Lecture 13

Agenda: * Entropic compression review

* Connection b/w compression and probability estimation/assignment

* Add-α estimator and its Bayesian interpretation

$\boxed{A}$ Entropic compression review (See notes of IT)

Consider an iid sequence $X_1, \ldots, X_n$ with common distribution P.

If this distribution P is known,

- we can use Huffman codes to achieve avg. length

$$\mathbb{E}[L] \leq H(P)+1,$$

even for $n=1$.

- in fact, arithmetic coding allows us to attain codewords with length $\ell(\underline{x}) = \lceil -\log Q^{(n)}(\underline{x}) \rceil, \forall \underline{x} \in \mathcal{X}^n$.

* The advantage here is that we do not require the knowledge of $Q^{(n)}(\underline{x})$ at the outset, but only requires $Q(x_i | x^{i-1})$ when encoding x_i, $1 \leq i \leq n$.

When P is unknown, and we used $Q^{(n)}$ to assign the lengths to arithmetic encoder, the average length achieved satisfies

$$\mathbb{E}[L] \leq \mathbb{E}_{P^n}\left[\log \frac{1}{Q^{(n)}(X^n)}\right] + 1 \tag{1}$$

$$= D\left(P^n \| Q^{(n)}\right) + nH(P) + 1.$$

On the other hand, for any prefix-free code (or even uniquely decodeable code), the lengths $\{\ell(\underline{x}),\ \underline{x} \in \mathcal{X}^n\}$ satisfy Kraft's inequality

$$k_n = \sum_{\underline{x}} 2^{-\ell(\underline{x})} \leq 1.$$

Consider $P_\ell(\underline{x}) = \dfrac{2^{-\ell(\underline{x})}}{k_n}$. Then,

$$\mathbb{E}[L(X^n)] = \mathbb{E}\left[-\log k_n + \log \frac{1}{P_\ell(\underline{x})}\right]$$

$$= -\log k_n + D(P^n \| P_\ell) + nH(P)$$

(2) $\Rightarrow$ $\mathbb{E}[L(X^n)] - nH(P) \geq D(P^n \| P_\ell)$.

(1) and (2) yield an interesting duality b/w compression and probability assignment: the compression problem is equivalent (upto 1 bit) to assigning probabilities to sequencies $x^n \in \mathcal{X}^n$.

$\boxed{B}$ Regret formulation

We now focus on the probability assignment problem, where the cost $D(P^n \| Q^{(n)})$ shows naturally.

Regret (instead of reward)

An individual sequence formulation:

$\ell(\underline{x}, Q) = \log \frac{1}{Q(\underline{x})} \equiv$ log-loss function ③

$\mathcal{E} \equiv$ class of experts

Instead of seeking the absolute optimal scheme, we seek schemes that compete with the class $\mathcal{E}$ of experts.

$\mathcal{E}_0 \equiv$ memoryless experts $\equiv \{P^n, \ P \in \mathcal{P}_k\}$.

Then, the "regret" for using Q is given by

$$r(\underline{x}, Q) = \log \frac{1}{Q(\underline{x})} - \min_{P \in \mathcal{P}_k} \log \frac{1}{P^n(\underline{x})}$$

$$= \max_{P \in \mathcal{P}_k} \log \frac{P^n(\underline{x})}{Q(\underline{x})}$$

The worst-case regret is given by

$$r_{k,n}(Q) = \max_{\underline{x} \in \mathcal{X}^n} r(\underline{x}, Q)$$

and the minimax regret is given by

$$r(k,n) = \min_{Q \in \mathcal{P}_{[k]^n}} r_{k,n}(Q)$$

Average regret formulation

$$\bar{r}_{k,n}(Q) = \mathbb{E}_{P^n}\left[\log \frac{P(X^n)}{Q(X^n)}\right]$$

$$= D(P^n \| Q)$$

$$\bar{r}(k,n) = \min_{Q \in \mathcal{P}_{[k]^n}} R_{k,n}(Q) \leq r(k,n)$$

In fact, we will see

$$r(k,n) = \bar{r}(k,n) = \frac{k-1}{2}\log n + O_k(1).$$

$\boxed{C}$ Bayesian Estimators: Uniform prior

$\pi \equiv \mathrm{unif}(\mathcal{P}_k)$.

$Q \sim \pi$, denote by $\mathcal{Q}_\pi$ the overall measure.

Then, $\mathcal{Q}_\pi(X_{\ell+1} = x \mid X^\ell = x^\ell)$

$$= \frac{\mathcal{Q}_\pi(X^\ell = x^\ell,\ X_{\ell+1} = x)}{\mathcal{Q}_\pi(X^\ell = x^\ell)}$$

$$\mathcal{Q}_\pi(X^\ell = x^\ell) = \mathbb{E}_{Q\sim\pi}\left[Q^\ell(X^\ell = x^\ell)\right]$$

We first consider the case $\boxed{k=2}$. Then,

$Q^\ell(X^\ell = x^\ell)$

$$= \prod_{i=1}^{\ell} Q(X_i = x_i) = q^{n(1|\underline{x})}(1-q)^{n(0|\underline{x})}$$

To evaluate $\mathcal{Q}_\pi$, we can evaluate

$$\underbrace{\mathbb{E}_{q\sim \mathrm{unif}[0,1]}\left[q^i(1-q)^{\ell-i}\right]}_{I_{i,\ell}} = \int_0^1 q^i(1-q)^{\ell-i}\,dq$$

$$= -\frac{q^i(1-q)^{\ell-i+1}}{(\ell-i+1)}\Bigg|_0^1 + \frac{i}{\ell-i+1}\int (1-q)^{\ell-i+1} q^{i-1}\,dq$$

$$= \begin{cases} \dfrac{1}{\ell+1}, & i = 0, \ell, \\[2ex] \dfrac{i}{\ell-i+1}\displaystyle\int (1-q)^{\ell-i+1} q^{i-1}\,dq, & 1 \le i \le \ell-1. \end{cases}$$

Thus, $$I_{i,\ell} = \frac{i}{\ell-i+1}\cdot\frac{i-1}{\ell-i+2}\cdots\frac{1}{\ell+1} = \frac{i!\,(\ell-i)!}{(\ell+1)!}$$

and so, denoting $i = n(1|x^\ell)$,

$$\frac{Q_\pi(X^\ell = x^\ell, X_{\ell+1} = x)}{Q_\pi(X^\ell = x^\ell)} = \begin{cases} \dfrac{1}{\ell+2}\cdot\dfrac{(i+1)!\,(\ell-i)!}{i!\,(\ell-i)!} & x = 1 \\[2ex] \dfrac{1}{\ell+2}\cdot\dfrac{i!\,(\ell+1-i)!}{i!\,(\ell-i)!}, & x = 0 \end{cases}$$

$$= \begin{cases} \dfrac{i+1}{(\ell+2)}, & x = 1, \\[2ex] \dfrac{\ell+1-i}{\ell+2}, & x = 0. \end{cases}$$

$$Q_\pi(X_{\ell+1} = x \mid X^\ell = x^\ell) = \frac{n(x|x^\ell)+1}{\ell+2}.$$

For $k=3$,

$$I_{i,j,\ell} = \int_0^1 q_1^i \int_0^{1-q_1} q_2^j\,(1-q_1-q_2)^{\ell-i-j}\,dq_2\,dq_1$$

Note that the inner integral

$$\int_0^\theta q^j\,(\theta-q)^{\ell-j}\,dq = \frac{j}{\ell-j+1}\cdot\frac{j-1}{\ell-j+2}\cdot\cdots\cdot\frac{1}{\ell}\cdot\int_0^\theta(\theta-q)^\ell\,dq$$

$$= \frac{\theta^{\ell+1}}{\ell+1}\cdot\frac{1}{\binom{\ell}{j}}$$

Thus,

$$I_{i,j,\ell} = \frac{1}{(\ell-i+1)} \cdot \frac{1}{\binom{\ell-i}{j}} \int_0^1 q^i (1-q)^{\ell-i+1}\, dq$$

$$= \frac{1}{(\ell-i+1)} \cdot \frac{1}{\binom{\ell-i}{j}} \cdot \frac{1}{(\ell+2)\binom{\ell+1}{i}}$$

$$= \frac{(\ell-i-j)!\, j!}{(\ell-i)!} \cdot \frac{1}{(\ell+2)} \cdot \frac{(\ell-i)!\ \ i!}{(\ell+1)!}$$

$$= \frac{1}{(\ell+2)(\ell+1)} \cdot \frac{1}{\binom{\ell}{i,j}}$$

Thus, in general we get that a uniform prior yields an add-1 estimator.

$\boxed{D}$ Jeffrey's prior / Dirichlet prior

k=2 case

Consider $\pi(q) \propto \frac{1}{\sqrt{q(1-q)}}$

As before, we need to evaluate

$$\int_0^1 q^{i-1/2} (1-q)^{\ell-i-1/2}\, dq = \frac{(i-\frac{1}{2})}{(\ell-i-\frac{1}{2}+1)} \cdots \frac{\frac{3}{2}}{(\ell-3/2)} \int_0^1 \sqrt{q}\, (1-q)^{\ell-3/2}\, dq$$

Thus, $Q_\pi(X_{\ell+1}=0 \mid X^\ell = x^\ell)$ equals

$$\frac{(\ell-i+1/2)}{(\ell-1/2)} \left[\int_0^1 \sqrt{q}\, (1-q)^{\ell-1/2}\, dq \Big/ \int_0^1 \sqrt{q}\, (1-q)^{\ell-3/2}\, dq \right]$$

Note that

$$I_{\ell+1} = \int_0^1 \sqrt{q}\,(1-q)^{\ell-1/2}\,dq = I_\ell - \frac{3}{2(\ell-1/2)} I_{\ell+1}$$

$$\Leftrightarrow \left(\frac{2\ell+2}{2\ell-1}\right) \cdot I_{\ell+1} = I_\ell$$

Therefore, $Q_\pi(X_{\ell+1}=0 \mid X^\ell = x^\ell) = \dfrac{\ell - i + 1/2}{\ell+1}$

$$= \frac{n(0 \mid x^\ell) + 1/2}{n+1}.$$

Similarly, can be extended to arbitrary k.

Finally, for any $Beta(a,b)$ prior,

$$\pi(q) \propto \frac{1}{q^{a-1}(1-q)^{b-1}}.$$

add-α corresponds to $Beta(\alpha,\alpha)$.

→ In the next class, we will analyse the redundancy corresponding to $\alpha = \frac{1}{2}$, namely the Krichevsky–Trofimov (KT) estimator.