10-704 Homework 2
Due: Thursday 2/26/2015

Instructions: Turn in your homework in class on Thursday 2/26/2015

1. **Maximum Entropy**

   (a) Suppose we want to maximize the entropy of a distribution supported on the non-negative integers ($\mathbb{N} \cup \{0\}$) subject to a mean constraint:

$$p^\star = \max_p - \sum_{i \in \mathbb{N} \cup \{0\}} p_i \log p_i \qquad \text{s.t.} \sum_{i \in \mathbb{N} \cup \{0\}} p_i = 1 \sum_{i \in \mathbb{N} \cup \{0\}} i p_i = \alpha$$

   Verify that the solution to this program (the MaxEnt distribution) is a Geometric distribution (i.e. $p_k = (1-\lambda)^k \lambda$ for some parameter $\lambda > 0$).

   Solution: We know that the solution will be from the Gibb's family:

$$p^\star(k) = p_k^\star = \exp(\lambda_0 + \lambda_1 k)$$

   Plugging into the mean constraint gives:

$$\sum_{k=0}^{\infty} k \exp(\lambda_1 k) = \alpha \exp(-\lambda_0)$$

   The series on the left hand side converges to $\frac{e^{\lambda_1}}{(e^{\lambda_1}-1)^2}$, provided that $\lambda_1 < 0$ (To see this, let S denote the series and consider $(1-e^{\lambda_1})S$). Meanwhile the sum to 1 constraint says:

$$\sum_{k=0}^{\infty} \exp(\lambda_1 k) = \frac{1}{e^{\lambda_1} - 1} = \exp(-\lambda_0)$$

   Combining, we have:

$$\frac{e^{\lambda_1}}{e^{\lambda_1}-1} = \alpha \Leftrightarrow e^{\lambda_1} = \frac{\alpha}{\alpha - 1}, e^{\lambda_0} = \frac{1}{\alpha - 1}$$

   Plugging back into our expression for $p^\star$ we have:

$$p_k^\star = \frac{1}{\alpha - 1} \left(\frac{\alpha}{1 - \alpha} \right)^k$$

   And the result follows by setting $\lambda = \frac{1}{\alpha - 1}$.

(b) Consider the following MaxEnt of a joint distribution $p(x_1, \dots x_d) = p(\mathbf{x})$:

$$p^\star = \max_p -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \qquad \text{s.t.} \quad \int p(\mathbf{x}) d\mathbf{x} = 1, \forall i \in [d] p_i(\cdot) = \int p(\mathbf{x}) d\mathbf{x}_{-i} = f_i(\cdot)$$

Specifically, given marginal functions $f_i : \mathcal{X}_i \to \mathbb{R}$ (assume $\int f_i(x_i) dx_i = 1$ and $f_i(x_i) \geq 0$), we want the actual marginals of p to match. The notation $d\mathbf{x}_{-i}$ denotes integration over all but the ith variable, which is an argument to $p_i(\cdot)$.

Solution: We write down the marginal constraints more explicitly in terms of expectations. For each $i \in [d]$, and for each value $x_i \in \mathcal{X}_i$, we have a constraint of the form:

$$p_i(x_i) = \mathbb{E}_{\mathbf{x} \sim p} \mathbf{1}[x_i = \mathbf{x}_i] = f_i(x_i)$$

So we will introduce a lagrange function $\lambda_i : \mathcal{X}_i \to \mathbb{R}$ to encode this constraints. As before, we start with the Gibbs family:

$$p^\star(\mathbf{x}) = \exp\left(\lambda_0 + \sum_{i=1}^d \lambda_i(x_i)\right) = \exp(\lambda_0) \prod_{i=1}^d \exp(\lambda_i(x_i))$$

We must make the marginals match, and we know that the ith marginal is:

$$\begin{aligned} p_i^\star(\cdot) &= \exp(\lambda_i(\cdot)) \int \exp(\lambda_0) \prod_{j \neq i} \exp(\lambda_j(x_j)) d\mathbf{x}_{-i} \\ &= \frac{\exp(\lambda_i(\cdot))}{\int_{\mathcal{X}_i} \exp(\lambda_i(x_i)) dx_i} = f_i(\cdot) \end{aligned}$$

The transition from the first line to the second line uses the fact that $\int p(\mathbf{x}) d\mathbf{x} = 1$ so that integrating out everything else makes the normalization just over x_i. At this point, we can set $\lambda_i(\cdot) = \log f_i(\cdot)$ and we know that since f_i is a valid marginal function, it integrates to one.

Since this is true for all i, the maximum entropy distribution is just the product of the marginals.

$$p^\star(\mathbf{x}) = \prod_{i=1}^d f_i(x_i).$$

2. **Relative Entropy**

(a) Show that relative entropy $D(p\|q)$ is convex in p.

Solution: Since:

$$D(p\|q) = -H(p) - \int p \log q$$

and the second term is *linear* in p, we must simply show that the negative entropy is *convex*.

The first and second derivatives are:

$$\frac{\partial - H(p)}{\partial p(x)} = \log p(x) + 1 \qquad \frac{\partial^2 - H(p)}{\partial p(x)^2} = \frac{1}{p(x)}$$

and the cross terms in the Hessian operator are zero. Since p is a distribution, it must be the case that $p(x) \geq 0$ which means that the Hessian operator is positive semi-definite. This implies that the negative entropy is convex.

(b) Derive the conjugate function of D w.r.t. p

Solution: By definition, the conjugate function is:

$$\psi(f) = \sup_p \langle f, p \rangle - D(p\|q) = \sup_p \int p(x) \left(f(x) - \log p(x) + \log q(x) \right)$$

This is a concave function in p from part (a) above, so we can maximize by setting the derivative equal to zero. This gives:

$$\frac{\partial(\cdot)}{\partial p(x)} = f(x) - \log p(x) - 1 + \log q(x) = 0 \Leftrightarrow p(x) = \exp\left(f(x) + \log q(x) - 1 \right)$$

Technically, you should also add in Lagrange parameter on the constraint that $\int p(x) = 1$, but all this does is force you to normalize, so the optimizing p is:

$$p(x) = \frac{\exp\left(f(x) + \log q(x) \right)}{\int \exp\left(f(x) + \log q(x) \right) dx}$$

Plugging this in above we get:

$$\begin{aligned} \psi(f) &= \frac{\int q(x) f(x) \exp(f(x)) - q(x) \exp(f(x)) \log\left(\frac{\exp f(x)}{\int \exp(f(x) + \log q(x)) dx} \right)}{\int \exp\left(f(x) + \log q(x) \right) dx} \\ &= \log \left(\int q(x) \exp(f(x)) dx \right) \end{aligned}$$

which is known as the *log partition function*.

3. **Source Coding**

(a) A set of symbols have a distribution p. You encode the symbols so that the length ℓ of a symbol x is $\ell(x) = \lceil \log \frac{1}{q(x)} \rceil$ for some other distribution q. Show that:

$$H(p) + D(p\|q) \leq \mathbb{E}\ell(x) < H(p) + D(p\|q) + 1.$$

Solution: Both directions are based on the fact that:

$$\begin{aligned}\sum_x p(x)\log\frac{1}{q(x)} &= \sum_x p(x)\log\frac{p(x)}{p(x)q(x)} \\ &= \sum_x p(x)\log\frac{1}{p(x)} + \sum_x p(x)\log\frac{p(x)}{q(x)} \\ &= H(p) + D(p\|q)\end{aligned}$$

The lower bound follows since:

$$\mathbb{E}_{x\sim p}l(x) = \sum_x p(x)\lceil\log\frac{1}{q(x)}\rceil \geq \sum_x p(x)\log\frac{1}{q(x)}$$

and the upper bound follows since:

$$\mathbb{E}_{x\sim p}l(x) = \sum_x p(x)\lceil\log\frac{1}{q(x)}\rceil < \sum_x p(x)\left(\log\frac{1}{q(x)}+1\right)$$

(b) Consider the following method for generating a code for a random variable X on p symbols $\{1,2,\ldots,m\}$ with probabilities $p_1 \geq p_2 \geq \ldots p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_k$$

The codeword for i is the number $F_i \in [0,1]$ rounded off to ℓ_i bits where $\ell_i = \lceil\log\frac{1}{p_i}\rceil$. (E.g. for the symbols $\{a,b,c,d\}$ with probabilities $\{0.5, 0.25, 0.125, 0.125\}$ the codeword assignment would be $\{0, 10, 110, 111\}$.) Show that

i. That this code is a prefix code

ii. The code satisfies $H(X) \leq \mathbb{E}\ell_i \leq H(X)+1$

Solution:

For claim 1, we will show that c_i and c_j differ somewhere in the first l_i locations, which means that p_i is not a prefix of p_j and vice versa. Since we have arranged the p_i's in decreasing order, F_j for $j > i$ differs from F_i by at least $p_i \geq 2^{-l_i}$. This means that the binary representation of F_i differs from F_j in at least one place in the first l_i bits. As $l_j > l_i$, c_i will not be a prefix for c_j.

Claim 2 is trivial since the code words have length $l_i = \lceil\log\frac{1}{p_i}\rceil$. As we saw in class, and essentially the same argument as part (a) above, this means that:

$$H(p) = \sum_i p_i\log\frac{1}{p_i} \leq \sum_i p_i\lceil\log\frac{1}{p_i}\rceil < \sum_i p_i\log\frac{1}{p_i} + 1 = H(p)+1$$