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

Lecture 8: Feb 5

Lecturer: Aarti Singh *Scribe: Siheng Chen*

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.*

8.1 Generalized Maximum Entropy

Instead of maximizing entropy with respect to some constraints, generalized maximum entropy minimizes the relative entropy and puts the constraint into the objective function. The primal problem of generalized maximum entropy is

$$\text{(Primal)} \qquad \min_{p\in\Delta} D(p||p_0) + U(\mathbb{E}_p[r]),$$

where p_0 is a base distribution, $U(\cdot)$ is some penalty function, and $\mathbb{E}_p[r] \in \mathbb{R}^d$ is a set of features, corresponding to d constraints. Then, the dual problem is

$$\text{(Dual)} \qquad \max_{\lambda\in\mathbb{R}^d} -\ln Z_\lambda - U^\star(\lambda),$$

where Z_λ is the partition function of a Gibbs distribution ($q \propto p_0(x)\exp\{-\lambda^T r(x)\}$) with features r and the based distribution p_0, and $U^\star(\cdot)$ is the conjugate function for U. Denote

$$\begin{aligned} Q(\lambda) &= -\ln Z_\lambda - U^*(\lambda) \\ &= L_t(0) - L_t(\lambda) - U_t^*(\lambda), \end{aligned}$$

where $L_t(\lambda) = -\mathbb{E}_t[\ln p_\lambda]$ and t is any distribution, the dual problem can be reformulated as

$$\min_\lambda L_t(\lambda) + U_t^*(\lambda). \tag{8.1}$$

If $t = \widehat{p}$, then (8.1) is actually maximum likelihood with regularization because

$$L_{\widehat{p}}(\lambda) = -\frac{1}{n}\sum_{i=1}^n \ln p_\lambda(X_i).$$

Note that

$$L_t(\lambda) = -\mathbb{E}_t[\ln p_\lambda] = D(t||p_\lambda) + H(t),$$

equation (8.1) is then

$$\min_\lambda D(t||p_\lambda) + U_t^*(\lambda), \tag{8.2}$$

where t is typically $\widehat{p}$.

Recall that the dual (or conjugate) function of $\Psi(u) = I(u \in A)$ is

$$\begin{aligned} \Psi^*(\lambda) &= \sup_u[\lambda\cdot u - \Psi(\lambda)] \\ &= \sup_{u\in A} \lambda\cdot u. \end{aligned}$$

We then look at some examples of the penalty functions.

Examples:

1. $U(\mathbb{E}_p[r]) = I(\mathbb{E}_p[r] = \mathbb{E}_{\widehat{p}}[r])$. Then,

$$U^*_{\widehat{p}}(\mathbb{E}_p[r]) = I(\mathbb{E}_p[r] = 0). \tag{8.3}$$
$$U^*_{\widehat{p}}(\lambda) = 0. \tag{8.4}$$

   The problem thus gets back to the basic maximum entropy duality.

2. $U(\mathbb{E}_p[r]) = I(|\mathbb{E}_p[r_j] - \mathbb{E}_{\widehat{p}}[r_j]| \leq \beta_j, \forall j)$.
   Then,

$$U^*_{\widehat{p}}(\mathbb{E}_p[r]) = I(|\mathbb{E}_p[r_j]| \leq \beta_j, \forall j). \tag{8.5}$$
$$U^*_{\widehat{p}}(\lambda) = \sum_j \beta_j |\lambda|, \tag{8.6}$$

   which corresponds to the maximum likelihood with ℓ_1 regularization.

3. $U(\mathbb{E}_p[r]) = ||(|\mathbb{E}_p[r] - \mathbb{E}_{\widehat{p}}[r]||_2^2/2\alpha$.
   Then,

$$U^*_{\widehat{p}}(\mathbb{E}_p[r]) = ||\mathbb{E}_p[r]||_2^2/2\alpha. \tag{8.7}$$
$$U^*_{\widehat{p}}(\lambda) = \alpha||\lambda||_2^2/2, \tag{8.8}$$

   which corresponds to the maximum likelihood with ℓ_2^2 regularization.

8.2 Entropy Rate of Stochastic Processes

Entropy of random variable X is $H(X)$, the joint entropy of $X_1 \ldots X_n$ is then

$$\begin{aligned} H(X_1, \ldots, X_n) &= \sum_{i=1}^n H(X_i|X_{i-1} \ldots X_1) \quad \text{chain rule} \\ &\leq \sum_{i=1}^n H(X_i) \quad \text{since conditioning does not increase entropy} \\ &= nH(X) \quad \text{if the variables are identically distributed} \end{aligned}$$

If the random variables are also independent, then the joint entropy of n random variables increases with n. How does the joint entropy of a sequence of n random variables with possibly arbitrary dependencies scale?

To answer this, we consider a stochastic process which is an indexed sequence of random variables with possibly arbitrary dependencies. We define

Entropy rate of a stochastic process $\{X_i\} =: \mathcal{X}$ as

$$H(\mathcal{X}) := \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}$$

i.e. the limit of the per symbol entropy, if it exists.

Stationary stochastic process: A stochastic process is stationary if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts:

$$p(X_1, \ldots, X_n) = p(X_{1+l}, \ldots, X_{n+l}) \qquad \forall l, \ \forall n$$

Theorem 8.1 *For a stationary stochastic process, the following limit always exists*

$$H(\mathcal{X}) := \lim_{n\to\infty} \frac{H(X_1,\ldots,X_n)}{n}$$

i.e. limit of per symbol entropy, and and is equal to

$$H'(\mathcal{X}) := \lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1)$$

i.e. the limit of the conditional entropy of last random variable given past.

For stationary first order Markov processes:

$$H(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1)$$

Theorem 8.2 Burg's Maximum Entropy Theorem
The max entropy rate stochastic process $\{X_i\}$ satisfying the constraints

$$E[X_iX_{i+k}] = \alpha_k \qquad \textit{for } k = 0,1\ldots m \quad \forall i \quad (\star)$$

is the Gauss-Markov process of the p^{th} order, having the form:

$$X_i = -\sum_{i=1}^{m} a_k X_{i-k} + Z_i,$$

where $Z_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$, a_k and σ^2 are parameters chosen such that constraints $\star$ are satisfied.

Note: The process $\{X_i\}$ is NOT assumed to be (1) zero-mean, (2) Gaussian or (3) stationary.
Note: The theorem states that $AR(m)$ auto-regressive Gauss-Markov process of order m arise as natural solutions when finding maximum entropy stochastic processes under second-order moment constraints up to lag m.

Proof: Let $\{X_i\}$ be a stochastic process that satisfies constraints $\star$, $\{Z_i\}$ be a Gaussian process that satisfies constraints $\star$, and $\{Z_i'\}$ be a m^{th} order Gauss-Markov process with the same some distribution for all orders up to p. (Existence of such a process will be established after the proof.)

Since the multivariate normal distribution maximizes entropy over all vector-valued random variables under a covariance constraint, we have:

$$\begin{aligned}
H(X_1,\ldots,X_n) &\le H(Z_1,\ldots,Z_n) \\
&= H(Z_1,\ldots,Z_m) + \sum_{i=m+1}^{n} H(Z_i|Z_{i-1},\ldots,Z_1) \quad \text{(chain rule)} \\
&\le H(Z_1,\ldots,Z_m) + \sum_{i=m+1}^{n} H(Z_i|Z_{i-1},\ldots,Z_{i-m}) \quad \text{(conditioning does not increase entropy)} \\
&= H(Z_1',\ldots,Z_m') + \sum_{i=m+1}^{n} H(Z_i'|Z_{i-1}',\ldots,Z_{i-m}') \\
&= H(Z_1',\ldots,Z_m')
\end{aligned}$$

$$\Rightarrow \lim_{m\to\infty} \frac{1}{m} H(X_1\ldots X_m) \le \lim_{m\to\infty} \frac{1}{m} H(Z_1'\ldots Z_m')$$

Existence: Does a m^{th} order Gaussian Markov process exists s.t. $(a_1 \dots a_m, \sigma^2)$ satisfy $\star$?

$$\begin{aligned} X_i X_{i-l} &= -\sum_{k=1}^{m} a_k X_{i-k} X_{i-l} + Z_i X_{i-l} \\ E[X_i X_{i-l}] &= -\sum_{k=1}^{m} a_k E[X_{i-k} X_{i-l}] + E[Z_i X_{i-l}] \end{aligned}$$

Let $R(l) = E[X_i X_{i-l}] = E[X_{i-l} X_i] = \alpha_l$ be the given $m+1$ constraints. Then we obtain *The Yule-Walker equations - m+1 equations in m+1 variables* $(a_1 \dots a_p, \sigma^2)$*:*

$$\text{for } l = 0 \qquad R(0) = -\sum_{k=1}^{m} a_k R(-k) + \sigma^2$$

$$\text{for } l > 0 \qquad R(l) = -\sum_{k=1}^{m} a_k R(l-k) \quad (\text{since } Z_i \perp X_{i-l} \text{ for } l > 0.)$$

The solution to the Yule-Walker equations will determine the m^{th} order Gaussian Markov process. ■

8.3 Data Compression / Source coding

Figure 8.1 shows the coding scheme.

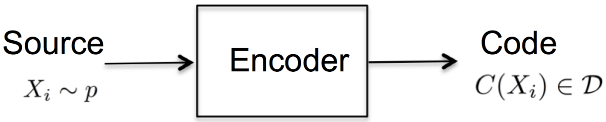

Figure 8.1: Coding schema.

Source code: A source code C is a mapping from the range of a random variable or a set of random variables to finite length strings of symbols from a $\mathcal{D}$-ary alphabet, that is

$$C : \mathcal{X} \rightarrow \mathcal{D}^\star,$$

and the code $C(X)$ for a symbol X is an element in $\mathcal{D}^\star$.

Instead of encoding an individual symbol, we can also encode blocks of symbols together. A length n **block code** encodes n length strings of symbols together and is denotes by $C(X_1, \cdots, X_n) =: C(X^n)$. We then define the extension of the code using concatenation as $C(X^n) = C(X_1)...C(X_n)$

Expected length of a source code denoted by $L(C)$ is given as follows:

$$L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$$

where $l(x)$ is the length of codeword $c(x)$ for a symbol $x \in \mathcal{X}$, and $p(x)$ is the probability of the symbol.

Several classes of symbol codes have appealing properties that are widely used. For example,

- Non-singularity : $\forall X_1 \neq X_2 \Rightarrow C(X_1) \neq C(X_2)$
- Unique-decodability : $\forall X_1^n \neq X_2^m \Rightarrow C(X_1^n) \neq C(X_2^m)$
- Self-punctuating (Prefix) : $\forall X_1 \neq X_2 \Rightarrow C(X_1) \notin \text{Prefix}(C(X_2))$

Note that unique decodability implies non-singularity and self-punctuating implies unique decodability. Self-punctuating codes are also called **instantaneous or prefix codes**. For unique decodability, we may need to see the entire sequence to decode it uniquely, but for instantaneous ones, you can decode a symbol as soon as you've seen its encoding.

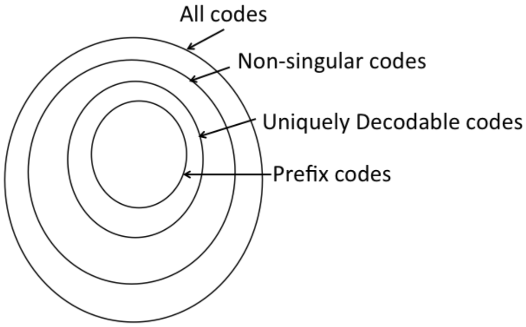

Figure 8.2: Codes.