Stability, regularization, and generalization

Chicheng Zhang

CSC 588, University of Arizona

Motivation

So far, we have seen generalization error analyses by establishing "uniform convergence" on hypothesis classes, assuming $\hat{h} \in \mathcal{H}$, where the key step is:

$$L_D(\hat{h}) - L_S(\hat{h}) \leq \sup_{h \in \mathcal{H}} \left(L_D(h) - L_S(h) \right)$$

Can we establish generalization error bounds on models output by learning algorithms that do not use fixed hypothesis classes?

Stability: abstract definition

- Algorithm $\mathcal{A}$ is stable, if small changes in input dataset does not change the output model by much.
- $\mathcal{A}$ is stable $\implies$ $\mathcal{A}$ is unlikely to capture the idiosyncrasies of individual datasets, but rather property of the distribution
- E.g. regularized loss minimization:

$$\hat{w} \leftarrow \operatorname*{argmin}_{w} \underbrace{\lambda \cdot R(w)}_{\text{complexity regularizer}} + \underbrace{\sum_{i=1}^{m} \ell(w, z_i)}_{\text{empirical risk}}$$

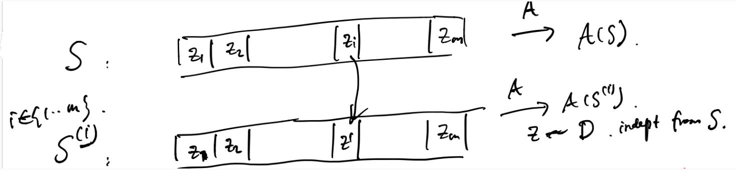
$\lambda \uparrow \implies \hat{w}$ less affected by individual training examples $\implies$ more stable

Formal setting

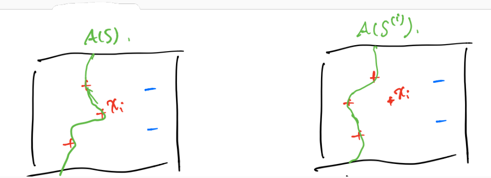
- Training dataset $S = (z_1, \ldots, z_m) \overset{iid}{\sim} D$
- Learning model parameterized by $w \in \mathbb{R}^d$
- Loss function ℓ: $\ell(w, z) \in \mathbb{R}$ (e.g. 0-1 loss, hinge loss, ...)
- Generalization loss of model w: $L_D(w) = \mathbb{E}_{z \sim D} \ell(w, z)$
- Training (empirical) loss of model w:
  $L_S(w) = \mathbb{E}_{z \sim S} \ell(w, z) = \frac{1}{|S|} \sum_{z \in S} \ell(w, z)$
- learning algorithm $\mathcal{A}$; output model $\hat{w} = \mathcal{A}(S)$
- Goal: bound $\hat{w}$'s expected generalization loss:

$$\mathbb{E}_{S \sim D^m}\left[L_D(\hat{w})\right] = \underbrace{\mathbb{E}_{S \sim D^m}\left[L_S(\hat{w})\right]}_{\text{expected empirical loss}} + \underbrace{\mathbb{E}_{S \sim D^m}\left[L_D(\hat{w}) - L_S(\hat{w})\right]}_{\text{expected generalization gap}}$$

Stability: an intuitive formulation



- Compare $\ell(\mathcal{A}(S), z_i)$ vs. $\ell(\mathcal{A}(S^{(i)}), z_i)$
- If the former is much smaller, than $\mathcal{A}$ “overfits” on z_i



Stability: formal definition

Definition

Learning algorithm $\mathcal{A}$ is on-average-replace-one (OARO) stable with rate function $g : \mathbb{N} \to \mathbb{R}$, if for any distribution D, any sample size m,

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

where $[m] := \{1, \ldots, m\}$.

Remarks:

- Usually denote by $\hat{w} = \mathcal{A}(S)$ and $\hat{w}^{(i)} = \mathcal{A}(S^{(i)})$
- Intuitively, $\mathcal{A}$ more stable $\implies$ can choose g to be smaller

Stability implies generalization

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

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

Proof.
It suffices to show

$$\begin{aligned} & \mathbb{E}_{S\sim D^m}\left[L_D(\mathcal{A}(S)) - L_S(\mathcal{A}(S))\right] \\ = \; & \mathbb{E}_{(S,z')\sim D^{m+1}, i\sim \mathrm{Unif}([m])}\left[\ell(\mathcal{A}(S^{(i)}), z_i) - \ell(\mathcal{A}(S), z_i)\right] \end{aligned}$$

We will look at the first and the second terms on the LHS / RHS respectively.

Stability implies generalization (cont'd)

Proof (cont'd).
For the second term:

$$\mathbb{E}_{S\sim D^m}\left[L_S(\mathcal{A}(S))\right] \stackrel{?}{=} \mathbb{E}_{(S,z')\sim D^{m+1}, i\sim \mathrm{Unif}([m])}\left[\ell(\mathcal{A}(S), z_i)\right]$$

Observe:

$$\begin{aligned}
& \mathbb{E}_{(S,z')\sim D^{m+1}, i\sim \mathrm{Unif}([m])}\left[\ell(\mathcal{A}(S), z_i)\right] \\
= \; & \mathbb{E}_{S\sim D^{m+1}}\left[\mathbb{E}_{i\sim \mathrm{Unif}([m])}\left[\ell(\mathcal{A}(S), z_i)\right]\right] \\
= \; & \mathbb{E}_{S\sim D^{m+1}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\mathcal{A}(S), z_i)\right] \\
= \; & \mathbb{E}_{S\sim D^{m+1}}\left[L_S(\mathcal{A}(S))\right]
\end{aligned}$$

Stability implies generalization (cont'd)

Proof (cont'd).
For the first term:

$$\mathbb{E}_{S\sim D^m}\left[L_D(\mathcal{A}(S))\right] \stackrel{?}{=} \mathbb{E}_{(S,z')\sim D^{m+1}, i\sim \mathrm{Unif}([m])}\left[\ell(\mathcal{A}(S^{(i)}), z_i)\right]$$

Observe: for every i, $(S^{(i)}, z_i) \stackrel{d}{=} (S, z') \stackrel{d}{=} D^{m+1}$,

Therefore,

$$\begin{aligned}
&\mathbb{E}_{(S,z')\sim D^{m+1}, i\sim \mathrm{Unif}([m])}\left[\ell(\mathcal{A}(S^{(i)}), z_i)\right] \\
=\;& \mathbb{E}_{i\sim \mathrm{Unif}([m])}\left[\mathbb{E}_{(S,z')\sim D^{m+1}}\left[\ell(\mathcal{A}(S^{(i)}), z_i)\right]\right] \\
=\;& \mathbb{E}_{i\sim \mathrm{Unif}([m])}\left[\mathbb{E}_{(S,z')\sim D^{m+1}}\left[\ell(\mathcal{A}(S), z')\right]\right] \\
=\;& \mathbb{E}_{i\sim \mathrm{Unif}([m])}\left[\mathbb{E}_{S\sim D^m}\left[L_D(\mathcal{A}(S))\right]\right]
\end{aligned}$$

□

ℓ_2-Regularization gives stability

Assume:

- $\ell(w, z)$ is ρ-Lipschitz in w wrt ℓ_2 norm,
    - i.e. for any z, any w_1, w_2,

$$\left|\ell(w_1, z) - \ell(w_2, z)\right| \leq \rho \|w_1 - w_2\|_2$$

    - A sufficient condition: ℓ is differentiable in w and $\|\nabla \ell(w, z)\|_2 \leq \rho$
- $\ell(w, z)$ is convex in w
    - e.g. ℓ is hinge / logistic / exponential loss
    - does not capture 0-1 loss: $\ell(w, (x, y)) = I(y \langle w, x \rangle \leq 0)$
- $\mathcal{A}$ takes input S, outputs

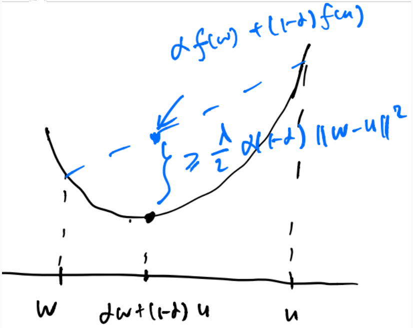
$$\hat{w} = \operatorname*{argmin}_w \left(\frac{\lambda}{2} \|w\|_2^2 + L_S(w) \right)$$

We will show that, $\mathcal{A}$ is $g(m) := \frac{2\rho^2}{\lambda m}$-OARO-stable.

Key tool: strong convexity

Definition
Function f in convex domain $C \subset \mathbb{R}^d$ is said to be λ-strongly convex (SC) with respect to norm $\|\cdot\|$, if $\forall w, u \in C, \alpha \in (0,1)$:

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



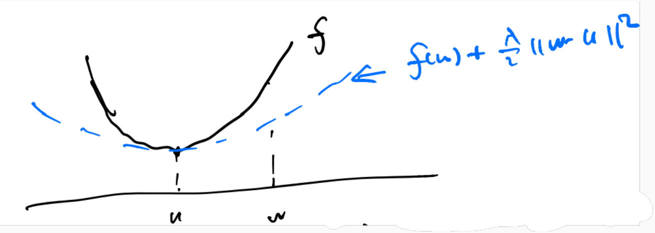
Strong convexity: Useful properties

- f is 0-SC $\iff$ f is convex
- For $f(x) = \langle a, x \rangle + b$, is f λ-SC with $\lambda > 0$?
- $f(w) = \frac{\lambda}{2}\|w\|_2^2$ is λ-SC wrt $\|\cdot\|_2$
- If f is λ-SC wrt $\|\cdot\|$, g is convex, then $h = f + g$ is λ-SC wrt $\|\cdot\|$

Strong convexity: behavior around minimizer

Lemma
If f is λ-SC wrt $\|\cdot\|$ and $u = \mathrm{argmin}_{w \in C} f(w)$, then for all w,

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



Strong convexity: behavior around minimizer

Proof.
We only show the special case when f is differentiable and $C = \mathbb{R}^d$ (the general proof needs to use *subgradient*, introduced later in the course)

1. u is the minimizer $\implies \nabla f(u) = 0$
2. f is λ-SC $\implies \forall w, \alpha,$

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

3. Letting $\alpha \to 0$:
   - $\mathrm{RHS} \to f(w) - f(u) - \frac{\lambda}{2}\|w - u\|^2$
   - $\mathrm{LHS} = \frac{g(\alpha) - g(0)}{\alpha - 0}$, where $g(\alpha) = f(u + \alpha(w - u))$.
     $\mathrm{LHS} \to g'(\alpha)\big|_{\alpha=0} = \langle \nabla f(u + \alpha(w - u)), w - u \rangle\big|_{\alpha=0} = \langle \nabla f(u), w - u \rangle = 0$

$\square$

ℓ_2-Regularization gives stability (cont'd)

Theorem
If $\ell(w, z)$ is convex, and ρ-Lipschitz in w wrt ℓ_2 norm, then algorithm $\mathcal{A}$ that outputs

$$\hat{w} = \operatorname*{argmin}_{w} \left(\frac{\lambda}{2} \|w\|_2^2 + L_S(w) \right)$$

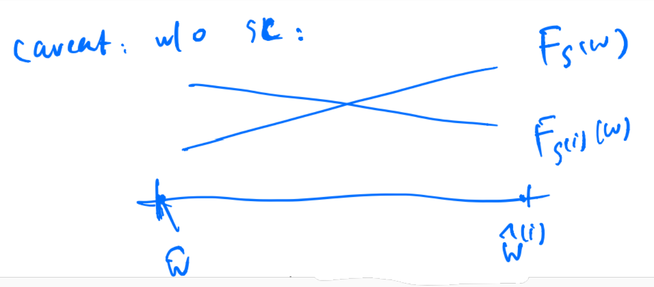
is $\frac{2\rho^2}{\lambda m}$-OARO-stable.

Intuition:

$$\hat{w} = \operatorname*{argmin}_{w} F_S(w), \text{ where } F_S(w) := \frac{\lambda}{2} \|w\|_2^2 + L_S(w)$$

$$\hat{w}^{(i)} = \operatorname*{argmin}_{w} F_{S^{(i)}}(w), \text{ where } F_{S^i}(w) := \frac{\lambda}{2} \|w\|_2^2 + L_{S^{(i)}}(w)$$

Why would $\hat{w}$ and $\hat{w}^{(i)}$ be close?



F_S, F_{S^i} λ-SC $\implies$ can rule out pathological cases where $\hat{w} - \hat{w}^{(i)}$ is large

ℓ_2-Regularization gives stability (cont'd)

Proof.
Local property of strong convexity $\implies$

$$F_S(\hat{w}^{(i)}) - F_S(\hat{w}) \geq \frac{\lambda}{2}\|\hat{w}^{(i)} - \hat{w}\|_2^2$$

$$F_{S^{(i)}}(\hat{w}) - F_{S^{(i)}}(\hat{w}^{(i)}) \geq \frac{\lambda}{2}\|\hat{w}^{(i)} - \hat{w}\|_2^2$$

Summing up the two inequalities and regrouping,

$$\left(F_S(\hat{w}^{(i)}) - F_{S^{(i)}}(\hat{w}^{(i)})\right) - \left(F_S(\hat{w}) - F_{S^{(i)}}(\hat{w})\right) \geq \lambda\|\hat{w}^{(i)} - \hat{w}\|_2^2$$

Note

$$\begin{aligned}
\text{LHS} &= \left(\frac{1}{m}\ell(\hat{w}^{(i)}, z_i) - \frac{1}{m}\ell(\hat{w}^{(i)}, z')\right) - \left(\frac{1}{m}\ell(\hat{w}, z_i) - \frac{1}{m}\ell(\hat{w}, z')\right) \\
&= \left(\frac{1}{m}\ell(\hat{w}^{(i)}, z_i) - \frac{1}{m}\ell(\hat{w}, z_i)\right) - \left(\frac{1}{m}\ell(\hat{w}^{(i)}, z') - \frac{1}{m}\ell(\hat{w}, z')\right) \\
&= \frac{2\rho}{m}\|\hat{w} - \hat{w}^{(i)}\|_2
\end{aligned}$$

ℓ_2-Regularization gives stability (cont'd)

Proof cont'd.
Therefore,

$$\frac{2\rho}{m}\|\hat{w} - \hat{w}^{(i)}\|_2 \geq \lambda\|\hat{w}^{(i)} - \hat{w}\|_2^2,$$

and consequently,

$$\|\hat{w}^{(i)} - \hat{w}\|_2 \leq \frac{2\rho}{m\lambda}.$$

Hence, for all i,

$$\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \leq \rho\|\hat{w}^{(i)} - \hat{w}\|_2 \leq \frac{2\rho^2}{m\lambda}.$$

Taking expectation over $i \sim \mathrm{Unif}([m])$ and $S, z' \sim D^{m+1}$, we conclude that $\mathcal{A}$ is $g(m) = \frac{2\rho^2}{m\lambda}$-OARO-stable. $\square$

Stability-fitting tradeoff

For

$$\hat{w} = \operatorname*{argmin}_{w} F_S(w), \text{ where } F_S(w) := \frac{\lambda}{2}\|w\|_2^2 + L_S(w),$$

$\hat{w}$ has guarantee:

$$\begin{aligned}
\underbrace{\mathbb{E}_{S\sim D^m}\left[L_D(\hat{w})\right]}_{\text{expected generalization loss}} &= \underbrace{\mathbb{E}_{S\sim D^m}\left[L_S(\hat{w})\right]}_{\text{expected empirical loss}} + \underbrace{\mathbb{E}_{S\sim D^m}\left[L_D(\hat{w}) - L_S(\hat{w})\right]}_{\text{expected generalization gap}}, \\
&\leq \mathbb{E}_{S\sim D^m}\left[L_S(\hat{w})\right] + \frac{2\rho^2}{m\lambda} \\
&\leq \mathbb{E}_{S\sim D^m}\left[F_S(\hat{w})\right] + \frac{2\rho^2}{m\lambda} \\
&\leq \mathbb{E}_{S\sim D^m}\left[F_S(w^*)\right] + \frac{2\rho^2}{m\lambda}, \quad \forall w^* \\
&\leq \mathbb{E}_{S\sim D^m}\left[L_S(w^*) + \frac{\lambda}{2}\|w^*\|_2^2\right] + \frac{2\rho^2}{m\lambda}, \quad \forall w^* \\
&\leq L_D(w^*) + \frac{\lambda}{2}\|w^*\|_2^2 + \frac{2\rho^2}{m\lambda}, \quad \forall w^* \\
&\leq L_D(w^*) + \frac{\lambda}{\ }\|w^*\|^2 + \quad \frac{2\rho^2}{\ } \quad \forall w^*
\end{aligned}$$

Tuning 1: competing with fixed bounded hypothesis class

- Suppose we would like $\hat{w}$ to compete with hypothesis class $\mathcal{H} = \left\{ w \in \mathbb{R}^d : \|w\|_2 \le B \right\}$
- Recall:

$$
\begin{aligned}
\mathbb{E}_{S \sim D^m}\left[L_D(\hat{w}) \right] &\le L_D(w) + \frac{\lambda}{2}\|w\|_2^2 + \frac{2\rho^2}{m\lambda}, \quad \forall w \in \mathcal{H}, \\
&\le L_D(w) + \frac{\lambda B^2}{2} + \frac{2\rho^2}{m\lambda}, \quad \forall w \in \mathcal{H},
\end{aligned}
$$

i.e.

$$
\mathbb{E}_{S \sim D^m}\left[L_D(\hat{w}) \right] \le \min_{w \in \mathcal{H}} L_D(w) + \left(\frac{\lambda B^2}{2} + \frac{2\rho^2}{m\lambda} \right)
$$

Choosing $\lambda = \frac{2\rho}{B\sqrt{m}} \implies$

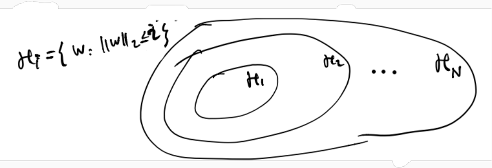
$$
\mathbb{E}_{S \sim D^m}\left[L_D(\hat{w}) \right] \le \min_{w \in \mathcal{H}} L_D(w) + \rho B \sqrt{\frac{4}{m}}.
$$

Tuning 2: competing with unbounded hypothesis class

Choosing $\lambda = \Theta(\frac{1}{\sqrt{m}}) \implies$

$$\begin{aligned} \mathbb{E}_{S\sim D^m}\left[L_D(\hat{w})\right] &\leq L_D(w^*) + \frac{\lambda}{2}\|w^*\|_2^2 + \frac{2\rho^2}{m\lambda}, \quad \forall w^* \in \mathbb{R}^d \\ &\leq L_D(w^*) + O\left(\frac{\|w^*\|_2^2 + \rho^2}{\sqrt{m}}\right), \quad \forall w^* \in \mathbb{R}^d \end{aligned}$$

This yields a model selection guarantee – competing with all hypothesis classes $\mathcal{H}_i$ simultaneously



What have we learned?

- Stability provides another view of generalization, complementary to uniform convergence
- Through strong convexity, regularized convex loss minimization enjoys stability guarantees
- Tuning of regularization parameter results in stability-fitting tradeoff