Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter — Lecture 5
Not scribed — Th 2/22

1 Uniform Convergence of Empirical Risks

In the previous lecture, we established that if the empirical risk converges uniformly to the true risk, then minimizing empirical risk minimizes true risk asymptotically, which is what we desire. We also saw examples where uniform convergence occurs trivially: in finite and finite-range parameterizations. In today's lecture we will show a powerful generalization, which gives a sufficient (and in fact necessary, which we won't show) condition for uniform convergence, based on the complexity (annealed entropy) of the model class and the underlying distribution. We will also introduce a technique which provides a conservative sufficient condition (finite VC dimension) which is distribution-independent. We will elaborate on the latter in the next lecture.

1.1 Notation

The following are the notational conventions and basic defintitions we have been using:

- Our data sample, or sample set, is denoted by $z^\ell = (z_1, \cdots, z_\ell)$, where we have $z_i = (x_i, y_i)$.
- Our model class is $\mathcal{C} = \{f(\cdot, \alpha) : \alpha \in \Lambda\}$.
- We consider only indicator loss functions: $L(z_i, \alpha) = \begin{cases} 0 & \text{if } y_i = f(x_i, \alpha), \\ 1 & \text{otherwise.} \end{cases}$
- Given the underlying distribution, the true risk associated with model α is $R(\alpha) = \mathbf{E}[L(Z;\alpha)]$. The empirical risk for the same model given data z^ℓ is $R^\ell_{emp}(\alpha) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(z_i;\alpha)$.
- The set of all achievable loss ℓ-tuples given data z^ℓ is $B^\Lambda(z^\ell) = \{(L(z_1,\alpha), \cdots, L(z_\ell,\alpha)) : \alpha \in \Lambda\}$, itself a subset of all possible zero-one ℓ-tuples, i.e. $\{0,1\}^\ell$.
- The cardinality of $B^\Lambda(z^\ell)$ is denoted by $N^\Lambda(z^\ell) = |B^\Lambda(z^\ell)|$.
- We define the *annealed entropy* as $H^\Lambda_{annl}(\ell) = \ln \mathbf{E}[N^\Lambda(Z^\ell)]$. It is a quantity that we will interpret as a measure of the complexity of the model class together with the underlying distribution, and it will appear in the bound we are about to derive.

1.2 Statement of Main Results

The following theorem establishes an upper bound on the probability of the uniform deviation of the empirical risk from the true risk. The sufficient condition for uniform convergence is then just a matter of restating the theorem in a corollary.

Theorem 1. *Let $\mathcal{C}$ be a model class parametrized by $\alpha \in \Lambda$, and $L(z;\alpha)$ the corresponding indicator loss functions, then for all $\epsilon > 0$ the following holds:*

$$\mathbf{P}\left\{\sup_\alpha \left|\frac{1}{\ell}\sum_{i=1}^{\ell} L(Z_i;\alpha) - \mathbf{E}[L(Z;\alpha)]\right| > \epsilon\right\} \le 4\exp\left\{\left[\frac{H^\Lambda_{annl}(2\ell)}{\ell} - \left(\epsilon - \frac{1}{\ell}\right)^2\right]\ell\right\}.$$

We will prove this theorem in the next section. Meanwhile, the aforementioned corollary immediately follows.

Corollary. *If*

$$\frac{H^{\Lambda}_{annl}(2\ell)}{\ell} \longrightarrow 0 \qquad \text{as} \qquad \ell \to \infty,$$

then the property of uniform convergence of the empirical risk holds. The converse is also true.

Proof. We will not show the converse. For the forward direction, simply note that if the condition holds, then by choosing ℓ large enough, the exponent of the bound in Theorem 1 becomes negative, and therefore the bound converges to zero. □

1.3 Data, Precision, and Confidence

An alternative perspective into these results is to ask the question of what precision does a given amount ℓ of data provide within a certain desired probabilistic confidence $1-\eta$.

More concretely we have that:

$$\left|R^{\ell}_{emp}(\alpha) - R(\alpha)\right| < \epsilon(\ell, \eta),$$

with probability $1-\eta$, where

$$\epsilon(\ell, \eta) = \frac{1}{\ell} + \sqrt{\frac{H^{\Lambda}_{annl}(2\ell)}{\ell} - \frac{1}{\ell}\ln\frac{\eta}{4}}.$$

This follows from writing:

$$\eta = 4\exp\left\{\left[\frac{H^{\Lambda}_{annl}(2\ell)}{\ell} - \left(\epsilon - \frac{1}{\ell}\right)^2\right]\ell\right\},$$

which we can rearrange as

$$\frac{H^{\Lambda}_{annl}(2\ell)}{\ell} - \frac{1}{\ell}\ln\frac{\eta}{4} = \left(\epsilon - \frac{1}{\ell}\right)^2.$$

As a result, one can potentially solve for the appropriate value of ℓ to provide a given precision ϵ with confidence $1-\eta$.

2 Proof of the Main Theorem

We will divide the proof of Theorem 1 into three stages. In the first two stages, two different symmetries are exploited to transform the problem and derive bounds. In the final stage, a key idea is used to break the infinite parameter problem in a way similar to the finite case. Combined with the bounds derived in the first two stages, this yields the statement of the theorem.

2.1 Symmetrization with Ghost Samples

We begin with some notations and definitions. Let $z^{(1)}$ be a sample set of ℓ points, and let $z^{(2)}$ be an independent sample set of the same size (a 'ghost' sample set). Let $z^{2\ell} = z^{(1)}z^{(2)}$ be the extended sample set of 2ℓ points, formed by concatenating the original and ghost sample sets. Define the following:

$$\nu(z^{(1)}, \alpha) = \frac{1}{\ell}\sum_{i=1}^{\ell} L(z_i; \alpha),$$

$$\nu(z^{(2)}, \alpha) = \frac{1}{\ell}\sum_{i=\ell+1}^{2\ell} L(z_i; \alpha),$$

$$\Pi(z^{(1)}) = \sup_{\alpha}\left|\nu(z^{(1)}, \alpha) - \mathbf{E}[L(Z; \alpha)]\right|,$$

and

$$\rho(z^{2\ell}) = \sup_{\alpha}\left|\nu(z^{(1)}, \alpha) - \nu(z^{(2)}, \alpha)\right|.$$

Observe that we are interested in saying something about Π. Much in the spirit of Cauchy convergence, the following lemma tells us that we can alternatively look at the symmetrized version, i.e. ρ.

Lemma 1. *We have:*

$$\mathbf{P}\left\{\Pi(Z^{(1)}) > \epsilon\right\} \leq 2\mathbf{P}\left\{\rho(Z^{2\ell}) > \epsilon - \frac{1}{\ell}\right\}.$$

Proof. Let $\theta(u) = \mathrm{I}_{u>0}(u)$ be the indicator function of $\{u|u > 0\}$, and write:

$$\begin{aligned}
\mathbf{P}\left\{\rho(Z^{2\ell}) > \epsilon - \frac{1}{\ell}\right\} &= \int_{\mathcal{Z}^{2\ell}} \theta\left(\rho(z^{2\ell}) - \epsilon + \frac{1}{\ell}\right) \mathrm{d}F(z^{2\ell}) \\
&= \int_{\mathcal{Z}^{(1)}} \underbrace{\int_{\mathcal{Z}^{(2)}} \theta\left(\rho(z^{2\ell}) - \epsilon + \frac{1}{\ell}\right) \mathrm{d}F(z^{(2)})}_{I(z^{(1)})} \mathrm{d}F(z^{(1)}) \\
&\geq \int_{\mathcal{Q}} I(z^{(1)}) \mathrm{d}F(z^{(1)}),
\end{aligned}$$

where $\mathcal{Q} = \{z^{(1)} : \Pi(z^{(1)}) > \epsilon\} \subset \mathcal{Z}^{(1)}$.

Restricting our attention henceforth to $z^{(1)} \in \mathcal{Q}$, it suffices to show that $I(z^{(1)}) \geq \frac{1}{2}$, since we will then have:

$$\mathbf{P}\left\{\rho(Z^{2\ell}) > \epsilon - \frac{1}{\ell}\right\} \geq \frac{1}{2}\int_{\mathcal{Q}} \mathrm{d}F(z^{(1)}) = \frac{1}{2}\mathbf{P}\mathcal{Q} = \frac{1}{2}\mathbf{P}\left\{\Pi(Z^{(1)}) > \epsilon\right\}.$$

For that, first observe that for all $z^{(1)} \in \mathcal{Q}$ there must exist at least one $\alpha^\star$ such that $|\nu(z^{(1)}, \alpha^\star) - \mathbf{E}[L(Z;\alpha^\star)| > \epsilon$, since otherwise $\Pi = \sup_\alpha(\cdot) \leq \epsilon$, which means $z^{(1)} \notin \mathcal{Q}$.

Next, write:

$$\begin{aligned}
\rho(z^{2\ell}) &\geq \left|\nu(z^{(2)}, \alpha^\star) - \nu(z^{(1)}, \alpha^\star)\right| \\
&\geq \left|\nu(z^{(2)}, \alpha^\star) - \mathbf{E}[L(Z;\alpha^\star)]\right| + \left|\nu(z^{(1)}, \alpha) - \mathbf{E}[L(Z;\alpha^\star)]\right| \\
&\geq \left|\nu(z^{(2)}, \alpha^\star) - \mathbf{E}[L(Z;\alpha^\star)]\right| + \epsilon \\
&\geq \nu(z^{(2)}, \alpha^\star) - \mathbf{E}[L(Z;\alpha^\star)] + \epsilon
\end{aligned}$$

Finally, it follows that:

$$\begin{aligned}
I(z^{(1)}) &\geq \int_{\mathcal{Z}^{(2)}} \theta\left(\nu(z^{(2)}, \alpha^\star) - \mathbf{E}[L(Z;\alpha^\star)] + \epsilon - \epsilon + \frac{1}{\ell}\right) \mathrm{d}F(z^{(2)}) \\
&= \mathbf{P}\left\{\nu(Z^{(2)}, \alpha^\star) > \mathbf{E}[L(Z;\alpha^\star)] - \frac{1}{\ell}\right\} \\
&= \mathbf{P}\left\{\sum_{i=\ell+1}^{2\ell} L(Z_i;\alpha) > \ell\mathbf{E}[L(Z;\alpha^\star)] - 1\right\} \\
&= \mathbf{P}\left\{Y > \ell q - 1\right\}, \qquad\qquad Y \sim \mathsf{Binomial}(\ell, q) \\
&= \sum_{\ell q - 1 < k \leq \ell} \begin{pmatrix} \ell \\ k \end{pmatrix} q^k (1-q)^{1-k} \geq \frac{1}{2}.
\end{aligned}$$

□

2.2 Symmetrization with Random Swapping

The second stage of the proof is based on two techniques: symmetrizing with respect to swaps between the original and ghost sample points, and conditioning on a given extended sample set.

Let $\varsigma_1, \cdots, \varsigma_\ell$ be i.i.d. Bernoulli random variables with $p = \frac{1}{2}$. To formalize swaps, we define the swap vector random variable $\boldsymbol{\sigma}$ as follows:

$$\boldsymbol{\sigma}_i = \begin{cases} i + \ell\varsigma_i & i = 1, \cdots, \ell \\ i - \ell\varsigma_{i-\ell} & i = \ell+1, \cdots, 2\ell. \end{cases}$$

The swapped extended sample set is then defined to be:

$$Z^{2\ell}_{\boldsymbol{\sigma}} = (Z_{\boldsymbol{\sigma}_1}, \cdots, Z_{\boldsymbol{\sigma}_{2\ell}}).$$

This consists of effectively independently swapping pairs of elements between the original sample set and the ghost sample set. To see the usefulness of the construction, notice that, conditioned on a given swap vector σ, the probability of any event on the swapped extended sample set is the same as the probability of the event on the original extended sample set, because of the i.i.d. nature of the sample points:

$$\mathbf{P}\left\{\mathcal{E}(Z_{\boldsymbol{\sigma}_1}, \cdots, Z_{\boldsymbol{\sigma}_{2\ell}}) | \boldsymbol{\sigma} = \sigma\right\} = \mathbf{P}\left\{\mathcal{E}(Z_1, \cdots, Z_{2\ell})\right\}.$$

Thus, taking the expectation over $\boldsymbol{\sigma}$, we have that the probability of an event on the swapped extended sample set is the same too:

$$\mathbf{P}\left\{\mathcal{E}(Z_{\boldsymbol{\sigma}_1}, \cdots, Z_{\boldsymbol{\sigma}_{2\ell}})\right\} = \mathbf{E}\left[\mathbf{P}\left\{\mathcal{E}(Z_{\boldsymbol{\sigma}_1}, \cdots, Z_{\boldsymbol{\sigma}_{2\ell}}) | \boldsymbol{\sigma}\right\}\right] = \mathbf{P}\left\{\mathcal{E}(Z_1, \cdots, Z_{2\ell})\right\}.$$

Recall that what we have yet to accomplish is to bound $\mathbf{P}\left\{\rho(Z^{2\ell}) > \bar{\epsilon}\right\}$, where $\bar{\epsilon} = \epsilon - \frac{1}{\ell}$. By introducing swaps, we can equivalently bound $\mathbf{P}\left\{\rho(Z^{2\ell}_{\boldsymbol{\sigma}}) > \bar{\epsilon}\right\}$, as we will do in the final stage of the proof. The ability that we will need is to condition on a given extended sample set $z^{2\ell}$. Swaps allow us to do that in a meaningful fashion, via the following lemma.

Lemma 2. *We have:*

$$\mathbf{P}\left\{\left|\nu(Z^{(1)}_{\boldsymbol{\sigma}}, \alpha) - \nu(Z^{(2)}_{\boldsymbol{\sigma}}, \alpha)\right| > \bar{\epsilon} \middle| Z^{2\ell} = z^{2\ell}\right\} < 2e^{-\bar{\epsilon}^2\ell}.$$

Proof. We start by observing the following:

$$\begin{aligned} &\mathbf{P}\left\{\left|\nu(Z^{(1)}_{\boldsymbol{\sigma}}, \alpha) - \nu(Z^{(2)}_{\boldsymbol{\sigma}}, \alpha)\right| > \bar{\epsilon} \middle| Z^{2\ell} = z^{2\ell}\right\} \\ &= \mathbf{P}\left\{\left|\frac{1}{\ell}\sum_{i=1}^{\ell} L(Z_{\boldsymbol{\sigma}_i}; \alpha) - \frac{1}{\ell}\sum_{i=\ell+1}^{2\ell} L(z_{\boldsymbol{\sigma}_i}; \alpha)\right| > \bar{\epsilon} \middle| Z^{2\ell} = z^{2\ell}\right\} \\ &= \mathbf{P}\left\{\left|\frac{1}{\ell}\sum_{i=1}^{\ell} \underbrace{\left[L(Z_{\boldsymbol{\sigma}_i}; \alpha) - L(Z_{\boldsymbol{\sigma}_{i+\ell}}; \alpha)\right]}_{\zeta_i}\right| > \bar{\epsilon} \middle| Z^{2\ell} = z^{2\ell}\right\}. \end{aligned}$$

First, notice that, conditioned on $Z^{2\ell} = z^{2\ell}$, each ζ_i is independent, because each then depends only on ς_i. The distribution of each ζ_i may vary. If $L(z_i; \alpha) - L(z_{i+\ell}; \alpha) = 0$ then ζ_i is deterministically 0. If $L(z_i; \alpha) - L(z_{i+\ell}; \alpha) = 1$ or -1, then $\zeta_i = \pm 1$ with equal probability $\frac{1}{2}$, because of the random swapping. Hoeffding's inequality for independent real valued random variables ξ_i each bounded in an interval of width Δ_i, can be written as follows:

$$\mathbf{P}\left\{\left|\sum \xi_i - \mathbf{E}\left[\sum \xi_i\right]\right| > \epsilon\right\} < 2\exp\left(\frac{-2\epsilon^2}{\sum \Delta_i}\right).$$

In our case, we have that all ζ_i are zero mean and that $\Delta_i \leq 2$, and the statement of the lemma follows from:

$$\mathbf{P}\left\{\left|\frac{1}{\ell}\sum \zeta_i\right| > \bar{\epsilon} \middle| Z^{2\ell} = z^{2\ell}\right\} < 2e^{-\bar{\epsilon}^2\ell}.$$

□

2.3 Equivalence Classes and Union Bound

The final ingredient of the proof consists of using the results so far to break the problem in a way similar to the finite case. Recall that for every $z^{2\ell}$, the set $B^{\Lambda}(z^{2\ell})$ defines an equivalence relationship over Λ, which partitions it into a finite number of equivalence classes, mainly $N^{\Lambda}(z^{2\ell})$ of them. More precisely, we can define the relationship as $\alpha_1 \sim \alpha_2$ if and only if $(L(z_1;\alpha_1), \cdots L(z_{2\ell};\alpha_1)) = (L(z_1;\alpha_2), \cdots L(z_{2\ell};\alpha_2))$. Let us construct the finite set $\Lambda^{\star}(z^{2\ell})$ by choosing exactly one element from each equivalence class. Based on this key idea, we can use union bounds together with the bounds we obtained by random swapping and conditioning, to write:

$$\begin{aligned}
\mathbf{P}\left\{\rho(Z^{2\ell}) > \bar{\epsilon}\right\} &= \mathbf{P}\left\{\rho(Z^{2\ell}_{\boldsymbol{\sigma}}) > \bar{\epsilon}\right\} \\
&= \mathbf{E}\left[\mathbf{P}\left\{\rho(Z^{2\ell}_{\boldsymbol{\sigma}}) > \bar{\epsilon}\Big| Z^{2\ell} = z^{2\ell}\right\}\right] \\
&= \mathbf{E}\left[\mathbf{P}\left\{\sup_{\alpha\in\Lambda}\left|\nu(Z^{(1)}_{\boldsymbol{\sigma}},\alpha) - \nu(Z^{(2)}_{\boldsymbol{\sigma}},\alpha)\right| > \bar{\epsilon}\Big| Z^{2\ell} = z^{2\ell}\right\}\right] \\
&= \mathbf{E}\left[\mathbf{P}\left\{\sup_{\alpha\in\Lambda^{\star}(Z^{2\ell})}\left|\nu(Z^{(1)}_{\boldsymbol{\sigma}},\alpha) - \nu(Z^{(2)}_{\boldsymbol{\sigma}},\alpha)\right| > \bar{\epsilon}\Big| Z^{2\ell} = z^{2\ell}\right\}\right] \\
&= \mathbf{E}\left[\mathbf{P}\bigcup_{\alpha\in\Lambda^{\star}(Z^{2\ell})}\left\{\left|\nu(Z^{(1)}_{\boldsymbol{\sigma}},\alpha) - \nu(Z^{(2)}_{\boldsymbol{\sigma}},\alpha)\right| > \bar{\epsilon}\Big| Z^{2\ell} = z^{2\ell}\right\}\right] \\
&\leq \mathbf{E}\left[\sum_{\alpha\in\Lambda^{\star}(Z^{2\ell})}\mathbf{P}\left\{\left|\nu(Z^{(1)}_{\boldsymbol{\sigma}},\alpha) - \nu(Z^{(2)}_{\boldsymbol{\sigma}},\alpha)\right| > \bar{\epsilon}\Big| Z^{2\ell} = z^{2\ell}\right\}\right] \\
&\leq \mathbf{E}\left[\sum_{\alpha\in\Lambda^{\star}(Z^{2\ell})} 2e^{-\bar{\epsilon}^2\ell}\right] \\
&= \mathbf{E}\left[N^{\Lambda}(Z^{2\ell})2e^{-\bar{\epsilon}^2\ell}\right] = 2e^{\ln\mathbf{E}\left[N^{\Lambda}(Z^{2\ell})\right]-\bar{\epsilon}^2\ell} = 2e^{H^{\Lambda}_{annl}(2\ell)-\bar{\epsilon}^2\ell}
\end{aligned}$$

Finally, by combining the above with the bound we have for the ghost sample set, we obtain the statement of Theorem 1:

$$\mathbf{P}\left\{\Pi(Z^{(1)}) > \epsilon\right\} \leq 2\mathbf{P}\left\{\rho(Z^{2\ell}) > \epsilon - \frac{1}{\ell}\right\} \leq 4\exp\left\{\left[\frac{H^{\Lambda}_{annl}(2\ell)}{\ell} - \left(\epsilon - \frac{1}{\ell}\right)^2\right]\ell\right\}.$$

3 Vapnik-Chervonenkis (VC) Dimension

The bound obtained in section 2, via its corollary stated in section 1, provides a direct link between uniform convergence and the growth of the complexity of a set. The complexity, however, is measured with respect to the annealed entropy, which depends not only on the hyopthesis class, but also on the underlying distibution.

However, it is desirable to obtain bounds that are distribution independent. Vapnik and Chervonenkis suggested the following bound on the annealed entropy. Define the growth function as follows:

$$G^{\Lambda}(\ell) = \max_{z^{2\ell}} \ln N^{\Lambda}(z^{2\ell}).$$

Clearly:

$$H^{\Lambda}_{annl}(\ell) = \ln\mathbf{E}[N^{\Lambda}(Z^{\ell})] \leq \ln\max_{z^{2\ell}} N^{\Lambda}(z^{2\ell}) = G^{\Lambda}(\ell).$$

It follows that if $\frac{G^{\Lambda}(2\ell)}{\ell} \longrightarrow 0$ as $\ell\to\infty$ so does $\frac{H^{\Lambda}_{annl}(2\ell)}{\ell}$, and by the corollary in section 1, the uniform convergence of empirical risk follows.

The growth function is a combinatorial bound that depends on the model class alone, and not on the underlying distribution. The price we pay is looseness: the condition on the growth function is sufficient but not necessary, as can be attested by counterexamples (c.f. homework).

In the next lecture, we will show that its behavior is linear $G^{\Lambda}(\ell) = \ell \ln 2$ up to a certain point, which may be infinite, beyond which it grows logarithmically.

We define the VC dimension of a model class as the the largest value of ℓ where the growth function is still linear:

$$h = \max\left\{\ell : G^{\Lambda}(\ell) = \ell \ln 2\right\}.$$

From the behavior we just described and which is yet to prove, it follows that a finite VC dimension implies uniform convergence. Infinite VC dimension, on the other hand does not necessarily preclude that, because of the looseness of the bound.

An interesting alternative characterization of the VC dimension, which we will also prove next time, is stated in terms of the ability of a model class to *shatter* data points.

Consider a model class $\mathcal{C}$ parametrized by $\alpha \in \Lambda$, and the corresponding indicator loss functions $L(\cdot;\alpha)$. $\mathcal{C}$ generates a family of sets $A_\alpha = \{z | L(z;\alpha) = 1\}$. Consider a given z^ℓ, and let $B = \{z_1, \cdots, z_\ell\}$. We say that $\{A_\alpha\}$ shatters B if $\forall S \subset B \ \exists \alpha \in \Lambda$ such that $S = B \cap A_\alpha$. In other words, $\mathcal{C}$ shatters z^ℓ if it can 'select' any subset of points.

It turns out that:

$$h = \max\left\{|B| : B \text{ shattered by } \{A_\alpha\}\right\}.$$