Generalization Bounds and Stability

9.520 Class 06, 26 February 2003

Alex Rakhlin

Plan

- Generalization Bounds
- Stability
- Generalization Bounds Using Stability

Algorithms

We define an algorithm $\mathcal{A}$ to be a mapping from a training set $S = \{z_1, \ldots, z_\ell\}$ to a function f_S. Here, $z_i \equiv (\mathbf{x}_i, y_i)$.

Throughout the next several lectures, we assume that $\mathcal{A}$ is deterministic, and that A does not depend on the ordering of the points in the training set. These assumptions are not very restrictive, but greatly simplify the math.

How can we measure "goodness" of f_S?

Risks

Recall that in Lecture 2 we've defined the true (expected) risk:

$$I[f_S] = \mathbb{E}_{(\mathbf{x},y)}\left[V(f_S(\mathbf{x}),y)\right] = \int V(f_S(\mathbf{x}),y)d\mu(\mathbf{x},y)$$

and the empirical risk:

$$I_S[f_S] = \frac{1}{\ell}\sum_{i=1}^{\ell} V(f_S(\mathbf{x}_i),y_i).$$

Note: the true and empirical risks are denoted in Bousquet & Elisseeff as $R(\mathcal{A},S)$ and $\hat{R}(\mathcal{A},S)$, respectively, to emphasize the algorithm that produced f_S.

Note: the loss is sometimes written as $c(f,z) = V(f(\mathbf{x}),y)$, where $z = (\mathbf{x},y)$.

Generalization Bounds

Our goal is to choose an algorithm $\mathcal{A}$ so that $I[f_S]$ will be small. This is difficult because we can't measure $I[f_S]$.

We can, however, measure $I_S[f_S]$. A **generalization bound** is a (probabilistic) bound on how big the defect

$$D[f_S] = I[f_S] - I_S[f_S]$$

can be. If we can bound the defect and we can observe that $I_S[f_S]$ is small, then $I[f_S]$ must be small.

Note that this is **consistency**, as we've defined in Lect. 2: $D[f_S] \to 0$, as $\ell \to \infty$.

Properties of Generalization Bounds, I

What will a generalization bound depend on? A generalization bound is a way of saying that the performance of a function on the training set has to be similar to its performance on future examples. For this reason, generalization bounds are always **probabilistic**: they hold with some (high) probability, to take into account the (low) chance that you'll see a very unrepresentative training set.

Properties of Generalization Bounds, II

Generalization bounds depend on some measure of the size of the hypothesis space we allow ourselves to choose from. As the hypothesis space gets smaller, the generalization bound will get tighter (but the empirical performance will often go down). As the hypothesis space gets bigger, the generalization bound will get looser.

The bound will depend on the number of samples we have. In general, we would like the bounds to get tighter at least as fast as $\frac{1}{\sqrt{\ell}}$.

Properties of Generalization Bounds, III

A good generalization bound will **not** depend on the probability distribution P from which the examples are drawn. If it did, we couldn't measure it, since P is unknown.

Generalization Bounds By Bounding the Hypothesis Space

In 9.520, we discuss two different ways to obtain generalization bounds:

One way is to explicitly bound the size of the hypothesis space $\mathcal{H}$. For example, functions in an RKHS with $||f||_K^2 \leq M$ form a bounded hypothesis space whose "size" can be measured and used to obtain generalization bounds (recall uGC classes of functions).

$$\mathbb{P}_S \left(\sup_{f \in \mathcal{H}} |I_S[f] - I[f]| > \epsilon \right) < \delta$$

This approach will be discussed in future lectures.

Generalization Bounds By Stability

The other approach is to use **stability** of algorithms. Here, the basic idea is that we bound how much the function produced by an algorithm can change when we modify the training set slightly. In this class and the next class, we will explain and develop this approach to generalization bounds, and show that Tikhonov reguarization in an RKHS exhibits the necessary stability.

Note that in this approach we are not concerned with "good performance" of **all** functions, but only the one produced by our algorithm:

$$\mathbb{P}_S \left(|I_S[f_S] - I[f_S]| > \epsilon \right) < \delta$$

Uniform Stability

Given a training set S, we define $S^{i,z}$ to be the new training set obtained when point i of S is replaced by the new point $z \in \mathcal{Z}$. Given this definition, we say that an algorithm $\mathcal{A}$ has **uniform stability** β (is β-stable) if

$$\forall (S, z) \in \mathcal{Z}^{\ell+1},\ \forall i,\ \sup_u |c(f_S, u) - c(f_{S^{i,z}}, u)| \leq \beta.$$

An algorithm is β-stable if, for any possible training set, we can replace an arbitrary training point with any other possible training point, and the loss at any point will change by no more than β.

Uniform Stability Cont'd

Uniform stability is a strong requirement, because it ignores the fact that the points are drawn from a probability distribution. For uniform stability, the function still has to change very little even when a very unlikely ("bad") training set is drawn.

In general, the stability β is a function of ℓ, and should perhaps be written β_ℓ.

Stability and Concentration Inequalities

Question: Given that an algorithm $\mathcal{A}$ has stability β, how can we get bounds on its performance?

Answer: Concentration Inequalities. In particular, we will use McDiarmid's Inequality.

Concentration Inequalities show how a variable is concentrated around its mean.

Michel Talagrand:

A random variable that depends (in a "smooth" way) on the influence of many independent variables (but not too much on any of them) is essentially constant.

McDiarmid's Inequality

Given random variables $v_1, \ldots, v_\ell$, and a function $F : v^\ell \to \mathbb{R}$ satisfying

$$\sup_{v_1,\ldots,v_\ell,v_i'} |F(v_1, \ldots, v_\ell) - F(v_1, \ldots, v_{i-1}, v_i', v_{i+1}, \ldots, v_\ell)| \leq c_i,$$

the following statement holds:

$$\mathbb{P}\left(|F(v_1, \ldots, v_\ell) - \mathbb{E}_S(F(v_1, \ldots, v_\ell))| > \epsilon\right) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{\ell} c_i^2}\right).$$

This is an example of the law of large numbers.

Example: Hoeffding's Inequality

Suppose each $v_i \in [a,b]$, and we define $F(v_1,\ldots,v_\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} v_i$, the average of the v_i. Then, $c_i = \frac{1}{\ell}(b-a)$. Applying McDiarmid's Inequality, we have that

$$\begin{aligned}
\mathbb{P}\left(|F(\mathbf{v}) - \mathbb{E}(F(\mathbf{v}))| > \epsilon\right) &\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{\ell} c_i^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{\ell}(\frac{1}{\ell}(b-a))^2}\right) \\
&= 2\exp\left(-\frac{2\ell\epsilon^2}{(b-a)^2}\right).
\end{aligned}$$

We have easily recovered the famous "Hoeffding's Inequality". (Of course, we did not prove McDiarmid's Inequality.)

Generalization Bounds via McDiarmid's Inequality

We will use β-stability to apply McDiarmid's inequality to the defect $D[f_S] = I[f_S] - I_S[f_S]$. To do this, we will need two things:

1. the expectation of the defect (we can't measure it, but we can bound its expectation) and

2. a bound on how much the defect can change when we replace a point.

In order to bound the deviation (the second quantity), we require that there exist an upper bound M on the loss.

Bounding The Expectation of The Defect

$$
\begin{aligned}
\mathbb{E}_S D[f_S] &= \mathbb{E}_S \left[I_S[f_S] - I[f_S] \right] \\
&= \mathbb{E}_{S,z} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} V(f_S(\mathbf{x}_i), y_i) - V(f_S(\mathbf{x}), y) \right] \\
&= \mathbb{E}_{S,z} \left[\frac{1}{\ell} \sum_{i=1}^{\ell} V(f_{S^{i,z}}(\mathbf{x}), y) - V(f_S(\mathbf{x}), y) \right] \\
&\leq \beta
\end{aligned}
$$

The second equality follows by exploiting the "symmetry" of expectation: The expected value of a training set on a training point doesn't change when we "rename" the points.

Bounding The Deviation of The Defect

$$
\begin{aligned}
|D[f_S] - D[f_{S^{i,z}}]| &= |I_S[f_S] - I[f_S] - I_{S^{i,z}}[f_{S^{i,z}}] + I[f_{S^{i,z}}]| \\
&\leq |I[f_S] - I[f_{S^{i,z}}]| + |I_S[f_S] - I_{S^{i,z}}[f_{S^{i,z}}]| \\
&\leq \beta + \frac{1}{\ell}|V(f_S(\mathbf{x}_i), y_i) - V(f_{S^{i,z}}(\mathbf{x}), y)| \\
&\quad + \frac{1}{\ell}\sum_{j \neq i} |V(f_S(\mathbf{x}_j), y_j) - V(f_{S^{i,z}}(\mathbf{x}_j), y_j)| \\
&\leq \beta + \frac{M}{\ell} + \beta \\
&= 2\beta + \frac{M}{\ell}
\end{aligned}
$$

Applying McDiarmid's Inequality

By McDiarmid's Inequality, for any ϵ,

$$\begin{aligned}\mathbb{P}\left(|D[f_S] - \mathbb{E}D[f_S]| > \epsilon\right) &\le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{\ell}(2(\beta + \frac{M}{\ell}))^2}\right) = \\ = 2\exp\left(-\frac{\epsilon^2}{2\ell(\beta + \frac{M}{\ell})^2}\right) &= 2\exp\left(-\frac{\ell\epsilon^2}{2(\ell\beta + M)^2}\right)\end{aligned}$$

Note that

$$\begin{aligned}\mathbb{P}(D[f_S] > \beta + \epsilon) &= \mathbb{P}(D[f_S] - \mathbb{E}D[f_S] > \epsilon) \\ &\le \mathbb{P}(|D[f_S] - \mathbb{E}D[f_S]| > \epsilon)\end{aligned}$$

Hence,

$$\mathbb{P}(I_S[f_S] - I[f_S] > \beta + \epsilon) \le 2\exp\left(-\frac{\ell\epsilon^2}{2(\ell\beta + M)^2}\right)$$

A Different Form Of The Bound

If we define

$$\delta \equiv 2\exp\left(-\frac{\ell\epsilon^2}{2(\ell\beta + M)^2}\right).$$

Solving for ϵ in terms of δ, we find that

$$\epsilon = (\ell\beta + M)\sqrt{\frac{2\ln(2/\delta)}{\ell}}.$$

By varying δ (and ϵ), we can say that for any $\delta \in (0, 1)$, with probability $1 - \delta$,

$$I[f_S] \leq I_S[f_S] + \beta + (\ell\beta + M)\sqrt{\frac{2\ln(2/\delta)}{\ell}}.$$

Fast Convergence

Note that if $\beta = \frac{k}{\ell}$ for some k, we can restate our bounds as

$$P\left(|I[f_S] - I_S[f_S]| \geq \frac{k}{\ell} + \epsilon\right) \leq 2\exp\left(-\frac{\ell\epsilon^2}{2(k+M)^2}\right),$$

and with probability $1-\delta$,

$$I[f_S] \leq I_S[f_S] + \frac{k}{\ell} + (2k+M)\sqrt{\frac{2\ln(2/\delta)}{\ell}}.$$

Fast Convergence, Cont'd

For the uniform stability approach we've described, $\beta = \frac{k}{\ell}$ (for some constant k) is "good enough". Obviously, the best possible stability would be $\beta = 0$ — the function can't change at all when you change the training set. An algorithm that always picks the same function, regardless of its training set, is maximally stable and has $\beta = 0$. Using $\beta = 0$ in the last bound, with probability $1 - \delta$,

$$I[f_S] \leq I_S[f_S] + M\sqrt{\frac{2\ln(2/\delta)}{\ell}}.$$

The convergence is still $O\left(\frac{1}{\sqrt{\ell}}\right)$. So once $\beta = O(\frac{1}{\ell})$, further increases in stability don't change the rate of convergence.

Other kinds of stabilities

Notation: $c(f,z) = V(f(x),y)$ for $z=(x,y)$. $\forall^\delta$ means "for all except a set of measure δ.

An algorithm $\mathcal{A} : Z^\ell \to \mathcal{F}$ is

uniformly β hypothesis stable:

$$\forall i, (S,u) \in \mathcal{Z}^{\ell+1},\ \sup_{z\in\mathcal{Z}}\{|c(f_S,z) - c(f_{S^{i,u}},z)|\} \leq \beta.$$

(β,δ) leave-one-out stable:

$$\forall^\delta S, \forall i,\ \left|c(f_S,z_i) - c(f_{S^i},z_i)\right| \leq \beta.$$

(β,δ) error stable:

$$\forall^\delta (S,u), \forall i,\ \left|I[f_S] - I[f_{S^{i,u}}]\right| \leq \beta.$$

(β,δ) cross-validation stable:

$$\forall^\delta S \in \mathcal{Z}^\ell, \forall i, u \in \mathcal{Z},\ \left|c(f_S,u) - c(f_{S^{i,u}},u)\right| \leq \beta.$$

Thoughts on stability and open questions

Stability is a new research area – many things to be done.

The "right" definition of stability is still an open question.

Good generalization bounds can be proved for specific algorithms if certain types of stabilities can be shown.

There might be a way to apply other concentration inequalities to get $O\left(\frac{1}{\ell}\right)$ convergence.

Summary

We used McDiarmid's inequality to prove a generalization bound for a uniformly β-stable algorithm. Note that this bound cannot tell us that the expected error will be low *a priori*, it can only tell us that with high probability, *the expected error will be close to the empirical error.* We have to actually observe a low empirical error to conclude that we have a low expected error.

Uniform stability of $O\left(\frac{1}{\ell}\right)$ seems to be a strong requirement. Next time, we will show that Tikhonov regularization possesses this property.