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

Homework 2: Solution

Lecturer: Aarti Singh

Acknowledgement: *The TA graciously thanks Rafael Stern for providing most of these solutions.*

2.1 Problem 1

$$D(q||p) = \int q \log\left(\frac{q}{p}\right) dx$$

Hence,

$$\nabla D(q||p) = 1 + \log\left(\frac{q}{p}\right)$$

Similarly, $h_i(q) = E[r_i(X)] = \int r_i(x) q dx$. Thus:

$$\nabla h_i(q) = r_i(x)$$

Finally $h_0(q) = \int q dx$ and hence, $\nabla h_0 = 1$. Since $D(p|q)$ is convex and the equality restrictions are linear, we wish to solve a convex optimization problem. The Lagrangian of this problem is:

$$L(q, \lambda_i) = \nabla D(q||p) + \sum_{i=0}^{m} \lambda_i \nabla h_i(q)$$

Solving for $L(q^*, \lambda_i) = 0$, obtain:

$$1 + \log(q^*) - \log(p) + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x) = 0$$

Calling $\lambda_0^* = \lambda_0 - 1$, obtain:

$$q^* = p e^{\lambda_0^* + \sum_{i=1}^{m} \lambda_i r_i}$$

Taking λ_0^* such that $\int q dx = 1$, obtain:

$$q^* = \frac{p e^{\sum_{i=1}^{m} \lambda_i r_i}}{\sum_x p e^{\sum_{i=1}^{m} \lambda_i r_i}}$$

Assume there exist unique values for each λ_i such that the equality constraints are satisfied. In this case, (q^*, λ) clearly satisfy stationarity and primal feasibility. Since there are no inequality conditions, dual

feasibility and complementary slackness are also satisfied. Hence, the KKT conditions are satisfied and q^* minimizes $D(q||p)$.

2.2 Problem 2

By results from class, we need only find constants λ_0, λ_1, λ_2 such that the distribution $p(x) = \exp(\lambda_0 + \lambda_1 x + \lambda_2 x^2)$ satisfy the moment constraints.

We inspect the Gaussian pdf with first moment μ and second moment $\sigma^2 - \mu^2$

$$\begin{aligned}\phi(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2} + \frac{x}{\sigma^2} - \frac{\mu^2}{2\sigma^2})\end{aligned}$$

And we conclude immediately that $\lambda_1 = \frac{1}{\sigma^2}$ and $\lambda_2 = -\frac{1}{2\sigma^2}$ and λ_0 is whatever constant required to normalize the distribution.

2.3 Problem 3

Recall that, by HW1 2(b):

$$H(P_1, \ldots, P_n) = \sum_{i=1}^{n} H(P_i|P_{i-1}, \ldots, P_1) \leq \sum_{i=1}^{n} H(P_i)$$

The right side is completely determined by the marginals and corresponds exactly to the joint distribution of independent variables. Hence, the result is proven.

2.4 Problem 4

2.4.1 4.1

Let $r(X)$ be the entropy rate of a stocastic process X. Recall that:

$$r(X) = \lim_{n\to\infty} \frac{H(X_1, \ldots, X_n)}{n}$$

by HW1 2(b):

$$H(X_1, \ldots, X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$

By the Markovian property, X_i is conditionally independent of $(X_{i-2}, \ldots, X_1)$ given X_{i-1}. Hence:

$$\sum_{i=2}^{n} H(X_i|X_{i-1},\ldots,X_1) = \sum_{i=2}^{n} H(X_i|X_{i-1})$$

Since the Markov chain is homogeneous and stationary, for all i, $H(X_i|X_{i-1}) = H(X_2|X_1)$. Thus:

$$r(X) = \lim_{n\to\infty} \frac{H(X_1) + (n-1)H(X_2|X_1)}{n} = H(X_2|X_1)$$

Finally,

$$H(X_2|X_1) = \sum_i P(X_1 = i) \sum_j P(X_2 = j|X_1 = i) \log\left(\frac{1}{P(X_2 = j|X_1 = i)}\right)$$

Call P_i the $i-th$ row of P. Observe that:

$$\sum_j P(X_2 = j|X_1 = i) \log\left(\frac{1}{P(X_2 = j|X_1 = i)}\right) = H(P_i)$$

Hence, by stationarity:

$$H(X_2|X_1) = \sum_i P(X_1 = i)H(P_i) = \sum_i \mu(i)H(P_i) = \mu^{'}(H(P_i))_i$$

Observe that $r(X) = H(X_2|X_1) \leq H(X_2)$. If we take the variables to be i.i.d. $H(X_2|X_1) = H(X_2)$. Finally, $H(X_2)$ is maximized taking the uniform distribution on the support of the Markov chain. Hence, the $r(X)$ is maximized taking P as having all rows equal to $\frac{1}{|S|}$, were S is the support of the Markov chain.

2.4.2 4.2

The invariant measure is obtained solving for $\mu(1) = p\mu(0)$ and $\mu(0) + \mu(1) = 1$, which lead to $\mu(0) = \frac{1}{1+p}$ and $\mu(1) = \frac{p}{1+p}$. From the last item, the entropy rate of the Markov chain is $\mu^{'}(H(P_i))_i$. Observe that P_1 is degenerate and, therefore, $H(P_1) = 0$. Hence, r(X) = $\frac{-1}{1+p}((1-p)\log(1-p) + p\log(p))$.

$$\frac{dr}{dp} = \frac{1}{(1+p)^2}((1-p)\log(1-p) + p\log(p)) + \frac{-1}{1+p}((-\log(1-p) + \log(p)) =$$

$$= \frac{1}{(1+p)^2}(2\log(1-p) - \log(p))$$

Setting $\frac{dr}{dp} = 0$, obtain:

$$2\log(1-p) - \log(p) = 0$$

$$p = (1-p)^2$$

$$p^2 - 3p + 1 = 0$$

Obtain: $p = \frac{3\pm\sqrt{5}}{2}$. Since $0 \leq p \leq 1$ and $r(X) = 0$ for $p = 0$ and $p = 1$, by Weiestrass's theorem: $p = \frac{3-\sqrt{5}}{2}$ maximizes the entropy rate of this Markov chain. On one hand, Reducing p increases the weight $H(X_2|X_1 = 0)$ contributes to the entropy, which helps increase the entropy. On the other hand, reducing p decreases the value of $H(X_2|X_1 = 0)$. The optimum value is the sweet spot between these tendencies.

5. $I(X;Y) = H(X) - H(X|Y)$. In class we proved that $H(X) = 0.5\log(2\pi e\sigma^2)$. Hence, it suffices to find $H(X|Y)$. Recall that $X|Y$ is a normal random variable with variance $\sigma^2 - \rho\sigma^2\frac{1}{\sigma^2}\rho\sigma^2 = (1-\rho^2)\sigma^2$, which does not depend on Y. Hence $H(X|Y) = 0.5\log(2\pi e(1-\rho^2)\sigma^2)$ if $|\rho| < 1$. Thus,

$$I(X;Y) = H(X) - H(X|Y) = -0.5\log((1-\rho^2))$$

This value is minimized when $\rho = 0$. In this case, the variables are independent and, therefore, there is no mutual information. When $\rho = 1$ or $\rho = -1$, X is completely determined by Y, and therefore $H(X|Y) = 0$. Hence, in this case, $I(X;Y) = H(X)$ and is the maximum value obtainable.

2.5 Problem 5

$I(X;Y) = H(X) - H(X|Y)$. In class we proved that $H(X) = 0.5\log(2\pi e\sigma^2)$. Hence, it suffices to find $H(X|Y)$. Recall that $X|Y$ is a normal random variable with variance $\sigma^2 - \rho\sigma^2\frac{1}{\sigma^2}\rho\sigma^2 = (1-\rho^2)\sigma^2$, which does not depend on Y. Hence $H(X|Y) = 0.5\log(2\pi e(1-\rho^2)\sigma^2)$ if $|\rho| < 1$. Thus,

$$I(X;Y) = H(X) - H(X|Y) = -0.5\log((1-\rho^2))$$

This value is minimized when $\rho = 0$. In this case, the variables are independent and, therefore, there is no mutual information. When $\rho = 1$ or $\rho = -1$, X is completely determined by Y, and $I(X,Y) = \infty$

2.6 Problem 6

$$-H(Y|X) = \sum_x p(x)\sum_y p(y|x)\log(p(y|x))$$

Hence,

$$\nabla - H(Y|X)(p) = p(x)(\log(p(y|x)) + 1)$$

Similarly, $h_i(q) = E[r_i(X)Y] = \sum_x r_i(x)p(x)\sum_y yp(y|x)$. Thus:

$$\nabla h_i(p) = r_i(x)p(x)y$$

Finally $h_{0,x}(p) = \sum_y p(y|x)$ and hence, $\nabla h_{0,x} = I_x$. Since $-H(Y|X)$ is convex and the equality restrictions are linear, we wish to solve a convex optimization problem. The Lagrangian of this problem is:

$$L(p,\lambda) = p(x)(\log(p(y|x)) + 1) + \sum_i \lambda_i r_i(x) p(x) y + \sum_x \lambda_{0,x} I_x$$

Call $\sum_x \lambda_{0,x} I_x = f(x)$ and obtain:

$$L(p,\lambda) = p(x)(\log(p(y|x)) + 1) + \sum_i \lambda_i r_i(x) p(x) y + f(x)$$

Solving for $L(p^*, \lambda) = 0$:

$$p^*(y|x) = \exp\left(\frac{-\sum_i \lambda_i r_i(x) p(x) y + f(x) - p(x)}{p(x)}\right) =$$

Call $g(x) = \frac{f(x)-p(x)}{p(x)}$:

$$p^*(y|x) = \exp\left(-\sum_i y \lambda_i r_i(x) + g(x)\right)$$

Since $p^*(0|x) + p(1|x) = 1$:

$$p^*(y|x) = \frac{\exp\left(-\sum_i y \lambda_i r_i(x)\right)}{1 + \exp\left(-\sum_i y \lambda_i r_i(x)\right)}$$

Note that we can cancel out the $g(x)$ from the numerator and the denominator.

Observe that p^* clearly satisfies stationarity. Hence, if there exist λ_i's such that p^* satisfies the constraints, it also satisfies primal feasiblity. Finally, since the solution follows the inequalities but did not use them as a constraint, dual feasibility and complementary slackness are also satisfies. Hence, since the KKT conditions are satisfied, p^* maximizes $H(Y|X)$.