# Stability and Generalization (cont.)

## Recap:

data $Z_1, Z_2, \ldots \overset{iid}{\sim} P$ $\quad P \in \mathcal{P}$, on $\mathcal{Z}$

$A : Z^n \mapsto A(Z^n) \in \mathcal{F}$

so:
$A : \mathcal{Z}^* \to \mathcal{F}$ $\quad \mathcal{Z}^* := \bigcup_{n \geq 1} \mathcal{Z}^n$

loss: $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$

$\mathcal{F}$ : closed, convex subset of H.S. $\mathcal{Y}$


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

r.v.!

$$= \mathbb{E}[\ell(A(Z^n), Z_{n+1}) \mid Z^n]$$

$$\mathbb{E}[L(A(Z^n))] = \mathbb{E}[\ell(A(Z^n), Z_{n+1})]$$

$$L_n(A(Z^n)) = \frac{1}{n} \sum_{i=1}^{n} \ell(A(Z^n), Z_i) \quad \text{(empirical risk)}$$

**Note:** $\mathbb{E}[L_n(A(Z^n))] \neq \mathbb{E}[L(A(Z^n))]$

## Generalization gap of $A$:

$$g_n(A) := \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ L_n(A(Z^n)) - L(A(Z^n)) \right]$$

A generalizes if $\lim_{n\to\infty} g_n(A) = 0$

(— as sample gets large enough, $L_n(A(Z^n))$ is a good proxy for $L(A(Z^n))$ ).

$Z^n \longrightarrow \boxed{A} \longrightarrow A(Z^n)$

$(Z_{n+1}, \dots, Z_{2n}) : \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(A(Z^n), Z_{n+i}) \,\middle|\, Z^n\right] = L(A(Z^n))$

**Consistency:** $L^* := \inf_{f\in\mathcal{F}} L(f)$

$= \inf_{f\in\mathcal{F}} \mathbb{E}_P[\ell(f,Z)] = L(f^*)$

A is consistent if

$$c_n(A) := \sup_{P\in\mathcal{P}} \mathbb{E}\underbrace{[L(A(Z^n)) - L^*]}_{\geq 0} \longrightarrow 0$$

as $n\to\infty$

---

**Asymptotically ERM algorithms:**

$L_n^* := \inf_{f\in\mathcal{F}} L_n(f) = \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f, Z_i)$

AERM: $\ell_n(A) := \sup_{P\in\mathcal{P}} \mathbb{E}[L_n(A(Z^n)) - L_n^*] \to 0$

as $n\to\infty$.

ERM $\Rightarrow$ AERM (trivially)

## Generalization on average:

$$\bar{g}_n(A) := \sup_{P \in \mathcal{P}} \left| \mathbb{E}\left[ L_n(A(Z^n)) - L(A(Z^n)) \right] \right|$$

$A$ generalizes on avg if $\bar{g}_n(A) \to 0$ as $n \to \infty$

$$\bar{g}_n(A) \le g_n(A) \quad \forall n$$

**Thm** For any $A$,

$$c_n(A) \le \underbrace{e_n(A)}_{\to 0 \text{ (AERM)}} + \underbrace{\bar{g}_n(A)}_{\to 0 \text{ (gen on avg)}}$$

$\Rightarrow$ AERM + (gen on avg) $\Rightarrow$ Consistent.

**Proof**

$$\begin{aligned}
&\mathbb{E}\, L(A(Z^n)) \\
&= \mathbb{E}\left[ L(A(Z^n)) - L_n(A(Z^n)) \right] + \mathbb{E}\left[ L_n(A(Z^n)) \right] \\
&\le \bar{g}_n(A) + \mathbb{E}\left[ L_n(A(Z^n)) \right] \\
&= \bar{g}_n(A) + \mathbb{E}\left[ L_n(A(Z^n)) - L_n^* \right] + \mathbb{E}\left[ L_n^* \right] \\
&\le \bar{g}_n(A) + e_n(A) + \mathbb{E}\left[ L_n^* \right]
\end{aligned}$$

$\Rightarrow$ $\mathbb{E}[L(A(Z^n))] \le \mathbb{E}[L_n^*] + \bar{g}_n(A) + e_n(A)$

where $\mathbb{E}[L_n^*] = \mathbb{E}[\underbrace{L_n^* - L_n(f^*)}_{\le 0}] + \mathbb{E}[\underbrace{L_n(f^*)}_{= L(f^*)}]$

where $f^*$ achieves $L^*$

$\therefore$ $\mathbb{E}[L_n(A(Z^n)) - L^*] \le e_n(A) + \bar{g}_n(A)$. ∎

AERM + (gen on avg) $\Rightarrow$ consistency

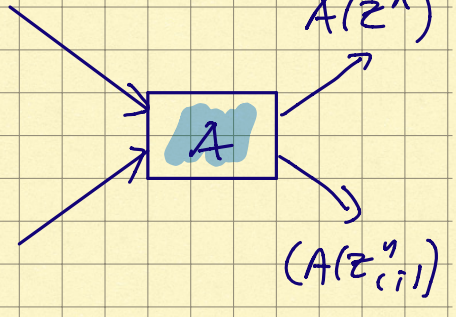
gen on avg.: $\bar{g}_n(A) = \sup_{P \in \mathcal{P}} \left| \mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))] \right|$

## Stability

$Z^n = (Z_1, \dots, Z_{i-1}, Z_i, Z_{i+1}, \dots, Z_n)$

$Z'^n = (Z_1', \dots, Z_{i-1}', Z_i', Z_{i+1}', \dots, Z_n')$

$Z^n_{(i)} = (Z_1, \dots, Z_{i-1}, Z_i', Z_{i+1}, \dots, Z_n)$




Stability: $A(Z^n)$ "$\approx$" $A(Z^n_{(i)})$ for each $i$

$\mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))]$

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

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

- for each $i \in [n]$, $(A(Z^n), Z_i) \overset{d}{=} (A(Z^n_{(i)}), Z_i')$

$$\Rightarrow \mathbb{E}[\ell(A(Z^n), Z_i)] = \mathbb{E}[\ell(A(Z^n_{(i)}), Z_i')] \quad !$$

So

$$\mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))]$$

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

- if $A(Z^n_{(i)})$ "$\approx$" $A(Z^n)$ $\forall i$, then

  $\mathbb{E}[\ell(A(Z^n_{(i)}), Z_i') - \ell(A(Z^n), Z_i')]$ is "small".

e.g. $O\left(\frac{1}{\sqrt{n}}\right)$

$$\Rightarrow \mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))] = O\left(\frac{1}{\sqrt{n}}\right)$$

- Stability $\leftrightarrow$ also not too "sensitive" to modifications of individual samples

$I \sim \text{Unif}([n])$, $I \perp (Z^n, Z'^n)$

$$\frac{1}{n} \sum_{i=1}^{n} \left[\ell(A(Z^n_{(i)}), Z_i') - \ell(A(Z^n), Z_i')\right]$$

$$= \mathbb{E}\left[\ell(A(Z^n_{(I)}), Z_I') - \ell(A(Z^n), Z_I') \,\middle|\, Z^n, Z'^n\right]$$

Overall,

$$\mathbb{E}\left[L_n(A(Z^n)) - L(A(Z^n))\right]$$

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

## Stability on Avg.:

$A$ is stable on avg if

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

$\to 0$ as $n \to \infty$

## Key points:

1) $\overline{gen}_n(A) = s_n(A)$ (gen on avg $\Leftrightarrow$ stab. on avg)

2) in many cases, $s_n(A)$ can be easily estimated

- sufficient condition: uniform stability

$$\sup_{z \in \mathcal{Z}} \left|\ell(A(Z^n), z) - \ell(A(Z^n_{(i)}), z)\right| = o(n)$$

3) For any $A$, $\epsilon_n(A) \le \underbrace{\text{en}(A)}_{\to 0 \text{ AERM}} + \underbrace{s_n(A)}_{\to 0 \text{ stable on avg}}$

**Thm** Suppose $0 \le \ell(f, z) \le 1$ for all $f \in \mathcal{F}, z \in \mathcal{Z}$.

Then, for any $A$,

$$g_n(A) \le \bar{g}_n(A) + 2\left(\varepsilon_n(A) + \frac{1}{\sqrt{n}}\right).$$

**Remark:**

$$\bar{g}_n(A) = \sup_P \left| \mathbb{E} L_n(A(Z^n)) - \mathbb{E} L(A(Z^n)) \right|$$

$$g_n(A) = \sup_P \mathbb{E} \left| L_n(A(Z^n)) - L(A(Z^n)) \right|$$

$$\bar{g}_n(A) \le g_n(A)$$

— AERM + (gen on avg) $\Rightarrow$ gen.

---

## Stochastic Gradient Descent (SGD)

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

- $f \in \mathcal{F}$, closed convex subset of a Hilbert space $\mathcal{H}$
- $f \mapsto \ell(f, z)$ is differentiable for each $z \in \mathcal{Z}$

  $\nabla \ell(f, z)$ : gradient (Fréchet derivative)

$$\ell(f', z) = \ell(f, z) + \underbrace{\nabla \ell(f, z)(z' - z)}_{\text{linear map } \mathcal{H} \to \mathbb{R}} + \underbrace{o(\|z' - z\|)}_{\text{error} \to 0 \text{ as } z' \to z}$$

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

alternative (approx. ERM): GD

$f_0 \in \mathcal{F}$ (init)

at $t = 1, 2, \ldots$

$$f_t = \Pi\left(f_{t-1} - \alpha_t \nabla L_n(f_t)\right)$$

where $\Pi$ is the projection onto $\mathcal{F}$:

$$\Pi(g) := \operatorname{argmin}_{f \in \mathcal{F}} \|g - f\|$$

$\alpha_t > 0$ - stepsize

$t = 1, \ldots, T$: $\hat{f}_n = f_T$

bound $L_n(f_T) - L_n^*$ (AERM)

<u>However:</u> computation of $\nabla L_n(f)$ has $O(n)$ complexity

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

SGD: $L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell_i(f)$ $\quad \ell_i(f) := \ell(f, z_i)$

$I_1, I_2, I_3, \ldots I_t, \ldots$: stochastic process in $[n]$ indep. of $Z^n$

$$f_t = \Pi\left(f_{t-1} - \alpha_t \nabla \ell_{I_t}(f_{t-1})\right)$$

each step has now $O(1)$ complexity

## Examples of $(I_t)$:

- sampling w/ replacement: $I_t \overset{iid}{\sim} \text{Unif}([n])$
- cyclic coordinate descent:
  $\pi(1), \dots, \pi(n)$ : random perm of $[n]$

$$(I_1, I_2, \dots, I_n, I_{n+1}, \dots)$$
$$= (\pi(1), \pi(2), \dots, \pi(n), \pi(1), \pi(2), \dots, \pi(n), \dots)$$

- new perm. every $n$ rounds

**Next lecture:** SGD is stable (under regularity assumptions)