10-704: Information Processing and Learning **Fall 2016**

Lecture 8: September 26

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

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 Review

8.1.1 Maximum Entropy and Information Projection

Last time we discussed that the problem of finding the maximum entropy distribution constrained to lie in a subset $Q \subset \mathcal{P}$ is essentially equivalent to finding the information projection of the uniform distribution onto Q, i.e. the distribution in Q that is closest to uniform in KL sense [1]

$$\max_{p \in Q} H(p) = \min_{p \in Q} D(p||u)$$

If the set of constraints in Q are linear in p, i.e. of the form $\mathbb{E}_p[f_j(X)]$ equal to or bounded by some constant, then the maximum entropy distribution belongs to the exponential family:

$$p^\star(x) = \frac{\exp(\sum_j \lambda_j f_j(x))}{Z_{\boldsymbol{\lambda}}}$$

where the Lagrange parameters $\boldsymbol{\lambda} = \{\lambda_j\}$ are chosen so that $p^\star$ meets the constraints.

The information projection can be defined more generally with respect to any given base distribution $p_0(x)$ (instead of uniform):

$$\min_{p \in Q} D(p||p_0)$$

If the set of constraints in Q are linear in p, i.e. of the form $\mathbb{E}_p[f_j(X)]$ equal to or bounded by some constant, then the information projection distribution belongs to the Gibbs family:

$$p^*(x) = p_0(x)\frac{\exp(\sum_j \lambda_j f_j(x))}{Z_{\boldsymbol{\lambda}}}$$

where the normalizing constant is the partition function:

$$Z_{\boldsymbol{\lambda}} = \sum_x p_0(x) e^{\sum_j \lambda_j f_j(x)}.$$

[1] Here the uniform distribution is defined such that all distributions in Q are absolutely continuous with respect to it.

8.2 Maximum Entropy Duality with Maximum Likelihood Estimation

So far, we haven't talked about data in the discussion of maximum entropy. Often the constraints on the distribution are actually specified using the data. For example, when we seek Maximum likelihood model in the exponential (Gibbs) family then we are essentially seeking the Maximum Entropy distribution (Information Projection) given empirical constraints based on data. We will show this connection next.

Consider the maximum likelihood model given data $X_1, \ldots, X_n$

$$\begin{aligned}
p^*_{ML}(x) &= \operatorname*{argmax}_{p_{\boldsymbol{\lambda}}} \prod_{i=1}^{n} p_{\boldsymbol{\lambda}}(X_i) \\
&= \operatorname*{argmin}_{p_{\boldsymbol{\lambda}}} \sum_{i=1}^{n} \log \frac{1}{p_{\boldsymbol{\lambda}}(X_i)} \\
&= \operatorname*{argmin}_{p_{\boldsymbol{\lambda}}} \mathbb{E}_{\hat{p}} \left[\log \frac{1}{p_{\boldsymbol{\lambda}}(X)} \right] \\
&= \operatorname*{argmin}_{p_{\boldsymbol{\lambda}}} \mathbb{E}_{\hat{p}} \left[\log \frac{\hat{p}(X)}{p_{\boldsymbol{\lambda}}(X)} \right] + \mathbb{E}_{\hat{p}} \left[\log \frac{1}{\hat{p}(X)} \right] \\
&= \operatorname*{argmin}_{p_{\boldsymbol{\lambda}}} D(\hat{p}||p_{\boldsymbol{\lambda}}) + H(\hat{p}) \\
&= \operatorname*{argmin}_{p_{\boldsymbol{\lambda}}} D(\hat{p}||p_{\boldsymbol{\lambda}}),
\end{aligned}$$

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 estimation in exponential family with base distribution p_0 to information projection of p_0 onto a set of distributions with constraints specified by the empirical mean of the sufficient statistics:

Theorem 8.1 *Duality Theorem*

Let $\alpha_j = \mathbb{E}_{\hat{p}}[f_j(X)]$, then

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

The theorem states that the distribution belonging to the exponential family (with sufficient statistics $f_j(x)$ 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 $f_j(x)$) that are given by data.

Proof: Since we know the information projection lies in the exponential family, all we need to show is that the λ's in the maximum likelihood model satisfy the empirical linear constraints. So lets analyze the λ's that achieve the maximum likelihood of the data. Recall that

$$Z_{\boldsymbol{\lambda}} = \sum_x p_0(x) \exp[\sum_j \lambda_j f_j(x)] \qquad \text{and}$$

$$\begin{aligned}
\lambda^{**} &= \operatorname*{argmax}_{\lambda} \prod_{i=1}^{n} p_\lambda(X_i) = \operatorname*{argmax}_{\lambda} \prod_{i=1}^{n} \log p_\lambda(X_i) \\
&= \operatorname*{argmax}_{\lambda} \sum_{i=1}^{n} [\log p_0(X_i) + \sum_j \lambda_j f_j(X_i) - \log Z_{\boldsymbol{\lambda}}] \qquad .
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial \lambda_j} \prod_{i=1}^{n} \log p_{\boldsymbol{\lambda}}(X_i) &= \sum_{i=1}^{n} f_j(X_i) - n \frac{\partial}{\partial \lambda_j} \log Z_{\boldsymbol{\lambda}} \\
&= \sum_{i=1}^{n} f_j(X_i) - \frac{n}{Z_{\boldsymbol{\lambda}}} \frac{\partial Z_{\boldsymbol{\lambda}}}{\partial \lambda_j} \\
&= \sum_{i=1}^{n} f_j(X_i) - \frac{n}{Z_{\boldsymbol{\lambda}}} \sum_{x} p_0(x) f_j(x) \exp[\sum_{k} \lambda_k f_k(x)] \\
&= \sum_{i=1}^{n} f_j(X_i) - n \sum_{x} [\frac{p_0(x) \exp[\sum_k \lambda_k f_k(x)]}{Z_{\boldsymbol{\lambda}}}] f_j(x) \\
&= \sum_{i=1}^{n} f_j(X_i) - n \sum_{x} p_{\lambda}(x) f_j(x)
\end{aligned}$$

At the maximizing λ_{ML}^{**} the derivative is equal to 0, so we get:

$$\begin{aligned}
\implies \sum_{x} p_{\lambda_{ML}^{**}}(x) f_j(x) &= \frac{1}{n} \sum_{i=1}^{n} f_j(X_i) \\
\implies \mathbb{E}_{p_{\lambda^{**}ML}}[f_j(X)] &= \mathbb{E}_{\hat{p}}[f_j(X)]
\end{aligned}$$

■

8.3 Maximum Entropy Generalization and Duality with regularized Maximum Likelihood

We can consider a generalization of the maximum entropy (information projection) problem [DPS08]

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

where $U(\mathbb{E}_p[\mathbf{f}])$ is a regularizer and $\mathbf{f} = [f_1(X) \ldots f_m(X)]^\top$. Here are three example regularizers:

Example 8.2 *Standard Maximum entropy/Information projection is obtained with*

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

Notice that for the equivalence to hold the indicator function is defined so that 1_A is 0 if A is true and ∞ otherwise. This penalty requires the true constraints to match the empirical constraints exactly.

Example 8.3 *L1 Norm Regularizer*

$$U(\mathbb{E}_p[\mathbf{f}]) = 1(|\mathbb{E}_p[f_j] - \mathbb{E}_{\hat{p}}[f_j]| \le \beta_j) \quad \forall j$$

Here also we use the same definition of indicator function as above. This penalty requires the true constraints to match the empirical constraints in an ℓ_1 sense.

Example 8.4 *L2 Norm Regularizer*

$$U(\mathbb{E}_p[\mathbf{f}]) = \frac{||\mathbb{E}_p[\mathbf{f}] - \mathbb{E}_{\hat{p}}[\mathbf{f}]||^2}{2\alpha}$$

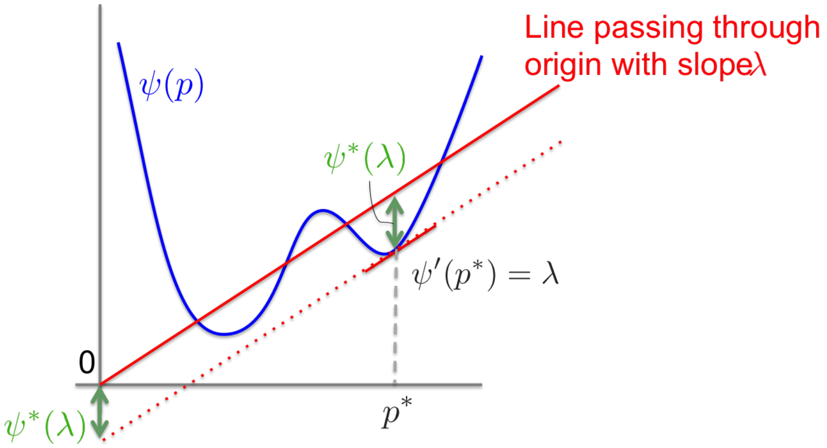
This penalty requires the true constraints to match the empirical constraints in ℓ_2 sense.

To find the solution to the generalized MaxEnt problem, we could consider taking the derivative of the regularized objective with respect to p, however notice that some of the regularizations are not differentiable. So far, we have mostly ignored such issues, assuming differentiability. But lets consider a more formal treatment via Fenchel duality (instead of Lagrangian duality) that allows us to handle convex but non-differentiable functions.

First, lets define the convex conjugate or Fenchel dual of a function $\psi(p)$ as

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

It is essentially the largest difference between a line through the origin with slope λ and the graph of the function. If the function is differentiable, the largest difference happens at a point p^* where the gradient of the function $\psi'(p^*) = \lambda$. See the image below, for example.



For a convex function, the conjugate is just a characterization of the function in terms of (intercept values of) its supporting hyperplanes corresponding to different slopes λ.

The following theorem relates a primal optimization problem of closed, proper and convex function(s) to the dual optimization problem specified in terms of convex conjugate of the function(s). Recall that a function is proper if it is not infinite everywhere.

Definition 8.5 *Fenchel's Duality*

Let ψ, φ be closed, proper, and convex, and A is any matrix. ***Fenchel's Duality*** *states that*

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

Returning to the previous maximum entropy generalization problem. We can consider $p(x) \equiv p_x$ as a vector, which may be an infinite-dimensional object. Lets define a matrix F with entries $F_{jx} = f_j(x)$. Then, $Fp = \sum_x f_j(x)p(x) = \mathbb{E}[f_j(X)]$ and we have the primal

$$\min_{p \in \mathcal{P}} \mathcal{D}(p||p_0) + U(Fp)$$

where U will be closed, convex, and proper. Let $\psi(p) = D(p||p_0)$ if $p \in \mathcal{P}$ and ∞ otherwise, which is closed, proper and convex in p. To apply Fenchel duality, we first derive the conjugate of $\psi(p)$ as

$$\psi^*(\lambda) = \ln(\sum_x p_0(x)e^{\lambda_x}).$$

For closed, convex and proper functions, the conjugate of a conjugate is the function itself, hence we instead evaluate

$$\psi^{**}(p) = \sup_\lambda[\lambda^\top p - \ln(\sum_x p_0(x)e^{\lambda_x})]$$

Taking derivative with respect to λ_x and setting it equal to 0, we get that optimal λ_x satisfies $p_x = \frac{p_0(x)e^{\lambda_x}}{\sum_x p_0(x)e^{\lambda_x}}$. Plugging this value of λ we get:

$$\begin{aligned}\psi^{**}(p) &= \lambda^\top p - \ln(\sum_x p_0(x)e^{\lambda_x}) = \sum_x p_x(\lambda_x - \ln(\sum_x p_0(x)e^{\lambda_x})) \\ &= \sum_x p_x \ln \frac{e^{\lambda_x}}{\sum_x p_0(x)e^{\lambda_x}} = \sum_x p_x \ln \frac{p_x}{p_0(x)} = D(p||p_0) = \psi(p)\end{aligned}$$

So, using Fenchel duality we have the dual problem

$$\begin{aligned}\sup_\lambda[-\psi^*(F^\top\lambda) - U^*(-\lambda)] &= \sup_\lambda[-\ln\sum_\lambda p_0(x)e^{(F^\top\lambda)_x}] - U^*(-\lambda) \\ &= \sup_\lambda[-\ln\sum_\lambda p_0(x)e^{\sum_j \lambda_j f_j(x)} - U^*(-\lambda) \\ &= \sup_\lambda[-\ln Z_\lambda - U^*(-\lambda)]\end{aligned}$$

We will show that this dual problem is essentially finding the regularized Maximum Likelihood model under exponential family with base distribution $p_0(x)$. Before we can do that, we need one more notion - that of a shifted regularizer.

8.3.1 Shifted regularization

Define a shifted regularizer with respect to any distribution t as follows

$$U_t(u) = U(\mathbb{E}_t[\mathbf{f}] - u),$$

Then the dual of the shifted regularizer is

$$\begin{aligned}U_t^*(\lambda) &= \sup_u[\lambda^\top u - U_t(u)] \\ &= \sup_u[\lambda^\top u - U(\mathbb{E}_t[\mathbf{f}] - u)] \\ &= \sup_{u'}[\lambda^\top \mathbb{E}_t[\mathbf{f}] - \lambda^\top u' - U(u')] \\ &= \lambda^\top \mathbb{E}_t[\mathbf{f}] + U^*(-\lambda)\end{aligned}$$

8.3.2 Dual as Regularized Maximum Likelihood

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

$$\begin{aligned}
Q(\lambda) &= -\ln Z_\lambda - U^*(-\lambda) \\
&= -\ln Z_\lambda - U_t^*(\lambda) + \lambda^\top \mathbb{E}_t[\mathbf{f}] \\
&= -\mathbb{E}_t[\ln p_0] + \mathbb{E}_t[\ln p_0 + \lambda^\top \mathbf{f} - \ln Z_\lambda] - U_t^*(\lambda) \\
&= -\mathbb{E}_t[\ln p_0] + \mathbb{E}_t[\ln \frac{p_0 \exp(\sum_j \lambda_j f_j(x))}{Z_\lambda}] - U_t^*(\lambda) \\
&= -L_t(0) - L_t(\lambda) - U_t^*(\lambda),
\end{aligned}$$

where $L_t(\lambda) := -\mathbb{E}_t[\ln p_\lambda]$ is the loss of the exponential family model $p_\lambda(x) = \frac{p_0 \exp(\sum_j \lambda_j f_j(x))}{Z_\lambda}$ with respect to distribution t. If $t = \hat{p}$, then this it just the negative log likelihood of the data under the model p_λ. Therefore, the dual problem is

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

and when $t = \hat{p}$, this is just regularized Maximum likelihood estimation.

We then look at some examples of how the regularization on maximum entropy/information projection transforms to regularization term in maximum likelihood solution.

Examples:

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

$$U_{\widehat{p}}(\mathbb{E}_p[\mathbf{f}]) = 1(\mathbb{E}_p[\mathbf{f}] = 0). \tag{8.1}$$
$$U_{\widehat{p}}^*(\lambda) = \sup_u [\lambda^\top u - 1(u = 0)] = 0. \tag{8.2}$$

   The last step follow since $1(u = 0)$ is infinity everywhere except when $u = 0$. The problem thus gets back to the basic maximum entropy duality with unregularized maximum likelihood.

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

   Then,

$$U_{\widehat{p}}(\mathbb{E}_p[\mathbf{f}]) = 1(|\mathbb{E}_p[f_j]| \leq \beta_j, \forall j). \tag{8.3}$$
$$U_{\widehat{p}}^*(\lambda) = \sup_u [\lambda^\top u - 1(|u_j| \leq \beta_j)] = \sum_j \beta_j |\lambda_j|, \tag{8.4}$$

   The last step follow since $1(|u_j| \leq \beta_j)$ is infinity everywhere except when $|u_j| \leq \beta_j$. Thus the expression is maximized when $u_j = sign(\lambda_j)\beta_j$. This corresponds to maximum likelihood with ℓ_1 regularization.

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

   Then,

$$U_{\widehat{p}}(\mathbb{E}_p[\mathbf{f}]) = ||\mathbb{E}_p[\mathbf{f}]||_2^2/2\alpha. \tag{8.5}$$
$$U_{\widehat{p}}^*(\lambda) = \sup_u [\lambda^\top u - u^\top u/2\alpha] = \alpha||\lambda||_2^2/2, \tag{8.6}$$

   The last step follows since the maximizing $u = \alpha\lambda$ (in this case, penalty is differentiable - simply take derivative wrt u and set to zero). This corresponds to maximum likelihood with ℓ_2^2 regularization.

References

[DPS08] M. Dudik, S.J. Phillips and R. Schapire, “Maximum Entropy Density Estimation with Generalized Regularization and an Application to Species Distribution Modeling,” *Journal of Machine Learning Research 8*, 2007, pp. 1217–1260.