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

Lecture 15: Oct 24

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.*

15.1 Review: Sequential prediction | Universal coding with log-loss

Definition 15.1 *Regret*

In an adversarial setting, the regret of using a predictor q instead of p for a sequence of n symbols x_1^n is:

$$Reg(q,p) = \sup_{x_1^n} \sum_{i=1}^{n} \frac{1}{\log q(x_i|x_1^{i-1})} - \frac{1}{\log p(x_i|x_1^{i-1})}$$

where $x_1^i = \{x_1, \ldots, x_i\}$.

Theorem 15.2 *The minimax regret for a class $\{P_\theta\}_{\theta\in\Theta}$ takes the following form:*

$$\mathcal{R}_n := \inf_q \sup_{p\in\{P_\theta\}_{\theta\in\Theta}} Reg(q,p) = Comp_n(\Theta) := \log \int \sup_{\theta\in\Theta} p_\theta(x^n) d\mu(x_1^n)$$

where $Comp_n(\Theta)$ is a measure of the complexity of the class $\{P_\theta\}$, and μ is a base measure. If $Comp_n(\Theta) < \infty$, then the normalized maximum likelihood distribution is uniquely minimax optimal. This distribution is defined as:

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

Note that the normalized likelihood distribution is not a sequential predictor, since we must know the entire sequence to compute $q(x_1^n)$.

Definition 15.3 *Redundancy*

In the stochastic setting, when x_1^n is drawn from some $p \in \mathcal{P}$, we define the redundancy as

$$Red(q,p) = \mathbb{E}_{x_1^n\sim p}\left[\sum_{i=1}^{n} \frac{1}{\log q(x_i|x_1^{i-1})} - \frac{1}{\log p(x_i|x_1^{i-1})}\right] = D(p^n||q^n)$$

where $D(p^n||q^n)$ is the KL divergence between p and q based on n samples.

We considered mixture strategies, where q is a mixture over the $\{p_\theta\}_{\theta\in\Theta}$, i.e.

$$q^\pi(x_1^n) = \int_\Theta \pi(\theta) p_\theta(x_1^n) d\theta$$

where $\pi(\theta)$ is a prior over Θ. This yields a sequential predictor via the exponential weighting approach as follows.

Definition 15.4 *Exponential weights update*

Given a prior $\pi(\theta)$, at the i^{th} iteration, the posterior is

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

which is exponentially proportional to p_θ's loss on x_1^{i-1}. Then we can use the sequential update for the i^{th} symbol:

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

In the adversarial setting, for a finite collection Θ, the following theorem tells us that the universal regret is constant:

Theorem 15.5 *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 $|\Theta| = d$.

In the stochastic setting, the following theorem tells us that the mixture strategy for the worst case prior is minimax optimal for redundancy:

Theorem 15.6 *Redundancy-Capacity Duality*

$$\sup_\pi \inf_Q \int KL(P_\theta||Q) d\pi(\theta) = \sup_\pi I_{\theta\sim\pi}(\theta; X) = \inf_Q \sup_\theta KL(P_\theta||Q)$$

The right-most term is the minimax redundancy. The left-most term is the Bayesian redundancy for the worst-case prior, where the inf is achieved by the Bayes optimal predictor which is a mixture distribution with weights $\pi(\theta)$. The theorem states that the minimax redundancy is same as worst-case Bayesian redundancy, and the minimax optimal strategy for redundancy is the mixture distribution corresponding to worst-case prior. The term in the middle is the largest mutual information between a random variable θ drawn according to prior π and is known as the capacity of the channel with θ as input and X as output (we will talk more about channel capacity in a few lectures).

15.2 Sequential prediction: Beyond log-loss

We now extend these results to a general loss function.

Definition 15.7 *Loss Function*

*Let l be our **loss function** if $l : \hat{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}_+$, where $\mathcal{X}$ is a space of symbols and $\hat{\mathcal{X}}$ is the space of our predictions.*

Example 15.8 *0-1 Loss*

$$l_{0/1}(\hat{x}, x) = \mathbb{1}\{\hat{x} \cdot x \leq 0\}$$

Example 15.9 *Squared Loss*

$$l_{sq}(\hat{x}, x) = (\hat{x} - x)^2$$

15.2.1 Redundancy bounds

The stochastic setting for general loss is analogous to redundancy (which uses the negative log likelihood loss), we want to minimize

$$\sum_{i=1}^{n} \mathbb{E}_{X_i \sim p} l(\hat{X}_i, X_i)$$

where p is the data-generating distribution. This is the same as the risk from last few lectures, but now we are making predictions in a sequential fashion. For example, X_i is whether it rains and $\hat{X}_i$ is our prediction of whether it rains. Another example, X_i represents whether rock, paper, or scissors was chosen and $\hat{X}_i$ is what we guessed would be chosen.

If we knew p, then the Bayes optimal predictor is

$$\begin{aligned} x_i^* &= \operatorname*{argmin}_{x \in \hat{\mathcal{X}}} \mathbb{E}_{X_i \sim p}[l(x, X_i)|X_i^{i-1}] \\ &= \operatorname*{argmin}_{x \in \hat{\mathcal{X}}} \int_{\mathcal{X}} l(x, x_i) dp(x_i|X_1^{i-1}) \end{aligned}$$

However, we usually do not know p, so we use a different distribution q, which gives the prediction

$$\begin{aligned} \hat{x}_i &= \operatorname*{argmin}_{x \in \hat{\mathcal{X}}} \mathbb{E}_{X_i \sim q}[l(x, X_i)|X_i^{i-1}] \\ &= \operatorname*{argmin}_{x \in \hat{\mathcal{X}}} \int_{\mathcal{X}} l(x, x_i) dq(x_i|X_1^{i-1}) \end{aligned}$$

Given the predictor using q, the performance of q against p is the loss-based redundancy

$$Red_n(q, p, l) = \mathbb{E}_{X_1^n \sim p}[\sum_{i=1}^{n} (l(\hat{X}_i, X_i) - l(X_i^*, X_i))].$$

Theorem 15.10 *If $|l(\hat{x}, x) - l(x^*, x)| \leq L \ \forall x, \hat{x}, x^*$, then*

$$\frac{1}{n} Red_n(q, p_\theta, l) \leq L \sqrt{\frac{2}{n} Red_n(q, p_\theta)}$$

This implies that if we use exponential weighting approach based on log-loss, then we have $\sup_\theta \frac{1}{n} Red_n(q, p_\theta, l) \to 0$.

Note: This bound is not always tight, however. Under l_{sq}*, linear predictors can get* $\frac{\log n}{n}$ *instead of* $\sqrt{\frac{\log n}{n}}$ *rate.*

Remarks 15.11

1. The loss assumption holds for $l_{0/1}$.
2. This holds for other loss functionals, such as l_{sq}, if $x, \hat{x}, x^* \in \mathcal{X}$, where $\mathcal{X}$ is a compact set.

For the proof of theorem 15.10, we will need to use the Pinsker's and Holder's inequalities.

Theorem 15.12 *(Pinsker's Inequality)*

$$\frac{1}{2}\int |p(x) - q(x)| = ||P - Q||_{TV} \leq \sqrt{\frac{1}{2}\mathcal{D}(P||Q)}$$

Theorem 15.13 *(Hölder's Inequality) The following holds for* $p, q \in \mathbb{R}$ *s.t.* $1/r + 1/s = 1$

$$\langle f, g\rangle \leq \|f\|_r \|g\|_s$$

Proof: (Proof of theorem 15.10)

We will begin with redundancy and transform the equation so that we can apply Pinsker's inequality. Combined with Holder's inequality, this will give us our bound.

$$\begin{aligned}
Red_n(q, p_\theta, l) &= \sum_{i=1}^n \mathbb{E}_\theta[l(\hat{x}_i, x_i) - l(x_i^*, x_i)] \\
&= \sum_{i=1}^n \int_{\mathcal{X}_1^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}_i} p_\theta(x_i|x_1^{i-1})[l(\hat{x}_i, x_i) - l(x_i^*, x_i)]dx_i dx_1^{i-1} \\
&= \sum_{i=1}^n \int_{\mathcal{X}_1^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}_i} (p_\theta(x_i|x_1^{i-1}) - q(x_i|x_1^{i-1}))(l(\hat{x}_i, x_i) - l(x_i^*, x_i))dx_i dx_1^{i-1} \\
&\quad + \sum_{i=1}^n \int_{\mathcal{X}_1^{i-1}} p_\theta(x_1^{i-1}) \int_{\mathcal{X}_i} q(x_i|x_1^{i-1}))(l(\hat{x}_i, x_i) - l(x_i^*, x_i))dx_i dx_1^{i-1}
\end{aligned}$$

Note that

$\int_{\mathcal{X}} q(x_i|x_1^{i-1})(l(\hat{x}_i, x_i) - l(x_i^*, x_i))dx_i = \mathbb{E}_Q[l(\hat{x}_i, x_i) - l(x_i^*, x_i)|x_1^{i-1}]$, and since $\hat{x}_i$ is the argmin under q,

$$\leq 0$$

which implies that, continuing from the previous equation and Holder's inequality,

$$\begin{aligned}
Red_n(q, p_\theta, l) &\leq \sum_{i=1}^{n} \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) ||p_\theta(\cdot|x_1^{i-1}) - q(\cdot|x_1^{i-1})||_1 \sup_x |l(\hat{x}_i, x_i) - l(x_i^*, x_i)| \\
&\leq L \sum_{i=1}^{n} \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) ||p_\theta(\cdot|x_1^{i-1}) - q(\cdot|x_1^{i-1})||_1, && \text{the loss bound condition,} \\
&\leq L \sum_{i=1}^{n} \left(\int p_\theta(x_1^{i-1}) \right)^{\frac{1}{2}} \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) ||p_\theta(\cdot|x_1^{i-1}) - q(\cdot|x_1^{i-1})||_1^2 \right)^{\frac{1}{2}}, && \text{by Holders,} \\
&= L \sum_{i=1}^{n} \left(\int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) ||p_\theta(\cdot|x_1^{i-1}) - q(\cdot|x_1^{i-1})||_1^2 \right)^{\frac{1}{2}}, \\
&= L\sqrt{n} (\sum_{i=1}^{n} \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) ||p_\theta(\cdot|x_1^{i-1}) - q(\cdot|x_1^{i-1})||_1^2)^{\frac{1}{2}}, && \text{by Holders,} \\
&= L\sqrt{n} (2 \sum_{i=1}^{n} \int_{\mathcal{X}^{i-1}} p_\theta(x_1^{i-1}) KL(p_\theta(\cdot|x_1^{i-1}) || q(\cdot|x_1^{i-1})))^{1/2}, && \text{by Pinsker's inequality,} \\
&= L\sqrt{2n} \sqrt{KL(p_\theta^n || q^n)} \\
&= L\sqrt{2n Red_n(q, p_\theta)}
\end{aligned}$$

The second last step is essentially a chain rule for relative entropy (aka KL divergence). Dividing both sides by n gives us the desired inequality. ■

15.2.2 Regret bounds

Now let us consider the adversarial setting, where examples or symbols are not generated according to any distribution. We will consider the case where there are finitely many candidate models (also called experts) we compete against. This setting can be recast as follows. We want to predict a sequence of variables $y_1, \ldots, y_n \in \mathbb{R}$ given a collection of experts $\{1, ..., d\}$ (you may think of the experts as $\{P_\theta\}_{\theta \in \Theta}$, except that with a general loss function and adversarial setting the experts don't need to be probability distributions, and $d = |\Theta|$). In round i, expert θ makes a prediction $x_{i\theta} \in \mathbb{R}$ (again, you may think of $x_{i\theta}$ as $P_\theta(x_i|x_1^{i-1})$, though the experts can now be more general). At each round, we see the expert predictions $x_{i\theta}$ and make our choice.

To make this setting concrete, consider if we used log-loss, and for example $y_i \in \{0, 1\}$ and $x_{i\theta} \in \{0, 1\}$ are Bernoulli random variables. Then when $y_i = 1$, the loss incurred is $\log \frac{1}{x_{i\theta}}$ and when $y_i = 0$, the loss incurred is $\log \frac{1}{1-x_{i\theta}}$. We can write this as

$$l(x_{i\theta}, y_i) = y_i \log \frac{1}{x_{i\theta}} + (1 - y_i) \log \frac{1}{1 - x_{i\theta}}$$

and use exponential weights strategy, with $\pi_0(\theta) = \frac{1}{d}$, given as

$$\begin{aligned}
\pi_i(\theta|Y_1^{i-1}) &\propto \pi_0(j) \prod_{t=1}^{i} x_{t\theta}^{y_t} (1 - x_{t\theta})^{1-y_t} \\
&= \pi_0(j) e^{-\sum_{t=1}^{i} l(x_{t\theta}, y_t)}
\end{aligned}$$

Such a weighted update strategy can work for general losses if the losses behave similar to log loss. To continue with our analysis for general losses, we will need to define weakly-exponential concave losses.

First, we start by defining exponential concave losses. A loss l is exp-concave if for some $\eta > 0$ if $f(z) = e^{-\eta l(z,y)}$ is concave for all y. This implies that l satisfies

$$f(\sum_\theta \pi_\theta z_\theta) \geq \sum_\theta \pi_\theta f(z_\theta)$$

or equivalently,

$$e^{-\eta l(\sum_\theta \pi_\theta z_\theta, y)} \geq \sum_\theta \pi_\theta e^{-\eta l(z_\theta, y)}.$$

Notice that the log-loss satisfies this for $\eta = 1$.

We now generalize this notion to weakly-exponential concave.

Definition 15.14 *Weakly-Exponential Concave*

A loss function is weakly-exponential concave if $\exists c, \eta$ such that for any $\pi \in \mathbb{R}_+^d$, $\sum_\theta \pi(\theta) = 1$, then there is some way to chose $\hat{y}_i$ using $x_{i\theta}$ such that $\forall y_i$

$$l(\hat{y}_i, y_i) \leq -c \log(\sum_\theta \pi(\theta) e^{(-\eta l(x_{i\theta}, y_i))}),$$

which is equivalent to

$$e^{(-\frac{1}{c} l(\hat{y}, y))} \geq \sum_\theta \pi(\theta) e^{(-\eta l(x_{i\theta}, y))}.$$

Then we say that l is (c, η)-realizable.

Such losses behave enough like the log loss that a Bayesian updating of the experts (playing a mixture of experts) works.

Example 15.15

Log-loss is $(1, 1)$-realizable, with $\hat{y}_i = \sum_\theta \pi(\theta) x_{i\theta}$.

Example 15.16 *0-1 loss*

0-1 loss is (c, η)-realizable with any c such that $c^{-1} \leq \log \frac{2}{1+e^{-\eta}}$ with

$$\hat{y} = \sum_{j=1}^{d} \pi(j) sign(x_j)$$

i.e. we take a majority vote of the expert predictions under distribution π.

To see this, notice that when $\hat{y}$ has the correct sign, then loss is 0 while $-c \log(\sum_\theta \pi(\theta) e^{(-\eta l(x_{i\theta}, y_i))})$ is positive since $\sum_\theta \pi(\theta) e^{(-\eta l(x_{i\theta}, y_i))} \leq 1$.

If $\hat{y}$ has wrong sign, then we know at least (weighted by π_θ) half of the values $x_{i,\theta}$ have incorrect sign. Thus,

$$\sum_\theta \pi(\theta) e^{(-\eta l(x_{i\theta}, y_i))} = \sum \theta : x_{i,\theta} y_i \leq 0 \pi_\theta e^{-\eta} + \sum \theta : x_{i,\theta} y_i > 0 \pi_\theta \leq \frac{1}{2} e^{-\eta} + \frac{1}{2}.$$

Thus, to attain

$$l(\hat{y}_i, y_i) = 1 \leq -c \log(\sum_\theta \pi(\theta) e^{(-\eta l(x_{i\theta}, y_i))})$$

it is sufficient that

$$1 \leq -c \log\left(\frac{1+e^{-\eta}}{2}\right)$$

or

$$1/c \leq \log\left(\frac{2}{1+e^{-\eta}}\right).$$

Now we can state a version of the exponential weights strategy for general losses. We initialize $w_j = 1 \; \forall j \in [d]$. Over each turn $t = 1, \ldots, n$, we update our weights

$$w_j^t = e^{-\eta \sum_{i=1}^t l(x_{ij}, y_i)}$$
$$W^t = \sum_j w_j^t$$
$$\pi_j^t = \frac{w_j^t}{W^t}$$

Then choose $\hat{y}_t$ satisfying definition 15.14, with $\pi = \pi^t$ and expert values $\{x_{tj}\}_{j=1}^d$. That is, we choose $\hat{y}_t$ and suffer loss $l(\hat{y}_t, y_t)$, which can be used to update the weights again.

That brings us to our second goal of obtaining low regret.

Theorem 15.17 *[HKW 98] Consider any weakly-exponentially concave loss that is (c, η)-realizable. For any $j \in [d]$ and any sequence $y_1^n \in \mathbb{R}^n$,*

$$\sum_{i=1}^n l(\hat{y}_i, y_i) \leq c \log d + c\eta \sum_{i=1}^n l(x_{ij}, y_i).$$

That is, our total loss is logarithmic in d, and as the number of rounds goes to infinity, our per round loss goes to zero.

Before we prove the theorem, let us go through the two examples:

Example 15.18 *log-loss*

Let $y \in \{0, 1\}$, $x_{ij} \in [0, 1]$, $\hat{y}_i = \sum_j \pi(j) x_{ij}$, then the theorem holds for $c = \eta = 1$, so the regret $\sum_{i=1}^n l(\hat{y}_i, y_i) - \sum_{i=1}^n l(x_{ij}, y_i)$ w.r.t all experts is $\leq \log d$. Thus, even in the adversarial setting, using exponential weighting yields a constant in n and logarithmic in d regret for log loss.

Example 15.19 $0 - 1$ *loss*

Notice that $\log \frac{2}{1+e^{-\eta}} \approx \frac{\eta}{2} - \frac{\eta^2}{8}$, and we can set $c^{-1} = \log \frac{2}{1+e^{-\eta}}$ to find a bound on our regret. From the theorem, for any sequence and for any set of experts,

$$\begin{aligned} Regret &= [\sum_{i=1}^n l(\hat{y}_i, y_i) - \sum_{i=1}^n l(x_{ij}, y_i)] \\ &\leq \frac{\log d}{\log \frac{2}{1+e^{-\eta}}} + \frac{\eta - \log \frac{2}{1+e^{-\eta}}}{\log \frac{2}{1+e^{-\eta}}} \sum_{i=1}^n l(x_{ij}, y_i) \\ &= O(\frac{\log d}{\eta} + \eta \sum_{i=1}^n l(x_{ij}, y_i)) \\ &= O(\sqrt{n \log d}) \end{aligned}$$

where the last step follows since the loss $l \leq 1$ and choosing $\eta \approx \sqrt{\frac{\log d}{n}}$.

Thus, for $0-1$ loss, we have a per round regret of $O(\sqrt{(\log d)/n})$ in the adversarial setting using a purely sequential strategy. Thus, exponential weighting is a good strategy.

Proof: Proof of theorem 15.17. By definition of weakly-exponentially concave loss, $l(\hat{y}_t, y_t) \leq -c\log(\sum_j \pi(j)e^{(-\eta l(x_{tj}, y_t))}) = -c\log(W^{t+1}/W^t)$.

Summing over $t = 1$ to n and using the fact that $W^1 = d$, we have

$$\begin{aligned}
\sum_{t=1}^{n} l(\hat{y}_t, y_t) &\leq -c\log(\frac{W^{n+1}}{W^1}) \\
&= c\log d - c\log(\sum_{j=1}^{d} e^{-\eta \sum_{t=1}^{n} l(x_{tj}, y_t)}) \\
&\leq c\log d - c\log e^{-\eta \sum_{t=1}^{n} l(x_{tj}, y_t)} \\
&= c\log d + c\eta \sum_{t=1}^{n} l(x_{tj}, y_t)
\end{aligned}$$

The second inequality follows since we are just lower bounding sum over j with one of the terms (all terms are positive). ■