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

Lecture 7: February 3

Lecturer: Aarti Singh *Scribes: Erik Louie*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

7.1 Review

7.1.1 Maximum Entropy

In the Maximum Entropy (MaxEnt) problem we seek to maximize entropy over all probability densities p

$$p^*(x) = \underset{p \in \Delta}{\operatorname{argmax}} - \int p(x) \log p(x) dx,$$

given constraints

$$\begin{aligned} \text{subject to} \qquad & p \geq 0, \int p = 1 \\ & \mathbb{E}_p[r_i(x)] = \alpha_i \qquad\qquad i \in 1, \ldots, n, \\ & \mathbb{E}_p[s_j(x)] \leq \beta_j \qquad\qquad j \in 1, \ldots, m \end{aligned}$$

where $r_i, s_j : \mathbb{R} \to \mathbb{R}$.

Since entropy is a concave function over a convex set, we can differentiate the Lagrangian (shown in the previous notes) and obtain the form for maximizing density:

$$p^\star(x) \in e^{1-\lambda_0^\star - \sum_i \lambda_i^\star r_i(x) - \sum_{j=1}^m \nu_j^\star s_j(x)},$$

For Lagrange parameters $\lambda^\star, \nu^\star$ that are chosen so that $p^\star$ meets the constraints. Note that if the constraints, r_i are linear, then these distributions belong to the exponential family.

Several 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 7.1 *Multivariate Gaussian distribution*

Given the constraints

$$\begin{aligned} \mathbb{E}_p[X_i X_j] &= K_{ij} \qquad\qquad \forall i, j \in 1, \ldots, p \\ \mathbb{E}_p[X_i] &= 0 \end{aligned}$$

then this is the multivariate Gaussian distribution and the MaxEnt density is

$$p^*(x) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} e^{-\frac{1}{2} x^T K^{-1} x} \tag{7.1}$$

Example 7.2 *Ising graphical model*

Graphical models are a special case of exponential families. In the Ising graphical model, we consider interacting spins of electrons, represented by $X_i \in 0, 1$. Given constraints on the first and second moments

$$\begin{aligned} \mathbb{E}_p[X_i X_j] &= x_i x_j & \forall i, j \in 1, \ldots, n \\ \mathbb{E}_p[X_i] &= x_i & \forall i \in 1, \ldots, n \end{aligned}$$

then we obtain the probability density

$$p(x) \propto e^{\sum_{i,j} \lambda_{ij}(x_i * x_j + (1-x_i)(1-x_j))}.$$

Example 7.3 *Discrete distributions (Gibb's distribution)*

The solution to the linear constraints can be rewritten to simplify entry conditions for different sets of problems. λ_i^* is such that p^* satisfies the constraints. In the case of the discrete distribution, p^* is proportional to the normalized solution, commonly known as the Gibb's distribution:

$$p^*(x) = \frac{e^{\sum_i \lambda_i^* r_i(x)}}{Z_{\lambda^*}}$$

where the normalizing constant is the partition function:

$$Z_{\lambda^*} = \sum_{j=1}^{n} e^{\sum_i \lambda_i^* r_i(x_j)}$$

7.1.2 Information Projection

Definition 7.4 *Information Projection*

We define information projection of a distribution p onto a set of distributions P as

$$p^*(x) = \operatorname*{argmin}_p \mathcal{D}(p||p_0)$$

If all distributions in P have bounded support, p is the dominating uniform distribution and P has linear constrains, ie. $\mathbb{E}_p[r_i(X)] = \alpha_i$, then the information projection is the maximum entropy distribution in P. We can show that the probability density estimator for P with linear constraints is

$$p^*(x) = \frac{p(x)e^{\sum_i \lambda_i r_i(x)}}{\sum_x p(x)e^{\sum_i \lambda_i r_i(x)}}$$

i.e. it is in the exponential family.

7.1.3 I-Geometry

Information projection has a nice geometric interpretation captured by the following Pythagoras theorem:

Theorem 7.5 *Pythagorean Theorem for KL-Divergence*

Let $\mathcal{P}$ be closed and convex, $p_0 \notin \mathcal{P}$, and $p^* = \underset{p \in P}{\operatorname{argmin}} \mathcal{D}(p||p_0)$, then

$$\mathcal{D}(p||p_0) \geq \mathcal{D}(p||p^*) + \mathcal{D}(p^*||p_0)$$

Note that this does not satisfy the triangle inequality.

$$\implies \mathcal{D}(\cdot||\cdot) \equiv (Euclideandistance)^2$$

Equality holds if $\mathcal{P}$ is linear.

See 7.1.3 for an intuitive graphical explanation of the Pythagoras 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.)

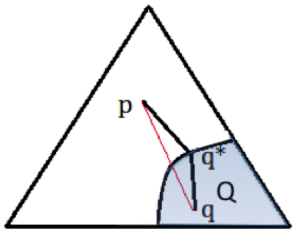

Figure 7.1: *Triangle depicts the simplex of all probability distributions. The angle between segments qq^* and qp is necessarily obtuse if Q is convex. If we think of $D(q||p)$ as distance squared, then Pythagoras Theorem states that, in a triangle with an obtuse angle, the square of the distance of the side opposite to the obtuse angle is greater than the sum of the squared-distance of the other two sides.*

Example 7.6 *n tosses of a fair coin*

What is the probability that the average of n fair coin tosses is greater than $\frac{3}{4}$? Let us consider the set of all distributions that have the same empirical distribution as the sequence we observe:

$$P = \{p: \quad p_H \geq \frac{3}{4}n\}$$

then we can show that if our distribution is a fair coin, $p = (\frac{1}{2}, \frac{1}{2})$, then

$$\begin{aligned} Pr(x^n : \textit{empirical distribution of } x^n \textit{ is in } P) &\approx 2^{-n \underset{p \in \mathcal{P}}{\operatorname{argmax}} \mathcal{D}(p||p_0)} \\ &\approx 2^{-n\mathcal{D}((\frac{3}{4},\frac{1}{4})||(\frac{1}{2},\frac{1}{2}))} \end{aligned}$$

7.2 Maximum Entropy Duality

Maximum likelihood estimate is in the exponential family given empirical constraints. Given

$$\exp p_\lambda(x) = \frac{p_0(x) \exp \sum_i \lambda_i r_i(x)}{Z_\lambda} = \mathcal{P},$$

then

$$\begin{aligned} P^*_{ML}(x) &= \underset{p\in\lambda}{\operatorname{argmax}} \prod_{i=1}^{n} p_\lambda(x_i) \\ &= \underset{p\in\lambda}{\operatorname{argmin}} \sum_{i=1}^{n} \log \frac{1}{p_\lambda(x_i)} \\ &= \underset{p\in\lambda}{\operatorname{argmin}} \mathcal{D}(\hat{p}||p_\lambda) + H(\hat{p}) \\ &= \underset{p\in\lambda}{\operatorname{argmin}} \mathcal{D}(\hat{p}||p_\lambda), \end{aligned}$$

since the solution is equivalent without $H(\hat{p})$. Note that the final solution is not the same as the projection.

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

Theorem 7.7 *Duality Theorem*

Let $\alpha_i = \mathbb{E}_{\hat{p}}[r_i(x)]$, *then*

$$\begin{aligned} p^*_{ML}(x) &= \underset{p\in\lambda}{\operatorname{argmin}} \mathcal{D}(\hat{p}||p_\lambda) \\ &= \underset{\substack{p\in\mathcal{P} \\ \mathbb{E}_p[r_i(x)]=\alpha_i}}{\operatorname{argmin}} \mathcal{D}(p||p_0) \\ &= p^*_{IP}(x) \end{aligned}$$

The theorem states that the distribution belonging to the exponential family (with sufficient statistics $r_i(x)$ and base distribution $p_0(x)$) whose parameters maximize the likelihood of data, is the same as the information projection of $p_0(x)$ on to a set of distributions with linear equality constraints (specified by $r_i(x)$) that are given by data.

Proof: We must show that the λ's come from the distribution and satisfy linear constraints. Let $r_i(x)$ be sufficient statistics, $p_0(x)$ the base distribution as above, and

$$\begin{aligned} Z_\lambda &= \sum_\lambda p_0(x) \exp[\sum_i \lambda_i r_i(x)] \qquad \text{and} \\ \lambda^{**} &= \underset{\lambda}{\operatorname{argmax}} \prod_{i=1}^{n} p_\lambda(x_i) \\ &= \underset{\lambda}{\operatorname{argmax}} \sum_{i=1}^{n} [\log p_0(x) + \sum_i \lambda_i r_i(x) - \log Z_\lambda] \qquad . \end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial \lambda_j}\lambda^{**} &= \sum_{i=1}^{n} r_j(x_i) - n\frac{\partial}{\partial \lambda_j}\log Z_\lambda \\
&= \sum_{i=1}^{n} r_i(x_i) - \frac{n}{Z_\lambda}\frac{\partial Z_\lambda}{\partial \lambda_j} \\
&= \sum_{i=1}^{n} r_i(x_i) - \frac{n}{Z_\lambda}\sum_{\lambda} p(x) r_j(x) \exp[\sum_i \lambda_i r_i(x)] \\
&= \sum_{i=1}^{n} r_i(x_i) - n\sum_{x}[\frac{p(x) r_j(x) \exp[\sum_i \lambda_i r_i(x)]}{Z_\lambda}] r_i(x) \\
&= \sum_{i=1}^{n} r_i(x_i) - n\sum_{x}[p_\lambda(x)] r_i(x) \\
|_{\lambda=\lambda^{**}_{ML}} &= 0
\end{aligned}$$

taking λ^{**}_{ML} for λ, then we find the expectation to be equal:

$$\begin{aligned}
&\implies \sum_x p_{\lambda^{**}_{ML}}(x) r_i(x) = \frac{1}{n}\sum_{i=1}^{n} r_j(x_i) \\
&\implies \mathbb{E}_{p_{\lambda^{**}ML}}[r_j(x)] = \mathbb{E}_{\hat{p}}[r_i(x)]
\end{aligned}$$

■

We can also use Lagrange duality to show the above.

7.2.1 Maximum Entropy Generalization

Given the mutual information estimator

$$\min_{p\in\Delta} \mathcal{D}(p||p_0) \quad \text{s.t.} \quad \mathbb{E}_p[r] = \mathbb{E}_{\hat{p}}[r],$$

then the primal is

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

where $U(\mathbb{E}_p[r])$ is a regularizer. Any $U(p)$ can be used, but to obtain linear constraints, we use the regularizer with respect to expectation. Here are three example regularizers:

Example 7.8 *Regularizer*

$$U(p) = 1(\mathbb{E}_p[r] = \mathbb{E}_{\hat{p}}[r])$$

Example 7.9 *L1 Norm Regularizer*

$$U(p) = 1(|\mathbb{E}_p[r_j] - \mathbb{E}_{\hat{p}}[r_j]| \leq \beta_j) \quad \forall j$$

Example 7.10 *L2 Norm Regularizer*

$$U(p) = \frac{||\mathbb{E}_p[r_j] - \mathbb{E}_{\tilde{p}}[r_j]||}{2\alpha} \quad \forall j$$

Returning to the generalized maximum entropy problem, the dual is

$$\psi(p) = \left\{ \begin{array}{ll} \mathcal{D}(p||p_0) & p \in \Delta \\ \infty & p \notin \Delta \end{array} \right\},$$

which is closed, convex, and proper. A function is proper if it is not always infinity. Then we have the conjugate, which can be thought of as a gradient function

$$\psi^*(\lambda) = \sup_p [\lambda p - \psi(p)].$$

Definition 7.11 *Fenchel's Duality*

Let ψ, φ be closed, proper, and convex, and A is any matrix. We define ***Fenchel's Duality*** *as*

$$\inf_p \psi(p) + \varphi(A_p) = \sup_\lambda -\psi^*(A^t\lambda) - \varphi^*(-\lambda)$$

This definition is useful when the function is convex, but not differentiable. For example, at a corner in a function, we can consider all tangent lines at that point.

Returning to the previous maximum entropy generalization problem. Let R be a matrix, p the density as a vector. Within ML we can consider $p(x) \equiv p_x$, which may be an infinity object. Then, we have the primal

$$\min_{p \in \Delta} \mathcal{D}(p||p_0) + U(R_p)$$

and U is closed, convex, and proper. We can think of R_{jx} as $r_j(x)$. Let $\psi(p) = D(p||p_0)$. If there is a convex divergence, then $\psi^*(\lambda) = \ln(\sum_x p_0(x)e^{\lambda_x})$. So,

$$\begin{aligned} \sup_\lambda [-\psi^*(A^t\lambda) - U^*(-\lambda)] &= \sup_\lambda [-\ln \sum_\lambda p_0(x) e^{(R^t\lambda)x}] - U^*(-\lambda) \\ &= \sup_\lambda [-\ln Z_\lambda - U^*(-\lambda)] \end{aligned}$$

7.2.2 Shifts

With respect to any data, let us choose

$$U_t[\mathcal{U}] = U(\mathbb{E}_t[r] - \mathcal{U}),$$

where $\mathbb{E}_t[r] - \mathcal{U}$ is the difference of moments of t and p. Then the dual of the shift from the mean is

$$U_t^*(\lambda) = U^*(-\lambda) + \lambda t[r]$$

Example 7.12 $U_t[\mathcal{U}] = 1(\mathcal{U} = t[r] - \tilde{\pi}[r])$

Let $Q(\lambda) = -\ln Z_\lambda - U^*(-\lambda)$*, then*

$$\begin{aligned} Q(\lambda) &= -\ln Z_\lambda - U^*(-\lambda) \\ &= -\ln Z_\lambda - U_t^*(\lambda) + \lambda t[r] \\ &= -\mathbb{E}_t[\ln p_0] + \mathbb{E}_t[\ln p_0 + \lambda t[r] - \ln Z_\lambda] - U_t^*(\lambda) \\ &= -L_t(0) - L_t(\lambda) - U_t^*(\lambda), \end{aligned}$$

where $L_t(\lambda) := -\mathbb{E}_t[\ln p_\lambda]$*, which, if* t *is with respect to the empirical data, is just the log likelihood of the data,* $t = \hat{p}$*.* t *can also be* L_1 *or* L_2*, for example.*

Then the dual is

$$\sup_\lambda Q = \min_\lambda L_t(\lambda) + U_t^*(\lambda).$$

p_λ^* *corresponds to the minimizer at the end.*

Example 7.13 $t[r] = \mathbb{E}_t[r]$

Then the dual is

$$U_t^*(\lambda) = U^*(-\lambda) + \lambda \mathbb{E}_t[r].$$