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

Lecture 4: Maximum Entropy Distributions and Exponential Family

Lecturer: Aarti Singh *Scribes: Min Xu*

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.*

4.1 Optimizing for Maximum Entropy Distributions

The following optimization solves for a maximum entropy distribution that satisfies linear constraints (typically moment constraints):

$$\min_f - H(f) \tag{4.1}$$

$$\text{s.t. } f(x) \geq 0 \tag{4.2}$$

$$\int f(x)dx = 1 \tag{4.3}$$

$$\int f(x)r_i(x)dx = \alpha_i \text{ for } i = 1, ..., n \tag{4.4}$$

$$\int f(x)s_i(x)dx \leq \beta_i \text{ for } i = 1, ..., m \tag{4.5}$$

First, we will derive the form of the maximum entropy distribution subject to these constraints using a crude argument. Then we will present a formal proof that the derived form indeed is the maximum entropy distribution subject to given constraints.

Notice that the Lagragian is

$$L(f, \lambda) = -H(f) + \lambda_0 \int f(x)dx + \sum_{i=1}^{n} \lambda_i \int f(x)r_i(x)dx + \sum_{i=n+1}^{n+m} \lambda_i \int f(x)s_i(x)dx \tag{4.6}$$

where $\lambda_{n+1}, ..., \lambda_{n+m} \geq 0$.

For the rest of the derivation, we will use the crude argument that we can think of a function f as an infinite-dimensional continuous vector with $f(x)$ as the value at each coordinate. Under this simplification, $\int f(x)dx$ is similar to $\sum_x f_x$.

We can take derivative of $L(f, \lambda)$ with respect to $f(x) \equiv f_x$ treating f is a vector and all integrals as just summations.

$$\frac{\partial L(f, \lambda)}{\partial f(x)} = \frac{f(x)}{f(x)} + \log f(x) + \lambda_0 + \sum_{i=1}^{n} \lambda_i r_i(x) + \sum_{i=n+1}^{n+m} \lambda_i s_i(x)$$

Setting $\frac{\partial L(f,\lambda)}{\partial f(x)} = 0$ for all x, we get that

$$f^*(x) = \exp\left[-1 - \lambda_0^* - \sum_{i=1}^{n} \lambda_i^* r_i(x) - \sum_{i=n+1}^{n+m} \lambda_i^* s_i(x)\right]$$

where the $\{\lambda^*\}$'s are chosen such that $f^*(x)$ satisfies the constraints. If the constraints cannot be satisifed for any values of λ^*'s, then the maximum entropy distribution does not exist.

Now we formally prove that f^*, as derived above, is indeed the maximum entropy distribution.

Theorem 4.1 *For all distributions f that satisfy the constraints, we have*

$$H(f^*) \geq H(f)$$

Proof:

$$\begin{aligned}
H(f) &= -\int f \log f \frac{f^*}{f^*} \\
&= -D(f\,||\,f^*) - \int f \log f^* \\
&\leq -\int f \log f^* \text{ (holds with inequality when } f = f^*) \\
&= -\int f \left(-1 - \lambda_0^* - \sum_{i=1}^{n} \lambda_i^* r_i(x) - \sum_{i=n+1}^{n+m} \lambda_i^* s_i(x)\right) \\
&\leq 1 + \lambda_0^* + \sum_{i=1}^{n} \lambda_i^* \alpha_i + \sum_{i=n+1}^{n+m} \lambda_i^* \beta_i \quad \text{(since } f \text{ satisfies the constrainst and } \lambda_{n+1}^*, \ldots, \lambda_{n+m}^* \geq 0) \\
&= -\int f^* \left(-1 - \lambda_0^* - \sum_{i=1}^{n} \lambda_i^* r_i(x) - \sum_{i=n+1}^{n+m} \lambda_i^* s_i(x)\right) \\
&\qquad \text{(since, by complementary slackness, } \lambda_i^* \left(\int f^* s_i(x) - \beta_i\right) = 0 \text{ for } i = n+1, \ldots, n+m.) \\
&= -\int f^* \log f^* = H(f^*)
\end{aligned}$$

■

Thus, distirbutions belonging to the exponential family arise as natural solutions to the maximum entropy problem subject to linear constraints. This provides a justification for the use of exponential family models.

4.2 Examples of Exponential Family

- $N(\mu, \sigma^2)$

  $f^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$

- $Exp(\lambda)$

  $f^*(x) = \lambda \exp(-\lambda x)$ where $x \geq 0$

- $Ber(\theta)$

  $f^*(x) = \theta^x (1-\theta)^{1-x} = \exp\{x \log \frac{\theta}{1-\theta} + \log(1-\theta)\}$

Several other parametric distributions that are used commonly for modeling belong to the exponential family, and arise as solutions to the maximum entropy problem under different linear moment constraints and support sets. We give some examples below.

Example:
In a lot of cases, it is possible that the maximum entropy distribution does not exist. For example, suppose the only constraint is $\mathbb{E}[X] = \int f(x)xdx = \mu$. We must choose λ_0 and λ_1 such that

$$\int_{-\infty}^{\infty} e^{\lambda_0+\lambda_1 x}dx = 1$$
$$e^{\lambda_0}\frac{e^{\lambda_1 x}}{\lambda_1}\,|_{-\infty}^{\infty} = 1$$

For all possible values of λ_1, the above equation does not hold.

Example:
If we take the above case and add a support restriction, then the maximum entropy distribution does exist. Suppose that we have two constraints: (1) $\int f(x)I_{[0,\infty)}(x)dx = 1$ and (2) $\mathbb{E}[X] = \int f(x)xdx = \mu$

In this case, we must find λ_0 and λ_1 such that

$$\int_{0}^{\infty} e^{\lambda_0+\lambda_1 x} = 1$$
$$e^{\lambda_0}\frac{e^{\lambda_1 x}}{\lambda_1}\,|_{0}^{\infty} = 1$$

From above, we see that $\lambda_1 = -e^{\lambda_0}$

We now use the second constraint:

$$\int_0^{\infty} xe^{\lambda_0+\lambda_1 x}dx = -\lambda_1\int_0^{\infty} xe^{\lambda_1 x}dx = \mu$$

Using integration by parts, we set $v = x, du = e^{\lambda_1 x}$ and get that

$$\mu = -\lambda_1\left[\frac{e^{\lambda_1 x}}{\lambda_1}x\bigg|_0^{\infty} - \int_0^{\infty}\frac{1}{\lambda_1}e^{\lambda_1 x}dx\right]$$

Solving, we get $\lambda_1 = -\frac{1}{\mu}$. Thus, the maximum entropy distribution with mean μ that is supported on the non-negative reals is the exponential distribution $f^*(x) = \frac{1}{\mu}e^{-x/\mu}$.

Example:
Suppose the support is $(-\infty,\infty)$ and we impose two constraints: $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2 - \mu^2] = \sigma^2$, then the maximum entropy distribution is a Gaussian with mean μ and variance σ^2. You will prove this in the Homework.

4.3 Information Projection

First, we will show that maximizing the entropy over a set of linearly constrained distributions, is equivalent to minimizing the relative entropy with respect to the uniform distribution supported on the largest support of any distribution in the set (if it exists). Let Q_{linear} be the set of all distributions that satisfy the linear constraints and let u be the dominating uniform distribution as described above.

$$\begin{aligned} q^* &= \arg \max_{q \in Q_{linear}} H(q) \\ &= \arg \min_{q \in Q_{linear}} -H(q) - \mathbb{E}_q[\log u] \\ &= \arg \min_{q \in Q_{linear}} D(q \,||\, u) \end{aligned}$$

Definition 4.2 *More generally, we define* ***information projection*** *of a distribution p onto a set of distributions Q as*

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

If all distributions in Q have bounded support, p is the dominating uniform distirbution and $Q = Q_{linear}$, then the information projection is the maximum entropy distribution in Q. If p is not uniform and $Q = Q_{linear}$, the information projection has the form:

$$q^*(x) = p(x) \exp \left[1 - \lambda_0^* - \sum_{i=1}^{n} \lambda_i^* r_i(x) - \sum_{i=n+1}^{n+m} \lambda_i^* s_i(x) \right]$$

You will show this in the Homework.

Information projection have a nice geometric interpretation captured by the following pythagoras theorem:

Theorem 4.3 *(Pythagoras Theorem for Information Projection)*
Suppose Q is closed and convex and suppose $p \notin Q$. Let $q^* = \min_{q \in Q} D(q \,||\, p)$, then we have

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

See figure 4.3 for an intuitive graphical explanation of the Pythagora's theorem. This implies that information divergence behaves as the square of euclidean distance since if the angle between two vectors AB and BC is obtuse, then $d_{AC}^2 \geq d_{AB}^2 + d_{BC}^2$. (Recall, however, that information divergence is not symmetric.)

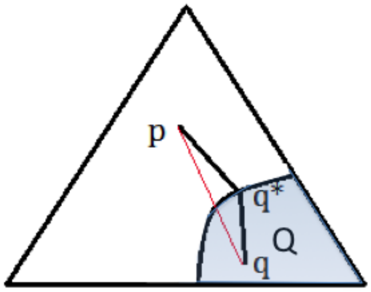

Figure 4.1: Triangle depicts the simplex of all probability distributions. The angle between segments qq^* and q^*p is necessarily obtuse if Q is convex. If we think of $D(q \,||\, p)$ as distance squared, then Pythagora's Theorem states that, in a triangle with an obtuse angle, the square of the distance of the side opposite to the obtuse angle is greater than the sum of the squared-distance of the other two sides.

Note: If the family of distributions Q is linear, that is, if $q_1, q_2 \in Q$ implies that mixtures of q_1, q_2 must be in Q as well, then set Q in this case corresponds to a straight line and the corresponding angle is a right angle. In this case, Pythagora's theorem for information projection holds with equality, and q^* is the exponential family with base distribution p. We will discuss this more in next class.

4.4 Maximum Likelihood Estimation in Exponential Family

We first define the exponential family distributions more formally:

Definition 4.4 *Let $\{r_i(x)\}_{i=1,...,m}$ be a collection of statistics. Let $p(x)$ be a distribution. We say that a distribution $q(x)$ is in the exponential family of (r,p), denoted $q \in E_{(r,p)}$, if*

$$q(x) \propto p(x)\exp\{\lambda_0 + \sum_{i=1}^{m}\lambda_i r_i(x)\}$$

where $\lambda_0, ..., \lambda_m$ are real numbers that parametrize the exponential family $E_{(r,p)}$.

Given i.i.d. data $X_1, ..., X_n$, the maximum likelihood problem for the exponential family is the following:

$$\begin{aligned}
q^{**} &= \arg\max_{q\in E_{(r,p)}} \prod_{i=1}^{n} q(X_i) \\
&= \arg\min_{q\in E} \sum_{i=1}^{n}\log\frac{1}{q(X_i)} \\
&= \arg\min_{q\in E} \mathbb{E}_{\widehat{p}}\left[\log\frac{1}{q(X_i)}\right] \\
&= \arg\min_{q\in E} Risk_{\widehat{p}}(q) \\
&= \arg\min_{q\in E} Risk_{\widehat{p}}(\hat{p}) + D(\widehat{p}\,||\,q) \quad \text{from lecture 1 about negative log likelihood loss} \\
&= \arg\min_{q\in E} D(\widehat{p}\,||\,q)
\end{aligned}$$

Note that $\arg\min_{q\in E} D(\widehat{p}\,||\,q)$ is **NOT** the information projection of $\widehat{p}$ onto $E_{(r,p)}$ because we have $D(\widehat{p}\,||\,q)$ instead of $D(q\,||\,\hat{p})$.

The following theorem relates maximum likelihood parameter estimation in exponential family to information projection:

Theorem 4.5 *Let $r_i(x)$ be a family of statistics for $i = 1, ..., m$. Suppose $Q_{linear,(r,\widehat{p})}$ is the set of all distributions that satisfy $\mathbb{E}_q[r_i(x)] = \alpha_i := \mathbb{E}_{\widehat{p}}[r_i(x)]$. Let p be a fixed distribution. Then we have*

$$q^{**} := \arg\min_{q\in E_{(r,p)}} D(\widehat{p}\,||\,q) = \arg\min_{q\in Q_{linear,(r,\widehat{p})}} D(q\,||\,p) =: q^*$$

The theorem states that the distribution belonging to the exponential family (with sufficient statistics $r_i(x)$ and base distrbution $p(x)$) whose parameters maximize the likelihood of data, is same as the information projection of $p(x)$ on to a set of distributions with linear equality constraints (specified by $r_i(x)$) that are given by data.

Proof: Recall the form of q^* from Section 1 and notice that we can rewrite the distribution as

$$q^* = p(x)\frac{\exp\left(-\sum_{j=1}^{m}\lambda_j^* r_j(x)\right)}{\Psi(\lambda_1^*, ..., \lambda_m^*)}$$

where $\Psi(\lambda_1^*, ..., \lambda_m^*) = \sum_x p(x)\exp\left(-\sum_{j=1}^{m}\lambda_j^* r_j(x)\right)$ is the normalization constant. Here λ_j^* are chosen such that $q^* \in Q_{linear,(r,\widehat{p})}$, i.e. $\sum_x q^*(x) r_j(x) = \mathbb{E}_{\widehat{p}}[r_j(X)]$.

We will show that the maximum likelihood distribution q^{**} in $E(r,p)$ has parameters λ^{**} that satisfy the same constraints. The maximum likelihood parameters for $E(r,p)$ are given by

$$\begin{aligned}\lambda_1^{**},...,\lambda_m^{**} &= \arg\max_{\lambda_1,...,\lambda_m} \prod_{i=1}^{n} p(X_i)\frac{\exp\left(\sum_{j=1}^{m}\lambda_j r_j(X_i)\right)}{\Psi(\lambda_1,...,\lambda_m)} \\ &= \arg\max_{\lambda_1,...,\lambda_m} \sum_{i=1}^{n}\left[\log p(X_i) + \sum_{j=1}^{m}\lambda_j r_j(X_i) - \log\Psi(\lambda_1,...,\lambda_m)\right]\end{aligned}$$

Taking derivative with respect to $\lambda_1,...,\lambda_m$ of the log likelihood function, we get that

$$\begin{aligned}\frac{\partial}{\partial\lambda_j} &= \sum_{i=1}^{n} r_j(X_i) - n\frac{\partial}{\partial\lambda_j}\log\Psi(\lambda_1,...,\lambda_m) \\ &= \sum_{i=1}^{n} r_j(X_i) - \frac{n}{\Psi(\lambda_1,...,\lambda_m)}\frac{\partial}{\partial\lambda_j}\Psi(\lambda_1,...,\lambda_m) \\ &= \sum_{i=1}^{n} r_j(X_i) - \frac{n}{\Psi(\lambda_1,...,\lambda_m)}\sum_x p(x)r_j(x)\exp\left(\sum_j \lambda_j r_j(x)\right) \\ &= \sum_{i=1}^{n} r_j(X_i) - n\sum_x\left[p(x)\frac{\exp(\sum_j\lambda_j r_j(x))}{\Psi(\lambda_1,...,\lambda_m)}\right]r_j(x)\end{aligned}$$

Since the derivative is zero for λ_j^{**}, we have:

$$\sum_x\left[p(x)\frac{\exp(\sum_j\lambda_j^{**}r_j(x))}{\Psi(\lambda_1,...,\lambda_m)}\right]r_j(x) = \frac{1}{n}\sum_{i=1}^{n} r_j(X_i)$$

Or equivalently,

$$\sum_x q^{**}(x)r_j(x) = \frac{1}{n}\sum_{i=1}^{n} r_j(X_i) = \mathbb{E}_{\widehat{p}}[r_j(X)]$$

Thus, q^{**} is same as q^*. ■