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

Lecture 14: Feb26

Lecturer: Akshay Krishnamurthy *Scribes: Erik Louie*

Note: *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.*

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

Definition 14.1 *Regret*

The regret of using a predictor Q instead of P for a sequence of n symbols is:

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

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

Definition 14.2 *Redundancy*

We define the redundancy as the expected regret

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

where $D(P_n||Q_n)$ is the KL divergence between P and Q based on n samples.

Theorem 14.3 *The minimax regret for a class $\{P_\theta\}_{\theta\in\Theta}$ takes the following form:*

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

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:

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

Remarks 14.4

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

- $Comp_n(\Theta)$ is often infinite.

The regret framework is used to study the adversarial setting where the sequence x_1^n can be adversarially chosen. We also talked about the setting where the sequence is random X_1^n, drawn from some unknown distribution P. We will focus on the case when $P \in \{P_\theta\}_{\theta\in\Theta}$, though it can be shown that the mixture strategy is quite robust to model mis-specification. For this setting, we considered mixture strategies, where Q is a mixture over the $\{P_\theta\}_{\theta\in\Theta}$. In terms of densities,

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

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

Definition 14.5 *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:

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

Theorem 14.6 *Consistency theorem*

Under some conditions, for $\Theta \subseteq \mathbb{R}^d$

$$KL(P_\theta^n||Q_n^\pi) - \frac{d}{2}\log\frac{n}{2\pi e} \to \log\frac{1}{\pi(\theta)} + \frac{1}{2}\log\det(I_\theta)$$

where I_θ is a quantity known as the Fisher information of θ (we will see this later in the course).

The following theorem tells us that the mixture strategy for the worst case prior is minimax optimal for redundancy:

Theorem 14.7 *Redundancy-Capacity Duality*

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

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 T drawn according to prior π i.e. $P(T = \theta) = \pi(\theta)$, and is known as the capacity of the channel with T as input and X as output (we will talk more about channel capacity after spring break).

14.2 Sequential prediction: Beyond log-loss

Stemming from our discussion in the prior class are two questions:

Question: Can we leverage log loss results with other loss functions?

Answer: Yes. → In particular, we can relate other loss functions to redundancy.

Question: Can we get interesting results for the adversarial case?

Answer: Yes. → We can use exponential weights to obtain low regret.

In response to these two questions, we will prove two related theorems and show examples of their application.

Definition 14.8 *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 14.9 *0-1 Loss*

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

Example 14.10 *Squared Loss*

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

The average case setting for general loss is analogous to redundancy (which is 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 an online fashion. For example, X_i is whether it rains and $\hat{X}_i$ is our prediction of whether it rained. 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 14.11 *If $Red_n(Q,P_\theta) \leq R_n(\theta)$, where $R_n(\theta)$ is the log-loss based redundancy, and $|l(\hat{x},x) - l(x^*,x)| \leq L \ \forall x, \hat{x}, x^*$, then*

$$\frac{1}{n}Red_n(Q,P_\theta,l) \leq L\sqrt{\frac{2}{n}R_n(\theta)}$$

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

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

Remarks 14.12

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.

Example 14.13 *(Classification with side information)*

Let our loss function be the 0-1 loss, ie. $l(\hat{y},y) = \mathbb{1}\{\hat{y} \cdot y \leq 0\}$. We want to predict y based on a vector $x \in \mathbb{R}^d$ or side information. Each p_θ specifies y:

$$y_i = \theta^T x_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0,\sigma^2)$$

We choose π with $\theta \sim \mathcal{N}(0,\tau^2 I_d)$. If we have seen the Y_i variables in addition to the loss:

$$\begin{aligned}\hat{Y}_{i+1} &= \operatorname*{argmin}_{y\in\mathbb{R}} \ \mathbb{E}_{Q^\pi}[l_{01}(y,Y_{i+1})|Y_1^i] \\ &= \operatorname*{argmin}_y \int_{-\infty}^{\infty} P_\theta(sign(y_{i+1}) \neq sign(y))\pi(\theta|Y_i^i)d\theta\end{aligned}$$

So, we need the posterior $\pi(\theta|x_1^i, y_1^i)$. Then,

$$\theta|y_1^i, x_1^i \sim \mathcal{N}(K_i^{-1}\sum_{j=1}^{i} x_j y_j, K_i^{-1}) \text{ with } K_i = \frac{1}{\tau^2}I + \frac{1}{\sigma^2}\sum_{j=1}^{i} x_j x_j^T$$

Note that $K_i^{-1}\sum_{j=1}^{i} x_j y_j$ looks like ridge regression. We get

$$\hat{Y}_{i+1} = \langle \theta, x_{i+1}\rangle + \epsilon_{i+1} > |y_1^i, x_1^i \sim \mathcal{N}(x_{i+1}^\top K_i^{-1}\sum_{j=1}^{i} x_j y_j, \ \ x_{i+1}^\top K_i^{-1} x_{i+1} + \sigma^2)$$

So, we can use the posterior mean $\hat{Y}_{i+1} = \langle x_{i+1}, K_i^{-1}\sum x_j y_j\rangle$.

From last class, we saw:

$$Red_n(Q^\pi, P_{\theta_0}) \leq d\log n + d\log\tau + \frac{||\theta_0||_2^2}{\tau^2} + \log\det(\sigma^{-2}X^T X)$$

where $X^TX \in \mathbb{R}^{d\times d}$*, which implies* $\det(\sigma^{-2}X^TX)$ *for* $X \in \mathbb{R}^{nxd}$*.*

So, we get, by the theorem,

$$\frac{1}{n}Red_n(Q^\pi, P_{\theta_0}, l_{01}) \leq \frac{1}{\sqrt{n}}\sqrt{d\log n + d\log\tau + \frac{||\theta_0||_2^2}{\tau^2}} + \log I_\theta$$

Remarks 14.14

1. Consider the effect of τ. If θ_0 is large, then we want τ to also be large, but we pay for it in the $\log\tau$ term. If $||\theta_0||_2 \to \infty$ then the bound is vacuous.
2. We can also handle the situation where we only see the label of $sign(Y_i)$ instead of y_i, but this is much harder computationally as we have to do Monte Carlo simulation to compute/approximate the posterior.

For the proof of theorem 14.11, we will need to use the Cauchy-Schwartz, Pinsker's, and Holder's inequalities.

Theorem 14.15 *(Pinsker's Inequality)*

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

Theorem 14.16 *(Hölder's Inequality) The following holds for* $p, q \in \mathbb{R}$ *s.t.* $1/p + 1/q = 1$ *and* $p = 1, q = \infty$*.*

$$\langle f, g\rangle \leq \|f\|_p\|g\|_q$$

Proof: (Proof of theorem 14.11)

We will begin with redundancy and transform the equation so that we can apply Pinsker's inequality. Combined with Cauchy-Schwartz 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

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

$$\leq 0$$

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

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

Dividing both sides by n gives us the desired inequality. ■

14.2.1 Adversarial Online Learning

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 strategies. In this setting, by Redundancy-Capacity theorem, the log-loss redundancy is constant for any n:

$$\begin{aligned}
\inf_Q \sup_{\theta \in \Theta} Red_n(Q, P_\theta) &= \sup_\pi I_\pi(T; X_1^n) \\
&\leq \log |\Theta|
\end{aligned}$$

Can we achieve constant regret in an adversarial case? 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 j makes a prediction $x_{ij} \in \mathbb{R}$ (again, you may think of x_{ij} 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_{ij} and make our choice.

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

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

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

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

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.

Definition 14.17 *Weakly-Exponential Concave*

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

$$l(\hat{y}_i, y_i) \leq -c \log(\sum_j \pi(j) e^{(-\eta l(x_{ij}, y_i))}),$$

which is equivalent to

$$e^{(-\frac{1}{c} l(\hat{y}, y))} \geq \sum_j \pi(j) e^{(-\eta l(x_j, y))}.$$

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

Intuitively, weakly-exponential concave tells us that we are only suffering as much loss as the worst expert. Such losses behave enough like the log loss that a Bayesian updating of the experts (playing a mixture of experts) works.

Example 14.18

Log-loss is $(1, 1)$-realizable, with $\hat{y}_i = \sum_j \pi(j) x_{ij}$.

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

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

Then choose $\hat{y}_t$ satisfying definition 14.17, 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 14.19 *[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 two examples:

Example 14.20 *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 (recall this is similar to the redundancy which was constant in n and logarithmic in d for log loss in the random setting). Recall that this is achieved using a sequential strategy while the normalized likelihood approach which was minimax optimal was not sequential.

Example 14.21 *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. See suggested reading for a proof.

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

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

where the last step follows since the loss $l \leq 1$.

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 14.19. 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). ■