10-704: Information Processing and Learning **Spring 2012**

Lecture 19: Minimax Risk as Channel Capacity

Lecturer: Aarti Singh *Scribes: Patrick Foley*

19.1 Review

So far in this course, we have covered:

1. **Source Coding and Data Compression**, where we derived the source coding theorem and learned codes for lossless compression. Key result: you can perfectly recover a signal if your rate (bits transmitted per communication) is at least the entropy of the random variable. We established source coding theorem using typical set encoding which requires exponential table lookups. Practical codes we dicsussed include Shannon code, Huffman code (optimal prefix code), and Arithmetic code (adaptive to source characterisitcs since involves probabilistic modeling of the source). Arithmetic codes are example of universal codes. Another popular universal code is Lempel-Ziv encoding which does not involve probabilistic modeling, see Cover-Thomas Sec 13.4, 13.5.

2. **Channel Coding**, where we derived the Channel Coding Theorem and learned codes to perfectly recover a signal from a noisy channel. Key result: you can perfectly recover a signal from a noisy channel if the transmission rate is less than the capacity of the channel. We established Channel coding theorem using random coding and jointly typical decoding which requires exponential table lookups. Practical error-correcting codes include block codes such as Hamming code, Reed-Solomon code, Reed-Muller code, BCH (Bose-Chaudhuri-Hocquenghem) code, Goppa codes; and convolution or streaming codes such as Turbo codes, Digital Fountain and LDPC (Low Density Parity Check) codes. The class of convolution codes are able to achieve performance pretty close to capacity, however we won't be able to discuss any of these in detail in this class.

3. **Joint Source and Channel Coding**, where we combined source and channel coding and learned to perfectly recover a signal through a noisy channel with some compression. You can perfectly recover a signal from a noisy channel if the entropy of the original signal is less than the rate of communication which must not exceed the channel capacity.

4. **Rate Distortion Theory**, where we learned how to accept some loss in signal integrity and studied the tradeoffs between compression and accuracy.

   Key result: If the compression rate $R > R^{(I)}(D) = \min_{p(\hat{X}|X):E[d(X,\hat{X})]\leq D} I(X,\hat{X})$, the rate distortion function, then you can recover the signal with distortion $\leq D$. We did not prove this formally, but rate-distortion is the dual of channel coding and the proof mimics that of noisy channel coding theorem with the following differences: The proof is based on randomly generated 2^{nR} codewords drawn iid from $p(\hat{X})$. The message is encoded using a random codeword that is *distortion typical* with it and decoding simply involves a table look up. $\hat{x}^n$ is distortion typical with x^n if they are jointly typical and $|d(x^n,\hat{x}^n) - E[d(x^n,\hat{x}^n)]| < \epsilon$. The failure probability is the probability that none of the 2^{nR} codewords is distortion typical with the message x^n, this occurs with probability

$(1 - 2^{-nI(X,\hat{X})})^{2^{nR}} \leq e^{-2^{n(R-I(X,\hat{X}))}}$ which goes to zero as $n \to \infty$ if $R > I(X, \hat{X})$. The result follows by choosing $p(\hat{X})$ based on the $p(\hat{X}|X)$ which achieves the min in the definition of the rate-distortion function. For a rigorous proof, see Cover-Thomas Sections 10.4, 10.5.

Today, we study how statistical modeling can be viewed as data compression, and make explicit the analogy between lossy coding through a noisy channel and statistical modeling. The key result we derive is that the minimax excess risk (or redundancy) for estimating a parameter θ from a family Θ, the Bayes risk associated with the least favorable prior, and the channel capacity when statistal modeling is viewed as communication through a noisy channel are all equivalent.

19.2 Statistical Modeling as Data Compression

Today we look at how choosing a model from a family of models $\{P_\theta\}_{\theta \in \Theta}$, given some observed data X^n, can be viewed as a noisy channel problem.

Consider the following channel to model the data generation process:

$$\boxed{\text{Source: } \pi_\theta} \to \theta \to \boxed{\text{Channel}} \to X$$

where the channel is specified by the transisition probabilities given by $P_\theta(X)$, and our θ was drawn from Θ according to a source distribution or prior $\pi(\theta)$.

19.2.1 Channel Capacity is Minimum Excess Risk

The main theorem we will prove in this lecture is the following:

Theorem 19.1 (Redundancy-Capacity Theorem) *Suppose we have a prior $\pi(\theta)$, let $p(x)$ be the mixture model $\sum_\theta \pi(\theta) p_\theta(x)$. Then we have:*

$$\underbrace{\max_{\pi(\theta)} D_{KL}(p_\theta(X)\pi(\theta) \| \pi(\theta)p(X))}_{\text{channel capacity } \max_{\pi(\theta)} I(\theta;X)} = \underbrace{\min_{q \in Q} \max_{\theta \in \Theta} D_{KL}(p_\theta(X) \| q(X))}_{\text{minimax KL-risk}}$$

and the distribution q^* that achieves the minimax KL-risk is given as $q^*(x) = \sum_\theta \pi^*(\theta) p_\theta(x)$, where π^* is the least favorable prior distribution that achieves the capacity.

Under log loss, recall the excess risk is the Kullback-Leibler divergence,

$$\mathbb{E}_\theta[\log \frac{1}{q(X)}] - H(X) = D_{KL}(p_\theta, q).$$

We can also clearly see that this is the expected redundancy for coding one symbol using q, when the symbol is drawn from p_θ. Using q, we expect X to be coded in $\mathbb{E}_\theta[\log \frac{1}{q(X)}]$ bits, when only $H(X)$ are required.

We then want to find the minimax risk $R = \min_{q \in Q} \max_{\theta \in \Theta} D_{KL}(p_\theta || q)$.

The channel capacity is defined as $C = \max_{P_X} I(X, Y)$, which in the case of modeling, would be $C = \max_{\pi_\theta} I(\theta, X)$. We prove below that the capacity of this channel C is the minimax excess risk.

We will divide the proof into two steps.

Step 1: Channel Capacity is Worst Optimal Bayes Risk

Given a prior $\pi(\theta)$, we define the optimal *Bayes Risk* as

$$\min_{q \in Q} \sum_{\theta} \pi(\theta) D_{KL}(p_\theta \| q)$$

The prior π that maximizes the optimal Bayes risk is then known as the *least-favorable* prior and we say that it achieves the worst optimal Bayes risk.

We are ready then to state the proposition that constitute the first step in the proof of Theorem 19.1.

Proposition 19.2

$$\underbrace{\max_{\pi(\theta)} D_{KL}(p_\theta(x)\pi(\theta) \| \pi(\theta)p(x))}_{channel\ capacity} = \max_{\pi(\theta)} \min_{q \in Q} \sum_{\theta} \pi(\theta) D_{KL}(p_\theta \| q)$$

Proof: (Proof of Proposition 19.2)
We compute

$$\begin{aligned} C &= \max_{\pi} I(\theta, x) \\ C &= \max_{\pi} D_{KL}(p(\theta, x) || \pi(\theta)p(x)) \\ C &= \max_{\pi} \sum_{\theta, x} p(\theta, x) \log \frac{p(\theta, x)}{\pi(\theta)p(x)} \end{aligned}$$

Now, using $p(\theta, x) = \pi(\theta)p_\theta(x)$, we continue

$$\begin{aligned} C &= \max_{\pi} \sum_{\theta, x} p_\theta(x)\pi(\theta) \log \frac{p_\theta(x)\pi(\theta)}{\pi(\theta)p(x)} \\ C &= \max_{\pi} \sum_{\theta, x} p_\theta(x)\pi(\theta) \log \frac{p_\theta(x)}{p(x)} \\ C &= \max_{\pi} \sum_{\theta} \pi(\theta) D_{KL}(p_\theta(x) || p(x)) \end{aligned}$$

Now, we note that $p(x)$ is the posterior, and we express it as $p(x) = \sum_\theta p(\theta, x) = \sum_\theta \pi(\theta)p_\theta(x) \equiv q_\pi$, a mixture of the distributions in the class $\{P_\theta\}_{\theta \in \Theta}$ with mixture weights $\pi(\theta)$.

To finish the proof, we will show that for all priors π, for all distributions q, it holds that

$$\sum_{\theta} \pi(\theta) D_{KL}(p_\theta \| q_\pi) \leq \sum_{\theta} \pi(\theta) D_{KL}(p_\theta \| q)$$

Note that this implies that q_π minimizes the Bayes risk under prior π.

$$\begin{aligned}
\sum_\theta \pi(\theta) D_{KL}(p_\theta||q_\pi) &= \sum_\theta \pi(\theta) \sum_x p_\theta(x) \log \frac{p_\theta(x)}{q_\pi(x)} \frac{q(x)}{q(x)} \\
&= \sum_\theta \pi(\theta) D_{KL}(p_\theta||q) + \sum_\theta \pi(\theta) \sum_x p_\theta(x) \log \frac{q(x)}{q_\pi(x)} \\
&= \sum_\theta \pi(\theta) D_{KL}(p_\theta||q) + \sum_x \sum_\theta \pi(\theta) p_\theta(x) \log \frac{q(x)}{q_\pi(x)} \\
&= \sum_\theta \pi(\theta) D_{KL}(p_\theta||q) + \sum_x q_\pi(x) \log \frac{q(x)}{q_\pi(x)} \\
&= \sum_\theta \pi(\theta) D_{KL}(p_\theta||q) - D_{KL}(q_\pi||q) \\
&\leq \sum_\theta \pi(\theta) D_{KL}(p_\theta||q)
\end{aligned}$$

using Gibb's inequality (KL divergence ≥ 0).

Therefore, we arrive at the conclusion that

$$\begin{aligned}
C &= \max_\pi \sum_\theta \pi(\theta) D_{KL}(p_\theta \| q_\pi) \\
&= \max_\pi \min_{q \in Q} \sum_\theta \pi(\theta) D_{KL}(p_\theta \| q)
\end{aligned}$$

■

Step 2: Worst Optimal Bayes Risk is Minimax Risk

We will now show that

$$\max_\pi \min_{q \in Q} \sum_\theta \pi(\theta) D_{KL}(p_\theta \| q) = \min_{q \in Q} \max_\theta D_{KL}(p_\theta \| q)$$

We will use the minimax theorem.

Theorem 19.3 *(Minimax Theorem)*

For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y)$$

and the minimizer in LHS is equal to the minimizer in RHS for the value of y that achieves the maximum in RHS, provided that f satisfies the following conditions:

1. *f continuous*
2. *f convex in x for fixed y*
3. *f concave in y for fixed x*
4. *both $\mathcal{X}$ and $\mathcal{Y}$ compact and convex.*

We let $f(q, \pi) = \sum_\theta \pi(\theta) D_{KL}(p_\theta \| q)$. Since f is convex in q and linear in π and the set of distributions is the probability simplex and hence is compact and convex.

Therefore, we conclude that

$$\begin{aligned}\max_\pi \min_{q \in Q} \sum_\theta \pi(\theta) D_{KL}(p_\theta \| q) &= \min_{q \in Q} \max_\pi \sum_\theta \pi(\theta) D_{KL}(p_\theta \| q) \\ &= \min_{q \in Q} \max_\theta D_{KL}(p_\theta \| q)\end{aligned}$$

where the second step follows because for a fixed q, a prior which puts all probability mass on the worst case θ is a least favorable prior.

Moreover, the minimizer of minimax problem q^* is equal to $\sum_\theta \pi^*(\theta) p_\theta(x)$, where π^* achieves the maximum in maximin problem.

19.3 Example of Advantages of Mixture Models

Redundancy-Capacity theorem tells us that mixture models are good for inference, and the optimal mixture model corresponds to the least favorable prior. We briefly discuss an illustrative example to show why mixture models can be very useful in information theory. Consider X drawn from a Bernoulli distribution with parameter θ, $\theta \in \{0, 1\}$. Denote by P_0 and P_1 the distributions for X obtained from $\theta = 0$ and $\theta = 1$, respectively.

Note that $D_{KL}(P_0 || P_1) = \infty$. Thus the distributions are infinitely apart in KL divergence.

Now let Q be the mixture distribution $Q = \frac{1}{2} P_0 + \frac{1}{2} P_1$, and observe that $D(P_0 || Q) = D(P_1 || Q) = 1$ bit.

Thus, the mixture model approximates both distributions to within 1 bit. On the other hand, a plug-in model such as maximum likelihood model chosen based on data will yield either P_0 or P_1 which is bad in a universal or minimax sense since the chosen model will be arbitrarily bad in approximating the other candidate model.