Uniform Convergence and Consistency

9.520 Class 8, 5 March 2002

Sayan Mukherjee and Alex Rakhlin

Plan

- Mappings and hypothesis spaces
- A brief review of stability
- Uniform convergence
- Uniform convergence is necessary and sufficient for consistency
- Consistency when there is one function in the hypothesis space
- Consistency when there are a finite number of functions in the hypothesis space
- Consistency for RKHS and Ivanov regularization
- Covering numbers

Mappings and hypothesis spaces

We have discussed how the stability of the learning map $\mathcal{A} : S \to f_S$ can be used to obtain generalization bounds, i.e. consistency. We also noted that the uniform stability described is a strong condition and there are mappings that are consistent but do not have this strong notion of stability.

We now look from a different perspective: controlling the hypothesis space. The mapping and the hypothesis space can be related as follows:
the hypothesis space $\mathcal{H}$ is the space of all possible functions that the map $\mathcal{A} : S \to f_S$ can output given all possible sets S.

We will use properties of the hypothesis space to get generalization bounds, ie show consistency.

Generalization bounds by controlling the hypothesis space

We discussed how the stability of a map can be used to obtain generalization bounds. We now control the size of hypothesis spaces to obtain generalization bounds.

For example, functions in an RKHS with $||f||_K^2 \leq M$ form a totally bounded hypothesis space whose "size" can be measured and used to obtain generalization bounds. This approach will be discussed now.

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:

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

Generalization Bounds

Our goal is to choose a function f_S 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** for ERM, as we've defined in Lect. 2: $D[f_S] \to 0$, as $\ell \to \infty$.

Uniform convergence

By uniform convergence we mean for any $\varepsilon > 0$

$$\lim_{\ell\to\infty} \mathbb{P}\left\{\sup_{f\in\mathcal{H}} |I_S[f] - I[f]| \geq \varepsilon\right\} \to 0.$$

Note this is two-sided uniform convergence.

Function classes that satisfy this property are uGC classes.

We will show

$$\mathbb{P}\left\{\sup_{f\in\mathcal{H}} |I_S[f] - I[f]| \geq \varepsilon\right\} \leq \Phi(\mathcal{H}, \varepsilon, \ell)\,\mathrm{exp}(-C\varepsilon^2\ell).$$

Consistency

Empirical risk minimization is consistent if in probability

$$\lim_{\ell \to \infty} \left\{ I_S[f_S] \overset{p}{=} \inf_{f \in \mathcal{H}} I[f] \right\}$$

or for any given $\varepsilon > 0$

$$\lim_{\ell \to \infty} \mathbb{P} \left\{ \left| I_S[f_S] - \inf_{f \in \mathcal{H}} I[f] \right| > \varepsilon \right\} \to 0$$

The expected error of the minimizer of the empirical error converges to the best possible expected error in the hypothesis space $\mathcal{H}$.

Question: Is Tikhonov regularization consistent ?
Answer: Depends on how fast the regularization parameter λ decays.

A key Lemma

The following statements are equivalent:

- For any distribution in a set, the empirical risk minimization method is consistent on the set of functions $f \in \mathcal{H}$
- For any distribution in a set, the uniform convergence of the empirical error to the expected error takes place on the set of functions $f \in \mathcal{H}$.

Consistency of ERM $\Leftrightarrow$ uniform convergence $\Leftrightarrow$ uGC.

Uniform convergence for one function

If our function space consists of one function f_1, we can show

$$\mathbb{P}\left\{|I_S[f_1] - I[f_1]| \leq \epsilon\right\} \geq 1 - 2\exp(-\epsilon^2 \ell C).$$

If our loss function is bounded $0 \leq V(f_1(x), y) \leq B$ then we can use Hoeffding's inequality: let X be a set and D a distribution on X with functions $h : X \to [a, b]$ then

$$\mathbb{P}\left\{\left|\frac{1}{\ell}\sum_{i=1}^{\ell} h(x_i) - E_D h(x)\right| \geq \epsilon\right\} \leq 2\exp(-2\epsilon^2 \ell/(a-b)^2).$$

Applying the inequality to the loss results in

$$\mathbb{P}\left\{|I_S[f_1] - I[f_1]| \leq \epsilon\right\} \geq 1 - 2\exp(-2\epsilon^2 \ell/B^2).$$

Uniform convergence for k functions

If our RKHS consisted of k functions $f_1, ..., f_k$ we can show

$$\mathbb{P}\left\{\max_{f_i:i=1,..,k} |I_S[f_i] - I[f_i]| \leq \epsilon\right\} \geq 1 - 2k\exp(-\epsilon^2 \ell C).$$

We know that for each function f_i

$$\mathbb{P}\left\{|I_S[f_i] - I[f_i]| \leq \epsilon\right\} \geq 1 - 2\exp(-2\epsilon^2 \ell / B^2).$$

We need the above to hold for all k functions. So we apply the union bound (Bonferroni approximation)

$$P(a \cup b \cup c) \leq P(a) + P(b) + P(c)$$

so

$$\mathbb{P}\left\{\max_{f_i:i=1,..,k} |I[f_i] - I[f_i]| \leq \epsilon\right\} \geq 1 - 2k\exp(-2\epsilon^2 \ell / B^2).$$

Ivanov regularization

In the literature this is called *empirical risk minimization* within a restricted hypothesis space.

The functional we minimize has the form:

$$
\begin{aligned}
f_S \quad &= \quad \operatorname{arg}\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(x_i), y_i) \\
\text{s.t.} \quad & \quad \|f\|_K^2 \leq M.
\end{aligned}
$$

Tikhonov $\Rightarrow$ Ivanov regularization

We saw in the last class that for any Lipschitz loss function

$$\|f\|_K^2 \leq \frac{C_0}{\lambda},$$

where $\mathcal{Y} \leq C_0$.

So we can take our Tikhonov problem and solve instead

$$\begin{aligned} f_S \;&=\; \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(x_i), y_i) \\ \text{s.t.} \quad & \|f\|_K^2 \leq \frac{C_0}{\lambda}. \end{aligned}$$

Uniform convergence for totally bounded RKHS (intuition)

In general our RKHS are not a finite set of functions. Instead they have the form $f \in \mathcal{H}$ with $||f||_K^2 \leq M$.

We will count the number of functions in this space using an ϵ-net. That is we can pick some N functions $g_1, ..., g_N \in \mathcal{H}$ for which $\exists\, g_i$ such that $\forall f \;\; d(f, g_i) \leq \epsilon$.

We will use

$$d(f, g_i) = \|f - g_i\|_K^2 \text{ or } d(f, g_i) = \sup_x \|f - g_i\|.$$

Why is N finite for a totally bounded RKHS ?

Uniform convergence for totally bounded RKHS (more intuition)

The **covering number** $\mathcal{N}(\mathcal{H}, r)$ is the minimal $m \in \mathbb{N}$ such that there exists m disks in $\mathcal{H}$ with radius r covering $\mathcal{H}$.

Now we can show using the same argument as we used for k functions that

$$\mathbb{P}\left\{\sup_{f \in \mathcal{H}: \|f\|_K^2 \leq M} |I_S[f] - I[f]| \leq \epsilon\right\} \geq 1 - 2\mathcal{N}(\mathcal{H}(M), r(\epsilon), d) \exp(-2\epsilon^2 \ell / B^2).$$

Computing covering numbers in finite dimensional RKHS

For a finite dimensional bounded RKHS

$$\mathcal{H}_K = \left\{ f : f(x) = \sum_{p=1}^{m} c_p \phi_p(x) \right\},$$

with $\|f\|_K^2 \leq M$.

We want to compute $\mathcal{N}(\mathcal{H}, r, d)$.

Computing covering numbers in finite dimensional RKHS (cont)

Each function g_i can be written as

$$g_i(x) = \sum_{p=1}^{m} d_{ip}\phi_p(x)$$

so we now rephrase the problem as finding m vectors d_i for which

$$\sum_{p=1}^{m} \frac{(c_p - d_{ip})^2}{\lambda_p} < r^2.$$

Computing covering numbers in finite dimensional RKHS (cont)

This is equivalent asking how many balls of radius r are required to cover a ball of radius M in $\mathbb{R}^m$ using the Euclidean metric.

Instead of the covering numbers we will compute packing numbers.

N functions $g_1, ..., g_N$ are r-separated if $d(g_i, g_j) > r$ for $i \neq j$. The packing number $\mathcal{D}(\mathcal{H}, r, d)$ is the maximal cardinality of r-separated sets.

A fact

$$\mathcal{D}(\mathcal{H}, 2r, d) \leq \mathcal{N}(\mathcal{H}, r, d) \leq \mathcal{D}(\mathcal{H}, r, d).$$

Computing packing numbers in finite dimensional RKHS (cont)

How many balls of radius r are required to pack a ball of radius M in $\mathbb{R}^m$ using the Euclidean metric

$$\mathcal{D} \leq \left(\frac{3M}{r}\right)^m.$$

So we have

$$\begin{aligned}
\mathcal{N}(\mathcal{H}(M), r, \|\cdot\|_K) &\leq \left(\frac{3M}{r}\right)^m \\
\mathcal{N}(\mathcal{H}(M), r, \|\cdot\|_\infty) &\leq \left(\frac{3M}{r\kappa}\right)^m.
\end{aligned}$$

A quantity that will show up often is **metric entropy**

$$\log \mathcal{N}(\mathcal{H}(M), r).$$

Computing covering numbers in infinite dimension RKHS

In an infinite dimensional bounded RKHS we know our space is defined as

$$\mathcal{H}_K = \left\{ f : f(x) = \sum_{p=1}^{\infty} c_p \phi_p(x) \right\},$$

and $||f||_K^2 \leq M$.

Note the above results cannot be used since they are exponential in the dimensionality of the RKHS space.

The covering numbers are finite for the inclusion

$$I_K : \mathcal{H}_K \hookrightarrow \mathcal{C}(X)$$

which we will write $\overline{I_K(B_R)}$.

Computing covering numbers in infinite dimension RKHS (cont.)

We first state three results about the covering numbers for infinite dimensional RKHS.

If the eigenvalues of the kernel function decay exponentially then

$$\mathcal{N}(\mathcal{H}(M), r) \leq \left(\frac{CM}{r}\right)^{m+1}$$

where d is the dimensionality of the input space.

If the kernel is C_∞ then

$$\mathcal{N}(\mathcal{H}(M), r) \leq \exp\left(\left(\frac{C_h M}{r}\right)^{2m/h}\right)$$

where h is a finite integer.

Computing covering numbers in infinite dimension RKHS (cont.)

If the kernel is C_s (and s is odd) then

$$\mathcal{N}(\mathcal{H}(M), r) \leq \exp\left(\left(\frac{CM}{r}\right)^{2m/s}\right).$$

The following relation for metric entropy (the log of the covering number)

$$\ln(\mathcal{N}(\mathcal{H}(M), r)) = O\left(\left(\frac{CM}{r}\right)^{2m/s}\right)$$

is very classical and standard with the m/s tradeoff as the key aspect.

Computing covering numbers in infinite dimension RKHS (cont.)

The proof for the first result

$$\mathcal{N}(\mathcal{H}(M), r) \leq \left(\frac{CM}{r}\right)^{d+1}$$

is conceptually very similar to the Hilbert cube example (see Mathcamp 1).

The other two results require knowledge of Sobolev spaces and embedding in Sobolev spaces.

Sobolev spaces and embeddings

A function belongs to a Sobolev space $W^{m,p}(\mathbb{R}^d)$ if it and all its partial derivatives up to m belong to $L^P(\mathbb{R}^d)$. If the following norm is finite then a function is in $W^{m,2}(\mathbb{R}^d)$

$$||f||^2_{H^2_m(\mathbb{R}^d)} = \sum_{\alpha=1}^{m} \int_{\mathbb{R}^d} |\nabla^\alpha f|^2.$$

Zhou 2002: "Capacity of RKHS in Learning Thoery", Preprint

If a kernel that satisfies Mercer's theorem is C_s (for some odd $s > 0$) the RKHS associated with the kernel can be embedded into $C_{s/2}$. This allows us to upper bound the covering number of the RKHS using covering and packing number results for Sobolev spaces.

Covering numbers in Sobolev spaces

The following two results were computed using facts about covering numbers of Sobolev spaces.

If the kernel is C_∞ then

$$\mathcal{N}(\mathcal{H}(M), r) \leq \exp\left(\left(\frac{C_h M}{r}\right)^{2d/h}\right)$$

where h is an index of the Sobolev space embedded into.

If the kernel is C_s (and s is odd) then

$$\mathcal{N}(\mathcal{H}(M), r) \leq \exp\left(\left(\frac{C M}{r}\right)^{2d/s}\right).$$

Computing $r(\epsilon)$

For a variety of kernels we now know how to compute $\mathcal{N}(\mathcal{H}(M), r(\epsilon))$.

We now have to compute $r(\epsilon)$ to finish the bounds and this function will depend on the loss function used.

We fist look at the square loss case $V(f(x), y) = (f(x) - y)^2$ with the requirement that $|f(x) - y| < B' \quad \forall f(x)$ and y.

In this case $r(\epsilon) = \frac{\epsilon}{8B'}$.

Computing $r(\epsilon)$ for square loss

We first estimate the quantity

$$|I[f_1] - I_S[f_1] - I[f_2] + I_S[f_2]|.$$

$$|I[f_1]-I_S[f_1]-I[f_2]+I_S[f_2]| \leq |I[f_1]-I[f_2]|+|I_S[f_1]-I_S[f_2]|$$

and

$$\begin{aligned} |I[f_1] - I[f_2]| &= |\int (f_1(x) - f_2(x))(f_1(x) - f_2(x) - 2y)| \\ &\leq ||f_1 - f_2||_\infty \int |(f_1(x) - y) + (f_2(x) - y)| \\ &\leq 2B'||f_1 - f_2||_\infty. \end{aligned}$$

Computing $r(\epsilon)$ for square loss (cont)

Similarly

$$\begin{aligned}|I_S[f_1] - I_S[f_2]| &= \left|\frac{1}{\ell}\sum(f_1(x_i) - f_2(x_i))(f_1(x_i) - f_2(x_i) - 2y_i)\right| \\ &\leq 2B'||f_1 - f_2||_\infty.\end{aligned}$$

So

$$|I[f_1] - I_S[f_1] - I[f_2] + I_S[f_2]| \leq 4B'||f_1 - f_2||_\infty.$$

If we set $r = \frac{\epsilon}{4B'}$ then we know that

$$|I[f_1] - I_S[f_1] - I[f_2] + I_S[f_2]| \leq 4B'||f_1 - f_2||_\infty \leq 4B'\frac{\epsilon}{4B'} = \epsilon.$$

Computing $r(\epsilon)$ for square loss (cont)

From this we know within a cover D_j and f_j the j^{th} prototype function

$$\sup_{f \in D_j} |I[f] - I_S[f]| \geq 2\epsilon \Rightarrow |I[f_j] - I_S[f_j]| \geq \epsilon.$$

If we look at the $\mathcal{N}$ prototype functions $f_1, ..., f_{\mathcal{N}}$ used in the cover the following holds true for each cover:

$$\begin{aligned} \mathbb{P}\left\{\sup_{f \in D_j} |I[f] - I_S[f]| \geq 2\epsilon\right\} &\leq \mathbb{P}\left\{|I[f_j] - I_S[f_j]| \geq \epsilon\right\} \\ &\leq 2\exp(-\epsilon^2 \ell / B^2). \end{aligned}$$

Generalization bound for square loss (cont)

This gives us our relation

$$r(\epsilon) = \frac{\epsilon}{8B'}.$$

Which gives us the following bound

$$\mathbb{P}\left\{\sup_{f\in\mathcal{H}:\|f\|_K^2\le M} |I[f] - I_S[f]| \le \epsilon\right\} \;\le\; 1 - 2\mathcal{N}\left(\mathcal{H}(M), \frac{\epsilon}{8B'}\right) \exp(-\epsilon^2\ell/B^2).$$

This can be rewritten: with probability $1-\delta$

$$I[f] \le I_S[f] + \sqrt{\frac{B}{\ell}\left(\log\mathcal{N} + \log\left(2/\delta\right)\right)}.$$

Lipschitz Loss Functions

For Lipschitz loss functions we can also compute $r(\epsilon)$

Recall a loss function (over a possibly bounded domain $\mathcal{X}$) is Lipschitz with Lipschitz constant L if

$$\forall y_1, y_2, y' \in \mathcal{Y}, \; |V(y_1, y') - V(y_2, y')| \leq L|y_1 - y_2|.$$

Using much of the same algebra as for the square loss we can compute $r(\epsilon)$.

Sufficiency for uGC classes

If our hypothesis space is compact, has a finite cover, we have shown that we get uniform convergence and therefore consistency. Compactness is a sufficient condition for uGC classes.

Is compactness also necessary ?
How do we deal with $\{0, 1\}$ loss, which is not Lipschitz ?

Answer to these and other pressing issues next Monday.