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

Prof. Dahleh, Prof. Mitter — Lecture 3
Scribed by Rajiv Menjoge — Th 2/15

1 Example of Uniform Convergence of Empirical Risk: Discrete random variables with finite range

Let X be a Random Variable with finite range: $X : \Omega \to \{1, ..., M\}$ Let realization $X^\ell = x_1, x_2, ..., x_l$. We would like to estimate the density of X.

Assume we have picked some model class, C. In the case where we would impose no restrictions, our model class, C, would be the set of M nonnegative numbers that add to 1. Here, instead, we assume a general model class: $C = \{P_\alpha : P_\alpha$ is a probability mass function on $\{1, ..., M\}$, $\alpha \in \Lambda\}$

Recall that the risk function we used in density estimation was log likelihood. For any α, the likelihood function is given by:

$$R^\ell_{emp}(\alpha) = \frac{1}{\ell}\sum_{i=1}^{\ell} \log P_\alpha(x_i) = \frac{m_j}{\ell}\sum_{j=1}^{m} \log P_\alpha(j)$$

where m_j is the number of times $x_i = j$. The expected risk is then:

$$R(\alpha) = \mathbf{E}(\log P_\alpha(x)) = \sum_{i=1}^{m} P(i) \log P_\alpha(i)$$

Our problem would be to obtain $\alpha_{ML} = \text{argmax}(R^\ell_{emp}(\alpha))$

In the case where we impose no restrictions, we will leave it as an exercise to show that $\alpha = (\hat{p_1}, \hat{p_2}, ..., \hat{p_m})$, where $\hat{p}_{j,ML} = \frac{m_j}{\ell}$

We would like to show that the empirical risk converges almost surely to the risk uniformly over α. In other words, we would like to show that:

$$\mathbf{P}\left(\sup_\alpha \left| \frac{\sum_{i=1}^{\ell} \log P_\alpha(x_i)}{\ell} - \mathbf{E}[\log P_\alpha(x)] \right| \to 0 \right) \to 1$$

To do this, notice:

$$\begin{aligned}
\sup_\alpha \left| \frac{\sum_{i=1}^{\ell} \log P_\alpha(x_i)}{\ell} - \mathbf{E}[\log P_\alpha(x)] \right| &= \sup_\alpha \left| \sum_{j=1}^{m} \frac{m_j}{\ell} \log P_\alpha(j) - \sum_{j=1}^{m} P(j) \log P_\alpha(j) \right| \\
&\leq \sup_\alpha \left| \sum_{j=1}^{m} \log P_\alpha(j) \left(\frac{m_j}{\ell} - P(j) \right) \right| \\
&\leq \sup_\alpha \max_j |\log P_\alpha(j)| \times \sum_{j=1}^{m} \left| \frac{m_j}{\ell} - P(j) \right|
\end{aligned}$$

We know, by the strong law of large numbers that $|\frac{m_j}{\ell} - P(j)|$ converges almost surely to zero. Since m is finite, we have that $\sum_{j=1}^{m} |\frac{m_j}{\ell} - P(j)|$ converges almost surely to zero. Since $\max_j |\log P_\alpha(j)|$ stays constant as ℓ increases, we get that $\sum_{j=1}^{m} |\frac{m_j}{\ell} - P(j)| \max_j |\log P_\alpha(j)|$ converges almost surely to zero for any α. It follows that as long as C is such that $\forall j, P_\alpha(j) \neq 0$

when $\alpha \in C$, we have $sup_\alpha \max_j |\log P_\alpha(j)| \times \sum_{j=1}^m |\frac{m_j}{\ell} - P(j)|$ converges to zero as $\ell \to \infty$ almost surely, so that the empirical risk converges to the risk almost surely, uniformly over α.

We can then draw the conclusion that $\max_\alpha R^\ell_{emp}(\alpha) \to \max_\alpha R(\alpha)$ as $\ell \to \infty$, which is a consequence of uniform convergence. This is formalized in the next section. We can also show that no matter how we parameterize, we are minimizing the KL distance to the true distribution by maximizing the likelihood. This is shown below:

$$\begin{aligned} \operatorname*{argmax}_\alpha R(\alpha) &= \operatorname*{argmax}_\alpha \mathbf{E}[\log P_\alpha(x)] \\ &= \operatorname*{argmax}_\alpha (\mathbf{E}[\log P_\alpha(x)] - \mathbf{E}[\log P(x)]) \\ &= \operatorname*{argmax}_\alpha -D(P||P_\alpha) \\ &= \operatorname*{argmin}_\alpha D(P||P_\alpha) \end{aligned}$$

2 When Does the Empirical Risk Converge to the Risk Uniformly?

2.1 A proof that uniform convergence implies convergence of the minimum of the empirical risk

First we will formally prove that if the empirical risk converges to the risk, uniformly over α, then the minimum empirical risk converges to the minimal risk:

Theorem 1. *If $\forall \varepsilon > 0$ $\mathbf{P}(\sup_\alpha |R^\ell_{emp}(\alpha) - R(\alpha)| > \varepsilon) \to 0$ as $\ell \to \infty$, then $\forall \varepsilon > 0$ and $\eta > 0$, $\exists L$ such that with probability $1 - \eta$, $|\min_\alpha R^\ell_{emp}(\alpha) - \min_\alpha R(\alpha)| < \varepsilon$ for all $\ell > L$.*

Proof. Let $\alpha_l = \operatorname{argmin}_\alpha R^\ell_{emp}(\alpha)$ and let $\alpha^0 = \operatorname{argmin}_\alpha R(\alpha)$. Fix ε and choose L large enough, so that $P(sup_\alpha |R^\ell_{emp}(\alpha) - R(\alpha)| > \varepsilon) \leq \eta$ for all $\ell > L$. Then, for $\ell > L$, we have that with probability $1 - \eta$:

$$\begin{aligned} R^\ell_{emp}(\alpha_l) \leq &\quad R^\ell_{emp}(\alpha_0) \\ \leq &\quad R(\alpha_0) + \varepsilon \\ \leq &\quad R(\alpha_l) + \varepsilon \\ \leq &\quad R^\ell_{emp}(\alpha_l) + 2\varepsilon \end{aligned}$$

The first inequality follows from the definition of α_l and the third follows from the definition of α_0. The second and fourth inequalities follow from the fact that $\mathbf{P}(\sup_\alpha |R^\ell_{emp}(\alpha) - R(\alpha)| > \varepsilon) \leq \eta$. Now, subtracting the first, fourth, and sixth expression by $R^\ell_{emp}(\alpha_l) + \varepsilon$, we get that $-\varepsilon \leq R(\alpha_0) - R^\ell_{emp}(\alpha_l) \leq \varepsilon$. Hence, we have $|R^\ell_{emp}(\alpha_l) - R(\alpha_0)| < \varepsilon$, so we conclude that: $|\min_\alpha R^\ell_{emp}(\alpha) - \min_\alpha R(\alpha)| < \varepsilon$, which completes our proof. □

2.2 The Case of Finite Parametrization

Consider the case where $|\Lambda| = N < \infty$. Last lecture, we showed that $P(|R^\ell_{emp}(\alpha) - R(\alpha)| > \varepsilon) \leq 2e^{-\ell\varepsilon^2}$. Since the supremum over a finite set is less than or equal to the sum, we then get that

$$\mathbf{P}\left(\sup_{\alpha\in\Lambda} |R^\ell_{emp}(\alpha) - R(\alpha)| > \varepsilon\right) \leq \mathbf{P}\left(\sum_{j=1}^n |R^\ell_{emp}(\alpha_j) - R(\alpha_j)| > \varepsilon\right) \leq 2Ne^{-\ell\varepsilon^2}$$

Let $\eta = 2Ne^{-\ell\varepsilon^2}$ to complete the proof that we have uniform convergence in this case. We now have a way to determine L, given any η and ε, which is what we needed in the proof. If we want an expression for ε, we can solve the equation $\eta = 2Ne^{-\ell\varepsilon^2}$, to determine that $\varepsilon = (\frac{\ln 2N - \ln \eta}{\ell})^{\frac{1}{2}}$

2.3 The Case of Infinite Parametrization

Here, we will concentrate on the indicator function, so for simplicity, assume we are in the context of a classification problem, so that $L(x, y, \alpha) \in \{0, 1\}$.

Let $z^\ell = \{(x_1, y_1), ..., (x_l, y_l)\} = \{z_1, z_2, ..., z_l\}$ Hence, $L(z, \alpha) \equiv L(x, y, \alpha)$. Define the set of ℓ-dimensional vectors, $B^\Lambda(z^\ell)$, its cardinality $N^\Lambda(z^\ell)$, and the quantity $H_{annl}(\ell)$ as follows:

$$\begin{aligned} B^\Lambda(z^\ell) &= \{(L(z_1, \alpha), L(z_2, \alpha), ..., L(z_l, \alpha)) : \alpha \in \Lambda\} \\ N^\Lambda(z^\ell) &= |B^\Lambda(z^\ell)| \\ H_{annl}(\ell) &= \ln E(N^\Lambda(z^\ell)) \end{aligned}$$

Note that $B^\Lambda(z^\ell) \subset \{0, 1\}^\ell$, and thus $N^\Lambda(z^\ell) \leq 2^\ell$. $H_{annl}(\ell)$ is referred to as the annealed entropy of the set Λ.

Example: The Glivenko-Cantelli case

Recall the Glivenko-Cantelli Theorem from last lecture, which said that empirical frequencies converge almost surely to probabilities uniformly over $A \in A^*$, where $A^* = \{[0, x), x \in (0, 1]\}$. Now let A^* be our class. Suppose we have ℓ samples now, where x_i all lie between 0 and 1 according to some unspecified distribution. Let the corresponding y_i in each case be either 0 or 1. Let $I_A(x)$ be the indicator function which $= 1$ if $x \in A$, and is zero otherwise. Then, we have that $L(z, A)$ is the indicator function which $= 1$, when $I_A(x) \neq y$ and is zero otherwise. Hence, we have a classification problem, where our allowed set of functions is $I_A(x), A \in A^*$.

One can see from here that over all $A \in A^*$, there are only $\ell + 1$ possible distinct vectors of losses. This is because when the right endpoint of A occurs between $x_{(i)}$ and $x_{(i+1)}$, where it is between the two points does not change the vector of losses (Here, $x_{(j)}$ denotes the j^{th} smallest sample). Since there are $\ell + 1$ possible positions for the endpoint of A, we get our result.

It follows then that for any fixed sample of size ℓ, $N^\Lambda(z^\ell) \leq \ell + 1$. Hence, $H^\Lambda_{annl}(\ell) \leq \ln(\ell + 1)$. The theorem and the corollary below will show that this implies that the uniform convergence property is guaranteed.

Theorem 2.

$$\mathbf{P}(\sup_{\alpha \in \Lambda} \left| R^\ell_{emp}(\alpha) - R(\alpha) \right| > \varepsilon) \leq 4e^{\left[\frac{H^\Lambda_{annl}(2l)}{\ell} - (\varepsilon - \frac{1}{\ell})^2\right]\ell}$$

Corollary. *If $H^\Lambda_{annl}(2l)$ grows sub-linearly, i.e.:*

$$\frac{H^\Lambda_{annl}(2l)}{\ell} \to 0 \qquad as \qquad \ell \to \infty,$$

then uniform convergence is guaranteed.