Administrative Information

HW #1 is due tonight, by midnight. Please submit via Gradescope. The course access code is on HW#1 and Slack

The project description is now up. Feel free to email me if you'd like to chat and/or run something by me

The first deadline is April 9th, a one page project proposal

Stability, Regularization and Double Descent

Today we will study generalization thru the following framework:

> "If the hypothesis I learn does not depend on any one data point too much, then it must generalize"

Later we will see how stability comes about through explicit/implicit regularization

As usual we will be interested in the risk:

$$R(w) = \mathbb{E}_{z=(x,y)\sim D}[\ell(z,w)]$$

and the empirical risk:

$$R_S(w) = \frac{1}{N}\sum_i \ell(z_i,w)$$

and the generalization error:

$$\mathcal{E}_{gen}(w) \triangleq R(w) - R_S(w)$$

or alternatively $R(w) = R_S(w) + \mathcal{E}_{gen}(w)$

First, we will show that generalization error is an <u>average-case</u> notion of stability

Consider an algorithm $\mathcal{A}$ that we think of as a function

$$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^N \to \underset{\substack{\uparrow \\ \text{set of allowable params.}}}{\Omega}$$

Now, similar to what we did in the context of Rademacher complexity, can create <u>hybrid samples</u>

$$S = (z_1, \dots, z_N)$$

$$S' = (z_1', \dots, z_N') \quad \text{aka ``ghost samples''}$$

Now suppose we swap in the i^{th} sample from S':

$$S^{(i)} = (z_1, z_2 \dots z_{i-1}, z_i', z_{i+1}, \dots, z_N)$$

This leads us to:

<u>def:</u> The <u>average stability</u> of $\mathcal{A}$ is

$$\Delta(\mathcal{A}) = \mathbb{E}_{S, S'}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\ell(\mathcal{A}(S), z_i') - \ell(\mathcal{A}(S^{(i)}), z_i')\right)\right]$$

i.e. it is how much our loss changes on avg., when we swap out a single sample

This is just generalization error in disguise

<u>Theorem 1</u> $\mathbb{E}_S\left[\varepsilon_{gen}(\mathcal{A}(S))\right] = \Delta(\mathcal{A})$

<u>Proof:</u> Using the definition of generalization error, we have:

$$\mathbb{E}_S\left[\varepsilon_{gen}(\mathcal{A}(S))\right] = \mathbb{E}_S\left[R(\mathcal{A}(S)) - R_S(\mathcal{A}(S))\right]$$

Expanding each term:

$$\mathbb{E}_S\left[R_S(A(S))\right] = \mathbb{E}_S\left[\frac{1}{N}\sum_i \ell(A(S), z_i)\right] \quad (*)$$

$$\mathbb{E}_S\left[R(A(S))\right] = \mathbb{E}_{S,S'}\left[\frac{1}{N}\sum_i \ell(A(S), z_i')\right] \quad (1)$$

The key point is because z_i and z_i' are identically distributed, we have

$$\mathbb{E}_S\left[\ell(A(S), z_i)\right] = \mathbb{E}_{S,S'}\left[\ell(A(S^{(i)}), z_i')\right]$$

Thus rewriting (*) we have

$$\mathbb{E}_S\left[R_S(A(S))\right] = \mathbb{E}_{S,S'}\left[\frac{1}{N}\sum_i \ell(A(S^{(i)}), z_i')\right] \quad (2)$$

And combining (1) and (2), we're done

▨

It is usually challenging to work with average stability directly

Can instead use a stronger notion

def: The *uniform stability* of $\mathcal{A}$ is

$$\Delta_{unif}(\mathcal{A}) = \sup_{S,S'} \sup_{i\in[N]} \left| \ell(\mathcal{A}(S), z_i') - \ell(\mathcal{A}(S^{(i)}), z_i') \right|$$

This is a *worst-case* notion of stability

i.e. the risk never changes too much, regardless of the samples you were given

Since this is a strengthening, we have:

Corollary 1: $\mathbb{E}_{S}[\mathcal{E}_{gen}(\mathcal{A}(S))] \le \Delta_{unif}(\mathcal{A})$

You can actually get bounds that hold with high probability:

Theorem 2 [Feldman, Vondrak] Let $\gamma = \Delta_{unif}(\mathcal{A})$

Then

$$\mathbb{P}_{S}\left[\mathcal{E}_{gen}(\mathcal{A}(S)) \ge \gamma \log N \log \frac{N}{\delta} + \sqrt{\frac{\log 1/\delta}{N}} \right] \le \delta$$

where S is an iid sample of size N

Earlier bounds of Bousquet and Elisseeff were only meaningful when γ was somewhat large

In contrast, the bound above is best possible even when $\gamma = \frac{1}{N}$, i.e.

generalization error $\sim$ sampling error

Stability of Empirical Risk Minimization

An important algorithm to analyze is

$$A: S \longrightarrow \underset{w}{\operatorname{argmin}} \; R_S(w)$$

i.e. <u>empirical risk minimization</u>

As we discussed, we can't always solve this e.g. it is often nonconvex!

Next, we'll see that in some cases where we can, it is automatically stable

Following Shalev-Shwartz, Shamir, Srebro and Sridharan, we'll show:

Theorem 3 [Shalev-Shwartz et al]: Suppose the loss ℓ is α-strongly convex and L-Lipschitz. Then the ERM

$$\hat{w}_S \triangleq \underset{w \in \Omega}{\text{argmin}} \ \frac{1}{N} \sum_i \ell(w, z_i)$$

is uniformly stable, with $\Delta_{unif}(\text{ERM}) \leq \frac{4L^2}{\alpha N}$

Recall:

α-strongly convex:

$$\ell(w') \geq \ell(w) + \nabla \ell(w)^T (w' - w) + \frac{\alpha}{2} \|w' - w\|_2^2$$

L-Lipschitz:

$$|\ell(w') - \ell(w)| \leq L \|w' - w\|_2$$

First we'll need an elementary fact:

Fact 1: If $f(w)$ and $g(w)$ are both α-strongly convex then any convex combination

$$h(w) \triangleq \lambda f(w) + (1-\lambda)\, g(w)$$

is also α-strongly convex

Proof: This follows from the definition because

$$h(w') = \lambda f(w') + (1-\lambda)\, g(w')$$

$$\geq \lambda \left(f(w) + \nabla f(w)^T (w'-w) + \frac{\alpha}{2} \|w'-w\|_2^2 \right)$$

$$+ (1-\lambda) \left(g(w) + \nabla f(w)^T (w'-w) + \frac{\alpha}{2} \|w'-w\|_2^2 \right)$$

$$\overset{\text{collect terms}}{=} h(w) + \nabla h(w)^T (w'-w) + \frac{\alpha}{2} \|w'-w\|_2^2$$

■

Proof of Theorem 3: First consider

$$R_S(\hat{w}_{S(i)}) - R_S(\hat{w}_S)$$

Fact 1 tells us R_S is α-strongly convex,

$$R_S(\hat{w}_{S(i)}) - R_S(\hat{w}_S) \geq \cancelto{0}{\nabla R_S(w)\Big|_{\hat{w}_S}^{T}(\hat{w}_{S(i)} - \hat{w}_S)}$$

$$+ \frac{\alpha}{2}\|\hat{w}_{S(i)} - \hat{w}_S\|_2^2$$

From the definition of the ERM, we have:

$$\nabla R_S(w)\Big|_{\hat{w}_S} = 0$$

Thus we get

$$R_S(\hat{w}_{S(i)}) - R(\hat{w}_S) \geq \frac{\alpha}{2}\|\hat{w}_{S(i)} - \hat{w}_S\|_2^2 \quad (*)$$

Next we'll upper bound the difference in risk. Let's pull out the i^{th} term

$$R_S(\hat{w}_{S(i)}) - R_S(\hat{w}_S) =$$

$$\frac{1}{N}\left(\ell(\hat{w}_{S(i)}, z_i) - \ell(\hat{w}_S, z_i)\right) + \frac{1}{N}\sum_{j \neq i} \ell(\hat{w}_{S(i)}, z_j) - \ell(\hat{w}_S, z_j)$$

we want to get $R_{S(i)}(\hat{w}_{S(i)}) - R_{S(i)}(\hat{w}_S)$ into the expression, just need to add and subtract the appropriate terms

$$= \frac{1}{N}\left(\ell(\hat{w}_{S(i)}, z_i) - \ell(\hat{w}_S, z_i)\right)$$

bound by Lipschitzness

$$+ \frac{1}{N}\left(\ell(\hat{w}_S, z_i') - \ell(\hat{w}_{S(i)}, z_i')\right)$$

$$+ \left(\underbrace{R_{S(i)}(\hat{w}_{S(i)}) - R_{S(i)}(\hat{w}_S)}_{\leq 0 \text{ by the definition of ERM}}\right)$$

Now putting it all together, we have

$$\leq \frac{2L}{N} \|\hat{w}_{S(i)} - \hat{w}_S\|_2$$

$$\Rightarrow \frac{\alpha}{2} \|\hat{w}_{S(i)} - \hat{w}_S\|_2^2 \leq \frac{2L}{N} \|\hat{w}_{S(i)} - \hat{w}_S\|_2$$

$$\Rightarrow \|\hat{w}_{S(i)} - \hat{w}_S\|_2 \leq \frac{4L}{\alpha N}$$

using Lipschitzness again

$$|\ell(\hat{\omega}_{S^{(i)}}, z_i') - \ell(\hat{\omega}_S, z_i')| \le L \|\hat{\omega}_{S^{(i)}} - \hat{\omega}_S\|_2 \le \frac{4L^2}{\alpha N}$$

Since the l.h.s. holds for any S, S' and i, we have just shown the desired bound on uniform stability. ▨

To summarize, what we've shown is

$$\begin{matrix}\text{strongly} \\ \text{convex} \\ \text{loss}\end{matrix} \Rightarrow \begin{matrix}\text{uniform} \\ \text{stability}\end{matrix} \Rightarrow \begin{matrix}\text{average} \\ \text{stability}\end{matrix} \cong \begin{matrix}\text{general.} \\ \text{error}\end{matrix}$$

There is still one more part to the story

Regularization

what can you do if your loss is convex and Lipschitz, but not strongly convex?

"Adding a strongly convex regularizer to the objective still approximately minimize the original objective."

Let $F_S(w) \triangleq R_S(w) + \frac{\alpha}{2}\|w\|_2^2$, similarly for $F(w) \triangleq R(w) + \frac{\alpha}{2}\|w\|_2^2$

This is called ℓ_2-regularization, weight decay or Tikhonov regularization, etc

Moreover let

$$\hat{w}_S(\alpha) \triangleq \underset{w \in \Omega}{\operatorname{argmin}}\ F_S(w)$$

be the ERM of the regularized objective

Theorem 4 [Shalev-Shwartz et al] Suppose ℓ is convex and L-Lipschitz. And let

$$S = (z_1, \ldots, z_N)$$

be an i.i.d. sample from D. Then

$$R(\hat{w}_S(\alpha)) - \underset{\substack{\uparrow \\ \text{minimizer of } R}}{R^*} \leq 4\sqrt{\frac{L^2B^2}{\delta N}}\left(1 + \frac{8}{\delta N}\right)$$

with probability $\geq 1-\delta$, where B is an upper bound on the radius of Ω and we set

$$\alpha = \sqrt{\frac{16 L^2}{8 B^2 N}}$$

Proof: First notice that

$$f(w, z) \triangleq \ell(w, z) + \frac{\alpha}{2} \|w\|_2^2$$

is α-strongly convex and $L + \alpha B$-Lipschitz

Now using Theorem 3 we have

$$F(\hat{w}_S(\alpha)) - F_S(\hat{w}_S(\alpha)) \leq \frac{4 (L + \alpha B)^2}{\alpha N} \quad (1)$$

We claim that

$$\mathbb{E}_S \left[F_S(\hat{w}_S(\alpha)) \right] \leq F(\overset{\substack{\text{minimizer of } F \\ \downarrow}}{w^*}) \quad (2)$$

which is true because

$$F_S(\hat{w}_S(\alpha)) = \min_{w} F_S(w) \leq F_S(w^*)$$

and taking expectations on both sides

$$\mathbb{E}_S [F_S(\hat{w}_S(\alpha))] \leq \mathbb{E}_S [F_S(w^*)] = F(w^*)$$

Now combining (1) and (2) we have

$$\mathbb{E}_S [F(\hat{w}_S(\alpha)) - F(w^*)] \leq \frac{4(L+\alpha B)^2}{\alpha N}$$

Since the r.v. inside the expectation is nonnegative, we can apply Markov's bound

$$F(\hat{w}_S(\alpha)) - F(w^*) \leq \frac{4(L+\alpha B)^2}{\delta \alpha N}$$

with probability $\geq 1-\delta$

Furthermore

$$F(w^*) \leq \inf_{w \in \Omega} R(w) + \frac{\alpha B^2}{2}$$

And so we have

$$R(\hat{w}_S(\alpha)) \leq F(\hat{w}_S(\alpha))$$

$$\leq R^* + \frac{\alpha B^2}{2} + \frac{4(L + \alpha B)^2}{8 \alpha N}$$

Now plugging in our choice of α completes the proof. ∎

Implicit Regularization, Take 2

It turns out that explicit regularization is sometimes unneccessary

<u>Theorem 5</u> [Hardt, Recht, Singer] Suppose $\ell(w,z)$ is convex, β-smooth and L-Lipschitz for every z. Then if we run SGD with step sizes

$$\eta_t \leq \frac{2}{\beta}$$

for T timesteps, we have

$$\Delta_{unif}(SGD) \leq \frac{2L^2}{N} \sum_{t=1}^{T} \eta_t$$

In particular if we set $\eta_t \sim \frac{1}{\sqrt{N}}$ and $T = O(N)$ we get

$$\Delta_{unif}(SGD) \sim \frac{1}{\sqrt{N}} = \text{sampling error}$$

Once again we get some of the benefits of regularization without needing to do it explicitly

Note: SGD is a randomized algorithm so we need to modify the definition of uniform stability to

$$\sup_{S, S', i} \left| \mathbb{E}_{A} \left[\ell(A(S), z_i') - \ell(A(S^{(i)}), z_i') \right] \right|$$

Intuition: At each step, SGD run on S and $S^{(i)}$ has a $1 - \frac{1}{N}$ chance of picking the same sample, and if not can bound how much diverge

Connections to Privacy

There are strong parallels between uniform stability and differential privacy:

> "The output of an algorithm should not change much if change one datapoint."

Operationally, this means that you can't tell whether a particular sample belongs to the training set based on the output of the learner

In a seminal work, Dwork, McSherry, Nissim and Smith introduced this concept:

def: [Dwork et al] An algorithm $\mathcal{A}$

is (ε,δ)-differentially private if for any set $U \subseteq \Omega$ we have

$$\mathbb{P}[A(S) \in U] \leq e^{\varepsilon} \mathbb{P}[A(S^{(i)}) \in U] + \delta$$

Now suppose you want to distinguish between two hypotheses

H_0: model was learned on S

H_1: model was learned on $S^{(i)}$

Then we can define

"false alarm" $P_{FA} \triangleq$ Probability you rejected H_0 but it came from H_0

"missed detection" $P_{MD} \triangleq$ Probability you retained H_0, it came from H_1

Theorem [Oh, Viswanath] If $\mathcal{A}$ is (ε, δ)-differentially private

$$P_{FA} + e^{\varepsilon} P_{MD} \geq 1-\delta$$

$$e^{\varepsilon} P_{FA} + P_{MD} \geq 1-\delta$$

Thus when ε and δ are small, you cannot beat random guessing by much

Notice there is a subtle difference:

differential privacy	uniform stability
The parameters you learn donot depend much on any one data point	The loss on any point z does not change much, if you swap one data point

In particular, the nuissance params.

might not affect stability or generalization error, but might violate privacy!

There is a natural remedy:

"Inject noise into SGD"

Song, Chaudhuri and Sarwate showed

<u>Theorem</u> [Song et al] SGD when you inject a large enough amount of noise, is differentiably private

<u>Intuition</u>: You can couple the trajectories when you run SGD on S and $S^{(i)}$

Unfortunately:

reasonable level of privacy ⇒ large amount of noise injected ⇒ much worse learning performance

Double Descent

For square loss, we can decompose our prediction error into

bias $\cong$ are our predicted labels off in expectation?

variance $\cong$ are we overfitting the data?

In particular, if we want to predict the responses

$$y = f(x) + \varepsilon \swarrow \text{mean zero, variance } \sigma^2$$

and we learn a predictor $\hat{y} = \hat{f}(x; S)$ learned from the sample S

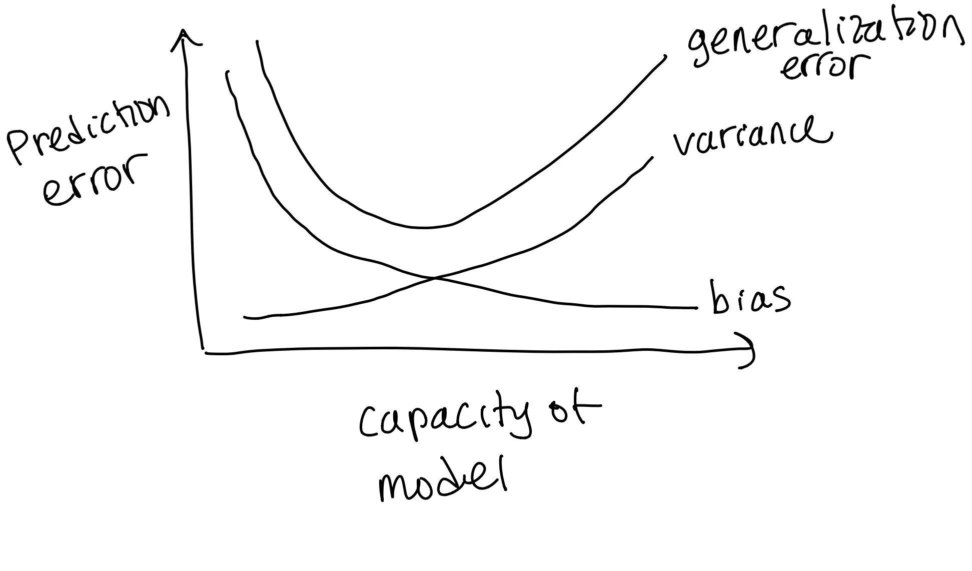
Then the square error can be written:

$$(y-\hat{y})^2 = \underbrace{\left(\mathbb{E}_S[\hat{f}(x;S)] - f(x)\right)^2}_{\text{bias}^2}$$

$$\underbrace{\mathbb{E}_S\left[\left(\mathbb{E}_{S'}[\hat{f}(x;S')] - \hat{f}(x;S)\right)^2\right]}_{\text{variance}} + \underset{\text{noise}}{\sigma^2}$$

This is called the <u>bias-variance decomposition</u>

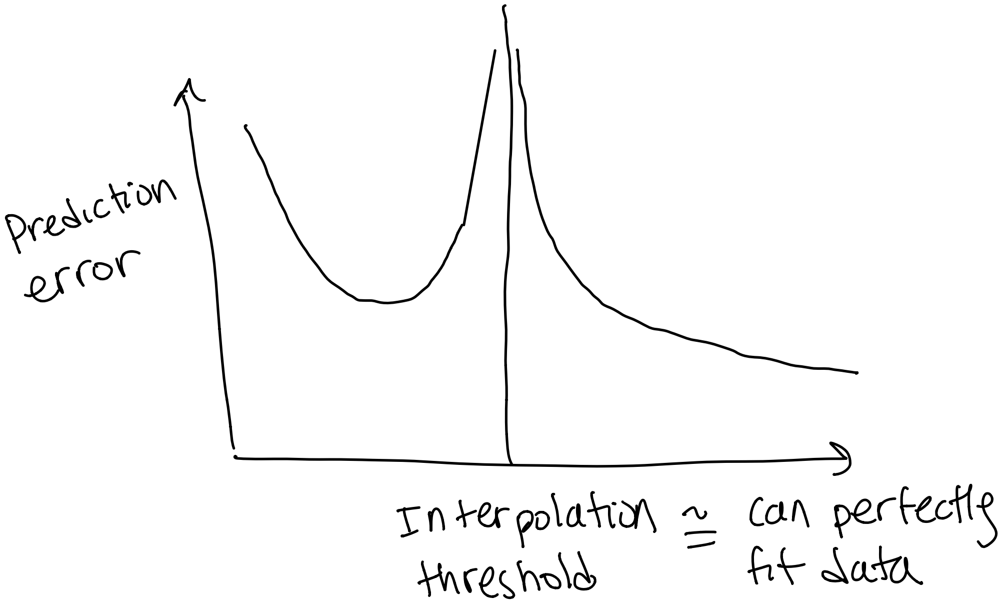
Classic generalization bounds suggest the following picture



But maybe things are more subtle...

An influential work of Belkin, Hsu, Ma and Mandal showed the generalization error curve is not always U-shaped

Instead they found



In particular, you are better off going beyond the point where the training error is zero

Hastie, Montanari, Rosset and Tibshirani gave a rigorous treatment of this phenomenon in the context of regression

The standard model is

$$y_i = \underset{\substack{\uparrow \\ \text{mean zero, covariance } \Sigma}}{X_i^T} \beta + \varepsilon_i \leftarrow \substack{\text{mean zero,} \\ \text{variance } \sigma^2}$$

$$\text{Let } \left.\begin{array}{l} p = \# \text{ of dimensions} \\ n = \# \text{ of data points} \end{array}\right\} \begin{array}{l} \text{usual stat.} \\ \text{notation} \end{array}$$

$$\text{Let } \hat{\beta} = \operatorname{argmin} \left\{ \|b\|_2 \;\middle|\; \begin{array}{l} b \text{ minimizes} \\ \|y - \underset{\uparrow}{X} b\|_2^2 \end{array} \right\}$$

Recall this is what gradient descent reaches when we initialize close to zero

Let's specialize to $\Sigma = I$

Theorem [Hastie et al] Let $\Sigma = I$ and $\frac{p}{n} \to \gamma \in (0, \infty)$. Then

$$R(\hat{\beta}, \beta) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1 \\ r^2\left(1 - \frac{1}{\gamma}\right) + \frac{\sigma^2}{\gamma - 1} & \text{for } \gamma > 1 \end{cases}$$

where $r^2 = \|\beta\|_2^2$ and

$$R(\hat{\beta}, \beta) = \mathbb{E}\left[\left(x^T \hat{\beta} - x^T \beta\right)^2\right]$$

Interpretation: Let $\text{SNR} \triangleq \frac{r^2}{\sigma^2}$

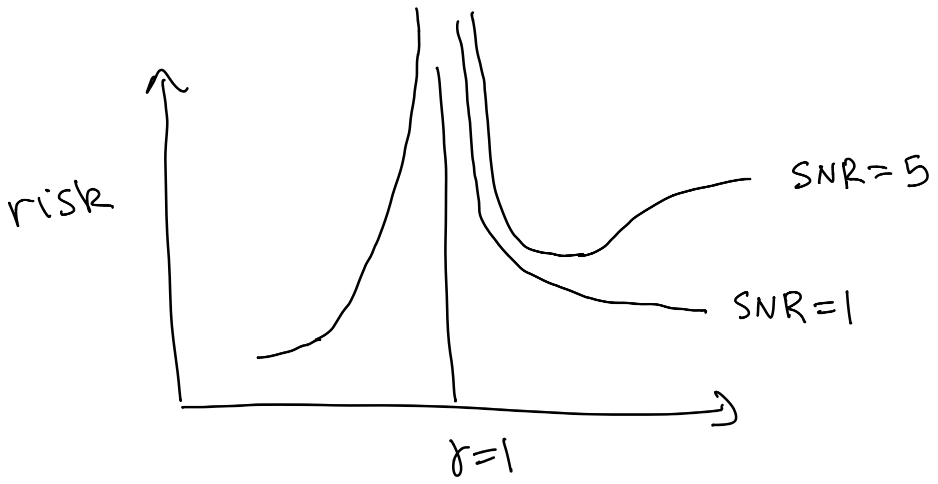
First, we can always choose $\hat{\beta} = 0$, in which case

$$R(0, \beta) = r^2 \cong \text{null risk}$$

Observation 1: Regardless of SNR, as $\gamma \to \infty$ the risk converges to the null risk

Observation 2: If $SNR < 1$, the risk is always a decreasing function of γ on $(1, \infty)$

In a picture



J