IFT 6085 - Lecture 11
(Stability and PAC Bayes)

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribe(S): [Amy Zhang, William Fedus] **Instructor:** Ioannis Mitliagkas

Summary

Sufficient condition: Given enough samples we can achieve a good enough generalization. However, typically in deep learning, we never have large enough data sets to get non-vacuous or meaningful bounds.

Last Time	Today
PAC Bounds Occam Bounds PAC Bayes Bounds Stability Bounds	Stability PAC Bayes (Practical) Generalization

How can we go from PAC Bayes to a non-vacuous generalization bound?

By sacrificing some data as part of a dedicated test set, we can measure test set generalization and achieve a tighter bound than the weak population bounds. See *Tutorial on Practical Prediction Theory for Classification* [1] for a comprehensive examination.

Stability

Definition 1 (Uniformly β-stable algorithm)**.**

$$h_s = \mathcal{A}(S), h_s \in \mathcal{H}$$

Algorithm $\mathcal{A}$ *is stable if* $\forall (s, z), \forall i = \{1, ..., n\}$

$$\sup_{z' \in \mathcal{Z}} |l(h_s, z') - l(h_{s^{i,z}}, z')| \leq \beta$$

where S *is the data set,* z *is an evaluation sample and* $S_{i,z}$ *refers to replacing the* i^{th} *element in* S *with* z*.*

Theorem 2.

$$R[h_s] \leq \hat{R}_s[h_s] + \beta + ... + (\beta n + \frac{M}{2})\sqrt{\frac{2\ln 2/\delta}{n}}$$

The term $(\beta n + \frac{M}{2})\sqrt{\frac{2\ln 2/\delta}{n}}$ is $O(\beta\sqrt{n})$. Informally, an algorithm is stable if $\beta = O(\frac{1}{n})$. If stability is $O(\frac{1}{\sqrt{n}})$, this term is $O(1)$ and we can no longer show decrease in generalization gap with with increase in n.

Empirical Risk Minimization + Regularization is Stable

Notation:

$$\hat{R}_S(w) \triangleq \hat{R}_S(h_w)$$

where h_w is a model parameterized by weights w.

$$l(h, z) \equiv l(h(x), y)$$
$$l(h_w, z) \equiv l(w, z)$$

Theorem 3 (ERM with regularization is β-stable)**.**

$$f_S(w) = \hat{R}_S(w) + \frac{\lambda}{2}||w||_2^2$$

Proof. Consider weights u, v for two different models.

$$f_S(v) - f_S(u) = [\hat{R}_S(v) + \frac{\lambda}{2}||v||_2^2] - [\hat{R}_S(u) + \frac{\lambda}{2}||u||_2^2]$$

We perturb the dataset by replacing the data point at i with z'_i. Now we get:

$$\begin{aligned} f_S(v) - f_S(u) &= \hat{R}_{S_{i,z'_i}}(v) + \lambda||v||_2^2 - (\hat{R}_{S_{i,z'_i}}(u) + \frac{\lambda}{2}||u||_2^2) + \frac{l(v, z_i) - l(v, z'_i)}{n} - \frac{l(u, z_i) - l(u, z'_i)}{n} \\ &= f_{S_{i,z'_i}}(v) - f_{S_{i,z'_i}}(u) + \frac{l(v, z_i) - l(v, z'_i)}{n} - \frac{l(u, z_i) - l(u, z'_i)}{n} \end{aligned}$$

Now we substitute $v = \mathcal{A}(S_{i,z'_i})$ and $u = \mathcal{A}(S)$.

$$\begin{aligned} f_S(\mathcal{A}(S_{i,z'_i})) - f_S(\mathcal{A}(S)) = & f_{S_{i,z'_i}}(\mathcal{A}(S_{i,z'_i})) - f_{S_{i,z'_i}}(\mathcal{A}(S)) \\ & + \frac{l(\mathcal{A}(S_{i,z'_i}), z_i) - l(\mathcal{A}(S_{i,z'_i}), z'_i)}{n} - \frac{l(\mathcal{A}(S), z_i) - l(\mathcal{A}(S), z'_i)}{n} \end{aligned}$$

Because

$$f_{S_{i,z'_i}}(\mathcal{A}(S_{i,z'_i})) = \min_w f_{S_{i,z'_i}}(w)$$
$$\implies \forall w f_{S_{i,z'_i}}(w) \geq f(S_{i,z'_i})(\mathcal{A}(S_{i,z'_i}))$$

Assumption 4. *$l(\cdot|z)$ is L-Lipschitz.*

$$\begin{aligned} f_S(\mathcal{A}(S_{i,z'_i})) - f_S(\mathcal{A}(S)) &\leq \frac{l(\mathcal{A}(S^{i,z'_i}), z_i) - l(\mathcal{A}(S), z_i)}{n} - \frac{l(\mathcal{A}(S^{i,z'_i}), z'_i) - l(\mathcal{A}(S), z'_i)}{n} \\ &\leq 2\frac{L}{n}||\mathcal{A}(S) - \mathcal{A}(S_{i,z'_i})||_2 \end{aligned} \quad (1)$$

Assumption 5. *$\hat{R}_S(w)$ is cvx.*

Which gives us $f_S(w)$ is λ-str cvx. Now we perform a Taylor expansion:

$$f_S(\mathcal{A}(S_{i,z'_i})) - f_S(\mathcal{A}(S)) \geq \frac{\lambda}{2}||\mathcal{A}(S_{i,z'_i}) - \mathcal{A}(S)||_2^2 \quad (2)$$

Since $\mathcal{A}(S)$ is the minimizer of f_s and λ-str cvx the first term disappears.
From 1 and 2 we get:

$$||\mathcal{A}(S) - \mathcal{A}(S_{i,z'_i})|| \leq \frac{4L}{\lambda n} \quad (3)$$

If we perturb the data by a single element, we learn $\mathcal{A}$ that can become arbitrarily close for large n.
We then use 3 and the L-Lipschitz property of $l(\cdot, z)$:

$$\implies \sup_z [l(\mathcal{A}(S), z) - l(\mathcal{A}(S_{i,z'_i}), z)| \leq \frac{4L^2}{\lambda n}$$

□

Stochastic Gradient Descent (SGD) is Stable

Stability Theorem

Recall the SGD update formula,

$$w_{t+1} = w_t - \alpha_t \nabla_w l(w_t, z_{i,t}), i_t \sim \text{uniform}(1, \cdots, n) \tag{4}$$

where w_t is the weight iterate at time t, α_t is an (annealing) learning rate at time t and $l(w_t, z_{i,t})$ is the computed loss for the current weight iterate for a particular example $z_{i,t}$.

Theorem 6. *If $f(\cdot, z)$ is γ-smooth, convex and L-Lipschitz, then*

$$\beta \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t$$

Analysis:
We are no longer requiring the function to be strongly convex. Additionally, this result holds for a finite number of steps T.

Stability Proof (Rough Outline)

We will consider two runs of the SGD algorithm. One run will be on the original data set S and the other run will be on the data set $S_{i,z_i'}$. Recall, this indicates the same data set S only now with the i^{th} element swapped with element z_i'. In order to compare the stability between the two runs, we maintain the same order of element selection (same random seed) for $t = 1, \cdots, T$.

Definition 7.

$$\delta_t = ||w_t - w_t'||$$

where w_t' denotes the iterate for the SGD algorithm on the data set $S_{i,z_i'}$.
We can write the expectation of the difference δ_{t+1} as the following:

$$E[\delta_{t+1}] = P(i_t = i)E[\delta_{t+1}|i_t = 1] + P(i_t \neq i)E[\delta_{t+1}|i_t \neq 1] \tag{5}$$

We introduce two Lemmas

Lemma 0.1. *We may use co-coercivity to show*

$$E[\delta_{t+1}|i_t \neq 1] \leq E[\delta_t]$$

Lemma 0.2. *And for the index that has been swapped*

$$E[\delta_{t+1}|i_t = 1] \leq E[\delta_t] + 2\alpha_t L$$

where L is the Lipschitz value.

Using Lemmas 0.1, 0.2, we may rewrite Equation 5 as:

$$E[\delta_{t+1}] \leq \left(1 - \frac{1}{n}\right) E[\delta_t] + \frac{1}{n}\left(E[\delta_t] + 2\alpha_t L\right) \tag{6}$$

which when recursively unrolled yields the following final δ_T

$$E[\delta_T] = E[||w_T - w_T'||] \leq \sum_{t=0}^{T-1} \frac{2\alpha_t L}{n} \tag{7}$$

SGD is therefore **stable** since $\sum_{t=0}^{T-1} \frac{2\alpha_t L}{n} \equiv \beta$ is $O(\frac{1}{n})$ for n data points.

References

[1] J. Langford. Tutorial on practical prediction theory for classification. *J. Mach. Learn. Res.*, 6:273–306, Dec. 2005. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1046920.1058111`.