10-704: Information Processing and Learning **Spring 2012**

Lecture 11: Universal redundancy bounds

Lecturer: Aarti Singh *Scribes: Rafael Izbicki*

11.1 Brief review

Let $\mathcal{P}$ be a family of probability distributions over the alphabet $\mathcal{X}$. Last time, we defined

$$\bar{R}_c = \sup_{p\in\mathcal{P}} E_p\left[\frac{L(X^n)}{n} - \left(-\frac{\log p(X^n)}{n}\right)\right]$$

to be the **worst expected redundancy** of a coding for a given family $\mathcal{P}$. We want $\bar{R}_c$ to be small. We also defined

$$R_c^* = \sup_{p\in\mathcal{P}} \max_{x^n}\left[\frac{L(x^n)}{n} - \left(-\frac{\log p(x^n)}{n}\right)\right]$$

to be the **worst maximum redundancy**. Note that $\bar{R}_c \leq R_c^*$.

Later in the course, we will show that when attempting to build a universal model for the distribution of the process X^n, mixtures of distributions over $\mathcal{P}$ are optimal in some sense. That is, one should estimate the distribution of the sequence as $q(x^n) = \sum_{p\in\mathcal{P}} \theta(p)p(x^n)$, where $\theta(p)$ is a prior measure probability over $\mathcal{P}$. Moreover, one can build efficient arithmetic coding using mixture distributions that are nearly optimal. Two examples are as follows:

Example 1 Let $\mathcal{P}$ be the class of all i.i.d. distributions over the (finite) alphabet $\mathcal{X}$. Note that each distribution in this class is characterized by a vector of probabilities $(p_1, \ldots, p_{|\mathcal{X}|})$. One can define the following predictive probabilities:

$$q^{iid}(x_t = j|x^{t-1}) = \frac{n(j|x^{t-1}) + \frac{1}{2}}{t-1+\frac{|\mathcal{X}|}{2}},$$

where x^{t-1} is used to indicate the first $t-1$ characters of the string and $n(j|x^{t-1})$ is the number of occurrences j in x^{t-1}, the first $t-1$ elements of the string. Today we will show that q^{iid} is a mixture over $\mathcal{P}$ and also that $\bar{R}_{q^{iid}} \leq \frac{|\mathcal{X}|-1}{2}\frac{\log n}{n} + \frac{K}{n}$ where $K > 0$ is a constant. Here $\bar{R}_{q^{iid}}$ is the worst expected redundancy of the arithmetic code associated with q^{iid}.

Example 2 Let $\mathcal{P}$ be the class of all m-order Markov processes over the (finite) alphabet $\mathcal{X}$. One can define the following predictive probabilities:

$$q^{markov}(x_t = j|x^{t-1}) = \frac{n((x_{t-m}^{t-1}, j)|x^{t-1}) + \frac{1}{2}}{n(x_{t-m}^{t-1}|x^{t-1}) + \frac{|\mathcal{X}|}{2}},$$

where $n((x_{t-m}^{t-1}, j)|x^{t-1})$ is the number of counts of the subsequence (x_{t-m}^{t-1}, j) in x^{t-1}. We have that $\bar{R}_{q^{markov}} \leq \frac{|\mathcal{X}|^m(|\mathcal{X}|-1)}{2}\frac{logn}{n} + \frac{K_m}{n}$ where $K_m > 0$ is a constant that depends only on m. Here $\bar{R}_{q^{markov}}$ is the worst expected redundancy of the arithmetic code associated with q^{markov}.

11.2 i.i.d Processes

We now develop Example 1, that is, i.i.d Processes. First, we show that q^{iid} is in fact a mixture of distribution on $\mathcal{P}$. Before that, let's define a Dirichlet distribution.

Definition 11.1 *Let $\alpha_1, \dots, \alpha_k > 0$. Let $\theta = (\theta_1, \dots, \theta_k) \in \Re^k$ be a random vector such that its probability density function is given by*

$$\pi(\theta) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma\left(\alpha_i\right)} \prod_i \theta_i^{\alpha_i - 1}$$

for all θ such that $\sum \theta_i = 1$, and 0 otherwise. We say that $\theta \sim$ Dirichlet$(\alpha_1, \dots, \alpha_k)$.

Note that when $k = 2$ we have a Beta distribution. Also, when all parameters $\alpha_1, \dots, \alpha_k$ are 1, we have a uniform distribution.

Proposition 11.2 *Every i.i.d process $p \in \mathcal{P}$ defined over the finite alphabet $\mathcal{X}$ can be associated with a vector of probabilities for each symbol $\theta = (p_1, \dots, p_{|\mathcal{X}|})$ where $\sum_i p_i = 1$. Let q^{iid} be as in Example 1. Then we will show that $q^{iid} = q$, where q is the distribution of the mixture of processes $p \in \mathcal{P}$ where the mixture weights for $p \equiv \theta$ are given by the prior $\pi(\theta)$, where the prior is Dirichlet$(1/2, \dots, 1/2)$.*

Proof: Notice that we can rewrite q^{iid} as

$$q^{iid}(x^n) = \prod_{t=1}^{n} q^{iid}(x_t|x^{t-1}) = \prod_{t=1}^{n} \frac{n(j|x^{t-1}) + \frac{1}{2}}{t - 1 + \frac{|\mathcal{X}|}{2}} = \frac{\prod_{x \in \mathcal{X}} (n_x - \frac{1}{2})(n_x - \frac{3}{2}) \dots (\frac{1}{2})}{(n + \frac{|\mathcal{X}|}{2} - 1)(n + \frac{|\mathcal{X}|}{2} - 2) \dots (\frac{|\mathcal{X}|}{2})}. \tag{11.1}$$

The last step follows by gathering together terms that refer to the same symbol. Denoting by π the prior density of the Dirichlet distribution and by using the Law of Total Probability, we can calculate the distribution q:

$$q(x^n) = \int_{p \in \mathcal{P}} p(x^n)\pi(p)dp = \int_{p \in \mathcal{P}} p(x^n) \frac{\Gamma\left(\sum_x \frac{1}{2}\right)}{\prod_x \Gamma\left(\frac{1}{2}\right)} \prod_x p_x^{\frac{1}{2} - 1} dp.$$

Now, by independence, we have that $p(x^n) = \prod_{k=1}^{n} p(x_k) = \prod_{x \in \mathcal{X}} p_x^{n_x}$ (we gather together term that refer to the same symbol). Hence

$$q(x^n) = \int_{p \in \mathcal{P}} \frac{\Gamma\left(\sum_x \frac{1}{2}\right)}{\prod_x \Gamma\left(\frac{1}{2}\right)} \prod_x p_x^{n_x + \frac{1}{2} - 1} dp =$$

$$= \frac{\Gamma\left(\sum_x \frac{1}{2}\right)}{\prod_x \Gamma\left(\frac{1}{2}\right)} \frac{\prod_x \Gamma\left(n_x + \frac{1}{2}\right)}{\Gamma\left(\sum_x n_x + \frac{1}{2}\right)} \int_{p \in \mathcal{P}} \prod_x p_x^{n_x + \frac{1}{2} - 1} \frac{\Gamma\left(\sum_x n_x + \frac{1}{2}\right)}{\prod_x \Gamma\left(n_x + \frac{1}{2}\right)} dp = \frac{\Gamma\left(\sum_x \frac{1}{2}\right)}{\prod_x \Gamma\left(\frac{1}{2}\right)} \frac{\prod_x \Gamma\left(n_x + \frac{1}{2}\right)}{\Gamma\left(\sum_x n_x + \frac{1}{2}\right)}, \tag{11.2}$$

where we use the fact that the integral is the integral of the density of a Dirichlet distribution over all values it assumes, and therefore is 1. Finally, using the fact that the Gamma distribution satisfies $\Gamma(s + 1) = s\Gamma(s) = s(s - 1)\Gamma(s - 1) = \dots$, we get that $\Gamma(n_x + \frac{1}{2}) = (n_x + \frac{1}{2} - 1)(n_x + \frac{1}{2} - 2) \dots (\frac{1}{2})\Gamma(\frac{1}{2})$. By using this and a similar expansion to $\Gamma\left(\sum_x n_x + \frac{1}{2}\right)$, and noting that $\sum_x n_x = n$, we get from 11.2 that

$$q(x^n) = \frac{\prod_{x \in \mathcal{X}} (n_x - \frac{1}{2})(n_x - \frac{3}{2}) \dots (\frac{1}{2})}{\prod_{x \in \mathcal{X}} (n + \sum_x \frac{1}{2})(n + \sum_x \frac{1}{2} - 1) \dots (\sum_x \frac{1}{2})},$$

which is the same as 11.1 (notice that $\sum_x 1 = |\mathcal{X}|$).

■

We will now prove a proposition that shows how well arithmetic codes generated using q^{iid} are for i.i.d. sequences. But, before that, here is a usefull lemma:

Lemma 11.3 *Let $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathcal{X}$, and denote $p_x = P(X_i = x)$, $\forall x \in \chi$. Let $\mathcal{P} = \{(p_x)_{x\in\mathcal{X}} : \sum p_x = 1, p_x \geq 0\}$. Then the maximum likelihood estimate for the sequence $x_1, \ldots, x_n$ is given as*

$$sup_{p\in\mathcal{P}}\, p(x_1, \ldots, x_n) = \prod_x \left(\frac{n_x}{n}\right)^{n_x}.$$

Proof: For any $p \in \mathcal{P}$, we have that $p(x_1, \ldots, x_n) = \prod_x p_x^{n_x}$. We want to find the supremum of this function with constrained to $\sum p_x = 1$. Equivalently, we want the supremum of $\log p(x_1, \ldots, x_n)$ subject to same constraints. The Lagrangian is given by

$$\sum_x n_x \log p_x + \lambda \sum_x p_x.$$

Taking the derivative and equating to zero, we get $p_x = -\frac{n_x}{\lambda}$. Plugging this into the constrains, we get $\lambda = -n$. The result follows from plugging the optimal p_x's on the target function.

■

Proposition 11.4 *Let $\mathcal{P}$ be the set of all i.i.d. distributions over the finite alphabet $\mathcal{X}$. Let q^{iid} be as in Example 1. Then $\bar{R}_{q^{iid}} \leq R^*_{q^{iid}} \leq \frac{|\mathcal{X}|-1}{2}\frac{\log n}{n} + \frac{K}{n}$.*

Proof: The first inequality is trivial. Now, by definition,

$$R^*_{q^{iid}} = sup_{p\in\mathcal{P}}\, max_{x^n} \log\left(\frac{p(x^n)}{q^{iid}(x^n)}\right).$$

For each x^n, and any $p \in \mathcal{P}$, we have using Lemma 11.3 that

$$p(x^n) \leq sup_{p\in\mathcal{P}}\, p(x^n) = \prod_x \left(\frac{n_x}{n}\right)^{n_x}.$$

We can also show that (by pairing each term on left side with a bounding term on right side, see e.g. pg 483 of Csiszar and Shields' Tutorial.):

$$\prod_x \left(\frac{n_x}{n}\right)^{n_x} \leq \frac{\prod_x (n_x - \frac{1}{2})(n_x - \frac{3}{2})\ldots(\frac{1}{2})}{(n-\frac{1}{2})(n-\frac{3}{2})\ldots(\frac{1}{2})}$$

Hence, by using this bound and also the explicit form of q^{iid} (which is in expression 11.1), we get (notice that the both numerators are the same)

$$\frac{p(x^n)}{q^{iid}(x^n)} \leq \frac{(n+\frac{|\mathcal{X}|}{2}-1)(n+\frac{|\mathcal{X}|}{2}-2)\ldots(\frac{|\mathcal{X}|}{2})}{(n-\frac{1}{2})(n-\frac{3}{2})\ldots(\frac{1}{2})} = \prod_{j=1}^{n} \frac{n+\frac{|\mathcal{X}|}{2}-j}{n+\frac{1}{2}-j}. \tag{11.3}$$

Now, assuming $|\mathcal{X}|$ is even (a similar argument can be worked out if $|\mathcal{X}|$ is odd), we can rewrite 11.3 as

$$\frac{(n+\frac{|\mathcal{X}|}{2}-1)!/(\frac{|\mathcal{X}|}{2}-1)!}{(2n-1)(2n-3)\ldots 1/2^n} = \frac{(n+\frac{|\mathcal{X}|}{2}-1)!2^n}{(\frac{|\mathcal{X}|}{2}-1)!(2n-1)(2n-3)\ldots 1}. \tag{11.4}$$

Now, notice that $(2n)! = (2n)(2n-1)(2n-2)\dots 1 = 2n(2n-2)(2n-4)\dots 2(2n-1)(2n-3)\dots 1 = 2^n(n-1)(n-2)\dots 1(2n-1)(2n-3)\dots 1 = 2^n n!(2n-1)(2n-3)\dots 1$. Hence

$$(2n-1)(2n-3)\dots 1 = \frac{(2n)!}{2^n n!}.$$

Plugging this into 11.4 yields

$$\frac{p(x^n)}{q^{iid}(x^n)} \leq \frac{(n+\frac{|\mathcal{X}|}{2}-1)!2^{2n}n!}{(\frac{|\mathcal{X}|}{2}-1)!(2n)!}$$

Now, using Stirling's approximation to the factorial ($n! \approx K\sqrt{n}n^n$), we get that

$$\frac{p(x^n)}{q^{iid}(x^n)} \leq Cn^{\frac{|\mathcal{X}|-1}{2}}.$$

By noticing that the result holds for all sequences x^n and all $p \in \mathcal{P}$, and by taking log we prove the proposition.

■

We note that a similar argument can be done for Example 2, that is, Markov Chains.

11.3 Stationary Processes

Now, let $\mathcal{P}$ be the class of all stationary distributions over the finite alphabet $\mathcal{X}$. Any distribution of this class can be approximated by a Markov process by letting the order of the Markov process $m \longrightarrow \infty$ with n. We have the following result

Proposition 11.5 *Let $p \in \mathcal{P}$ be a stationary process, and let $H_p(\mathcal{X})$ denote the entropy rate of p. Then if C^m is a universal code for Markov-m distributions,*

$$E_p[R_{p,C^m}] \leq H_m - H_p(\mathcal{X}) + \frac{|\mathcal{X}|^m(|\mathcal{X}|-1)}{2}\frac{\log n}{n} + \frac{K_m}{n},$$

where $H_m = H(X_{m+1}|X_1, \dots, X_m)$.

Note that we have a similar bound as before, except that now we have the extra term $H_m - H_p(\mathcal{X})$, which is the extra number of bits for allowing p to be any stationary measure. Also notice that the larger m is, the smaller the extra number of bits is. Also note that this bound is not uniform, because it depends on p. We will discuss this further in next class.