10-704: Information Processing and Learning **Fall 2016**

Lecture 14: Oct 19

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

In last lecture, we derived the **Normalized Maximum Likelihood** predictor that achieves minimax regret, defined as

$$\min_q \sup_{p\in\mathcal{P}} \sup_{x_1^n} \left[\log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)}\right],$$

in the adversarial setting where the sequence x_1^n can be arbitrary and we are competing against a collection predictors in a class $\mathcal{P} = \{p_\theta\}_{\theta\in\Theta}$. However, it is not a sequential predictor.

Given this drawback of the Normalized Maximum Likelihood estimator, we introduced a **Bayesian strategy** aka 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, given a prior π over Θ, we consider the mixture model

$$q^\pi(x_1^n) = \int_\Theta \pi(\theta) p_\theta(x_1^n) d\theta \tag{14.1}$$

To make it sequential, we start with some initial prior π and our algorithm will update the model and prior as we go as follows:

$$q^\pi(x_i|x_1^{i-1}) = \int_\Theta p_\theta(x_i|x_1^{i-1})\pi(\theta|x_1^{i-1}) d\theta \tag{14.2}$$

$$\pi(\theta|x_1^{i-1}) = \frac{\pi(\theta)p_\theta(x_1^{i-1})}{\int_\Theta \pi(\theta')p_{\theta'}(x_1^{i-1})d\theta'} \tag{14.3}$$

$$\propto \pi(\theta)\, e^{-\log \frac{1}{p_\theta(x_1^{i-1})}} \tag{14.4}$$

This is referred to as the exponential weights update algorithm since the prior is scaled exponentially by the loss on the data seen so far. It is a workhorse algorithm in online learning and we will see that it has good regret as well as redundancy guarantees.

14.1 Regret guarantees for Exponential weights algorithm

We now characterize the performance of the Exponential weights algorithm in the adversarial setting. We focus on a finite class where the number of competing models $|\Theta| = d$ and ask if it is possible to achieve constant regret (not scaling with n in this case). Notice that this would imply an average regret (i.e. when divided by n) of $O(1/n)$ which is the parametric rate.

Theorem 14.1 *For all $\theta \in \Theta$ and any sequence x_1^n the exponential weights algorithm with initial choice of uniform prior $\pi(\theta) = 1/d$ satisfies*

$$\log \frac{1}{q^\pi(x_1^n)} \leq \log \frac{1}{p_\theta(x_1^n)} + \log d$$

Thus, the regret is independent of n and scales only logarithmically with $|\Theta| = d$.

Proof: To ease analysis, lets define $W^i = \sum_\theta p_\theta(x_1^{i-1})$ with $W^1 = d$ and $\pi_i(\theta) = \pi(\theta|x_1^{i-1})$. We start by observing that

$$\begin{aligned}
\log \frac{1}{q^\pi(x_i|x_1^{i-1})} &= -\log \sum_\theta \pi_i(\theta) p_\theta(x_i|x_1^{i-1}) \\
&= -\log \sum_\theta \frac{\pi(\theta) p_\theta(x_1^{i-1}) p_\theta(x_i|x_1^{i-1})}{\sum_{\theta'} \pi(\theta') p_{\theta'}(x_1^{i-1})} \\
&= -\log \sum_\theta \frac{p_\theta(x_1^i)}{\sum_{\theta'} p_{\theta'}(x_1^{i-1})} = -\log \frac{W^{i+1}}{W^i}
\end{aligned}$$

where the second last step follows using the initial choice of prior $\pi(\theta) = 1/d$.

Now lets use this, to show the main result

$$\begin{aligned}
\log \frac{1}{q^\pi(x_1^n)} = \sum_{i=1}^n \log \frac{1}{q^\pi(x_i|x_1^{i-1})} &= -\sum_{i=1}^n \log \frac{W^{i+1}}{W^i} = -\log \frac{W^{n+1}}{W^1} \\
&= \log d - \log \sum_\theta p_\theta(x_1^n) \\
&\leq -\log p_\theta(x_1^n) + \log d = \log \frac{1}{p_\theta(x_1^n)} + \log d
\end{aligned}$$

where the last inequality follows by lower bounding the sum by any one of the terms since all terms are positive. ■

For countably infinite class of models Θ, the $\log d$ will be replaced by $\log 1/\pi(\theta)$ and you have for all sequences x_1^n

$$\log \frac{1}{q^\pi(x_1^n)} \leq \inf_{\theta \in \Theta} \log \frac{1}{p_\theta(x_1^n)} + \log \frac{1}{\pi(\theta)}.$$

Notice that this bound represents a tradeoff akin to bias-variance tradeoff as we have seen before, and the estimator is akin to a regularized estimator with regularizer log 1/prior which is the Shannon information content of model θ. In next class, we will extend this to general loss functions. And in homework, you will analyze a countably infinite class of models Θ for general bounded loss functions.

14.2 Minimax and Bayesian Redundancy

We now turn to the average setting where the sequence x_1^n is randomly drawn from a distribution p, and we care about minimax redundancy defined as

$$\min_q \sup_{p \in \mathcal{P}} \mathbb{E}_{x_1^n \sim p} \left[\log \frac{1}{q(x_1^n)} - \log \frac{1}{p(x_1^n)} \right] = \min_q \sup_{p \in \mathcal{P}} D(p^n || q^n),$$

over a class of distributions $\mathcal{P} = \{p_\theta\}_{\theta\in\Theta}$. Here p^n, q^n are used to indicate that the KL divergence is computed over a random sequence x_1^n. We will link the minimax redundancy to the Bayesian redundancy that we define next.

Suppose we knew that the parameter θ was drawn from some known prior π. The data $X = x_1^n$ is then drawn from p_θ. Then we can also define the Bayesian redundancy of model q as

$$\mathbb{E}_{\theta\sim\pi} D(p_\theta||q) = \int D(p_\theta||q)\pi(\theta)d\theta. \tag{14.5}$$

The mixture model q^π is the Bayes optimal predictor, i.e. the Bayesian redundancy of the mixture model q^π is the minimum Bayesian redundancy under prior π, i.e.

$$\mathbb{E}_{\theta\sim\pi} D(p_\theta||q^\pi) = \inf_q \mathbb{E}_{\theta\sim\pi} D(p_\theta||q). \tag{14.6}$$

To see this, notice that for any distribution q

$$\begin{aligned}
\mathbb{E}_{\theta\sim\pi} D(p_\theta||q^\pi) &= \int_\theta \pi(\theta) \int_x p_\theta(x) \log \frac{p_\theta(x)}{q^\pi(x)} dx d\theta && (14.7)\\
&= \int_\theta \pi(\theta) \int_x p_\theta(x) [\log \frac{p_\theta(x)}{q(x)} + \log \frac{q(x)}{q^\pi(x)}] dx d\theta && (14.8)\\
&= \int_\theta \pi(\theta) D(p_\theta||q) d\theta + \int_x [\int_\theta \pi(\theta) p_\theta(x) d\theta] \log \frac{q(x)}{q^\pi(x)} dx && (14.9)\\
&= \int \pi(\theta) D(p_\theta||q) d\theta - D(q^\pi||q) \le \int \pi(\theta) D(p_\theta||q) d\theta = \mathbb{E}_{\theta\sim\pi} D(p_\theta||q). && (14.10)
\end{aligned}$$

14.3 Redundancy Capacity Duality

We will now show that the Bayesian redundancy of the mixture model q^π over the worst possible prior π is the same as the minimax redundancy, and these are also the same as the capacity of a channel connecting the parameter θ to the data x_1^n. This is the redundancy-capacity theorem.

We haven't defined channel capacity yet, so we first give its definition. Consider the channel

$$\theta \to Channel \to x_1^n$$

The distribution characterizing the output $X = x_1^n$ of this channel for a given input θ is $p_\theta(X)$. The mutual information between the input and output of the channel is $I(\theta; X)$. If the input θ is distributed according to a prior π, then we define the capacity as the maximum mutual information that can be coupled between the input and output using best possible choice of input distribution π, i.e.

$$C = \sup_{\pi(\theta)} I(\theta; X)$$

Theorem 14.2 (Redundancy Capacity Theorem) *Let X be a random variable* [1]*, taking finite number of values. Let Θ be a measurable space. Then,*

$$\sup_\pi \inf_q \mathbb{E}_{\theta\sim\pi} D(p_\theta||q) = \sup_\pi I(\theta; X) = \inf_q \sup_\theta D(p_\theta||q) \tag{14.11}$$

Moreover, if the infimum on the right is uniquely achieved by some distribution q^* and if π^* achieves the supremum on the left, then $q^* = \int \pi^*(\theta) p_\theta = q^{\pi^*}$.

[1] You may think of X as a sequence x_1^n.

Proof: Our goal is to show:
(1) $\sup_\pi I(\theta; X) = \sup_\pi \inf_q \mathbb{E}_{\theta\sim\pi} D(p_\theta||q)$
(2) $\sup_\pi I(\theta; X) = \inf_q \sup_\theta D(p_\theta||q)$

To show (1), we first show that the Bayesian redundancy is essentially the mutual information between the parameter θ and the data X. To see this, note that the joint distribution of θ, X is $\pi(\theta)p_\theta(X)$, the marginal distribution of X is then $\int \pi(\theta)p_\theta(X)d\theta = q^\pi(X)$ and hence

$$\begin{aligned} I(\theta; X) = D(\pi(\theta)p_\theta(X)||\pi(\theta)q^\pi(X)) &= \int \pi(\theta)p_\theta(X)\log\frac{\pi(\theta)p_\theta(X)}{\pi(\theta)q^\pi(X)}d\theta \\ &= \int \pi(\theta)D(p_\theta||q^\pi)d\theta = \mathbb{E}_{\theta\sim\pi}D(p_\theta||q) \\ &= \inf_q \mathbb{E}_{\theta\sim\pi}D(p_\theta||q) \end{aligned}$$

The last step follows using Eq. 14.6. So we have (1)

$$\sup_\pi I(\theta; X) = \sup_\pi \inf_q \mathbb{E}_{\theta\sim\pi} D(p_\theta||q) \tag{14.12}$$

For (2), lets first show one direction. Bounding average by max we have for all π and q

$$\mathbb{E}_{\theta\sim\pi}D(p_\theta||q) \le \sup_\theta D(p_\theta||q)$$

Therefore, for all π and q

$$\inf_q \mathbb{E}_{\theta\sim\pi}D(p_\theta||q) \le \sup_\theta D(p_\theta||q)$$

Hence, for all π

$$\inf_q \mathbb{E}_{\theta\sim\pi}D(p_\theta||q) \le \inf_q \sup_\theta D(p_\theta||q)$$

So we get

$$\sup_\pi I(\theta, X) = \sup_\pi \inf_q \mathbb{E}_{\theta\sim\pi} D(p_\theta||q) \le \inf_q \sup_\theta D(p_\theta||q).$$

Now we need to show

$$\inf_q \sup_\theta D(p_\theta||q) \le C = \sup_\pi I(\theta; X). \tag{14.13}$$

Lets consider a q as follows: $q^{\pi^*} = \int \pi^*(\theta)p_\theta$ where π^* achieves supremum in definition of C. We will show that

$$D(p_\theta||q^{\pi^*}) \le C, \ \forall\theta \in \Theta \tag{14.14}$$

By contradiction: assume $\exists\theta$ such that this fails, call it θ^*. I.e. $D(p_{\theta^*}||q^{\pi^*}) > C$. Define a new prior and corresponding mixture

$$\pi_\lambda = (1-\lambda)\pi^* + \lambda\delta_{\theta^*}, \ q^{\pi^*,\lambda} = (1-\lambda)q^{\pi^*} + \lambda p_{\theta^*} \tag{14.15}$$

where $\lambda \in [0, 1]$, that generated the data. We will now consider mutual information under this new prior and mixture model $I_{\pi_\lambda}(\theta; X)$, and argue that, under the contradiction, the gradient of the mutual information is positive at $\lambda = 0$ and hence π^* does not achieve supremum in definition of C.

We have

$$H_{\pi_\lambda}(X|\theta) = (1-\lambda)H_{\pi^*}(X|\theta) + \lambda H(X|\theta^*) \tag{14.16}$$

Hence,

$$\begin{aligned} I_{\pi_\lambda}(\theta; X) &= H_{\pi_\lambda}(X) - H_{\pi_\lambda}(X|\theta) && (14.17) \\ &= H((1-\lambda)q^{\pi^*} + \lambda p_{\theta^*}) - (1-\lambda)H_{\pi^*}(X|\theta) - \lambda H(X|\theta^*) && (14.18) \end{aligned}$$

Now take the derivative with respect to λ

$$\frac{\partial}{\partial\lambda}H((1-\lambda)q^{\pi^*}+\lambda p_{\theta^*}) = -\sum(p_{\theta^*}(x)-q^{\pi^*}(x))\log((1-\lambda)q^{\pi^*}(x)+\lambda p_{\theta^*}(x)) \quad (14.19)$$

and

$$\begin{aligned}\frac{\partial}{\partial\lambda}I_{\pi_\lambda}(\theta;X)\Big|_{\lambda=0} &= -\int p_{\theta^*}(x)\log q^{\pi^*}(x)+\int q^{\pi^*}(x)\log q^{\pi^*}(x)+H_{\pi^*}(X|\theta)-H(X|\theta^*) && (14.20)\\ &= D(p_{\theta^*}||q^{\pi^*})+H(p^*_\theta)-H(q^{\pi^*})+H_{\pi^*}(X|\theta)-H(X|\theta^*) && (14.21)\\ &= D(p_{\theta^*}||q^{\pi^*})-C && (14.22)\end{aligned}$$

The last step follows by noticing that $H(X|\theta^*)=H(p_{\theta^*})$ and $H(q^{\pi^*})-H_{\pi^*}(X|\theta)=C$.

So if $D(p_{\theta^*}||q^{\pi^*})>C$, then π^* does not achieve the capacity since the gradient is not zero. Uniqueness follows since mutual information is strictly convex in q^π. ■

Remark 1: Notice that in the setting when Θ is finite, the capacity of the channel is $C\le\log|\Theta|$ and hence the minimax optimal redundancy is $\le\log|\Theta|$.

Remark 2: While the redundancy-capacity theorem shows interesting connections between worst-case Bayesian redundancy, capacity of channel relating input parameter to data, and minimax redundancy, often it is not easy to determine the worst-case prior and quantify the minimax redundancy. We will revisit this in a few lectures.