10-704: Information Processing and Learning **Spring 2015**

Lecture 13: Feb 24

Lecturer: Akshay Krishnamurthy *Scribes: Xuanchong Li*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

13.1 Universal Prediction

In the previous lecture, our results were based on the assumption that the source distribution is known. The goal of universal prediction is to relax this assumption. In universal prediction, the source distribution is not known.

In the homework we saw the following example: Given the data $X \sim p$, instead using the true distribution $p(x)$, we use another distribution $q(x)$ to encode the data, which means

$$l(x) = \lceil \log \frac{1}{q(x)} \rceil$$

. In the homework we proved the following bound:

$$H(p) + D(p||q) \leq \mathbb{E}_p l(x) \leq H(p) + D(p||q) + 1 \tag{13.1}$$

The goal in universal prediction is to find a q that has $D(p||q)$ small for all $p \in \mathcal{P}$. Such a coding distribution would be universal for $\mathcal{P}$.

There are two cases in universal prediction - the adversarial case and a more average case.

13.1.1 Adversarial Case

Given a sequence $x_1^n \in \mathcal{X}^n$, we define the regret of using distribution Q over P.

$$Reg(Q, P, x_1^n) \coloneqq \log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} = \sum_{i=1}^{n} \log \frac{1}{q(x_i|x_1^{i-1})} - \log \frac{1}{p(x_i|x_1^{i-1})} \tag{13.2}$$

We care about the worst-case regret with respect to a class of $\mathcal{P}$:

$$Reg_n(Q, \mathcal{P}) \coloneqq \sup_{P \in \mathcal{P}, x_1^n \in \mathcal{X}} Reg(Q, P, x_1^n) \tag{13.3}$$

13.1.2 Redundancy minimization

Here is a less adversarial case. We define the redundancy which is the expected regret under P.

$$Red_n(Q, P) \coloneqq \mathbb{E}_{x_1^n \sim P}\big[\log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} \big] = D(P||Q) \tag{13.4}$$

The worst-case redundancy with respect to a class $\mathcal{P}$ is

$$Red_n(Q, \mathcal{P}) \coloneqq \sup_{P \in \mathcal{P}} Red_n(Q, P) \tag{13.5}$$

13.1.3 Example

The problem in the HW2 that if we use Shannon code for Q instead of P is related to redundancy.

Let $l_P(x) = \lceil \log \frac{1}{p(x)} \rceil$, $l_q(x) = \lceil \log \frac{1}{q(x)} \rceil$. Then,

$$p(x) = 2^{-l_p(x)}, q(x) = 2^{-l_q(x)}$$

Then we have the redundancy.

$$\begin{aligned} Red_n(Q, P) &= \mathbb{E}_P\big[\log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} \big] && (13.6)\\ &= \sum_{i=1}^{n} \mathbb{E}_P l_q(x_i) - E_P l_p(x_i) && (13.7)\\ &= n[\mathbb{E}_P l_q(x) - E_P l_p(x)] && (13.8)\\ &= D(P^n || Q^n) && (13.9) \end{aligned}$$

13.2 Minimax Strategies for Regret

There are two questions here.

- How low regret can we hope for?
- How do we achieve this low regret?

Let's define the complexity of set Θ.

$$Comp_n(\Theta) \coloneqq \log \int_{\mathcal{X}^n} \sup_{\theta \in \Theta} p_\theta(x_1^n) d\mu(x_1^n) \tag{13.10}$$

where μ is some base measure on $\mathcal{X}^n$. Note that we may have $Comp_n(\Theta) = +\infty$. It turns out that the complexity equals to the minimax regret in adversarial setting.

Theorem 13.1 *The minimax regret for $\mathcal{P} = \{p_\theta\}, \theta \in \Theta$*

$$\inf_Q Reg_n(Q, \mathcal{P}) = Comp_n(\theta) \tag{13.11}$$

And if $Comp_n(\Theta) < +\infty$, then the normalized maximum likelihood estimator (as known as Shtarkov distribution) $\overline{Q}$, defined with density

$$\overline{q}(x_1^n) = \frac{\sup_{\theta \in \Theta} p_\theta(x_1^n)}{\int \sup_{\theta \in \Theta} p_\theta(x_1^n) dx_1^n} \tag{13.12}$$

is uniquely minimax optimal.

Proof: Assume $Comp_n(\Theta) < +\infty$. The normalized maximum likelihood distribution $\overline{Q}$ has constant regret:

$$\begin{aligned} Reg_n(\overline{Q}, \mathcal{P}) &= \sup_{x_1^n \in \mathcal{X}} \big[\log \frac{1}{\overline{q}(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \big] && (13.13) \\ &= \sup_{x_1^n \in \mathcal{X}} \big[\log \frac{\int \sup_{\theta \in \Theta} p_\theta(x_1^n) dx_1^n}{\sup_{\theta \in \Theta} p_\theta(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \big] && (13.14) \\ &= Comp_n(\mathcal{P}) && (13.15) \end{aligned}$$

Moreover, for any distribution Q on $\mathcal{X}^n$, we have

$$\begin{aligned} Reg_n(Q, \mathcal{P}) &= \int \big[\log \frac{1}{q(x_1^n)} - \log \frac{1}{\sup_\theta p_\theta(x_1^n)} \big] \overline{q}(x_1^n) d\mu(x_1^n) && (13.16) \\ &= \int \big[\log \frac{\overline{q}(x_1^n)}{q(x_1^n)} + Comp_n(\Theta) \big] \overline{q}(x_1^n) dx_1^n && (13.17) \\ &= D(\overline{Q} || Q) + Comp_n(\Theta) && (13.18) \end{aligned}$$

■

13.3 Mixture (Bayesian) Strategies and Redundancy

Here we move to the less adversarial case. Recall that

$$Red_n(Q, P) = D(P^n || Q_n) \tag{13.19}$$

The goal is to find distribution Q such that for any $\theta_0 \in \Theta$

$$\frac{1}{n} D(P_{\theta_0}^n || Q^n) \to 0, n \to 0 \tag{13.20}$$

We will use a mixture approach, which is based on choosing Q as convex combination (mixture) of all the possible source distribution P_θ for $\theta \in \Theta$.

In particular, we start with a prior π over Θ and compute the marginal

$$q_n^\pi(x_1^n) = \int_\Theta \pi(\theta) p_\theta(x_1^n) d\theta \tag{13.21}$$

Our algorithm will update the prior as we go and keep using the marginal.

$$q^\pi(x_i | x_1^{i-1}) = \int_\Theta p_\theta(x_i) \pi(\theta | x_1^{i-1}) d\theta \tag{13.22}$$

$$\begin{aligned} \pi(\theta | x_1^{i-1}) &= \frac{\pi(\theta) p_\theta(x_1^{i-1})}{\int_\Theta \pi(\theta') p_{\theta'}(x_1^{i-1}) d\theta'} && (13.23) \\ &\propto \pi(\theta) e^{-\log \frac{1}{p_\theta(x_1^{i-1})}} && (13.24) \end{aligned}$$

This is referred as the exponential weights update. It is a workhorse algorithm in online learning.

Theorem 13.2 *Let $\Theta \subseteq \mathbb{R}^d$, under some regularity conditions,*

$$D(P_{\theta_0}^n || Q_n^\pi) - \frac{d}{2} \log \frac{n}{2\pi e} \to \log \frac{1}{\pi(\theta_0)} + \frac{1}{2} \log det(I_{\theta_0}), n \to \infty \tag{13.25}$$

We do not give the rigorous proof here. The main point here is that we do get $\frac{1}{n}D(P_{\theta_0}^n||Q_n^\pi) \to 0$

Here we give an example. We have a Bernoulli distribution with Beta prior. Suppose $X_i \sim Ber(\theta)$ and let π be a $Beta(\alpha, \beta)$ distribution.

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{13.26}$$

where

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt \tag{13.27}$$

We have the fact that $\mathbb{E}_\pi[\theta] = \frac{\alpha}{\alpha+\beta}$. So what is the predictive distribution Q? Let $S_i = \sum_{j=1}^i X_j$ be the number of heads up to the i. Then,

$$\pi(\theta|x_1^i) \propto p_\theta(x_1^i)\pi(\theta) \propto \theta^{\alpha+S_i-1}(1-\theta)^{\beta+i-S_i-1} \tag{13.28}$$

$$\pi(\theta|x_1^i) \sim Beta(\alpha+S_i, \beta+i-S_i) \tag{13.29}$$

$$\implies Q(x_{i+1}=1|x_1^i) = \mathbb{E}_\pi[\theta|x_1^i] = \frac{S_i+\alpha}{i+\alpha+\beta} \tag{13.30}$$

13.4 Bayesian Redundancy

Suppose we knew that the parameter θ was drawn from some known prior π. The data is then drawn from P_θ. For a distribution Q_n that we choose, we have the Bayesian redundancy.

$$\mathbb{E}_{\theta\sim\pi}D(P_\theta^n||Q_n) \tag{13.31}$$

Now let $T \sim \pi$ denote the parameter. The mutual information between T to the data is

$$I(T;x_1^n) = \int \pi(\theta)D(P_\theta^n||Q_n^\pi)d\theta = \inf_Q \int \pi(\theta)D(P_\theta||Q)d\theta \tag{13.32}$$

The worst-case Bayesian redundancy is

$$\sup_\pi \inf_Q \int \pi(\theta)D(P_\theta||Q) = \sup_\pi I(T;x_1^n) \tag{13.33}$$

13.5 Redundancy Capacity Duality

We want to know if the worst-case Bayesian redundancy is the same as the minimax redundancy.

$$\sup_\pi \inf_Q \int \pi(\theta)D(P_\theta||Q) \stackrel{?}{=} \inf_Q \sup_\theta D(P_\theta||Q) \tag{13.34}$$

Clearly, the Bayesian redundancy $\leq$ minimax redundancy.

$$\sup_\pi I_\pi(T;x_1^n) \leq \inf_Q \sup_\theta D(P^n||Q) = \inf_Q Red(Q,\mathcal{P}) \tag{13.35}$$

It turns out that the other direction is also true. This is the redundancy-capacity theorem.

Theorem 13.3 *Let X be a random variable, taking finite number of values. Let Θ be a measurable space. Then,*

$$\sup_{\pi}\inf_{Q}\int D(P_\theta||Q)d\pi(\theta) = \sup_{\pi} I_\pi(T;x_1^n) = \inf_{Q}\sup_{\theta} D(P_\theta||Q) \tag{13.36}$$

Moreover, the infimum on the right is uniquely achieved by some distribution Q^* and if π^* achieves the supremum on the left, then $Q^* = \int P_\theta d\pi^*$

Proof: Recall the definitions:

$$Red(Q,\theta) = KL(P_\theta||Q) = \mathbb{E}_{P_\theta}[-\log\frac{1}{Q(x)} - \log\frac{1}{P_\theta(x)}] \tag{13.37}$$

$$Red(Q,\pi) = \int D(P_\theta, Q)d\pi(\theta) \tag{13.38}$$

Our goal is to show:
(1) $\sup_\pi I_\pi(T;X) = \sup_\pi \inf_Q Red(Q,\pi)$
(2) $\sup_\pi \inf_Q Red(Q,\pi) = \inf_Q \sup_\theta KL(P_\theta||Q)$

(1) is straight forward:

$$I_\pi(T;X) = \int \pi(T)p(X|T)\log\frac{\pi(T)p(X|T)}{\pi(T)p(X)} = \int \pi(T)D(P_T;\overline{P}) \tag{13.39}$$

where $\overline{P} = \int P_\theta d\pi(\theta)$ Need to show that

$$\int \pi(T)D(P_T;\overline{P}) \le \inf_Q \int \pi(T)D(P_T||Q) \tag{13.40}$$

$$\begin{aligned}
\int \pi(T)D(P_T;\overline{P}) &= \int_\theta\int_x \pi(\theta)P_T(x)\log\frac{P_T(x)}{\overline{P}(x)} && (13.41)\\
&= \int_\theta\int_x P_\theta(x)[\log\frac{P_\theta(x)}{Q(x)} + \log\frac{Q(x)}{\overline{P}(x)}]\pi(\theta) && (13.42)\\
&= \int_\theta \pi(\theta)D(P_\theta||Q) + \int_x[\int_\theta \pi(\theta)P_\theta(x)]\log\frac{q(x)}{\overline{p}(x)} && (13.43)\\
&= \int \pi(\theta)D(P_\theta||Q) - D(\overline{P}||Q) \le \int \pi(\theta)D(P_\theta||Q) && (13.44)
\end{aligned}$$

So we have (1)

$$\sup_\pi I_\pi(T;X) = \sup_\pi\inf_Q\int \pi(\theta)D(P_\theta||Q) = \sup_\pi\inf_Q Red(Q,\pi) \tag{13.45}$$

For (2), by (1) we already have one direction. i.e. we know that

$$\inf_Q\sup_\theta Red(Q,\theta) = \inf_Q\sup_\pi Red(Q,\pi) \ge \sup_\pi\inf_Q Red(Q,\pi) = \sup_\pi I_\theta(\theta;X) \tag{13.46}$$

So we need to show just

$$\inf_Q\sup_\theta Red(Q,\theta) \le C = \sup_\pi I_\pi(\theta;X) \tag{13.47}$$

We will exhibit a Q, $\overline{Q} = \int P_\theta d\pi(\theta)$ where π achieves supremum in definition of C. Now we will show that

$$\sum_x P_\theta(x)\log\frac{P_\theta(x)}{\overline{Q}(x)} \le C, \forall\theta\in\Theta \tag{13.48}$$

By contradiction: assume $\exists\theta$ such that this fails, call it θ^*. Define

$$\pi_\lambda = (1-\lambda)\pi + \lambda\delta_{\theta^*}, Q^{\pi,\lambda} = (1-\lambda)Q^\pi + \lambda P_{\theta^*} \tag{13.49}$$

We have

$$H(X|T) = (1-\lambda)H(X|T) + \lambda H(x|T=\theta^*) \tag{13.50}$$

$$\begin{aligned} I_{\pi_\lambda}(T;X) &= H_{\pi_\lambda}(X) - H_{\pi_\lambda}(X|T) && (13.51)\\ &= H((1-\lambda)Q^\pi + \lambda P_{\theta^*}) - (1-\lambda)H_\pi(X|T) - \lambda H(X|T=\theta^*) && (13.52)\end{aligned}$$

at $\lambda = 0$, both side are equal to capacity C, since $H_\pi(X|T) = H_\pi(X) - I_\pi(T;X)$.
Now take the derivative with respect to λ

$$\frac{\partial}{\partial\lambda}H((1-\lambda)Q^\pi + \lambda P_{\theta^*}) = -\sum(P_{\theta^*}(x) - Q^\pi(x))\log((1-\lambda)Q^\pi(x) + \lambda P_{\theta^*}(x)) \tag{13.53}$$

$$\begin{aligned} \frac{\partial}{\partial\lambda}I_{\pi,\lambda}(T;X)\Big|_{\lambda=0} &= -\sum(P_{\theta^*} - Q^\pi(x))\log(Q^\pi(x)) + H_\pi(x) - I_\pi(T;X) + \sum_x P_\theta(x)\log P_\theta(x) && (13.54)\\ &= D(P_\theta||\overline{Q}) - C && (13.55)\end{aligned}$$

So if inequality is violated, the π does not achieve the capacity since we can mix in some of θ^* to do better.
■