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

Lecture 5: Burg's Maximum Entropy Theorem

Lecturer: Aarti Singh *Scribe: Ina Fiterau*

5.1 Brief Review

5.1.1 Maximum Entropy Distributions under linear constraints

$$
\begin{aligned}
q^*_{ME} &= \arg\max_q H(q) \\
&\quad \text{s.t. } q \in Q_{linear} = \{q \in \mathcal{P} : E_q[r_i(X)] = \alpha_i\} \\
&\quad \text{where } \mathcal{P} \text{ is the set of all distributions.}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow q^*_{ME} &= \exp[\lambda^*_{0_{ME}} - 1 + \sum_i \lambda^*_{i_{ME}}] \\
&\quad \text{where } \lambda^*_i \text{ chosen s.t. } q^*_{ME} \in Q_{linear}.
\end{aligned}
$$

Normalizing to ensure $q^*_{ME} \in \mathcal{P}$,

$$
q^*_{ME} = \frac{\exp[\sum_i \lambda^*_{i_{ME}} r_i(x)]}{\sum_x \exp[\sum_i \lambda^*_{i_{ME}} r_i(x)]} \qquad \Rightarrow \qquad q^*_{ME} \in \text{ exponential family}
$$

More Examples:

1. Let's consider the multivariate maximum entropy distribution with $\mathbf{0}$ mean, $E[X_i X_j] = k_{ij}$ and unbounded support. Then $q^*_{ME} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ where $\mathbf{K} = \{k_{ij}\}$ is the covairance matrix.

2. Graphical models are a special case of exponential families, e.g. the graphical model known as the Ising model is used to model spin of electrons. The electron spin is modeled by a random variable $x_i \in \{0, 1\}$, neighboring spins are anti-parallel if $x_i \neq x_j$ and parallel if $x_i = x_j$. In ferromagnetic materials, configurations in which electron spins are parallel are favored and hence the probability of a spin is given as

$$
q^*_{\text{ISING}} \propto \exp\left[\sum_{ij} \lambda_{ij}(x_i x_j + (1 - x_i)(1 - x_j))\right]
$$

Notice that the probability of a spin alignment is higher if $x_i = x_j$, i.e. spins are parallel. Ising model is indeed the maximum entropy binary distribution that respects second moments between neighbors.

5.1.2 Information Projection (I-projection)

We define the *information projection* of a distribution p onto the family of distributions Q as:

$$q^*_{IP} \quad = \quad \arg\min_{q \in Q} D(q \,||\, p)$$

If $Q = Q_{linear}$, we can show that

$$q^*_{IP} = \frac{p(x)\exp[\sum_i \lambda^*_{i_{IP}} r_i(x)]}{\sum_x p(x)\exp[\sum_i \lambda^*_{i_{IP}} r_i(x)]},$$

i.e. it is in the exponential family with base distribution p.

If p is uniform and $Q = Q_{\text{linear}} \Rightarrow q^*_{IP} = q^*_{ME}$.

Examples of distributions from the exponential family with base distribution p:

- Poisson: $q^*(x) = \frac{1}{x!}\lambda^x e^{-\lambda}$
- Binomial: $\binom{n}{x}\theta^x(1-\theta)^{n-x}$

Reminder: The probability simplex. We're trying to find the point in Q that's closest to P.

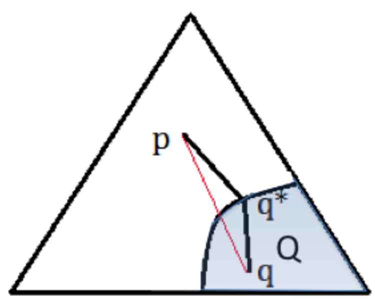

Figure 5.1: Triangle depicts the simplex of all probability distributions. The angle between segments qq^* and q^*p is necessarily obtuse if Q is convex, and is 90° if Q is linear.

5.1.3 Information Geometrically Orthogonal families

From figure 5.1.2, if we think of $D(q \,||\, p)$ as distance squared, then Pythagora's Theorem states that, in a triangle with an obtuse angle, the square of the distance of the side opposite to the obtuse angle is greater than the sum of the squared-distance of the other two sides.

Theorem 5.1 Pythagorean theorem for Information Projection
If Q is closed and convex and $p \notin Q$, and $q^* = \underset{q \in Q}{\operatorname{argmin}} D(q \,||\, p)$ then $\forall q \in Q$

$$D(q \,||\, p) \geq D(q \,||\, q^*) + D(q^* \,||\, p).$$

For what class of distributions, does the Pythagorean theorem hold with equality? Once again refering to figure 5.1.2, we expect that if the set Q corresponds to a line, then the angle between segments qq^* and q^*p is 90° and we have the pythagorean identity as follows. Also see figure 5.1.3

Theorem 5.2 Pythagorean identity for Information Projection
If $Q = Q_{linear}$

$$D(q \,||\, p) = D(q \,||\, q^*) + D(q^* \,||\, p).$$

Recall that the information projection q^* for Q_{linear} belongs to the exponential family. In fact, if we sweep through the constants in the linear constraints α_is, we get different linear families and the corresponding I-projections q^* are different distributions belonging to the exponential family. The same is true if we vary the base distribution p or the functions $r_i(x)$ specifying the linear constraints. Thus,

The exponential family is **"information geometrically orthogonal"** *to the linear family.*

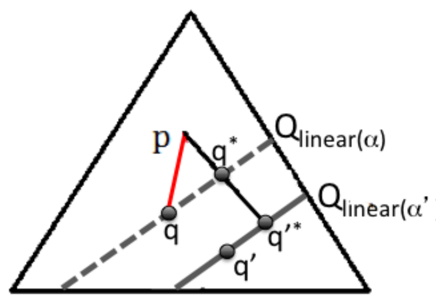

Figure 5.2: Information projections q^* and $q^{'*}$ for two linear families with different constraint parameters α and α'. All points along the line joining p to q or q' belong to the exponential family and are obtained by sweeping through different αs of the corresponding linear family. Thus, linear family and exponential family are information-geometrically orthogonal.

The notion of information projection will also be useful later when we talk about large deviation theory. Here is an example:

Example: ***Large deviation theory*** What's the probability that the average of n fair coin tosses (0,1) is greater than 3/4, i.e. more than 3n/4 tosses result in a 1? Solution: Consider the set of all distributions which have the same empirical distribution as the sequence we observe.

$$Q = \{q : \quad q(1) \geq 3/4\}$$

Then we will show that if $p = (1/2, 1/2)$ is the true distribution of the fair coin, then

$$\begin{aligned} Pr(Q) = Pr(x^n : \text{empirical distribution of } x^n \text{ is in } Q) &\approx 2^{-n \min_{q \in Q} D(q||p)} \\ &\approx 2^{-nD((1/4,3/4)||(1/2,1/2))} \end{aligned}$$

5.1.4 Maximum Likelihood Estimation under Exponential Families

Define the exponential family of distributions $E(r_i(x), p(x))$ as set of distributions of the form

$$q(x) \propto p(x) e^{\sum_i \lambda_i r_i(x)}$$

ML Estimation

$$\begin{aligned} q^*_{ML}(x) &= \underset{q \in E(r_i(x),\, p(x))}{\text{argmax}} \prod_{j=1}^{n} q(x_j) \\ &= \underset{q \in E(r_i(x),\, p(x))}{\text{argmin}} \mathbb{E}_{\hat{p}}[\log \frac{1}{q(x)}] \\ &= \underset{q \in E(r_i(x),\, p(x))}{\text{argmin}} D(\hat{p} \,||\, q) \end{aligned}$$

From previous lecture, we have seen that q^*_{ML} has the exponential family parametrization:

$$\begin{aligned} q^*_{ML}(x) &\propto p(x) e^{\sum_i \lambda^*_{i\,ML} r_i(x)} \\ \text{where } \lambda^*_{ML} \text{ chosen s.t.} & \\ \mathbb{E}_{q^*_{ML}}[r_i(X)] &= \mathbb{E}_{\widehat{p}}[r_i(X)] \\ \sum_x q^*_{ML}(x) r_i(x) &= \frac{1}{n} \sum_{j=1}^{n} r_i(x_j) \quad \forall i \end{aligned}$$

Define $Q_{linear} = \{q : \quad \mathbb{E}_q[r_i(X)] = \mathbb{E}_{\hat{p}}[r_i(X)]\}$ i.e. the linear constraints are given by the empirical moments of data. Then the maximum likelihood estimator is equivalent to the information projection of p onto Q_{linear}: $q^*_{IP} = \arg\min_{q \in Q_{linear}} D(q \,||\, p)$. Thus,

$$\begin{aligned} q^*_{ML_{Exp}} &= q^*_{IP} \quad \text{if } Q = Q_{linear} \text{ and } \alpha_i = \mathbb{E}_{\hat{p}}[r_i(X)] \\ &= q^*_{ME} \quad \text{if } Q = Q_{linear},\ \alpha_i = \mathbb{E}_{\hat{p}}[r_i(X)] \text{ and } p = u, \text{ the uniform distribution.} \end{aligned}$$

5.2 Max Entropy Rate Stochastic processes

Entropy of random variable $X : H(X)$

The joint entropy of $X_1 \dots X_n$:

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

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

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

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 5.3 *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 5.4 Burg's Maximum Entropy Theorem
The max entropy rate stochastic process $\{X_i\}$ satisfying the constraints

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

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

$$X_i = -\sum_{i=1}^{p} a_k X_{i-k} + Z_i \qquad Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

where 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 $AR(p)$ auto-regressive Gauss-Markov process of order p arise as natural solutions when finding maximum entropy stochastic processes under second-order moment constraints up to lag p.

Proof: Let $X_1 \ldots X_n$ be a stochastic process that satisfies constraints $\star$. Let $Z_1 \ldots Z_n$ be a Gaussian process that satisfies constraints $\star$.

Let $Z'_1 \ldots Z'_n$ be a p^{th} order Gauss-Markov process with the same some distribution as $Z_1 \ldots Z_n$ 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:

$$\begin{aligned}
H(X_1,\ldots,X_n) &\leq H(Z_1,\ldots,Z_n) \\
&= H(Z_1,\ldots,Z_p) + \sum_{i=p+1}^{n} H(Z_i|Z_{i-1},\ldots,Z_1) \quad \text{(chain rule)} \\
&\leq H(Z_1,\ldots,Z_p) + \sum_{i=p+1}^{n} H(Z_i|Z_{i-1},\ldots,Z_{i-p}) \quad \text{(conditioning does not increase entropy)} \\
&= H(Z'_1,\ldots,Z'_p) + \sum_{i=p+1}^{n} H(Z'_i|Z'_{i-1},\ldots,Z'_{i-p}) \\
&= H(Z'_1,\ldots,Z'_n)
\end{aligned}$$

$$\Rightarrow \lim_{n\to\infty} \frac{1}{n} H(X_1 \ldots X_n) \leq \lim_{n\to\infty} \frac{1}{n} H(Z'_1 \ldots Z'_n)$$

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

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

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

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

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

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