CSC 588: Machine learning theory **Spring 2021**

Lecture 16: Stability, strong convexity, and regularization

Lecturer: Chicheng Zhang *Scribe: Adrienne Kinney*

1 Stability

Stability provides another view of generalization that is complementary to the generalization theory we have seen using VC theory, Rademacher complexity, and Union Convergence Theorem.

We start with the regularized loss minimization. Consider learning algorithms of the form:

$$\hat{w} = \operatorname*{argmin}_w R(w) + \mathbb{E}_S \ell(w)$$

where $R(w)$ is a regularization term, S is the dataset, and $\ell(w)$ is the loss function.

We can show that adding the regularization term stabilizes the learning algorithm.

- *Intuition:* An algorithm is stable if small change in the input dataset does not change the output that much
- If an algorithm is stable, then you automatically guarantee good generalization properties

1.1 Formalized notion of stability

1.1.1 Setting

Start with the learning setting we're familiar with (generalization of the binary classification we have seen many times): we have distribution $\mathcal{D}$ and training set $S = (z_1, ..., z_m) \overset{iid}{\sim} \mathcal{D}$. Consider linear predictor learning model parameterized by linear coefficient w and loss function $\ell(w, z) \in \mathbb{R}$.

- Example loss functions:
  - 0-1 loss: $\ell(w, (x, y)) = I(y\langle w, x\rangle \leq 0)$
  - Hinge loss: $\ell(w, (x, y)) = \max(0, 1 - y\langle w, x\rangle)$

Then we use the generalization loss to measure the performance of the model wrt $\mathcal{D}$, $L_{\mathcal{D}}(w) = \mathbb{E}_{z\sim\mathcal{D}}\ell(w, z)$, and empirical loss $L_S(w) = \mathbb{E}_S\ell(w, z) = \frac{1}{|S|}\sum_{z\in S}\ell(w, z)$.

Learning algorithm $\mathcal{A}$ can be thought of as a deterministic mapping from training dataset to prediction model: $\mathcal{A}(S) = \hat{w}$. We want to bound the generalization gap:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_D(\hat{w}) - L_S(\hat{w})\right]$$

Note previously we wanted to give a high probability upper bound of generalization gap, but here we just want to bound the expected value.

1.1.2 Definition

Algorithm $\mathcal{A}$ will use training dataset S to produce a model, $\mathcal{A}(S)$. Now consider $S^{(i)}$ for $i \in \{1, ..., m\}$. All training examples in $S^{(i)}$ will be the same as the training examples in S with the exception of z_i, which will instead be $z' \overset{iid}{\sim} \mathcal{D}$ independent from S. We also generate $\mathcal{A}(S^{(i)})$.

Now compare performance of two models on sample z_i ($\ell(\mathcal{A}(S), z_i)$ versus $\ell(\mathcal{A}(S^{(i)}), z_i)$). If $\mathcal{A}(S)$ is overfitted, then $\ell(\mathcal{A}(S^{(i)}), z_i)$ could be quite large, as shown in Figure 1.

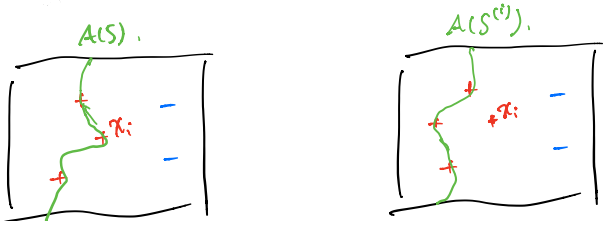

Figure 1: If $\mathcal{A}(S)$ overfits training data, then changing one training sample in $S^{(i)}$ could lead to large error: $\ell(\mathcal{A}(S), z_i) \ll \ell(\mathcal{A}(S^{(i)}), z_i)$.

Def: Learning alg $\mathcal{A}$ is on-average-replace-one (OARO) stable with rate function $g : \mathbb{N} \to \mathbb{R}$ if $\forall \mathcal{D}$ (distribution) and $\forall m$ (sample size), we have

$$\mathbb{E}_{(S,z')\sim\mathcal{D}^{m+1},i\sim\text{unif}(\{1,\ldots,m\})}\left[\ell(\mathcal{A}(S^{(i)}), z_i) - \ell(\mathcal{A}(S), z_i)\right] \leq g(m)$$

(If $\mathcal{A}(S^{(i)}) \approx \mathcal{A}(S)$, then this will be small)

1.1.3 Show OARO-stable algorithms have low expected generalization gap

Theorem 1. *If $\mathcal{A}$ is OARO-stable with rate g, then*

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))\right] \leq g(m).$$

Proof. Denote:

$$(*): \quad \mathbb{E}_{(S,z')\sim\mathcal{D},i\sim\text{unif}(\{1,\ldots,m\})}\left[\ell(\mathcal{A}(S^{(i)}), z_i) - \ell(\mathcal{A}(S), z_i)\right] \quad \text{(from definition)}$$
$$(\triangle): \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}(S)) - L_S(\mathcal{A}(S))\right] \quad \text{(from Theorem 1)}$$

We want to show $(*) = (\triangle)$.

Start with the first term in $(*)$:

$$\mathbb{E}_{(S,z)\overset{iid}{\sim}\mathcal{D}^{m+1},i\sim U(\{1,\ldots,m\})}\ell(\mathcal{A}(S^{(i)}), z_i).$$

Observe: for fixed i, $(S^{(i)}, z_i) \overset{d}{=} (S, z') \overset{d}{=} \mathcal{D}^{m+1}$. Then we see

$$\mathbb{E}_{(S,z)\overset{iid}{\sim}\mathcal{D}^{m+1}}\ell(\mathcal{A}(S^{(i)}), z_i) = \mathbb{E}_{S\sim\mathcal{D}^m,z'\sim\mathcal{D}}\ell(\mathcal{A}(S), z') = \mathbb{E}_{S\sim\mathcal{D}^m}L_{\mathcal{D}}(\mathcal{A}(S)).$$

Because this holds for every fixed i, the expectation over all i is the same value. Thus, the first terms of $(*)$ and $(\triangle)$ are equal.

Next consider the second term in $(*)$:

$$\mathbb{E}_{(S,z)\overset{iid}{\sim}\mathcal{D}^{m+1}}\mathbb{E}_{i\sim U(\{1,\ldots,m\})}\ell(\mathcal{A}(S), z_i).$$

Observe: $\mathbb{E}_{i\sim U(\{1,\ldots,m\})}\ell(\mathcal{A}(S), z_i) = \frac{1}{m}\sum_{i=1}^{m}\ell(\mathcal{A}(S), z_i) = L_S(\mathcal{A}(S))$. Thus,

$$\mathbb{E}_{(S,z)\overset{iid}{\sim}\mathcal{D}^{m+1}}\mathbb{E}_{i\sim U(\{1,\ldots,m\})}\ell(\mathcal{A}(S), z_i) = \mathbb{E}_{S\sim\mathcal{D}^m}L_S(\mathcal{A}(S)).$$

and the second terms of $(*)$ and $(\triangle)$ are equal.

Therefore, $(*) = (\triangle)$ and if an algorithm is OARO-stable, then the algorithm will produce a classifier with expected generalization gap at most $g(m)$. □

2 ℓ_2-regularization gives stability

Assumptions:

1. $\ell(w, z)$ is ρ-Lipschitz wrt w for any z
   - **Def:** ρ-Lipschitz means for $\forall w_1, w_2, |\ell(w_1, z) - \ell(w_2, z)| \leq \rho |w_1 - w_2|$
   - Note: for f differentiable $\rho = \max_z f'(z)$
2. $\ell(w, z)$ is convex in w for any z
   - Hinge loss, logistic loss, exponential loss, etc. are all convex
   - 0-1 loss is NOT convex
3. $\hat{w} = \mathcal{A}(S) = \text{argmin}_w \left(\frac{\lambda}{2} \|w\|_2^2 + L_S(w) \right)$

We will show $\mathcal{A}$ is $g(m) = \frac{2\rho^2}{\lambda m}$-OARO stable. Note, when λ is large, $g(m)$ is small which means the algorithm is more stable. Likewise when m is large.

2.1 Strong Convexity

Def: A function f is λ-strongly convex (λ-sc) if $\forall w, u$ and $\alpha \in (0, 1)$,

$$f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2} \alpha (1-\alpha) \|w - u\|_2^2$$

Note: 0-sc $\Leftrightarrow$ convex. Figure 2 shows an example of a strongly convex function.

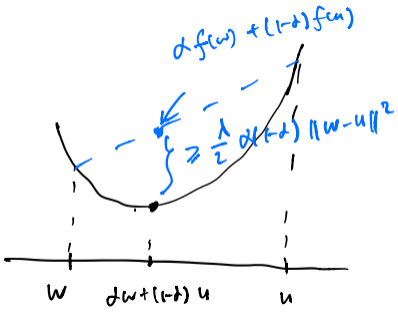

Figure 2: Linear functions are not strongly convex. We need curvature in order to find $\lambda > 0$ that satisfies the definition of strong convexity.

2.1.1 Key Properties of sc functions

1. $f(w) = \frac{\lambda}{2} \|w\|^2$ is λ-sc. (Exercise)
2. If f is λ-sc and g is 0-sc (cvx), then $h = f + g$ is λ-sc. (Write down the λ-sc and cvx defns of f and g and then sum)
3. If f is λ-sc and $u = \text{argmin}_w f(w)$, then $\forall w, f(w) - f(u) \geq \frac{\lambda}{2} \|w - u\|^2$. Figure 3 pictorially describes this.

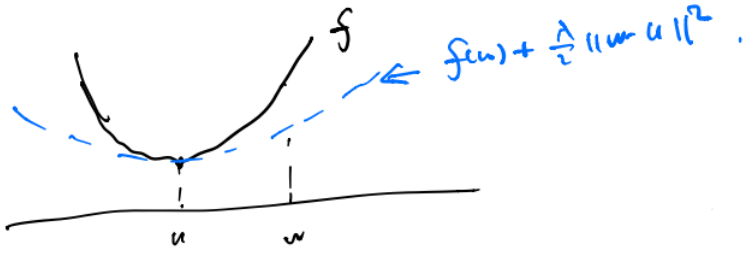

Figure 3: Consider $u = \text{argmin}_w f(w)$ and any point w, then f has a lower bound that looks like a quadratic function.

Proof. Consider special case when f is differentiable (for general f we need the definition of sub-gradient).

$$\nabla f(u) = 0$$

Start with rearranged definition of λ-sc:

$$\frac{f(u+\alpha(w-u)) - f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2$$

$$\lim_{\alpha\to 0^+}\left(\frac{f(u+\alpha(w-u)) - f(u)}{\alpha}\right) \leq \lim_{\alpha\to 0^+}\left(f(w) - f(u) - \frac{\lambda}{2}(1-\alpha)\|w-u\|^2\right)$$

$$\lim_{\alpha\to 0^+}\left(\frac{f(u+\alpha(w-u)) - f(u)}{\alpha}\right) \leq f(w) - f(u) - \frac{\lambda}{2}\|w-u\|^2$$

To solve the LHS, we see this looks like the definition of the derivative for $g(\alpha) = f(u+\alpha(w-u))$. The LHS is thus $\frac{g(\alpha)-g(0)}{\alpha-0}$ and LHS$\to g'(0) = 0$. Therefore

$$0 \leq f(w) - f(u) = \frac{\lambda}{2}\|w-u\|^2$$

and we have successfully shown the quadratic lower bound property

□

2.2 Use properties of sc-functions to show stability

Let $\hat{w} = \text{argmin}_w F_S(w)$ where $F_S(w) = L_S(w) + \frac{\lambda}{2}\|w\|^2$, and let $\hat{w}^{(i)} = \text{argmin}_w F_{S^{(i)}}(w)$ where $F_{S^{(i)}}(w) = L_{S^{(i)}}(w) + \frac{\lambda}{2}\|w\|^2$.

Recall definition of OARO-stable:

$$\mathbb{E}_{(S,z')\sim\mathcal{D}^{m+1},i\sim\text{unif}(\{1,\ldots,m\})}\left[\ell(\mathcal{A}(S^{(i)}), z_i) - \ell(\mathcal{A}(S), z_i)\right] \leq g(m)$$

where $\hat{w}^{(i)} = \mathcal{A}(S^{(i)})$ and $\hat{w} = \mathcal{A}(S)$. If we can show $\hat{w}$ and $\hat{w}^{(i)}$ are close to each other, we can show the algorithm is stable because of the Lipschitz property of ℓ. Caveat: without sc, the functions could be close, but the minimizers could be far (see Figure 4).

We know $F_S(w)$ and $F_{S^{(i)}}(w)$ are both λ-sc functions because $L_{S/S^{(i)}}$ is the average of convex functions and $\frac{\lambda}{2}\|w\|^2$ is sc, thus their sum is sc (Property 2 of sc functions). We want to show if two sc functions are similar, then their minimizers are also similar. Use Property 3 of sc functions (quadratic lower bound of sc functions):

$$F_S(\hat{w}^{(i)}) - F_S(\hat{w}) \geq \frac{\lambda}{2}\|\hat{w}^{(i)} - \hat{w}\|^2$$

$$F_{S^{(i)}}(\hat{w}) - F_{S^{(i)}}(\hat{w}^{(i)}) \geq \frac{\lambda}{2}\|\hat{w}^{(i)} - \hat{w}\|^2$$

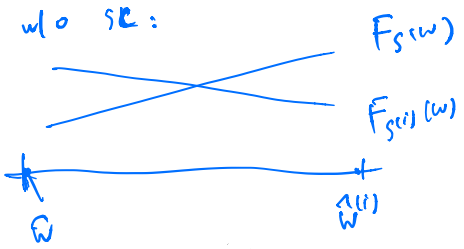

Figure 4: Without strong convexity, the functions could be close, but the minimizers could be far.

Next sum these two inequalities and cancel common terms:

$$\Big(F_S(\hat{w}^{(i)}) - F_{S^{(i)}}(\hat{w}^{(i)})\Big) - \Big(F_S(\hat{w}) - F_{S^{(i)}}(\hat{w})\Big) \geq \lambda\|\hat{w}^{(i)} - \hat{w}\|^2$$
$$\Big(\frac{1}{m}\ell(\hat{w}^{(i)}, z_i) - \frac{1}{m}\ell(\hat{w}^{(i)}, z')\Big) - \Big(\frac{1}{m}\ell(\hat{w}, z_i) - \frac{1}{m}\ell(\hat{w}, z')\Big) \geq \lambda\|\hat{w}^{(i)} - \hat{w}\|^2$$

Regroup wrt z_i or z':

$$\frac{\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i)}{m} - \frac{\ell(\hat{w}^{(i)}, z') - \ell(\hat{w}, z')}{m} \geq \lambda\|\hat{w}^{(i)} - \hat{w}\|^2$$

By Lipschitz property, we know $\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \leq \rho\|\hat{w}^{(i)} - \hat{w}\|$ and $\ell(\hat{w}^{(i)}, z') - \ell(\hat{w}, z') \leq \rho\|\hat{w}^{(i)} - \hat{w}\|$. Thus,

$$\frac{2\rho}{m}\|\hat{w} - \hat{w}^{(i)}\| \geq \lambda\|w^{(i)} - \hat{w}\|^2$$
$$\|\hat{w}^{(i)} - \hat{w}\| \leq \frac{2\rho}{m\lambda}$$

This shows that, pointwise, two predictors have similar losses:

$$\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \leq \rho\|\hat{w}^{(i)} - \hat{w}\| \quad \text{(Lipschitz property)}$$
$$\leq \frac{2\rho^2}{m\lambda}$$

Then if we take the expectation wrt all training samples S and replacement sample z', we conclude $\mathcal{A}$ is $g(m) = \frac{2\rho^2}{m\lambda}$-OARO-stable and strong convexity gives a tool of controlling the stability.