UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

Solutions to Problem Set 3
Fall 2011

Issued: Monday, October 10, 2011 **Due:** Monday, October 24, 2011

Reading: For this problem set: Chapters 8, 9.
Total: 40 points.

Problem 3.1
For each of the following problems, write out the maximum likelihood problem based on n i.i.d. samples $X_1, \ldots, X_n$, and compute the maximum likelihood estimate $\widehat{\theta}$.

(a) Let $p(x; \mu) = \mu^x (1-\mu)^{1-x}$ be a Bernoulli distribution, and consider estimating μ.

(b) Let $X \sim \text{Poi}(\lambda)$, and consider estimating the intensity parameter λ.

(c) Let $X \in \mathbb{R}^d$ be a zero-mean multivariate Gaussian, parametrized in canonical form in terms of a symmetric positive definite matrix $\Gamma \succ 0$ as $p(x; \Gamma) = \frac{1}{\sqrt{(2\pi)^d \det(\Gamma)^{-1}}} \exp\big(-\frac{1}{2} x^T \Gamma x\big)$, and consider estimating the matrix Γ.

Solution: For all parts of the problem, let X denote the collection of i.i.d. samples $x_1, \ldots, x_n$.

(a) $p(X; \mu) = \displaystyle\prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i}$, which leads to

$$l(X; \mu) = \sum_{i=1}^{n} (x_i \log(\mu) + (1-x_i)\log(1-\mu)).$$

Taking the derivative wrt μ and setting equal to 0, we obtain

$$0 = \sum_{i=1}^{n} \left(\frac{x_i}{\mu} - \frac{1-x_i}{1-\mu} \right) = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\mu(1-\mu)},$$

from which we obtain our estimate $\hat{\mu}_{MLE} = \frac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$, the sample mean. Note that

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{1}{\mu^2}\sum_i x_i - \frac{1}{(1-\mu)^2}\sum_i (1-x_i) < 0$$

for all μ, so this stationary point is indeed a maximum.

(b) $p(X; \lambda) = \displaystyle\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}$, which leads to

$$l(X; \lambda) = -n\lambda + \sum_{i=1}^{n} (x_i \log(\lambda) - \log(x_i!)).$$

Taking the derivative wrt λ and setting equal to 0, we obtain $0 = -n + \sum_{i=1}^{n} \frac{x_i}{\lambda}$, from which we obtain our estimate $\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$, the sample mean. Again, we can check that $\frac{\partial^2 l}{\partial \lambda^2} = -\frac{1}{\lambda^2} \sum_i x_i < 0$, so we have a maximum.

(c) $p(X; \Gamma) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d |\Gamma|^{-1}}} e^{-\frac{1}{2} x_i^T \Gamma x_i}$, which leads to

$$
\begin{aligned}
l(X; \Gamma) &= -\frac{nd}{2} \log(2\pi) + \frac{n}{2} \log(|\Gamma|) - \frac{1}{2} \sum_{i=1}^{n} x_i^T \Gamma x_i \\
&= -\frac{nd}{2} \log(2\pi) + \frac{n}{2} \log(|\Gamma|) - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}(x_i^T \Gamma x_i) \\
&= -\frac{nd}{2} \log(2\pi) + \frac{n}{2} \log(|\Gamma|) - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}(\Gamma x_i x_i^T),
\end{aligned}
$$

where we have used the facts that for any scalar $r, r = \operatorname{tr}(r)$, and for matrices A, B, C, $\operatorname{tr}(ABC) = \operatorname{tr}(CAB)$. Taking the derivative wrt Γ and using the facts that $\frac{\partial}{\partial A} \operatorname{tr}(BA) = B'$ and $\frac{\partial}{\partial A} \log(|A|) = (A^{-1})'$ (cf. Boyd's *Convex Optimization*), we obtain

$$
\frac{\partial l}{\partial \Gamma} = \frac{n}{2} \Gamma^{-1} - \frac{1}{2} \sum_{i=1}^{n} \operatorname{tr}(x_i x_i^T),
$$

so setting the derivative equal to 0 gives $\hat{\Gamma}_{MLE} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i^T\right)^{-1}$, the inverse of the sample covariance matrix. (Note that we are assuming here that the sample covariance matrix is invertible, which will occur with high probability, for instance, when $n \geq d$.)

Finally, to see that $l(X; \Gamma)$ is concave, note that the log det function is concave (cf. Boyd) and the trace function is affine in Gamma, so l is concave as well.

Problem 3.2

Maximum a posteriori (MAP) and MLE: Suppose that we adopt a Bayesian perspective, and view the parameter $\theta \in \mathbb{R}$ as a random variable, say distributed according to the prior distribution $\theta \sim \pi(\cdot)$. Given n i.i.d. samples $\{X_1, \ldots, X_n\}$, the MAP estimate is defined as the maximizer of the (rescaled) posterior likelihood $\frac{1}{n} \log p(\theta \mid X_1, X_2, \ldots, X_n)$.

(a) Suppose that $(X_i \mid \theta)$ is Gaussian with mean θ and fixed (known) variance $\sigma^2 > 0$, and let the prior $\pi(\cdot)$ distribution of θ be normal $N(\theta_0, \tau^2)$, where $\theta_0 \in \mathbb{R}$ and $\tau^2 > 0$ are fixed, known parameters. Compute the MAP estimate of θ.

(b) Compute the maximum likelihood estimate of θ.

(c) What happens to the MAP estimate as the number of samples n goes to infinity?

Solution:

(a) $\hat{\theta}_{MAP}$ is the value of θ that maximizes posterior distribution $p(\theta|X)$, where X denotes the collection of i.i.d. samples $x_1, \ldots, x_n$. Noting that $p(\theta|X) = \frac{p(X,\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}$, we conclude that $p(\theta|X) \propto p(X|\theta)p(\theta)$, where the constant of proportionality is independent of θ. Thus, we obtain

$$\begin{aligned}
p(\theta|X) &\propto \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2} \frac{1}{(2\pi\tau^2)^{1/2}} e^{-\frac{1}{2\tau^2}(\theta-\theta_0)^2} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2 - \frac{1}{2\tau^2}(\theta-\theta_0)^2\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2) - \frac{1}{2\tau^2}(\theta^2 - 2\theta_0\theta + \theta_0^2)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(-2\theta\sum_{i=1}^n x_i + n\theta^2) - \frac{1}{2\tau^2}(\theta^2 - 2\theta\theta_0)\right) \\
&= \exp\left(\frac{\tau^2 2n\theta\bar{x} - \tau^2 n\theta^2 - \theta^2\sigma^2 + 2\theta_0\theta\sigma^2}{2\sigma^2\tau^2}\right) \\
&= \exp\left(-\frac{1}{2}\frac{\theta^2 - 2\left(\frac{\tau^2 n\bar{x}+\theta_0\sigma^2}{n\tau^2+\sigma^2}\right)\theta}{\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}}\right),
\end{aligned}$$

where $\bar{x}$ is the sample mean of the x_i's. Maximizing the quadratic in the exponent with respect to θ, we obtain $\hat{\theta}_{MAP} = \frac{n\tau^2\bar{x}+\theta_0\sigma^2}{n\tau^2+\sigma^2}$.

(b) $p(X;\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}$, which leads to

$$l(X;\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^n \frac{(x_i-\theta)^2}{2\sigma^2}.$$

Taking the derivative wrt θ and setting equal to 0, we obtain

$$0 = -2\sum_{i=1}^n \frac{(x_i-\theta)(-1)}{\sigma^2} = \sum_{i=1}^n (x_i-\theta),$$

from which we obtain the estimate $\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^n x_i$, the sample mean.

(c) Letting $n \to \infty$, we see that $\hat{\theta}_{MAP} \to \bar{x} = \hat{\theta}_{MLE}$; i.e., as the number of samples increases, the estimate relies more and more on the data and not on the prior. Further note that by the Law of Large Numbers, both $\hat{\theta}_{MAP}$ and $\hat{\theta}_{MLE}$ converge to the true parameter θ.

Problem 3.3

Recall that a probability distribution in the exponential family takes the form

$$p(x;\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

for a parameter vector η, often referred to as the *natural parameter*, and for given functions T, A, and h.

(a) Determine which of the following distributions are in the exponential family, exhibiting the T, A, and h functions for those that are.

(a) $N(\mu, I)$—multivariate Gaussian with mean vector μ and identity covariance matrix.
Solution: The density for a d-dimensional Gaussian with mean μ and covariance matrix I is

$$p(x|\mu) = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}(x-\mu)^T(x-\mu)) = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}x^T x + \mu^T x - \frac{1}{2}\mu^T \mu)),$$

so we have an exponential family with parameters

$$\begin{aligned} h(x) &= \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}x^T x), \\ T(x) &= x, \\ \eta &= \mu, \\ A(\eta) &= \frac{1}{2}\eta^T \eta. \end{aligned}$$

(b) $\mathrm{Dir}(\alpha)$—Dirichlet with parameter vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$.
Solution: The Dirichlet density for $\theta \in \mathbb{R}^K$ is

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} = \prod_{i=1}^{K} \frac{1}{\theta_i} \exp\left(\sum_{i=1}^{K} \alpha_i \log \theta_i - \log B(\alpha) \right),$$

where $B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$. Hence, we have an exponential family with parameters

$$\begin{aligned} h(\theta) &= \prod_{i=1}^{K} \frac{1}{\theta_i}, \\ T(\theta) &= [\log(\theta_1), \cdots, \log(\theta_K)]^T, \\ \eta &= \alpha, \\ A(\eta) &= \log B(\eta). \end{aligned}$$

(c) $\mathrm{Mult}(\theta)$—multinomial with parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$. Use the fact that $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$ and express the distribution using a $(K-1)$-dimensional parameter η.
Solution: We assume the number of trials is fixed at n. Using the fact that $n = \sum_{i=1}^{K} x_i$,

we have the density

$$
\begin{aligned}
p(x|\theta) &= \binom{n}{x_1, x_2, \ldots, x_K} \prod_{i=1}^{K} \theta_i^{x_i} \\
&= \binom{n}{x_1, x_2, \ldots, x_K} \prod_{i=1}^{K} e^{x_i \log \theta_i} \\
&= \binom{n}{x_1, x_2, \ldots, x_K} \exp\left(\sum_{i=1}^{K} x_i \log \theta_i\right) \\
&= \binom{n}{x_1, x_2, \ldots, x_K} \exp\left(\sum_{i=1}^{K-1} x_i \log \theta_i + (n - \sum_{i=1}^{K-1} x_i) \log \theta_K\right) \\
&= \binom{n}{x_1, x_2, \ldots, x_K} \exp\left(\sum_{i=1}^{K-1} x_i (\log \theta_i - \log \theta_K) + n \log \theta_K\right).
\end{aligned}
$$

For $i = 1, \ldots, K$, take

$$\eta_i = \log \theta_i - \log \theta_K = \log \frac{\theta_i}{\theta_K}.$$

Note that

$$1 = \sum_{i=1}^{K} \theta_i = \theta_K \sum_{i=1}^{K} e_i^{\eta},$$

so

$$\theta_K = \left(\sum_{i=1}^{K} e^{\eta_i}\right)^{-1} = \left(1 + \sum_{i=1}^{K-1} e^{\eta_i}\right)^{-1}.$$

Hence, we have an exponential family with parameters

$$
\begin{aligned}
h(x) &= \binom{n}{x_1, x_2, \ldots, x_K}, \\
T(x) &= x, \\
\eta &= (\eta_1, \ldots, \eta_{K-1})^T, \\
A(\eta) &= -n \log \theta_K = -n \log(\sum_{i=1}^{K} e^{\eta_i})^{-1} = n \log(1 + \sum_{i=1}^{K-1} e^{\eta_i}).
\end{aligned}
$$

(d) the log normal distribution: the distribution of $Y = \exp(X)$, where $X \sim N(0, \sigma^2)$.

Solution: The log normal density has the form

$$p(y|\sigma) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log y)^2}{2\sigma^2}\right) = \frac{1}{y\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(\log y)^2 - 0.5 \log\left(\sigma^2\right)\right).$$

Hence, the log normal distributions over σ^2 are an exponential family with

$$\begin{aligned} h(y) &= \frac{1}{y\sqrt{2\pi}}, \\ T(y) &= (\log y)^2, \\ \eta &= -\frac{1}{2\sigma^2}, \\ A(\eta) &= -0.5\log(-2\eta). \end{aligned}$$

(e) the Ising model: an undirected graphical model $G = (V, E)$ involving a binary random vector X taking values in $\{0,1\}^n$ with distribution $p(x;\theta) \propto \exp\big\{ \sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t\big\}$.

Solution: Let $Z(\theta) = \sum_x \exp\big\{ \sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t\big\}$.
Then the Ising density has the form

$$\begin{aligned} p(x;\theta) &= \frac{1}{Z(\theta)}\exp\big\{ \sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t\big\} \\ &= \exp\big\{ \sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \log(Z(\theta))\big\}, \end{aligned}$$

and we have an exponential family with parameters

$$\begin{aligned} h(x) &= 1, \\ T(x)_s &= x_s, \forall s \in V, \\ T(x)_{(s,t)} &= x_{(s,t)}, \forall (s,t) \in E, \\ \eta_s &= \theta_s, \forall s \in V, \\ \eta_{(s,t)} &= \theta_{(s,t)}, \forall (s,t) \in E, \\ A(\eta) &= \log(Z(\eta)). \end{aligned}$$

(b) Recall that the function $A(\eta)$ has moment-generating properties: $\nabla_\eta A(\eta) = \mathbb{E}[T(X)]$. Demonstrate that this relationship holds for those examples that are in the exponential family in part (a).

Solution:

(a) Normal:

$$\frac{\partial}{\partial \eta_i}\frac{1}{2}\eta^T\eta = \frac{\partial}{\partial \eta_i}\frac{1}{2}\sum_{i=1}^{d}\eta_i^2 = \eta_i = \mu_i = E\left[X\right]_i,$$

so

$$\nabla \frac{1}{2}\eta^T\eta = E\left[X\right]$$

(b) Dirichlet:

For this distribution, we employ a general exponential family argument to derive the desired result:

$$1 = \int h(x)\exp\left(\eta^T T(x) - A(\eta)\right),$$

so

$$
\begin{aligned}
0 &= \nabla_\eta \int h(x) \exp\left(\eta^T T(x) - A(\eta)\right) dx \\
&= \int \nabla_\eta h(x) \exp\left(\eta^T T(x) - A(\eta)\right) dx \\
&= \int \left(T(x) - \nabla_\eta A(\eta)\right) h(x) \exp\left(\eta^T T(x) - A(\eta)\right) dx \\
&= E\left[T(X) - \nabla_\eta A(\eta)\right] = E\left[T(X)\right] - \nabla_\eta A(\eta),
\end{aligned}
$$

implying that

$$E\left[T(X)\right] = \nabla_\eta A(\eta).$$

Note that we exchange the order of integration and differentiation. There are cases in which this exchange is not valid. See Appendix A.9 in "Probability: Theory and Examples" by Durrett for sufficient conditions under which differentiation and integration can be exchanged.

One may also solve this by using the definition that $A(\eta)$ is the function which makes the density integrate to 1. That is,

$$A(\eta) = \log \int h(x) \exp(\eta^T T(x)).$$

(c) Multinomial:

$$\frac{\partial}{\partial \eta_i} n \log(1 + \sum_{i=1}^{K-1} e^{\eta_i}) = n \frac{e^{\eta_i}}{1 + \sum_{i=1}^{K-1} e^{\eta_i}} = n\theta_i = E\left[X\right]_i.$$

(d) Log normal:
Note that $E\left[(\log Y)^2\right] = E\left[X^2\right] = \sigma^2$, since $Y = \exp(X)$ for $X \sim N(0, \sigma^2)$. Furthermore,

$$\frac{d}{d\eta} A(\eta) = \frac{d}{d\eta}(-0.5 \log(-2\eta)) = \frac{-0.5}{\eta} = \sigma^2.$$

(e) Ising:

$$\nabla_\eta A(\eta) = \frac{\nabla_\eta \sum_x \exp\left\{x^T \eta\right\}}{Z(\eta)} = \frac{\sum_x \nabla_\eta \exp\left\{x^T \eta\right\}}{Z(\eta)} = \frac{\sum_x x \exp\left\{x^T \eta\right\}}{Z(\eta)} = E\left[X\right].$$

Problem 3.4

The course homepage has a data set named "lms.dat" that contains twenty rows of three columns of numbers. The first two columns are the components of an input vector x and the last column is an output y value. (We will not use a constant term for this problem; thus the input vector and the parameter vector are both two dimensional.)

(a) Solve the normal equations for these data to find the optimal value of the parameter vector. (I recommend using MATLAB or R.)

Solution:

The least squares objective is:

$$J(\theta) = (y - X\theta)^T (y - X\theta)$$

By solving the normal equations, we have:

$$\theta^* = (X^T X)^{-1} X^T y = \begin{pmatrix} 1.039 \\ -0.976 \end{pmatrix}$$

(b) Find the eigenvectors and eigenvalues of the covariance matrix of the input vectors and plot contours of the cost function J in the parameter space. These contours should of course be centered around the optimal value from part (a).

Solution:

The covariance matrix of the data is $C = \frac{1}{n} X^T X$, where $n = 20$ is the number of data points.

Note that the covariance matrix of a random vector x is defined as $E[x]\,x^T$. To make the link, let the distribution of x be uniform over the rows of X.

Eigenvectors and eigenvalues of C:

- $\lambda_1 = 2.4933$, $v_2 = \begin{pmatrix} 0.853064 \\ 0.521806 \end{pmatrix}$
- $\lambda_2 = 0.9754$, $v_1 = \begin{pmatrix} -0.521805 \\ -0.853064 \end{pmatrix}$

Note: if you define the covariance matrix as $C = \frac{1}{n-1} X^T X$ (as some programming languages such as numpy in python does), the resulting eigenvalues are 2.6246 and 1.0268, respectively. Also note that it is acceptable to define the covariance matrix (rescaled) as $C = X^T X$, and solve for the eigenvalues.

The contours (level sets) of J should be ellipses centered around θ^* with axes corresponding to the eigenvectors. Note that the larger eigenvector λ_1 should correspond to the minor axis and the smaller eigenvector λ_2 to the major axis.

(c) Initializing the LMS algorithm at $\theta = 0$ plot the path taken in the parameter space by the algorithm for three different values of the step size ρ. In particular let ρ equal the inverse of the maximum eigenvalue of the covariance matrix, one-half of that value, and one-quarter of that value.

Solution:

LMS is an online algorithm: pick up a point (x_i, y_i) and make the update:

$$\theta \leftarrow \theta + \rho(y_i - \theta^T x_i) x_i$$

To improve performance, it is advisable to choose a random order of the points rather than go in order.

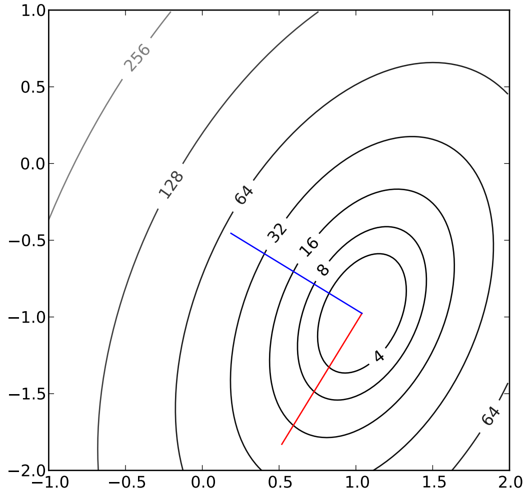

Figure 1: The contour of $J(\theta)$. The blue axes corresponds to λ_1 and the red λ_2.

Note that it may take many iterations for θ to approach θ^*. Even then, LMS is not guaranteed to converge at all, and in general, will not converge. The following batch update (which corresponds to gradient descent on J):

$$\theta \leftarrow \theta + \rho \sum_{i=1}^{n} (y_i - \theta^T x_i) x_i$$

does converge given an appropriate step size ρ.

For larger ρ, the algorithm takes bigger steps in the parameter space but tends to overshoot and be quite noisy. For smaller ρ, the algorithm takes smaller steps but is more stable. In practice, decreasing the step size ρ over time and monitoring the progress on the objective J is a good strategy.

Problem 3.5

(Properties of Kullback-Leibler divergence:) Given two probability distributions p and q (where the random variables take values in $\{0, 1, \ldots, k-1\}$), the Kullback-Leibler divergence is defined as $D(p\|q) = \sum_{x=0}^{k-1} p(x) \log \frac{p(x)}{q(x)}$.

(a) Show that $D(p \,\|\, q) \geq 0$ for all p, q, with equality if and only if $p = q$.

(b) Use part (a) to show that the $H(p) = -\sum_x p(x) \log p(x)$ satisfies $H(p) \leq \log k$ for all distributions p. When does equality hold?

Solution: Observe that

$$-D(p\|q) = -\sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right) = \sum_x p(x) \log \left(\frac{q(x)}{p(x)} \right). \tag{1}$$

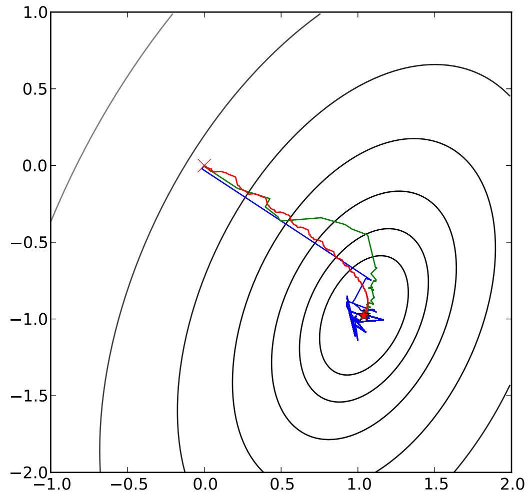

Figure 2: The LMS trajectory with learning rates ρ from large to small (in the order of blue, green and red). See the sample code for details.

Applying Jensen's inequality to the concave log function, we obtain

$$-D(p\|q) \leq \log\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) = \log\left(\sum_x q(x)\right) = 0,$$

so $D(p\|q) \geq 0$. Since log is strictly concave, equality holds in Jensen's inequality iff $p(x) = q(x)$ for all x.

Now define $q(x) = \frac{1}{k}\ \forall x$ (the uniform distribution). For an arbitrary density $p(x)$, we then have

$$D(p\|q) = \sum_x p(x)\log(kp(x)) = \log k + \sum_x p(x)\log p(x) = \log k - H(p).$$

Since $D(p\|q) \geq 0$, this implies that $H(p) \leq \log k$, as wanted. Equality holds iff p is the uniform distribution.