22. Stability and Generalization

Recap

- **So far.** Generalization bounds via the richness of the hypothesis space
    - Learning algorithm: ERM
        - Finding the minimum-risk hypothesis inside a bag of functions
        - Optimization aspect only indirectly affects the bound
            - constraining the hypothesis space (e.g., via norm control)
        - Relies on the uniform convergence of empirical mean

$$\sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)| \to 0$$

- **This week.** Generalization bounds via the stability of the algorithm
    - More recent-ish (2002 & after)
    - Direct reference to the learning algorithm itself
    - Does not rely on UCEM

Learning algorithm

- We formalize the learning algorithm as follows

Definition (Learning algorithm).

Given the training data $Z_i \in \mathcal{Z}, i \in \{1,2,\ldots,n\}$ and a hypothesis space $\mathcal{F}$, a learning algorithm is a mapping from the n-tuple sample space to the hypothesis space, i.e.,

$$A : \mathcal{Z}^{\otimes n} \to \mathcal{F}$$

$$(Z_1, Z_2, \ldots, Z_n) \longrightarrow \boxed{A} \longrightarrow \hat{f}$$

- Need not be an empirical risk minimizer

Toy example

- Consider the following learning algorithm:

$$A(Z^n) = f_0$$

  - That is, we select f_0 regardless of the training data
  - i.e., a strong "bias" toward f_0

- **Observation 1.** Clearly stupid
  - cannot achieve low training risk, no matter how many samples we have

$$\hat{R}(f_0) \gg 0$$

- **Observation 2.** It never overfits:
  - the expected generalization gap is zero, no matter how many samples we have

$$\mathbb{E}[\hat{R}(f_0) - R(f_0)] = 0$$

Intuition

- Generalize well, if the learning algorithm is insensitive to the input (i.e., training data)

- Various ways to formalize the insensitivity
  - **Stability.** Bousquet and Elisseeff (2002), Shalev-Shwartz et al., (2010), Hardt et al. (2015)
  - **Robustness.** Xu and Mannor (2008)
  - **Privacy.** Dwork et al., (2015)
  - **Information.** Russo and Zou (2015), Raginsky et al., (2016), Steinke and Zakynthinou (2020), …

- We'll discuss
  - the classic uniform stability,
  - and then move on to information-theoretic arguments

Stability

- Suppose that we have two datasets

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

  - The corresponding solutions are $\hat{f}$ and $\hat{f}^{(i)}$ (not necessarily ERM)

- Suppose that our algorithm has a (replace-one) stability property
  - For any (x, y), we have $|\ell(\hat{f}, z) - \ell(\hat{f}^{(i)}, z)| \leq \gamma$

- Then, we have the following lemma

Lemma 1.

We have

$$\mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] \leq \gamma$$

Proof sketch

- Consider all-replaced samples, drawn from the same distribution

$$Z'^n = (Z'_1, \ldots, Z'_n)$$

- Then, we have

$$\mathbb{E}[R(\hat{f})] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(\hat{f}, Z'_i)]$$

- On the other hand, since $\ell(\hat{f}, Z_i)$ and $\ell(\hat{f}^{(i)}, Z'_i)$ have the same distribution, we have

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

- Thus,

$$\begin{aligned}\mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(\hat{f}, Z'_i) - \ell(\hat{f}^{(i)}, Z'_i)] \\ &\leq \frac{1}{n}\sum_{i=1}^{n} \sup_{\tilde{Z}} |\ell(\hat{f}, \tilde{Z}) - \ell(\hat{f}^{(i)}, \tilde{Z})| \qquad \leq \gamma\end{aligned}$$

Convex optimization

- This result has straightforward implications for convex optimization
  - $\mathscr{F}$ is a convex subset of some Hilbert space $\mathscr{H}$
  - The mapping $f \mapsto \ell(f(x), y)$ is σ-strongly convex and L-Lipschitz

Theorem 1.

The ERM algorithm satisfies, with probability at least $1 - \delta$,

$$R(\hat{f}) - \inf_{f \in \mathscr{F}} R(f) \leq \frac{4L^2}{\delta \sigma n}$$

- Free of any "number of parameters"
- This is different from the guarantees in the optimization section
  - Rate in terms of n, not t
  - Rate for $R(\cdot)$, not $\hat{R}(\cdot)$

Proof sketch

- Consider a modified dataset

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

    - Let $\hat{R}^{(i)}$ be the training risk on $Z_{(i)}^n$
    - Let $\hat{f}^{(i)}$ be an ERM solution on $Z_{(i)}^n$

- Then, we have:

$$\begin{aligned}
\hat{R}(\hat{f}^{(i)}) - \hat{R}(\hat{f}) &= \frac{1}{n}\sum_{j=1}^{n}\left(\ell(\hat{f}^{(i)}, Z_j) - \ell(\hat{f}, Z_j)\right) \\
&= \frac{1}{n}\left(\ell(\hat{f}^{(i)}, Z_i) - \ell(\hat{f}, Z_i)\right) + \frac{1}{n}\sum_{j\neq i}\left(\ell(\hat{f}^{(i)}, Z_j) - \ell(\hat{f}, Z_j)\right) \\
&= \underbrace{\frac{1}{n}\left(\ell(\hat{f}^{(i)}, Z_i) - \ell(\hat{f}, Z_i)\right)}_{\leq \frac{L}{n}\|\hat{f}^{(i)} - \hat{f}\|} + \underbrace{\left(\hat{R}^{(i)}(\hat{f}^{(i)}) - \hat{R}^{(i)}(\hat{f})\right)}_{\leq 0} + \underbrace{\frac{1}{n}\left(\ell(\hat{f}, Z_i') - \ell(\hat{f}^{(i)}, Z_i')\right)}_{\leq \frac{L}{n}\|\hat{f}^{(i)} - \hat{f}\|}
\end{aligned}$$

Proof sketch

- Thus, we have:

$$\hat{R}(\hat{f}^{(i)}) - \hat{R}(\hat{f}) \leq \frac{L}{n}\|\hat{f}^{(i)} - \hat{f}\|$$

- On the other hand, by the σ-strong convexity, and as $\hat{f}$ is the minimizer of $\hat{R}(\cdot)$, we have

$$\hat{R}(\hat{f}^{(i)}) - \hat{R}(\hat{f}) \geq \frac{\sigma}{2}\|\hat{f} - \hat{f}^{(i)}\|^2$$

- Summing up, we have:

$$\|\hat{f} - \hat{f}^{(i)}\| \leq \frac{4L}{\sigma n}$$

- Thus, we have

$$|\ell(\hat{f}, z) - \ell(\hat{f}^{(i)}, z)| \leq L\|\hat{f} - \hat{f}^{(i)}\| \leq \frac{4L^2}{\sigma n}$$

- Apply the lemma and the Markov's inequality

Convex optimization

- It is quite straightforward to extend this to complexity-regularized ERM

Theorem 2.

Let $\mathcal{F}$ be a convex, norm-bounded subset of a Hilbert space $\mathcal{H}$, i.e., there exists some $B < \infty$ such that $\|f\| \leq B$ for all $f \in \mathcal{F}$. Suppose also that for each $z \in \mathcal{Z}$, the function $f \mapsto \ell(f, z)$ is convex and L-Lipschitz. For each $\lambda > 0$, consider the complexity-regularized ERM algorithm

$$\hat{f}_\lambda = A_\lambda(Z^n) := \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f, Z_i) + \frac{\lambda}{2} \|f\|^2 \right\}$$

Then, for $\lambda = L/B\sqrt{n}$, the following holds with probability at least $1 - \delta$

$$R(\hat{f}_\lambda) - \inf_{f \in \mathcal{F}} R(f) \leq \frac{LB}{2\sqrt{n}} + \frac{8LB}{\delta\sqrt{n}} + \frac{8LB}{\delta n \sqrt{n}}$$

Proof sketch

- Define the complexity-regularized loss function

$$\ell_\lambda(f, z) = \ell(f, z) + \frac{\lambda}{2}\|f\|^2$$

- Then, this is
    - λ-strongly convex
    - $(L + \lambda B)$-Lipschitz

- Applying the previous theorem, we have: with probability $1 - \delta$

$$\hat{R}_\lambda(\hat{f}_\lambda) - \inf_{f \in \mathcal{F}} \hat{R}_\lambda(f) \leq \frac{4(L + \lambda B)^2}{\delta \lambda n}$$

    - Here, $\hat{R}_\lambda(f) = \hat{R}(f) + \frac{\lambda}{2}\|f\|^2$

Proof sketch

- Therefore, with the same probability:

$$
\begin{aligned}
R(\hat{f}_\lambda) &\leq \inf_{f \in \mathcal{F}} R_\lambda(f) + \frac{4(L+\lambda B)^2}{\delta \lambda n} \\
&\leq R_\lambda(f^*) + \frac{4(L+\lambda B)^2}{\delta \lambda n} \\
&= R(f^*) + \frac{\lambda}{2}\|f^*\|^2 + \frac{4(L+\lambda B)^2}{\delta \lambda n} \\
&\leq R(f^*) + \frac{\lambda B^2}{2} + \frac{4(L+\lambda B)^2}{\delta \lambda n} \\
&\leq R(f^*) + \frac{\lambda B^2}{2} + \frac{8L^2}{\delta \lambda n} + \frac{8\lambda B^2}{\delta n}
\end{aligned}
$$

- Plug in the right λ

Formalisms

- A wonderful fact here is that we did not invoke any "uniform convergence of empirical means"
    - No size arguments on $\mathscr{F}$
- We can turn this into a more general, asymptotic results

Definition (Stability on average).

A learning algorithm A is stable on average (w.r.t. replace-one operation), whenever

$$\bar{s}_n(A) := \sup_P \left| \frac{1}{n} \sum_{i=1}^{n} [\ell(A(Z^n_{(i)}), Z'_i) - \ell(A(Z^n), Z'_i)] \right| \xrightarrow{n \to \infty} 0$$

- Needs to hold for a family of data-generating distributions P
    - i.e., many learning scenarios

Formalisms

- We also need the following definitions

Definition (Generalization).

A learning algorithm A generalizes, whenever:

$$g_n(A) := \sup_P \mathbb{E}\,|\,R(A(Z^n)) - \hat{R}(A(Z^n))\,| \xrightarrow{n\to\infty} 0$$

Definition (Generalization on Average).

A learning algorithm A generalizes on average, whenever:

$$\bar{g}_n(A) := \sup_P |\,\mathbb{E}[R(A(Z^n)) - \hat{R}(A(Z^n))]\,| \xrightarrow{n\to\infty} 0$$

- The latter is a much weaker condition

Learnability and stability

- We have already proved the following lemma.

Lemma 2.

For any learning algorithm, we have

$$\bar{g}_n(A) = \bar{s}_n(A)$$

In particular, A is stable on average if and only if it generalizes on average.

- **Proof idea.**
    - Recall what we did in the proof of Lemma 2

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

    - Take absolute value and supremum on both sides

Formalisms

- Also, we will need the following definitions

Definition (Consistency).

A learning algorithm A is consistent whenever:

$$c_n(A) := \mathbb{E}[R(A(Z^n)) - \inf_{f \in \mathcal{F}} R(f)] \xrightarrow{n \to \infty} 0$$

- Often, we expect this property to hold uniformly over a family of data-generating distributions P

Definition (Asymptotic ERM).

A learning algorithm A is an asymptotic ERM, if

$$e_n(A) := \mathbb{E}[\hat{R}(A(Z^n)) - \inf_{f \in \mathcal{F}} \hat{R}(f)] \xrightarrow{n \to \infty} 0$$

Learnability and stability

- Using the definitions, we can show the following result

Theorem 3.

For any algorithm A, we have

$$c_n(A) \leq \bar{s}_n(A) + e_n(A)$$

Therefore, a stable-on-average and AERM algorithm is consistent

- **Proof idea.** For any A and P, we have

$$\begin{aligned}\mathbb{E}[R(A(Z^n))] \leq \mathbb{E}[\hat{R}(A(Z^n))] + \bar{g}_n(A) &\leq \mathbb{E}[\inf_{f\in\mathcal{F}} \hat{R}(f)] + e_n(A) + \bar{g}_n(A) \\ &\leq \inf_{f\in\mathcal{F}} R(f) + e_n(A) + \bar{g}_n(A)\end{aligned}$$

  - Then apply Lemma 2

Learnability and stability

- Furthermore, we have the following:

Lemma 3.

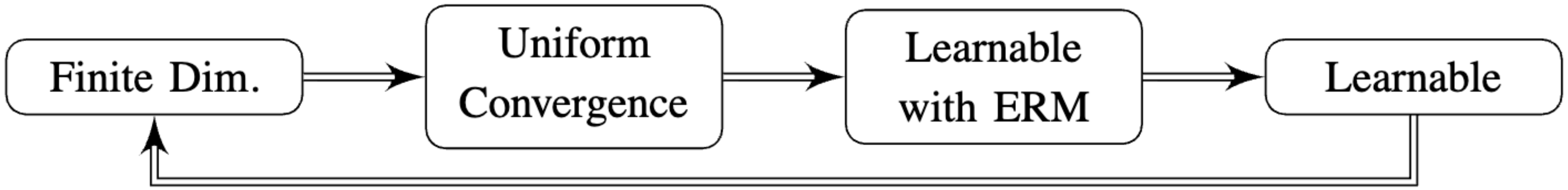
If A is AERM that generalizes on average, it generalizes. Moreover,

$$g_n(A) \leq \bar{g}_n(A) + 2e_n(A) + \frac{2}{\sqrt{n}}$$

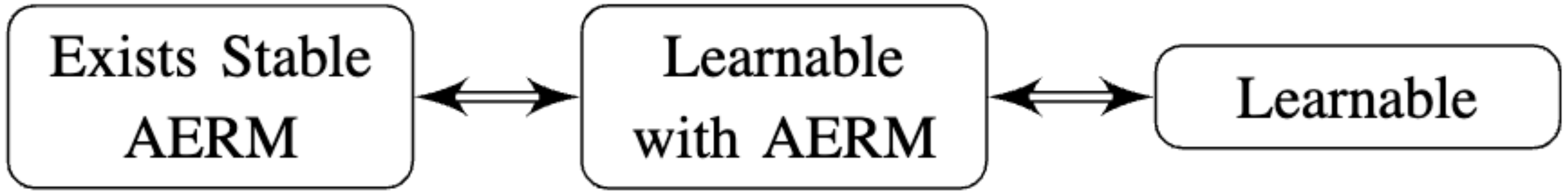
- **Proof.** Skipped.
  - https://jmlr.csail.mit.edu/papers/volume11/shalev-shwartz10a/shalev-shwartz10a.pdf

Summary

- Uniform convergence arguments show that



- Stability arguments show that



Next up

- Information-theoretic bounds