ECE 598 - Fall 2025

Lecture 13: Universal Compression

9th October, 2025

Instructor: Shubhanshu Shekhar

The key lesson from our previous lecture on (variable-rate) compression is that there exists a close relation between probability distributions and coding systems: for each probability distribution, we can construct a prefix-free code with lengths $\approx \log(1/p(x))$, and each prefix-free code is associated with a probability (sub-) distribution through Kraft's inequality. We continue this theme in this lecture and consider the problem of universal coding in two settings: stochastic and adversarial (or individual-sequence compression).

1 The Compression Game: Stochastic Formulation

Let $\mathcal{X}$ denote some observation space and let $\{P_\theta : \theta \in \Theta\}$ represent a family of distributions (that is, a statistical model) in $\mathcal{X}$ parametrized by Θ. Consider the following game between `Nature` and `SourceCoder`:

- `Nature` selects a distribution P_θ from the parametric family.
- `SourceCoder` selects a distribution $Q \in \mathcal{P}(\mathcal{X})$ (not necessarily from the parametric family)
- `SourceCoder` incurs a loss $D_{\mathrm{KL}}(P_\theta \parallel Q)$ (equivalently, `Nature` gets a reward equal to this value).

The value of this two-player zero-sum game is equal to

$$\mathrm{Red}(\Theta) = \inf_{Q} \sup_{P_\theta} D_{\mathrm{KL}}(P_\theta \parallel Q). \tag{1}$$

Remark 1.1. Due to the equivalence between the distributions and the source codes, we can think of $\mathrm{Red}(\Theta)$ as the minimum number of extra bits needed to encode any distribution of the family $\{P_\theta : \theta \in \Theta\}$ by a single code/distribution associated with Q^* (the code/distribution that achieves the minimax value above).

We begin by stating a key result in this area, called the *redundancy-capacity* theorem, which relates the minimax redundancy introduced in (1) to the channel capacity associated with $\{P_\theta : \theta \in \Theta\}$. For simplicity, we state the result for the case of finite alphabets $\mathcal{X}$.

Theorem 1.2. *Suppose $|\mathcal{X}|$ and $|\Theta|$ are finite. Then, we have:*

$$\inf_{Q \in \mathcal{P}(\mathcal{X})} \sup_{\theta \in \Theta} D_{KL}(P_\theta \parallel Q) \; = \; \sup_{\pi \in \mathcal{P}(\Theta)} \inf_{Q \in \mathcal{P}(\mathcal{X})} \mathbb{E}_{\theta \sim \pi} \left[D_{KL}(P_\theta \parallel Q) \right] = C(\Theta),$$

where $C(\Theta)$ is the capacity of the channel associated with the conditional distributions $\{P_\theta : \theta \in \Theta\}$, defined as follows (with $(X \mid \theta) \sim P_\theta$):

$$C(\Theta) = \sup_{\pi \in \mathcal{P}(\Theta)} I(\theta; X).$$

Proof. The proof of the result is essentially a consequence of the minimax theorem. The first step is to observe that redundancy can be rewritten as

$$\mathrm{Red}(\Theta) = \inf_{Q \in \mathcal{P}(\mathcal{X})} \sup_{\pi \in \mathcal{P}(\Theta)} \mathbb{E}_{\theta \sim \pi}[D_{\mathrm{KL}}(P_\theta \parallel Q)]. \tag{2}$$

This is simply due to the fact that $\sup_{\theta \in \Theta} D_{\mathrm{KL}}(P_\theta \parallel Q)$ is equal to $\sup_{\pi \in \mathcal{P}(\Theta)} \mathbb{E}_{\theta \sim \pi}[D_{\mathrm{KL}}(P_\theta \parallel Q)$, since expectation is a linear function of π.

Next, we observe the following: The objective in (2) is linear (hence concave) in π, and convex in Q, and both the domains $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\Theta)$ are compact (since we have assumed that both $\mathcal{X}$ and Θ are finite). Hence the conditions for minimax theorem are satisfied, which implies that we can interchange the order of $\inf$ and $\sup$ without changing the value of the game; that is,

$$\mathrm{Red}(\Theta) = \sup_{\pi \in \mathcal{P}(\Theta)} \inf_{Q \in \mathcal{P}(\mathcal{X})} \mathbb{E}_{\theta \sim \pi}[D_{\mathrm{KL}}(P_\theta \parallel Q)].$$

The final step is to observe that the inner minimization above defines the mutual information $I(\theta; X)$. To see this, consider $\theta \sim \pi$, and $(X \mid \theta) \sim P_\theta$, and $X \sim P_\pi = \sum_{\theta \in \Theta} \pi(\theta) P_\theta$. Then, we know that

$$\begin{aligned} I(\theta, X) &= D_{\mathrm{KL}}(P_{X|\theta} \parallel P_X \mid \pi) = \sum_\theta \pi(\theta) D_{\mathrm{KL}}(P_\theta \parallel P_\pi) \\ &= \sum_\theta \pi(\theta) D_{\mathrm{KL}}(P_\theta \parallel Q) - D_{\mathrm{KL}}(P_\pi \parallel Q) \\ &\leq \mathbb{E}_{\theta \sim \pi}[D_{\mathrm{KL}}(P_\theta \parallel Q)]. \end{aligned}$$

Thus, we know that $I(\theta; X) \leq \mathbb{E}_{\theta \sim \pi}[D_{\mathrm{KL}}(P_\theta \parallel Q)]$ and the inequality holds with an equality at $Q = P_\pi$; that mixture distribution associated with π. Using this fact in (2), we can conclude that

$$\mathrm{Red}(\Theta) = \sup_\pi I(\theta; X) = C(\Theta),$$

where we simply used the definition of channel capacity as the maximum (over all input distributions π) of the mutual information between input and output random variables. This completes the proof. □

Remark 1.3. An important consequence of the above result is that the optimal Q^* associated with a family $\{P_\theta : \theta \in \Theta\}$ is the mixture distribution with $\pi = \pi^* \in \mathrm{argmax}_\pi I(\theta; X)$. Hence, the distribution that achieves the minimum worst-case redundancy for the given family is the mixture according to the capacity achieving input distribution π^*.

In general, finding the optimal Q^*, or equivalently, finding the capacity achieving input distribution π^* is infeasible. Nevertheless, the broad principle of selecting Q to be a mixture of all $\{P_\theta : \theta \in \Theta\}$ is very useful, both in theory and practice. This is also closely related to the algorithmic idea of *exponential weights* or *hedging* in online learning. To see this, consider a sequential version of our coding problem:

- `Nature` selects a distribution P_θ
- For $t = 1, 2, \ldots, n$:
  - `SourceCoder` selects $Q_t(\cdot \mid X^{t-1})$
  - `SourceCoder` observes $X_t \sim P_\theta$ (assume for simplicity that the $X's$ are i.i.d. over time)
- The loss incurred by `SourceCoder` is

$$\mathrm{Red}_n(P_\theta, Q) = D_{\mathrm{KL}}\left(P_\theta^{\otimes n} \parallel \prod_{t=1}^{n} Q_t(\cdot \mid X^{t-1}) \right).$$

Note that selecting a distribution Q^n over the n-fold product space $\mathcal{X}^n$ is equivalent to sequentially choosing the conditional distributions: that is, $Q^n(x^n) = \prod_{t=1}^n Q_t(x_t \mid x^{t-1})$.

Mixture Distributions as Exponential Weights Scheme. Suppose we want to use the mixture distribution Q^π associated with π, and defined as

$$Q^\pi(E) = \int P_\theta^n(E)\pi(\theta)d\theta.$$

Then, given realizations $x_1, \ldots, x_{t-1}$, this gives the prediction scheme

$$Q_t^\pi(x_t \mid x^{t-1}) = \frac{Q_t^\pi(x^t)}{Q_{t-1}^\pi(x^{t-1})} = \frac{\int P_\theta^t(x^t)\pi(\theta)d\theta}{\int P_{\theta'}^{t-1}(x^{t-1})\pi(\theta')d\theta'} = \int P_\theta(x_t)\left(\frac{P_\theta^{t-1}(x^{t-1})\pi(\theta)}{\int P_{\theta'}^{t-1}(x^{t-1})\pi(\theta')d\theta'}\right)d\theta.$$

Now, if we think of π as a 'prior' distribution on the space of parameters, then the term inside the parenthesis in the last integral above can be interpreted as the 'posterior' distribution of θ given the first $t-1$ observations: $\pi(\theta \mid x^{t-1}) \propto \pi(\theta)P_\theta^{t-1}(x^{t-1})$. Furthermore, we can interpret this posterior distribution as

$$\pi(\theta \mid x^{t-1}) \propto \pi(\theta) \times \exp\left(-\log\left(1/P_\theta^{t-1}(x^{t-1})\right)\right).$$

From a coding perspective, $\log(1/P_\theta^{t-1}(x^{t-1})$ is the codeword length assigned to the sequence x^{t-1} by the code (distribution) P_θ (assuming i.i.d. observations). Hence the posterior $\pi(\theta \mid x^{t-1})$ is weighted exponentially according to this quantity: θ that assign longer codewords to x^{t-1} are down-weighted as compared to those parameters that assign shorter codewords.

Choice of mixture. Interestingly, any mixture distribution π that places nonzero mass on all parameters achieves zero average redundancy in the limit of large n. In fact, under very general situations, we can show that

$$D_{\text{KL}}(P_\theta^n \parallel Q_n^\pi) = \frac{|\mathcal{X}|-1}{2} \log \frac{n}{2\pi} + \log\left(\frac{\sqrt{\det I_\theta}}{\pi(\theta)}\right) + o(1). \tag{3}$$

This result above is an instance of a general result that says that asymptotically, the redundancy achieved by a mixture strategy is $\mathcal{O}\left(\frac{d}{2}\log\left(\frac{n}{2\pi}\right)\right)$, with d denoting the dimension of the parametric family of distributions. See (**?**, Theorem 16.4.1) and **?** for more details.

A consequence of (3), along with the variational definition of mutual information is that, for any π, we have

$$I_\pi(\theta; X^n) = \frac{|\mathcal{X}|-1}{2} \log\left(\frac{n}{2\pi}\right) + \int \log\left(\frac{\sqrt{\det I_\theta}}{\pi(\theta)}\right)\pi(\theta)d\theta + o(1).$$

This suggests that an asymptotically near-optimal mixture distribution is

$$\pi_J(\theta) = \frac{\sqrt{\det I_\theta}}{\int \sqrt{\det I_\theta} d\theta}.$$

This is the so-called Jeffreys prior.

Example. For the case of Bernoulli distributions, Jeffreys' prior is the $\text{Beta}(1/2, 1/2)$ distribution. To see why, observe that the Fisher information term I_θ for $\text{Bernoulli}(\theta)$ is $\mathbb{E}_\theta[s_\theta(X)^2]$, where $s_\theta = \partial \log \theta^X \bar{\theta}^{\bar{X}} / \partial\theta$. Hence, we have

$$I_\theta = \mathbb{E}_\theta\left[\left(\frac{X}{\theta} - \frac{\bar{X}}{\bar{\theta}}\right)^2\right] = \mathbb{E}_\theta\left[\frac{X^2}{\theta^2} + \frac{\bar{X}^2}{\bar{\theta}^2} - 2\frac{X\bar{X}}{\theta\bar{\theta}}\right] = \frac{1}{\theta} + \frac{1}{\bar{\theta}} = \frac{1}{\theta(1-\theta)}.$$

Thus, our discussion above suggests that when predicting binary-valued sequences, we should use the mixture strategy with the following prior:

$$\pi_J(\theta) \propto \sqrt{|I_\theta|} = 1/\sqrt{\theta(1-\theta)}.$$

This is exactly the definition of the $\text{Beta}(1/2, 1/2)$ distribution; that is, $\propto \theta^{1/2-1}(1-\theta)^{1/2-1}$. So, given $x_1, \ldots, x_{t-1}$, this strategy tells us to define $Q_t(\cdot \mid x^{t-1})$ as

$$Q_t(x \mid x^{t-1}) \propto \pi_J(\theta \mid x^{t-1})\theta^x(1-\theta)^{1-x}.$$

Let $S_{t-1} = \sum_{i=1}^{t-1} x_i$ denote the number of $1's$ seen in the first $t-1$ observations. Then, a standard fact about $\text{Beta}(a, b)$ distributions is that its posterior is also Beta with parameters $(a + S_{t-1}, b + S_{t-1})$. Hence, the prediction strategy $Q_t(1 \mid x^{t-1})$ in our case is the expected value of $\text{Beta}(1/2 + S_{t-1}, 1/2 + (t-1) - S_{t-1})$ which is equal to

$$Q_t(x_t = 1 \mid x^{t-1}) = \mathbb{E}_{\pi_J(\cdot|x^{t-1})}[\theta] = \frac{1/2 + S_{t-1}}{(1/2 + S_{t-1}) + (1/2 + t - 1 - S_{t-1})} = \frac{S_{t-1} + 1/2}{(t-1) + 1}.$$

This is an example of an "add-constant" estimator and is a special case of Krichevsky-Trofimov estimator.

2 Individual Sequence Formulation

In this section, we consider a fully adversarial version of the coding game, which is also referred to as individual sequence prediction under log loss. As before, we assume that there exists a collection of distributions $\{P_\theta : \theta \in \Theta\}$ on $\mathcal{X}^n$, and consider the following game:
For $t = 1, 2, \ldots, n$:

- SourceCoder selects a distribution Q_n over $\mathcal{X}^n$
- Nature selects a sequence $x^n \in \mathcal{X}^n$
- SourceCoder incurs a loss of $\log(1/Q_n(x^n)$, or equivalently, suffers a regret w.r.t. best constant code/distribution in $\{P_\theta : \theta \in \Theta\}$ equal to

$$\mathrm{Reg}_n(Q_n, \Theta, x^n) = \log\left(\frac{1}{Q_n(x^n)}\right) - \inf_{\theta\in\Theta} \log\left(\frac{1}{P_\theta(x^n)}\right)$$

The worst-case individual sequence regret incurred by Q_n is

$$\mathrm{Reg}_n(Q_n, \Theta) = \sup_{x^n\in\mathcal{X}^n, \theta\in\Theta} \log\left(\frac{1}{Q_n(x^n)}\right) - \log\left(\frac{1}{P_\theta(x^n)}\right) = \sup_{x^n,\, \theta} \log\left(\frac{P_\theta(x^n)}{Q_n(x^n)}\right).$$

Comparing this with the definition of redundancy, we see that regret measures the worst-case performance of a coding scheme Q_n or a per-sequence basis, as compared to the best coding scheme in hindsight selected from a class of codes $\{P_\theta : \theta \in \Theta\}$. This is unlike redundancy, which looks at the average case performance.

Our interest is in identifying the value of the minimax regret, defined as

$$\mathrm{Reg}_n(\Theta) = \inf_{Q_n} \sup_{\theta} \left[\sup_{x^n} \log\left(\frac{P_\theta(x^n)}{Q_n(x^n)}\right)\right] = \inf_{Q_n} \sup_{\theta} \left[\log\left(\frac{\sup_{x^n} P_\theta(x^n)}{Q_n(x^n)}\right)\right].$$

To state the minimax value, we need to introduce a term called the *Shtarkov sum*, defined as

$$S_n(\Theta) = \sum_{x^n} \sup_{\theta\in\Theta} P_\theta(x^n) = \sum_{x^n} P_{\widehat{\theta}_{ML}(x^n)}(x^n), \quad \text{where} \quad \widehat{\theta}_{ML}(x^n) \in \operatorname*{argmax}_{\theta\in\Theta} P_\theta(x^n)$$

is the maximum likelihood estimate using x^n from the family $\{P_\theta : \theta \in \Theta\}$. We will refer to $\log S_n(\Theta)$ as the (Shtarkov) *complexity* of the class $\{P_\theta : \theta \in \Theta\}$, and denote it by $\mathrm{Comp}_n(\Theta)$. Furthermore, introduce the Shtarkov distribution, Q_n^*, (also called the *normalized maximum likelihood* distribution) as

$$Q_n^*(x^n) = \frac{\sup_{\theta\in\Theta} P_\theta(x^n)}{S_n(\Theta)} = \frac{P_{\widehat{\theta}_{ML}(x^n)}(x^n)}{S_n(\Theta)}. \tag{4}$$

With these definitions available, we can now proceed to the main result of this section.

Theorem 2.1. *The minimax regret is equal to the complexity; that is,*

$$\mathrm{Reg}_n(\Theta) = \inf_{Q_n} \sup_{x^n} \log \left(\frac{\sup_\theta P_\theta(x^n)}{Q_n(x^n)} \right) = \log \left(\sum_{x^n} \sup_{\theta \in \Theta} P_\theta(x^n) \right) = \mathrm{Comp}_n(\Theta).$$

Furthermore, this optimal value is achieved by the NML or Shtarkov distribution Q_n^* defined in (4).

Proof. Throughout the proof, we will assume that the complexity term is finite (which is satisfied if Θ and $\mathcal{X}$ are both finite). One direction of the proof is easy.

$$\mathrm{Reg}_n(\Theta) \leq \sup_{\theta} \log \left(\frac{\sup_\theta P_\theta(x^n)}{Q_n^*(x^n)} \right) = \log S_n = \mathrm{Comp}_n(\Theta).$$

This is a crucial property of the Shtarkov distribution: it achieves equal regret at every sequence x^n, and that value is equal to the complexity. It turns out this "equalizing" property also makes it the minimax regret achieving distribution.

To show the other direction, observe that for any distribution Q_n, we have

$$\begin{aligned}\mathrm{Reg}_n(Q_n, \Theta) &= \sup_{x^n} \log \left(\frac{\sup_\theta P_\theta(x^n)}{Q_n(x^n)} \right) \geq \sum_{x^n} \log \left(\frac{\sup_\theta P_\theta(x^n)}{Q_n(x^n)} \right) Q_n^*(x^n) \\ &= \sum_{x^n} \log \left(\frac{Q_n^*(x^n) S_n(\Theta)}{Q_n(x^n)} \right) Q_n^*(x^n) \\ &= \log S_n(\Theta) + D_{\mathrm{KL}}(Q_n^* \parallel Q_n) \geq \log S_n(\Theta) = \mathrm{Comp}_n(\Theta).\end{aligned}$$

This completes the proof. □

One key drawback of the NML or Shtarkov distribution is that it changes with the horizon n and is thus not suitable in a sequential setting. This is in contrast to mixture distributions, which neatly factorize into one-step conditional distributions, and hence can be easily extended from n to $n+1$ observations. Furthermore, it can be proved that the optimal mixture distributions (with Jeffreys prior) are within $\mathcal{O}(1)$ of the optimal NML regret. More specifically, we can show that the optimal regret satisfies (with $|\mathcal{X}| = m$)

$$\mathrm{Reg}_n(\Theta) = \log S_n(\Theta) = \frac{m-1}{2} \log \frac{n}{2\pi} + \log \int \sqrt{\det I_\theta} d\theta + o(1),$$

while the Jeffreys mixture method (denoted by Q_J) satisfies

$$\mathrm{Reg}_n(Q_J, \Theta) = \frac{m-1}{2} \log \frac{n}{2\pi} + \log \left(\frac{\pi^{m/2}}{\Gamma(m/2)} \right) + \mathcal{O}(1).$$

That is, the mixture method with Jeffreys prior comes within a $\mathcal{O}(1)$ term of the optimal regret achieved by the NML method, while additionally being anytime valid and computationally efficient/feasible.

3 Connections to Gambling

So far in this lecture, we have seen that the problem of universal prediction can be reduced to that of sequentially assigning probabilities to symbols under logarithmic loss penalization. We now briefly touch upon how another task, that of gambling on *horse races*, also admits the same reduction.

Suppose a gambler has an initial capital (or wealth) of 1\$, and he wishes to bet on n sequential horse races, denoted by $(X_1, o_1), (X_2, o_2), \ldots, (X_n, o_n)$. Each horse race involving m horses can be represented by an outcome vector $X_i \in [m]$ (denoting the index of the winning horse), and and "odds ratio $o_i : [m] \to [0, \infty)$, with $o_i[j]$ indicating the multiplicative increase in wealth if horse j wins round i. The game can be described as follows.
For $t = 1, 2, \ldots, n$:

- The odds o_t are revealed to the gambler.
- `Gambpler` allocates his current capital W_{t-1} among the m horses according to $Q_t(\cdot \mid X^{t-1})$.
- The winner X_t is revealed.
- `Gambpler` updates his wealth as

$$W_t = W_{t-1} \times \sum_{j=1}^{m} \mathbf{1}_{X_t=j} o_t[j] Q_t(j \mid X^{t-1}) = W_{t-1} \times \left(Q_t(X_t \mid X^{t-1}) \times o_t[X_t] \right).$$

Now, let $\{P_\theta : \theta \in \Theta\}$ denote a collection of gambling strategies on n horse races (i.e., probability distributions on $[m]^n$). Then, the relative performance of the best strategy from this class, and the performance of the strategy Q_n, is equal to

$$\frac{\sup_{\theta \in P_\theta} W_n(\theta)}{W_n} = \frac{\sup_{\theta \in \Theta} \prod_{t=1}^{n} P_\theta(X_t \mid X^{t-1}) \times o_t[X_t]}{\prod_{t=1}^{n} Q_t(X_t \mid X^{t-1}) \times o_t[X_t]}.$$

Observe that the odds cancel out, and on taking logarithms, we get that

$$\log \left(\frac{\sup_{\theta \in \theta} P_\theta(X^n)}{Q_n(X^n)} \right) = \log \left(\frac{\sup_{\theta \in \Theta} W_n(\theta)}{W_n} \right) = \mathrm{Reg}_n(Q_n, X^n).$$

Hence, our earlier discussion on universal compression in both stochastic and individual sequence settings apply directly to the task of (universal) gambling on horse races when the performance criteria is to maximize the growth-rate (i.e., $\log W_n$).