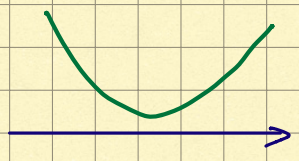
if $\tilde{\varphi}(f) := \varphi(f) - \frac{m}{2}\|f\|^2$ is convex

[note: not requiring φ to be differentiable]

Consequences of strong convexity:

- $\varphi(\lambda f + (1-\lambda) f') \le \lambda \varphi(f) + (1-\lambda)\varphi(f')$ (conv.)
  $\quad - \frac{m}{2}\lambda(1-\lambda)\|f - f'\|^2$ (A)

- unique minimizers



- if $\varphi(f^*) = \min_{f \in \mathcal{F}} \varphi(f)$, then

$$\varphi(f) \ge \varphi(f^*) + \frac{m}{2}\|f - f^*\|^2$$

(can be derived from (A))

- let $B(f)$ be an L-Lip. fcn, let

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \varphi(f)$$

$$\bar{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \{\varphi(f) + \overbrace{B(f)}^{\text{Lip. perturbation}}\}$$

Then $\|\bar{f} - f^*\| \le \frac{L}{m}$

(stability of minimizers under Lip. perturbations).

- if $\varphi(\cdot)$ is convex, then $\varphi(\cdot) + \frac{m}{2}\|\cdot\|^2$ is m-strongly convex

Back to learning:

$\ell(f, z)$: L-Lip., m-str. cvx in $f \in \mathcal{F}$ (uniformly in z)

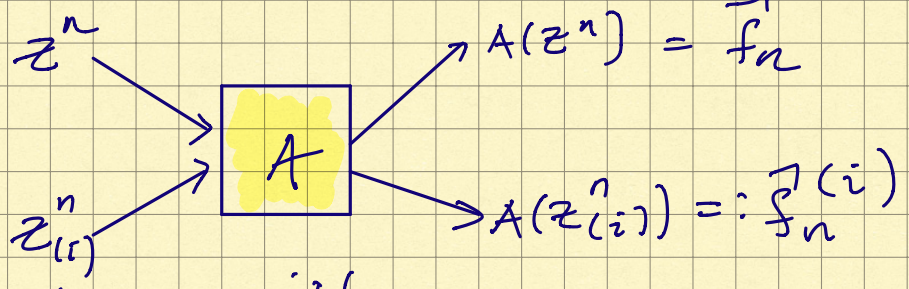
ERM: $\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i)$

Thm With pr. $\geq 1-\delta$,

$$L(\hat{f}_n) - \inf_{f \in \mathcal{F}} L(f) \leq \frac{2L^2}{\delta m n}.$$

Note: $\frac{1}{\delta}$ dependence instead of $\log(\frac{1}{\delta})$

Proof



$z'_1, z'_2, \ldots, z'_n \overset{iid}{\sim} P$, indep. of z^n

$z'^n_{(i)} := (z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)$

$z^n = (z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n)$

- Fix $f \in \mathcal{F}$

$$L_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i) \qquad \text{[loss on } z^n\text{]}$$

$$L_n^{(i)}(f) := \frac{1}{n}\ell(f, z'_i) + \frac{1}{n}\sum_{j: j \neq i} \ell(f, z_i) \qquad \text{[loss on } z'^n_{(i)}\text{]}$$

$$= \frac{1}{n}\ell(f, z'_i) + L_n(f) - \frac{1}{n}\ell(f, z_i)$$

$$= L_n(f) + \frac{1}{n}\left[\ell(f, z'_i) - \ell(f, z_i)\right]$$

$\therefore$ for each $i \in [n]$:

$$L_n^{(i)}(f) - L_n(f) = \frac{1}{n}\left[\ell(f, z'_i) - \ell(f, z_i)\right]$$

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} L_n(f)$$

$$\hat{f}_n^{(i)} = \operatorname*{argmin}_{f \in \mathcal{F}} L_n^{(i)}(f)$$

Claim: $\|\hat{f}_n - \hat{f}_n^{(i)}\| \le \frac{2L}{mn}$

Proof (of claim)

1) $f \mapsto L_n(f)$ is m-str. cvx. (b/c ℓ is)

2) $f \mapsto \frac{1}{n}(\ell(f, z_i') - \ell(f, z_i))$ is $\frac{2L}{n}$-Lip.

$$B(f) := \frac{1}{n}(\ell(f, z_i') - \ell(f, z_i))$$

$$|B(f) - B(f')| \le \frac{1}{n}|\ell(f, z_i') - \ell(f', z_i')| + \frac{1}{n}|\ell(f, z_i) - \ell(f', z_i)|$$

$$\le \frac{2L}{n}\|f - f'\| \qquad \text{(since } \ell \text{ is } L\text{-Lip. in } f\text{)}$$

$$\Rightarrow L_n^{(i)}(f) = \underbrace{L_n(f)}_{m\text{-s.c.}} + \underbrace{B(f)}_{\frac{2L}{n}\text{-Lip.}}$$

By stab. of minimizers, $\|\hat{f}_n - \hat{f}_n^{(i)}\| \le \frac{2L}{nm}$. ∎

- **Claim:** $\mathbb{E}[L(\hat{f}_n) - L_n(\hat{f}_n)] \le \frac{2L^2}{mn}$.

Proof (of claim) $\hat{f}_n := A(Z^n)$

$$\mathbb{E}[L(\hat{f}_n)] = \mathbb{E}[L(A(Z^n))]$$
$$= \mathbb{E}[\ell(A(Z^n), Z_i')] \quad \forall i$$
$$Z^n \perp\!\!\!\perp Z_i'$$

$$\Rightarrow \mathbb{E}[L(\hat{f}_n)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(A(Z^n), Z_i')]$$

$$\mathbb{E}[L_n(\hat{f}_n)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(A(Z^n), Z_i)]$$

$$\forall i \in [n]: \quad (A(Z^n), Z_i) \overset{d}{=} (A(Z^n_{(i)}), Z_i')$$

$$\Rightarrow \mathbb{E}[L_n(\hat{f}_n)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(A(Z^n_{(i)}), Z_i')]$$

$$\therefore \mathbb{E}[L(\hat{f}_n) - L_n(\hat{f}_n)]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(\hat{f}_n, Z_i') - \ell(\hat{f}_n^{(i)}, Z_i')]$$
$$\le \frac{1}{n}\sum_{i=1}^{n} L \cdot \mathbb{E}\|\hat{f}_n - \hat{f}_n^{(i)}\|$$
$$\le \frac{2L^2}{nm}. \qquad \square$$

- $\mathbb{E}[L(\hat{f}_n) - L(f^*)]$ $\qquad L(f^*) \ge \min_{f \in \mathcal{F}} L(f)$

$$= \mathbb{E}[L(\hat{f}_n) - L_n(\hat{f}_n)]$$
$$+ \underbrace{\mathbb{E}[L_n(\hat{f}_n) - L_n(f^*)]}_{\le 0} + \underbrace{\mathbb{E}[L_n(f^*) - L(f^*)]}_{=0}$$
$$\le \frac{2L^2}{nm}.$$

- By Markov's inequality,

$$\mathbb{P}\{L(\hat{f}_n) - L(f^*) \geq t\} \leq \frac{\mathbb{E}[L(\hat{f}_n) - L(f^*)]}{t}$$

$$\leq \frac{2L^2}{nmt}.$$

Let $\frac{2L^2}{nmt} = \delta \Rightarrow t = \frac{2L^2}{\delta nm}$. ∎

Key Points:

- $\| A(z^n) - A(z^n_{(i)}) \| \leq \frac{2L}{nm} \quad \forall i$
  - hypothesis stability

- $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell(A(z^n), z_i') - \ell(A(z^n_{(i)}), z_i')\right)\right]$

$$\leq \frac{2L^2}{nm}$$

  - replace-one stability.