ECE826 Lecture 5:

Stability of Empirical Risk Minimizers

Contents

- Parameter count bounds for ERM
- VC dim and Rademacher Complexity generalization bounds
- Do these bounds explain generalization in modern ML?
- What are we missing?

Some Definitions

- Our goal is to find a hypothesis (classifier) h_S with small expected risk

$$R[h_S] = \mathbb{E}_{(x,y)\sim\mathscr{D}}\left[\ell(h_S(x);y)\right]$$

- The loss measures the disagreement between predictions and reality

- Since we can't directly measure $R[h_S]$ (our true cost function), we can consider optimizing its sample-average proxy, i.e., the empirical risk

$$\hat{R}[h_S] = \frac{1}{n}\sum_{i=1}^{n}\ell(h_S(x_i);y_i)$$

- Our hope is that $\hat{R}[h_S]$ is close to $R[h_S]$

The generalization gap

- The gap of the true cost function from the one we have access to

$$\epsilon_{gen} = |R[h_S] - \hat{R}[h_S]|$$

- <u>Question:</u> When is it possible to bound ϵ_{gen} by a small constant?

- The answer must depend on:
  1) n, the sample size
  2) $\mathscr{H}$, the hypothesis class (and its geometry)
  3) $\mathscr{D}$, the data distribution
  [4) the optimization algorithm that outputs our classifier]

Previously: parameter/complexity bounds

- If Floats+parametric model => $n >>$ #params for good generalization (H.I.+Union bound over all classifiers)

- If Infinite class, then VC-dim can help in bounded the error, with not much better bound than $n >>$ #params for good generalization

- Compression arguments can lead to better results for nearly sparse/low-rank models

- RC not useful when model memorizes (happens in practice)

How to make the algorithm part of the equation?

Stability of Learning Algorithms

Algorithmic Stability

- Learning algorithm $A(S)$ is stable if:
  "the trained classifier does not depend too much on one data point"

- Let S^i = original data set, but with z_i data point replaced by z_i'

- *Def:* Stability*

$$\mathbb{E}_{S,z_i}\left|\text{loss}(A(S); z_i) - \text{loss}(A(S^i); z_i)\right| \leq \delta$$

- Thm: (Bousquet and Elisseeff 2002) [amazing paper, please read]

δ-stable algorithms achieve δ generalization gap

Many stability notions

- Replace-one stability:

$$\mathbb{E}_{S,z_i} \left| \text{loss}(A(S); z_i) - \text{loss}(A(S^i); z_i) \right| \leq \delta$$

- Hypothesis stability:

$$\mathbb{E}_{S,z} \left| \text{loss}(A(S); z) - \text{loss}(A(S^i); z) \right| \leq \delta$$

- Error stability:

$$\forall S, i \;\; \mathbb{E}_{z} \left| \text{loss}(A(S); z) - \text{loss}(A(S^i); z) \right| \leq \delta$$

- Uniform stability:

$$\forall S, i, z, \quad \left| \text{loss}(A(S); z) - \text{loss}(A(S^i); z) \right| \leq \delta$$

Many stability notions

[redacted]

Stability depends on: Algorithm, Data, Loss function!

[redacted]

- Error stability:

$$\forall S, i \;\; \mathbb{E}_z \left| \text{loss}(A(S); z) - \text{loss}(A(S^i); z) \right| \leq \delta$$

- Uniform stability:

$$\forall S, i, z, \;\; \left| \text{loss}(A(S); z) - \text{loss}(A(S^i); z) \right| \leq \delta$$

Many stability notions

[redacted]

Stability depends on: Algorithm, Data, Loss function!

[redacted]

Downside: it's tricky to establish

[redacted]

Stability <=> Generalization

Stability = Generalization

- Proof by renaming
- Let $S = \{z_1, \ldots, z_n\}, \quad S^j = \{z_1, \ldots, z_j', \ldots z_n\}$

$$
\begin{aligned}
\text{gen gap} &= (\text{empirical risk}) - (\text{true risk}) \\
&= \mathbb{E}_{S,A}\left[\frac{1}{n}\sum_{j=1}^{n} \text{loss}(A(S); z_j)\right] - \mathbb{E}_{S,A,z}\text{loss}(A(S); z) \\
&= \mathbb{E}_{S,A}\left[\frac{1}{n}\sum_{j=1}^{n} \text{loss}(A(S); z_j)\right] - \mathbb{E}_{S,A,z_j'}\text{loss}(A(S); z_j') \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{S,A,z_j'}\text{loss}(A(S^j); z_j') - \mathbb{E}_{S,A,z_j'}\text{loss}(A(S); z_j') \\
&= \mathbb{E}_{S,A,z_j'}\text{loss}(A(S^j); z_j') - \mathbb{E}_{S,A,z_j'}\text{loss}(A(S); z_j') \\
&= \mathbb{E}_{S,A,z_j'}\left[\text{loss}(A(S^j); z_j') - \text{loss}(A(S); z_j')\right]
\end{aligned}
$$

Boom, Stability

Stability = Generalization

- Proof by renaming
- Let $S = \{z_1, \ldots, z_n\}, \quad S^j = \{z_1, \ldots, z_j', \ldots z_n\}$

gen gap = (empirical risk) − (true risk)

[redacted]

Caveat: not a high probability result,
but possible to prove them with a bit more work

$$= \mathbb{E}_{S,A,z_j'} \text{loss}(A(S^j); z_j') - \mathbb{E}_{S,A,z_j'} \text{loss}(A(S); z_j')$$

$$= \mathbb{E}_{S,A,z_j'} \left[\text{loss}(A(S^j); z_j') - \text{loss}(A(S); z_j') \right]$$

Boom, Stability

Stable Algorithms generalize well

Q: Which algorithms are stable?

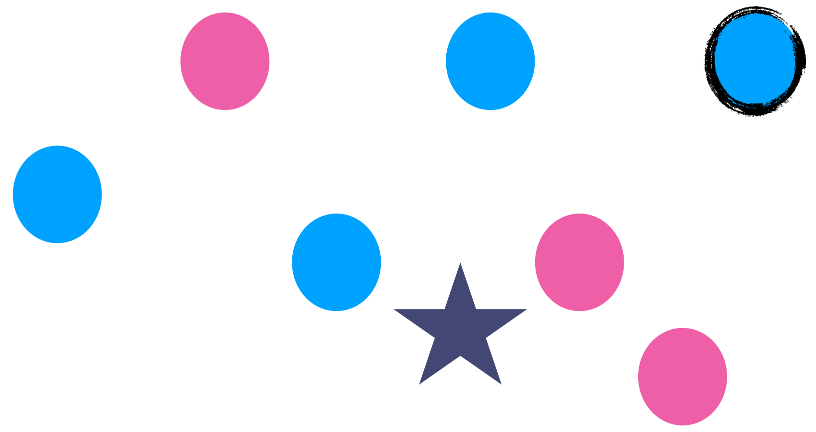
Example 0

- Trivial example of stable algorithm:

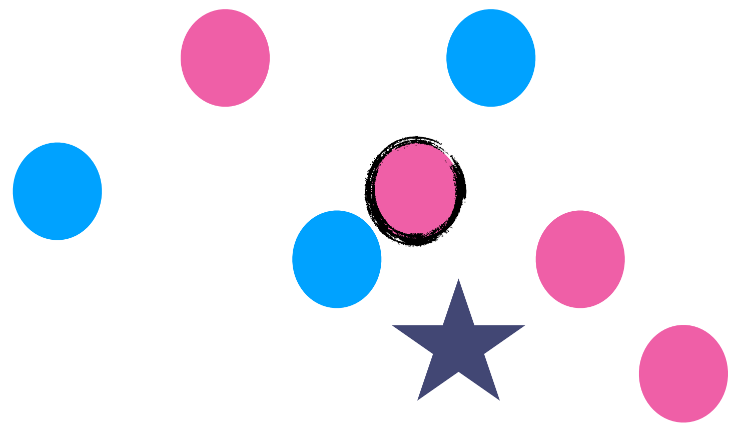
$$h(W; x) =$$

Example 1: k-NN

- Example training set:

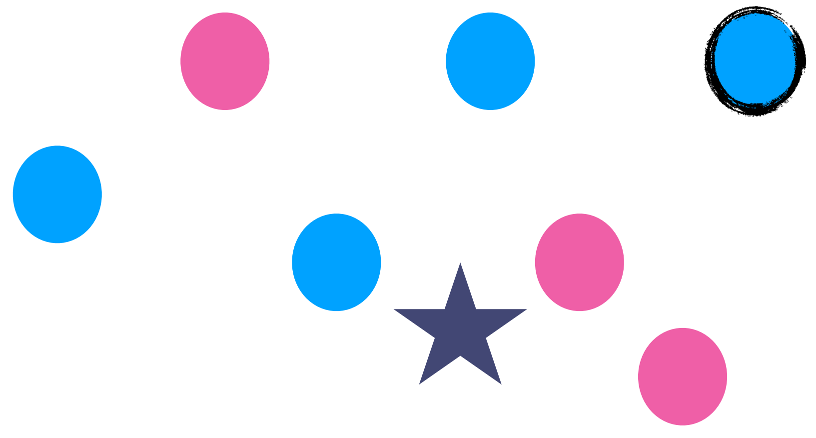


- Resampled training set:

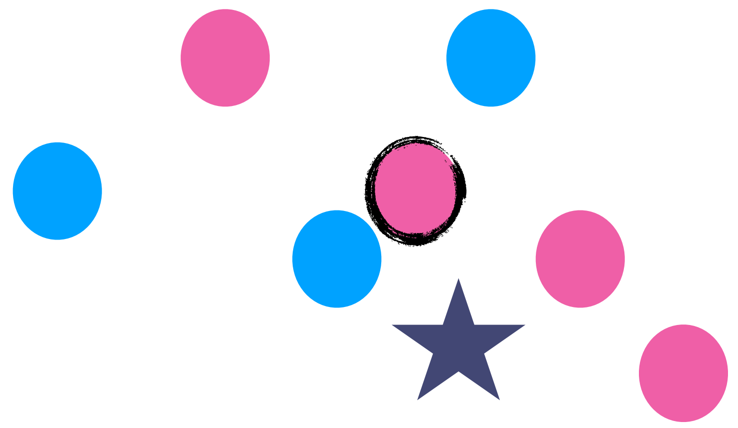


Example 1: k-NN

- Example training set:



- Resampled training set:

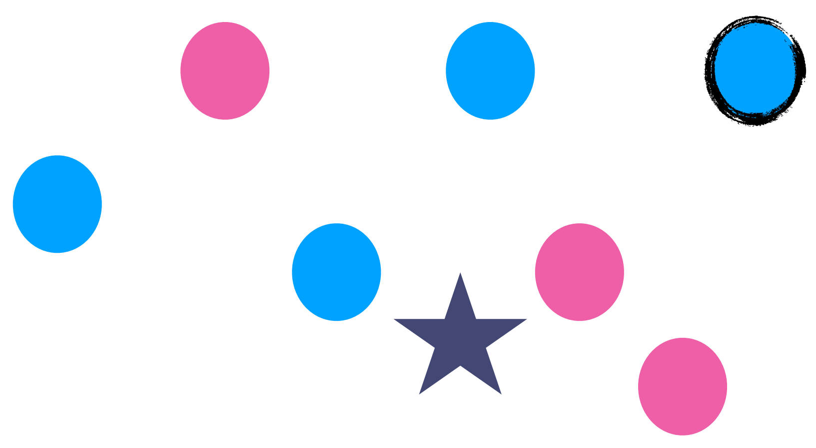


- Probability of difference in predictions:

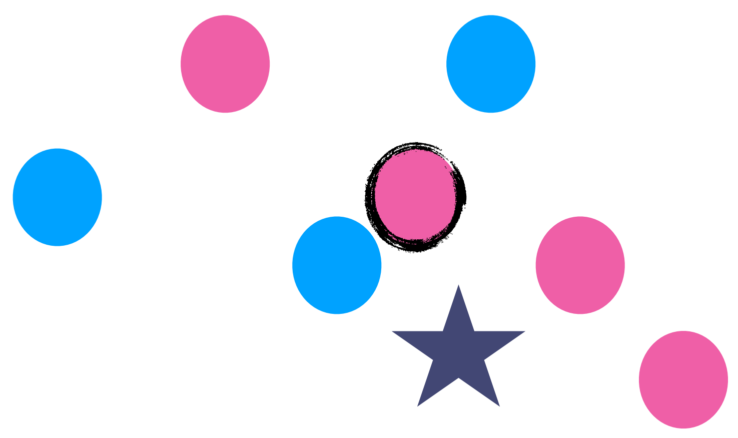
$$\Pr\left(h_S(x) \neq h_{S^i}(x)\right) \leq \Pr\left(\text{a neighbor of } x \text{ is resampled}\right)$$

Example 1: k-NN

- Example training set:



- Resampled training set:



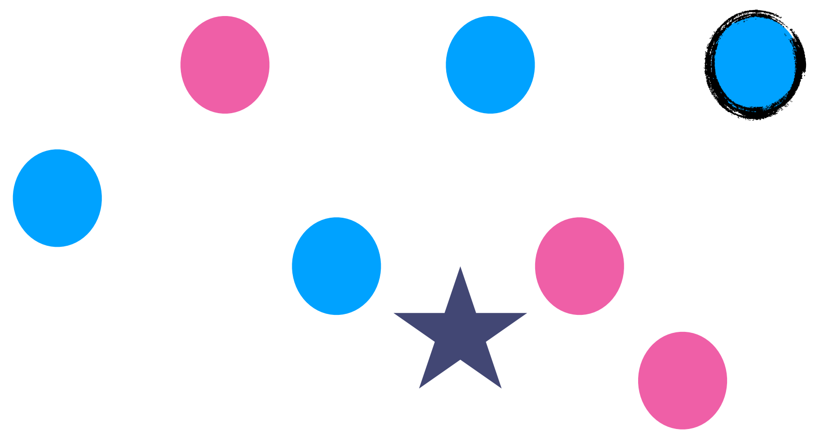
- Probability of difference in predictions:

$$\Pr\left(h_S(x) \neq h_{S^i}(x)\right) \leq \Pr\left(\text{a neighbor of } x \text{ is resampled}\right)$$

- Stability: $\text{loss}(h_S(x); y) - \text{loss}(h_{S^i}(x); y) = \Pr\left(h_S(x) \neq y\right) - \Pr\left(h_{S^i}(x) \neq y\right) =$

Example 1: k-NN

- Example training set:



- Resampled training set:

VC-dimension of kNN is infinite, yet it generalizes!

- Probability of difference in predictions:

$$\Pr\left(h_S(x) \neq h_{S^i}(x)\right) \leq \Pr\left(\text{a neighbor of } x \text{ is resampled}\right)$$

- Stability: $\text{loss}(h_S(x); y) - \text{loss}(h_{S^i}(x); y) = \Pr\left(h_S(x) \neq y\right) - \Pr\left(h_{S^i}(x) \neq y\right) =$

Before we move on: Loss functions

- The more information we have about the "loss landscape" easier the more we can say about stability/generalization AND optimization
- The "class" of the loss functions changes dramatically the guarantees one can get
- It can change things from learnable to non learnable, from poly-solvable to NP-hard
- Let's see some standard definitions

Lipschitzness & smoothness

- Lipschitz: "A function can't change too fast"

Def.:

A function $f(w)$ is L-Lipschitz on $\mathscr{W}$ if

$$|f(w) - f(w')| \leq L \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

Lipschitzness & smoothness

- Lipschitz: "A function can't change too fast"

Def.:

A function $f(w)$ is L-Lipschitz on $\mathscr{W}$ if

$$|f(w) - f(w')| \leq L \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

- Smooth: "A function whose gradients can't change too fast"

Def.:

A function $f(w)$ is β-Lipschitz on $\mathscr{W}$ if

$$\|\nabla f(w) - \nabla f(w')\| \leq \beta \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

Lipschitzness & smoothness

- Lipschitz: "A function can't change too fast"

Def.:

A function $f(w)$ is L-Lipschitz on $\mathscr{W}$ if

$$|f(w) - f(w')| \le L \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

- Smooth: "A function whose gradients can't change too fast"

Def.:

A function $f(w)$ is β-Lipschitz on $\mathscr{W}$ if

$$\|\nabla f(w) - \nabla f(w')\| \le \beta \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

Also, $f(w) \le f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\beta}{2}\|w - w\|^2$ (implying $f(w) \le f(w^*) + \frac{\beta}{2}\|w - w^*\|^2$)

Convexity

- "A function that looks like a bowl"

Def.:

A function $f(w)$ is convex on $\mathscr{W}$ if

$$f(a \cdot w + (1-a) \cdot w') \leq af(w) + (1-a)f(w')$$

Convexity

- "A function that looks like a bowl"

Def.:

A function $f(w)$ is convex on $\mathscr{W}$ if

$$f(a \cdot w + (1-a) \cdot w') \leq af(w) + (1-a)f(w')$$

- Convexity makes our lives much easier (more on next lecture).
- Most useful property (for us)

$$\langle \nabla f(w'), w' - w^* \rangle \geq f(w') - f(w^*)$$

gradient is always positively correlated with the right direction towards OPT

Strong Convexity

- "The best kind of convexity"

Def.:

A function $f(w)$ is λ-strongly convex on $\mathscr{W}$ if

$$f(w) \geq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2$$

Strong Convexity

- "The best kind of convexity"

Def.:

A function $f(w)$ is λ-strongly convex on $\mathscr{W}$ if

$$f(w) \geq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2$$

- A way to think of this: a fnct always lower bounded by a quadratic centered at OPT, i.e.,

$$f(w) \geq f(w^*) + \frac{\lambda}{2} \|w - w^*\|^2$$

Polyak Łojasiewicz (PL) functions

- "The best kind of non-convex function"

Def.:

A function $f(w)$ is μ-PL on $\mathscr{W}$ if

$$\frac{1}{2}\|\nabla f(w)\| \geq \mu \cdot (f(w) - f^*), \forall w \in \mathscr{W}$$

Polyak Łojasiewicz (PL) functions

- "The best kind of non-convex function"

Def.:

A function $f(w)$ is μ-PL on $\mathscr{W}$ if

$$\frac{1}{2}\|\nabla f(w)\| \geq \mu \cdot (f(w) - f^*),\ \forall w \in \mathscr{W}$$

- If the gradient is zero, you're at a global minimum (all local minima = global min)

Polyak Łojasiewicz (PL) functions

- "The best kind of non-convex function"

Def.:

A function $f(w)$ is μ-PL on $\mathscr{W}$ if

$$\frac{1}{2}\|\nabla f(w)\| \geq \mu \cdot (f(w) - f^*), \forall w \in \mathscr{W}$$

Back to Stability

Example 2: Minimizers of Str. Cvx Functions

- We would like to get a stability bound on

$$A(S) = w^* = \arg\min_w \left(R_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; z_i) \right)$$

Example 2: Minimizers of Str. Cvx Functions

- We would like to get a stability bound on

$$A(S) = w^* = \arg\min_w \left(R_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; z_i) \right)$$

- Assuming that $R_S(w) \geq R_S(w') + \langle \nabla R_S(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2$

Example 2: Minimizers of Str. Cvx Functions

- We would like to get a stability bound on

$$A(S) = w^* = \arg\min_w \left(R_S(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w; z_i) \right)$$

- Assuming that $R_S(w) \geq R_S(w') + \langle \nabla R_S(w'), w - w' \rangle + \frac{\lambda}{2}\|w - w'\|^2$
- What does str. convexity give us? Let's evaluate it at the opt

$$R_S(w) \geq R_S(w^*) + \langle \nabla R_S(w^*), w - w^* \rangle + \frac{\lambda}{2}\|w - w^*\|^2$$

Example 2: Minimizers of Str. Cvx Functions

- We would like to get a stability bound on

$$A(S) = w^* = \arg\min_w \left(R_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; z_i) \right)$$

- Assuming that $R_S(w) \geq R_S(w') + \langle \nabla R_S(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2$
- What does str. convexity give us? Let's evaluate it at the opt

$$R_S(w) \geq R_S(w^*) + \langle \nabla R_S(w^*), w - w^* \rangle + \frac{\lambda}{2} \|w - w^*\|^2$$

$$R_S(w) \geq R_S(w^*) + \frac{\lambda}{2} \|w - w^*\|^2$$

Example 2: Minimizers of Str. Cvx Functions

- We would like to get a stability bound on

$$A(S) = w^* = \arg\min_w \left(R_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; z_i) \right)$$

- Assuming that $R_S(w) \geq R_S(w') + \langle \nabla R_S(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2$
- What does str. convexity give us? Let's evaluate it at the opt

$$R_S(w) \geq R_S(w^*) + \langle \nabla R_S(w^*), w - w^* \rangle + \frac{\lambda}{2} \|w - w^*\|^2$$

$$R_S(w) \geq R_S(w^*) + \frac{\lambda}{2} \|w - w^*\|^2$$

$$R_S(w) - R_S(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

Example 2: Minimizers of Str. Cvx Functions

- We would like to get a stability bound on

$$A(S) = w^* = \arg\min_w \left(R_S(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w; z_i) \right)$$

- Assuming that $R_S(w) \geq R_S(w') + \langle \nabla R_S(w'), w - w' \rangle + \frac{\lambda}{2} \|w - w'\|^2$
- What does str. convexity give us? Let's evaluate it at the opt

$$R_S(w) \geq R_S(w^*) + \langle \nabla R_S(w^*), w - w^* \rangle + \frac{\lambda}{2} \|w - w^*\|^2$$

$$R_S(w) \geq R_S(w^*) + \frac{\lambda}{2} \|w - w^*\|^2$$

$$R_S(w) - R_S(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

we will use this

Example 2: Minimizers of Str. Cvx Functions

- Note that we can apply the str.cvx bound on both minimizers

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_S(w_i^*) - R_S(w^*)$$

$$+$$

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_{S^i}(w^*) - R_{S^i}(w_i^*)$$

Example 2: Minimizers of Str. Cvx Functions

- Note that we can apply the str.cvx bound on both minimizers

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_S(w_i^*) - R_S(w^*)$$

$$+$$

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_{S^i}(w^*) - R_{S^i}(w_i^*)$$

- This gives us

$$\lambda\|w^* - w_i^*\|^2 \leq (R_S(w_i^*) - R_S(w^*)) + (R_{S^i}(w^*) - R_{S^i}(w_i^*))$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{1}{n}\left(\sum_{z \in S} \ell(w_i^*; z) - \ell(w_i^*; z) + \sum_{z \in S^i} \ell(w^*; z) - \ell(w_i^*; z)\right)$$

Example 2: Minimizers of Str. Cvx Functions

- Note that we can apply the str.cvx bound on both minimizers

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_S(w_i^*) - R_S(w^*)$$

$$+$$

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_{S^i}(w^*) - R_{S^i}(w_i^*)$$

- This gives us

$$\lambda\|w^* - w_i^*\|^2 \leq (R_S(w_i^*) - R_S(w^*)) + (R_{S^i}(w^*) - R_{S^i}(w_i^*))$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{1}{n}\left(\sum_{z \in S} \ell(w_i^*; z) - \ell(w_i^*; z) + \sum_{z \in S^i} \ell(w^*; z) - \ell(w_i^*; z)\right)$$

Example 2: Minimizers of Str. Cvx Functions

- Note that we can apply the str.cvx bound on both minimizers

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_S(w_i^*) - R_S(w^*)$$

$$+$$

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_{S^i}(w^*) - R_{S^i}(w_i^*)$$

- This gives us

$$\lambda\|w^* - w_i^*\|^2 \leq (R_S(w_i^*) - R_S(w^*)) + (R_{S^i}(w^*) - R_{S^i}(w_i^*))$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{1}{n}\left(\sum_{z\in S} \ell(w_i^*; z) - \ell(w_i^*; z) + \sum_{z\in S^i} \ell(w^*; z) - \ell(w_i^*; z)\right)$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{1}{n}\left(\ell(w_i^*; z_i) - \ell(w^*; z_i) + \ell(w_i^*; z_i') - \ell(w^*; z_i')\right)$$

Example 2: Minimizers of Str. Cvx Functions

- Note that we can apply the str.cvx bound on both minimizers

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_S(w_i^*) - R_S(w^*)$$

$$+$$

$$\frac{\lambda}{2}\|w^* - w_i^*\| \leq R_{S^i}(w^*) - R_{S^i}(w_i^*)$$

- This gives us

$$\lambda\|w^* - w_i^*\|^2 \leq (R_S(w_i^*) - R_S(w^*)) + (R_{S^i}(w^*) - R_{S^i}(w_i^*))$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{1}{n}\left(\sum_{z \in S} \ell(w_i^*; z) - \ell(w_i^*; z) + \sum_{z \in S^i} \ell(w^*; z) - \ell(w_i^*; z)\right)$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{1}{n}\left(\ell(w_i^*; z_i) - \ell(w^*; z_i) + \ell(w_i^*; z_i') - \ell(w^*; z_i')\right)$$

$$\lambda\|w^* - w_i^*\|^2 \leq \frac{2L}{n}\|w^* - w_i^*\| \Rightarrow \lambda\|w^* - w_i^*\| \leq \frac{2L}{\lambda n}$$

- Now we're almost done.

Example 2: Minimizers of Str. Cvx Functions

- Strong convexity and Lipschitzness imply $\|w^* - w_i^*\| \leq \frac{2L}{\lambda n}$

Example 2: Minimizers of Str. Cvx Functions

- Strong convexity and Lipschitzness imply $\|w^* - w_i^*\| \leq \frac{2L}{\lambda n}$

- Reapplying L-Lipschitz, we obtain

$$|\ell(w^*; z) - \ell(w_i^*; z)| \leq L\|w^* - w_i^*\| \leq \frac{2L^2}{\lambda n}$$

Example 2: Minimizers of Str. Cvx Functions

- Strong convexity and Lipschitzness imply $\|w^* - w_i^*\| \leq \frac{2L}{\lambda n}$

- Reapplying L-Lipschitz, we obtain

$$|\ell(w^*; z) - \ell(w_i^*; z)| \leq L\|w^* - w_i^*\| \leq \frac{2L^2}{\lambda n}$$

Theorem:

Let the empirical risk be a strongly convex function for all data sets, the loss be bounded and Lipschitz. Then, $A(S) = \arg\min_w \hat{L}_S(w)$ is a $\frac{2L^2}{\lambda n}$-stable learning algorithm

Example 2: Minimizers of Str. Cvx Functions

[redacted]

What is strongly convex?

- Reapplying L-Lipschitz, we obtain

$$|\ell(w^*; z) - \ell(w_i^*; z)| \leq L\|w^* - w_i^*\| \leq \frac{2L^2}{\lambda n}$$

Theorem:

Let the empirical risk be a strongly convex function for all data sets, the loss be bounded and Lipschitz. Then, $A(S) = \arg\min_{w} \hat{L}_S(w)$ is a $\frac{2L^2}{\lambda n}$-stable learning algorithm

Example 2: Minimizers of Str. Cvx Functions

[redacted]

What is strongly convex?

- Reapplying L-Lipschitz, we obtain

[redacted]

any convex loss that has a $\lambda\|w\|^2$ penalty (eg.., regularized least squres/logistic regression etc)!

Theorem:

Let the empirical risk be a strongly convex function for all data sets, the loss be bounded and Lipschitz. Then, $A(S) = \arg\min_w \hat{L}_S(w)$ is a $\frac{2L^2}{\lambda n}$-stable learning algorithm

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

$$\left\| \nabla \frac{1}{n} \sum_i \ell(w; z_i) \right\|^2 \geq \mu \|w - w^*\|$$

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

$$\left\| \nabla \frac{1}{n} \sum_i \ell(w; z_i) \right\|^2 \geq \mu \|w - w^*\|$$

- Proof of stability:

$$|\ell(w^*; z) - \ell(w_i^*; z)| \leq L\|w^* - w_i^*\|$$

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

$$\left\| \nabla \frac{1}{n} \sum_i \ell(w; z_i) \right\|^2 \geq \mu \|w - w^*\|$$

- Proof of stability:

$$\begin{aligned} |\ell(w^*; z) - \ell(w_i^*; z)| &\leq L\|w^* - w_i^*\| \\ &\leq \frac{L}{\mu} \left\| \nabla \frac{1}{n} \sum_{z \in S} \ell(w_i^*; z) \right\|^2 \end{aligned}$$

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

$$\left\| \nabla \frac{1}{n} \sum_i \ell(w; z_i) \right\|^2 \geq \mu \|w - w^*\|$$

- Proof of stability:

$$\begin{aligned} |\ell(w^*; z) - \ell(w_i^*; z)| &\leq L\|w^* - w_i^*\| \\ &\leq \frac{L}{\mu} \left\| \nabla \frac{1}{n} \sum_{z \in S} \ell(w_i^*; z) \right\|^2 \\ &\leq \frac{L}{\mu} \left\| \nabla \frac{1}{n} \ell(w_i^*; z_i) \right\|^2 \end{aligned}$$

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

$$\left\| \nabla \frac{1}{n} \sum_i \ell(w; z_i) \right\|^2 \geq \mu \|w - w^*\|$$

- Proof of stability:

$$\begin{aligned}
|\ell(w^*; z) - \ell(w_i^*; z)| &\leq L\|w^* - w_i^*\| \\
&\leq \frac{L}{\mu} \left\| \nabla \frac{1}{n} \sum_{z \in S} \ell(w_i^*; z) \right\|^2 \\
&\leq \frac{L}{\mu} \left\| \nabla \frac{1}{n} \ell(w_i^*; z_i) \right\|^2 \\
&\leq \frac{L}{\mu n} \left\| \nabla \ell(w_i^*; z_i) \right\|^2
\end{aligned}$$

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

$$\left\| \nabla \frac{1}{n} \sum_i \ell(w; z_i) \right\|^2 \geq \mu \|w - w^*\|$$

Theorem:

Let the empirical risk be PL+Lipschitz+bounded gradients by.

Then, $A(S) = \arg\min_w \hat{L}_S(w)$ is a $\frac{2LD^2}{\mu n}$-stable learning algorithm

Example 3: Minimizers of PL Functions

- An empirical loss function is μ-PL if

[redacted]

Why is PL Interesting

Theorem:

Let the empirical risk be PL+Lipschitz+bounded gradients by.

Then, $A(S) = \arg\min_{w} \hat{L}_S(w)$ is a $\frac{2LD^2}{\mu n}$-stable learning algorithm

Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

Samet Oymak* and Mahdi Soltanolkotabi†

Loss landscapes and optimization in over-parameterized non-linear systems and neural networks

Chaoyue Liu[a], Libin Zhu[b,c], and Mikhail Belkin[c]

[a] *Department of Computer Science and Engineering, The Ohio State University*
[b] *Department of Computer Science and Engineering, University of California, San Diego*
[c] *Halicioğlu Data Science Institute, University of California, San Diego*

May 28, 2021

On the Convergence Rate of Training Recurrent Neural Networks

Zeyuan Allen-Zhu
`zeyuan@csail.mit.edu`
Microsoft Research AI

Yuanzhi Li
`yuanzhil@stanford.edu`
Stanford University
Princeton University

Zhao Song
`zhaos@utexas.edu`
UT-Austin
University of Washington
Harvard University

October 28, 2018

A Convergence Theory for Deep Learning via Over-Parameterization

Zeyuan Allen-Zhu
`zeyuan@csail.mit.edu`
Microsoft Research AI

Yuanzhi Li
`yuanzhil@stanford.edu`
Stanford University
Princeton University

Zhao Song
`zhaos@utexas.edu`
UT-Austin
University of Washington
Harvard University

Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

Samet Oymak* and Mahdi Soltanolkotabi†

Loss landscapes and optimization in over-parameterized non-linear systems and neural networks

Chaoyue Liu[a], Libin Zhu[b,c], and Mikhail Belkin[c]

[a]*Department of Computer Science and Engineering, The Ohio State University*
[b]*Department of Computer Science and Engineering, University of California, San Diego*
[c]*Halicioğlu Data Science Institute, University of California, San Diego*

May 28, 2021

On the Convergence Rate of Training Recurrent Neural Networks

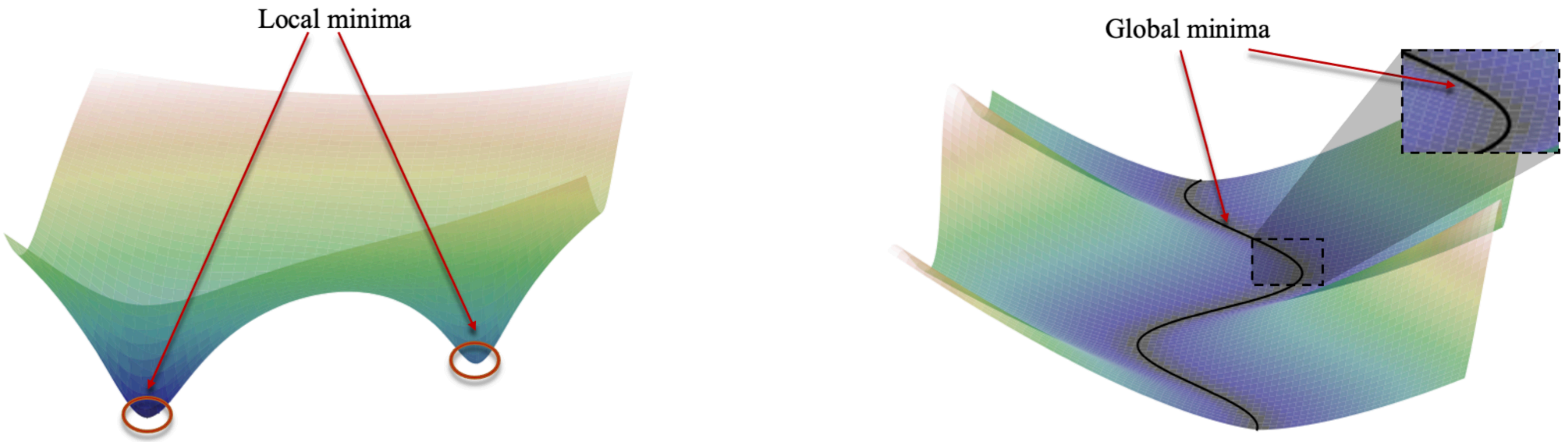
Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research AI

Yuanzhi Li
yuanzhil@stanford.edu
Stanford University
Princeton University

Zhao Song
zhaos@utexas.edu
UT-Austin
University of Washington

[redacted]

PL-like conditions hold in neighborhoods around initialization/optima.



(a) Loss landscape of under-parameterized models

(b) Loss landscape of over-parameterized models

Figure 1: Panel (a): Loss landscape is locally convex at local minima. Panel (b): Loss landscape incompatible with local convexity as the set of global minima is not locally linear.

Princeton University University of Washington
Harvard University

October 28, 2018

A Convergence Theory for Deep Learning via Over-Parameterization

PL-like conditions hold in neighborhoods around initialization/optima.

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research AI

Yuanzhi Li
yuanzhil@stanford.edu
Stanford University
Princeton University

Zhao Song
zhaos@utexas.edu
UT-Austin
University of Washington
Harvard University

Wrapping-up Generalization

Other Avenues to Generlization: PAC-Bayes bounds

- The training algorithm as a sampling distribution on $\mathscr{H}$

Theorem:

Let P be a prior distribution on $\mathscr{H}$. Let Q be the "trained" distribution for sampling a classifier. Then

$$\epsilon_{gen}[q] \leq O\left(\sqrt{\frac{\mathsf{KL}(Q||P)}{2m}}\right)$$

Other Avenues to Generlization: Information Theoretic Bounds

- The training algorithm as a sampling distribution on $\mathscr{H}$

Theorem (information):

Let $A(S)$ be a randomized learning algorithm. Then,

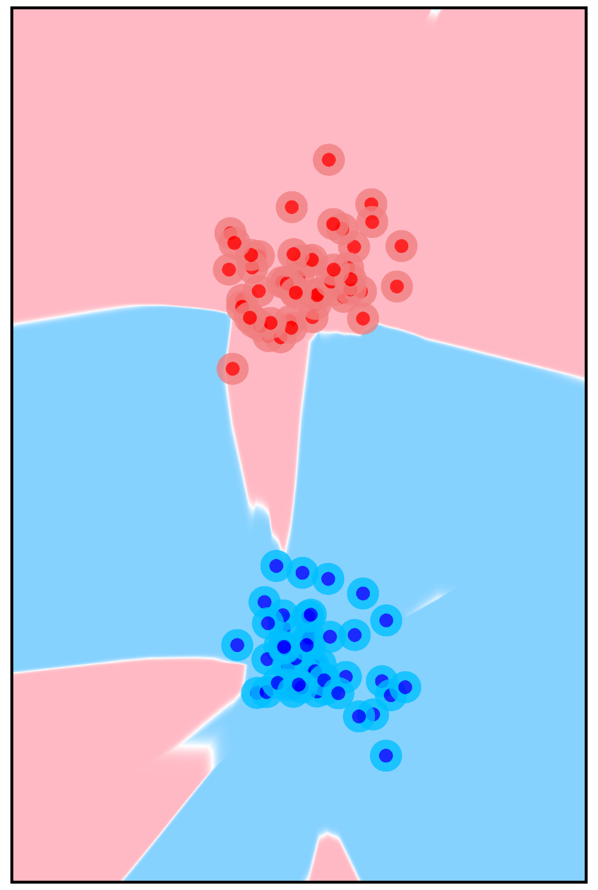
$$\epsilon_{gen}[A] \leq O\left(\sqrt{\frac{I(A(S); S)}{n}}\right)$$

- Algorithms that "leak" little information generalize better!
- Relates to stability/differential privacy

Wrapping up

- Generalization bounds = saying it will work without running it

- VC dim bounds ≈ naive parameter count bounds
- Parameter count bounds can get fancy with compression arguments
- Rademacher complexity doesn't always give interesting bounds in practice
- Stability begets generalization! Many interesting minimizers are stable

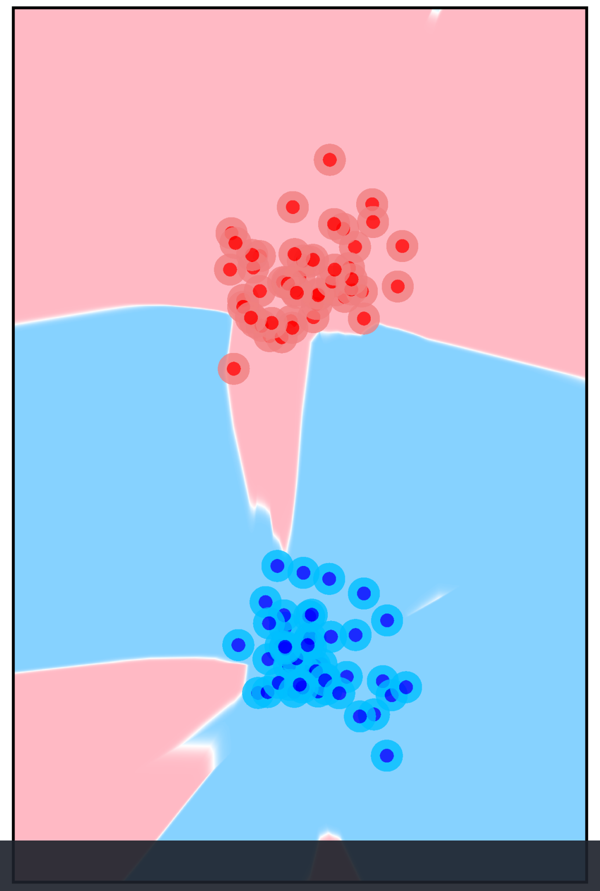


- Open Qs:
  - Are optimization algorithms like SGD stable?
  - Stability and loss geometry not well understood
  - Connections to implicit regularization?
  - Can we certify stability with limited access to data?
  - Combine with compression arguments?

Wrapping up

- Generalization bounds = saying it will work without running it

- VC dim bounds ≈ naive parameter count bounds
- Parameter count bounds can get fancy with compression arguments
- Rademacher complexity doesn't always give interesting bounds in practice
- Stability begets generalization! Many interesting minimizers are stable



- Open Qs:

[redacted]

Why do memorizing neural networks generalize?

Next Time: OPT Algorithms

Forget about the Why's, let's talk about the How's

reading list

Bousquet, Olivier, and André Elisseeff. "Stability and generalization." The Journal of Machine Learning Research 2 (2002): 499-526. https://www.jmlr.org/papers/volume2/bousquet02a/bousquet02a.pdf

(Stability Chapter) Understanding Machine Learning: From Theory to Algorithms, https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/copy.html

Hardt, M., Recht, B. and Singer, Y., 2016, June. Train faster, generalize better: Stability of stochastic gradient descent. In International conference on machine learning (pp. 1225-1234). PMLR. Vancouver, http://proceedings.mlr.press/v48/hardt16.pdf

Xu, A. and Raginsky, M., 2017. Information-theoretic analysis of generalization capability of learning algorithms. Advances in Neural Information Processing Systems, 30.
https://proceedings.neurips.cc/paper/2017/file/ad71c82b22f4f65b9398f76d8be4c615-Paper.pdf

McAllester, D.A., 1999, July. PAC-Bayesian model averaging. In Proceedings of the twelfth annual conference on Computational learning theory (pp. 164-170).
https://home.ttic.edu/~dmcallester/pac99.ps

Dziugaite, G.K. and Roy, D.M., 2017. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv preprint arXiv:1703.11008.
https://arxiv.org/pdf/1703.11008

Charles, Z. and Papailiopoulos, D., 2018, July. Stability and generalization of learning algorithms that converge to global optima. In International conference on machine learning (pp. 745-754). PMLR. http://proceedings.mlr.press/v80/charles18a/charles18a.pdf