CS281: HOMEWORK #3

MIKE SCHACHTER

Problem 1. Write out the maximum likelihood for each distribution and compute the ML estimate $\hat{\theta}$.

First I'll derive the maximum likelihood estimate for a general exponential family, and then apply that to (a) and (b). The likelihood function for an exponential family can be written as:

$$\begin{aligned} p(x_1, ..., x_n|\eta) &= \prod_{i=1}^{n} h(x_i) exp(< T(x_i), \eta > -A(\eta)) \\ &= \{\prod_{i=1}^{n} h(x_i)\} exp(\sum_{i=1}^{n} < T(x_i), \eta > -nA(\eta)) \end{aligned}$$

Taking the log gives:

$$l(x_1, ..., x_n|\eta) = \sum_{i=1}^{n} log(h(x_i)) + \sum_{i=1}^{n} < T(x_i), \eta > +nA(\eta)$$

And then taking the gradient with respect to η gives:

$$\nabla_\eta l = \sum_{i=1}^{n} \nabla_\eta < T(x_i), \eta > -n\nabla_\eta A(\eta)$$

Setting the gradient to zero gives the expression that the ML estimate has to satisfy:

$$\frac{1}{n} \sum_{i=1}^{n} \nabla_\eta < T(x_i), \eta >= \nabla_\eta A(\eta)$$

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

Let $x = (x_1, ..., x_n)$. First I'll rewrite the distribution as an exponential family. Taking the log-exponential of the expression gives:

$$\begin{aligned} p(x;\mu) &= exp\{\sum_{i=1}^{n} x_i log(\mu) + (1 - \sum_{i=1}^{n} x_i) log(1-\mu)\} \\ &= exp\{\sum_{i=1}^{n} x_i (log\frac{\mu}{1-\mu}) + log(1-\mu)\} \end{aligned}$$

From this expression we can see that:

$$T(x) = \sum_{i=1}^{n} x_i$$

$$\eta = log\frac{\mu}{1-\mu}$$

$$A(\eta) = log(1 + e^{\eta})$$

$$h(x) = 1$$

Using the expression for the ML estimate derived above implies that:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = (1 + e^{\eta})^{-1} e^{\eta}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$$

(b) Let $X \sim Poisson(\lambda)$, and estimate λ.

The likelihood of the Poisson distribution is given as:

$$p(x|\lambda) = \prod_{i=1}^{n} \lambda^{x_i} e^{-\lambda} \frac{1}{x_i!}$$

I'll rewrite this as an exponential family:

$$\begin{aligned} p(x|\lambda) &= exp\{\sum_{i=1}^{n} x_i log\lambda - \lambda - log(x_i!)\} \\ &= exp\{-\sum_{i=1}^{n} log(x_i!)\} exp\{\sum_{i=1}^{n} x_i log\lambda - \lambda\} \\ &= \prod_{i=1}^{n} x_i!\ exp\{\sum_{i=1}^{n} x_i log\lambda - \lambda\} \end{aligned}$$

Rewriting things this way shows that:

$$T(x) = \sum_{i=1}^{n} x_i$$

$$\eta = log\lambda$$

$$A(\eta) = e^{\eta}$$

$$h(x) = \prod_{i=1}^{n} x_i!$$

The ML criteria then gives:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = e^{\eta}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\lambda}$$

(c) Let $x_i \in R^d$ be distributed as a multivariate Gaussian with zero-mean with density function:

$$p(x_i|\Gamma) = \frac{1}{\sqrt{(2\pi)^d det(\Gamma)^{-1}}} exp(-\frac{1}{2}x_i^T \Gamma x_i)$$

Find the ML estimate $\hat{\Gamma} \in \mathbb{R}^{dxd}$.

Let $x = (x_1, ..., x_n)$. The likelihood function looks like this:

$$p(x|\Gamma) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d det(\Gamma)^{-1}}} exp(-\frac{1}{2}x_i^T \Gamma x_i)$$

Taking the log-likelihood gives:

$$l(x|\Gamma) = log\{((2\pi)^d det(\Gamma)^{-1})^{-n/2}\} - \frac{1}{2}\sum_{i=1}^{n} x_i^T \Gamma x_i$$

I'll break this up into two functions:

$$\begin{aligned} h(\Gamma) &= log\{((2\pi)^d det(\Gamma)^{-1})^{-n/2}\} \\ &= -\frac{n}{2}\{dlog2\pi + log\{det(\Gamma)^{-1}\} \\ g(x,\Gamma) &= -\frac{1}{2}\sum_{i=1}^{n} x_i^T \Gamma x_i \end{aligned}$$

If we differentiate $h(\Gamma)$ with respect to the matrix Γ, we get:

$$\begin{aligned} \frac{\partial h}{\partial \Gamma} &= \frac{n}{2}\frac{\partial}{\partial \Gamma}[log(det(\Gamma))] \\ &= \frac{n}{2}\Gamma \end{aligned}$$

Differentiating $g(x, \Gamma)$ with respect to Γ gives:

$$\begin{aligned} \frac{\partial g}{\partial \Gamma} &= -\frac{1}{2}\frac{\partial}{\partial \Gamma}[\sum_{i=1}^{n} Tr(x_i x_i^T \Gamma)] \\ &= -\frac{1}{2}\sum_{i=1}^{n} x_i x_i^T \end{aligned}$$

I used two tricks, one was taking the derivative of a log determinant with respect to a matrix, and differentiating the trace of a quadratic form. Both of these tricks were obtained from Ch. 13 of the course manual. Setting the log likelihood to zero and combining these results gives:

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T = \hat{\Gamma}$$

Problem 2. Assume the parameter $\theta \in \mathbb{R}$ is a random variable $\theta \sim \pi$, and the MAP estimate is given as the rescaled posterior likelihood $\frac{1}{n} log\ p(\theta|x_1, ..., x_n)$.

(a) Suppose $(x_i|\theta) \sim \mathcal{N}(\theta, \sigma^2)$, where σ is fixed, and let $\pi \sim \mathcal{N}(\theta_0, \tau^2)$, where both θ_0 and τ are fixed. Compute the MAP estimate of θ.

Let $x = (x_1, ..., x_n)$. The posterior is:

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

log posterior is given as:

$$log\, p(\theta|x) \propto log\, p(\theta) + log\, p(x|\theta)$$

where

$$\begin{aligned} log\, p(x|\theta) &= log\{(2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2)\} \\ &= log\{2\pi\sigma^2)^{-n/2}\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 \\ &= -\frac{1}{2}\, log\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 \end{aligned}$$

and

$$log\, p(\theta) = -\frac{1}{2}\, log\{2\pi\tau^2\} - \frac{1}{2\tau^2}(\theta-\theta_0)^2$$

Taking the derivative of the log-likelihood with respect to θ gives:

$$\frac{\partial}{\partial\theta} log\, p(x|\theta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta)$$

and the prior:

$$\frac{\partial}{\partial\theta}\, log\, p(\theta) = \frac{1}{\tau^2}(\theta-\theta_0)$$

Taking the derivative of the posterior and setting it to zero gives:

$$\begin{aligned} -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) &= \frac{1}{\tau^2}(\theta-\theta_0) \\ n\theta-\sum_{i=1}^{n}x_i &= \frac{\sigma^2}{\tau^2}(\theta-\theta_0) \\ n\theta-\frac{\sigma^2}{\tau^2}\theta &= \sum_{i=1}^{n}x_i-\frac{\sigma^2}{\tau^2}\theta_0 \\ \hat{\theta} &= \left(\frac{\tau^2}{\tau^2 n-\sigma^2}\right)\left(\sum_{i=1}^{n}x_i-\frac{\sigma^2}{\tau^2}\theta_0\right) \end{aligned}$$

(b) Compute the ML estimate of θ.

This is simply $\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i$

(c) What happens as $n\to\infty$?

It's supposed to converge to the ML estimate... There's a missing n in the numerator and I have no idea why.

Problem 3. Determine which of the following are exponential families, and show that $\nabla_\eta A(\eta)=\mathbb{E}[T(x)]$.

(a) Multivariate Gaussian with mean $\mu\in\mathbb{R}^d$ and identity covariance matrix $\Sigma=I$.

The density is given as:

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

where $\Sigma^{-1}=I$ and $|\Sigma|=d$. Bringing the coefficient into the exponential gives:

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

Then taking into account $\Sigma=I$ and expanding the quadratic term gives:

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

Reorganizing a bit:

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

Let $h(x)=exp\{-\frac{1}{2}d\,log(2\pi)-\frac{1}{2}log(d)-\frac{1}{2}x^Tx\}$, then we're left with:

$$p(x|\mu,\Sigma) = h(x)\,exp\{-x^T\mu-\frac{1}{2}\mu^T\mu\}$$

Then let $T(x)=x$, $\eta=-\mu$, and $A(\eta)=\frac{1}{2}\eta^T\eta$.

Moment Generation: For a dataset $x = (x_1, .., x_n)$, we have $T(x) = \sum_{i=1}^n x_i$, and $A(\eta) = \frac{n}{2}\eta^T\eta$, so that $\mathbb{E}[T(x)] = \sum_{i=1}^n \mathbb{E}[x_i] = n\mu$

$$\nabla_\eta A(\eta) = n\eta = n\mu = \mathbb{E}[T(x)]$$

(b) Dirichlet distribution with parameter $\alpha \in \mathbb{R}^K$.

Let $\alpha_0 = \sum_{i=1}^K \alpha_i$, and $x \in \mathbb{R}^K$ be such that $\sum_{i=1}^K x_i = 1$. The density of x under the Dirichlet distribution is:

$$p(x|\alpha) = \frac{1}{\beta(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1}$$

where

$$\beta(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$$

and

$$\Gamma(\alpha_i) = \int_0^\infty t^{\alpha_i - 1} e^t \, dt$$

We can rewrite the density as:

$$\begin{aligned} p(x|\alpha) &= exp\{log\{\frac{1}{\beta(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1}\}\} \\ &= exp\{-log\,\beta(\alpha) + \sum_{i=1}^K log(x_i^{\alpha_i - 1})\} \\ &= exp\{-log\,\beta(\alpha) + \sum_{i=1}^K (\alpha_i - 1) log(x_i)\} \end{aligned}$$

So the Dirichlet distribution is in the exponential family with $T(x) = log(x)$, $\eta = \alpha - 1$, $A(\eta) = log\,\beta(\alpha) = log\,\beta(1 + \eta)$, and $h(x) = 1$. **Moment-Generation**: No idea!

(c) The multinomial distribution with parameter $\theta = (\theta_1, ..., \theta_K)$.

Let $x = (x_1, ..., x_K)$ be distributed multi-nomially with N trials. Because $\sum_{i=1}^K \theta_i = 1$ and $\sum_{i=1}^K x_i = N$, let $\tilde{x} = (x_1, ..., x_{K-1})$ and $\tilde{\theta} = (\theta_1, ..., \theta_{K-1})$, and get rid of the Kth variables by writing $x_K = N - \sum_{i=1}^{K-1} x_i$ and $\theta_K =$

$1 - \sum_{i=1}^{K-1} \theta_i$. We can write the density function as:

$$\begin{aligned}
p(\tilde{x}|\tilde{\theta}) &= \frac{N!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i} \\
&= h(x)\, exp\{log\{\prod_{i=1}^{K} \theta_i^{x_i}\}\} \\
&= h(x)\, exp\{\sum_{i=1}^{K} log(\theta_i^{x_i})\} \\
&= h(x)\, exp\{\sum_{i=1}^{K} x_i log\theta_i\} \\
&= h(x)\, exp\{\sum_{i=1}^{K-1} x_i log\theta_i + (N - \sum_{j=1}^{K-1} x_i) log(1 - \sum_{j=1}^{K-1} \theta_j)\} \\
&= h(x)\, exp\{\sum_{i=1}^{K-1} x_i log\left(\frac{\theta_i}{1 - \sum_{j=1}^{K-1} \theta_j}\right) + N\, log(1 - \sum_{j=1}^{K-1} \theta_j)\}
\end{aligned}$$

where $h(x) = \frac{N!}{\prod_{i=1}^{K} x_i!}$. Let $\eta_i = log\left(\frac{\theta_i}{1-\sum_{j=1}^{K-1} \theta_j}\right)$, $T(x_i) = x_i$, both defined for $i = 1, ..., K-1$. The relationship between η_i and θ_i can be inverted:

$$\theta_i = \frac{e^{\eta_i}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}}$$

We can then set:

$$\begin{aligned}
A(\eta) &= -N\, log(1 - \sum_{i=1}^{K-1} \theta_i) \\
&= -N\, log\left(1 - \frac{\sum_{j=1}^{K-1} e^{\eta_j}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}}\right) \\
&= -N\, log\left(\frac{1 - \sum_{j=1}^{K-1} e^{\eta_j} + \sum_{j=1}^{K-1} e^{\eta_j}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}}\right) \\
&= N\, log(1 + \sum_{j=1}^{K-1} e^{\eta_j})
\end{aligned}$$

Moment-Generation: The expectation of the sufficient statistic is:

$$\mathbb{E}[x_i] = N\theta_i$$

because the parameter θ_i is the fraction of times that category i will occur. The partial derivative of $A(\eta)$ is:

$$\begin{aligned}\frac{\partial}{\partial\eta_i}A(\eta) &= N\left(\frac{1}{1+\sum_{j=1}^{K-1}e^{\eta_j}}\frac{\partial}{\partial\eta_i}[1+\sum_{j=1}^{K-1}e^{\eta_j}]\right)\\ &= N\left(\frac{e^{\eta_i}}{1+\sum_{j=1}^{K-1}e^{\eta_j}}\right)\\ &= N\theta_i\end{aligned}$$

(d) The log normal distribution of Y, where if $X \sim \mathcal{N}(0,\sigma^2)$, then $Y = e^X$.

The density function looks like this:

$$p(x|\theta) = exp(\frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{1}{2\sigma^2}x))$$

Although the logarithm of this density function is obviously in the exponential family, the lognormal distribution itself is not, because the terms inside the primary exponential can't be broken down into additive terms.

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

Let $a_{ij} = x_i x_j$, and $T(x) = [x_1 ... x_n\ a_{ij\in E}]$ be a vector of dimension $n + |E|$ that contains all values of x_i and products of x_i and x_j for all $(i,j) \in E$. Let $\eta = [\theta_1 ... \theta_n\ \theta_{ij\in E}]$ be another vector of dimension $n + |E|$. We can write the density of the Ising model as:

$$p(x|\eta) = \frac{1}{Z(\eta)}exp\{< T(x), \eta >\}$$

where $Z(\eta) = \sum_x exp\{< T(x), \eta >\}$ is the normalization function, giving $A(\eta) = -exp\{-log\, Z(\eta)\}$.

Moment-Generation: Taking the partial derivative of $A(\eta)$ gives:

$$\begin{aligned}\frac{\partial}{\partial\eta_i}A(\eta) &= -\frac{\partial}{\partial\eta_i}\frac{1}{Z(\eta)}\\ &= \frac{1}{Z(\eta)^2}\sum_x T(x)_i\end{aligned}$$

Not really sure where I'm going with this...

Problem 4. The file "lms.dat" contains data $\mathcal{D} = (x^i, y^i)$ where $x^i \in \mathbb{R}^2$ and $y^i \in \mathbb{R}$, and $N = |D|$ the superscript represents the sample #.

(a) Solve the normal equations to find the optimal value of the parameter vector.

Using the R code attached to the email with this homework, I solved the normal equations by first constructing a matrix of features $X = [x^1 ... x^N]^T \in$

$\mathbb{R}^{Nx2}$, and a vector of values $y = [y^1 ... y^N] \in \mathbb{R}^N$. The optimal parameter $\theta \in \mathbb{R}^2$ for the linear model $\hat{y} = X\hat{\theta}$ is given by solving:

$$X^T X \, \hat{\theta} = Xy$$

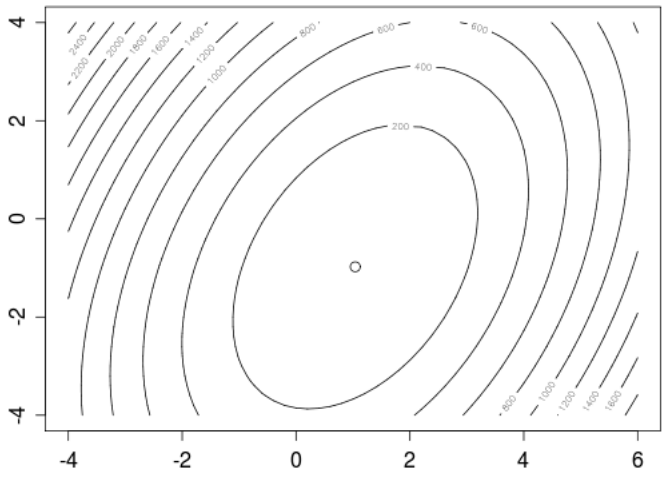
Using R, the optimal parameters are found to be $\hat{\theta} = [1.04, -0.98]$.

(b) Find the eigenvectors and eigenvalues of the covariance matrix, plot the contours of the cost function J in the parameter space, centered around the optimal value.

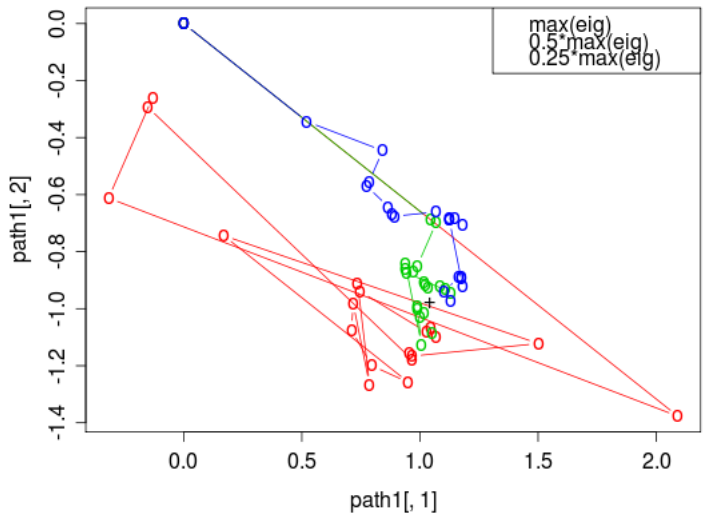
The eigenvectors were computed to be:

$$2.62 \begin{bmatrix} -0.85 \\ 0.52 \end{bmatrix} \; \textit{and} \; 1.03 \begin{bmatrix} -0.52 \\ -0.85 \end{bmatrix}$$

The contours look like this:



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



A plot of the three paths, with the plus character indicating the optimal solution found using the normal equations. All paths converge somewhere close.

Problem 5. Given two probability distributions p and q, defined on a discrete random variable $X = \{0, ..., k-1\}$, and the KL distance $D(p||q) = \sum_{x=0}^{k-1} p(x) log \frac{p(x)}{q(x)}$:

(a) Show that $D(p||q) \geq 0$ for all p, q, and $D(p||q) = 0$ iff $p = q$.

This proof is adapted from Theorem 2.6.3 in Cover and Thomas (2006). Taking the negative of $D(p||q)$ gives us the following form:

$$-D(p||q) = \sum_{x=0}^{k-1} p(x) log \left(\frac{q(x)}{p(x)} \right)$$

Let $Y_x = \frac{q(x)}{p(x)}$ be a random variable, so that $\mathbb{E}_p[log\, Y_x] = \sum_{x=0}^{k-1} p(x) log \left(\frac{q(x)}{p(x)} \right) = -D(p||q)$. Because log is concave, Jensen's inequality implies:

$$log\, (\mathbb{E}_p[Y_x]) \geq \mathbb{E}_p[log\, (Y_x)]$$

If we expand things out:

$$\begin{aligned} log\{\sum_{x=0}^{k-1} p(x) \frac{q(x)}{p(x)}\} &\geq \sum_{x=0}^{k-1} p(x) log \left(\frac{p(x)}{q(x)} \right) \\ log\{\sum_{x=0}^{k-1} q(x)\} &\geq -D(p||q) \\ log\, 1 &\geq -D(p||q) \\ 0 &\leq D(p||q) \end{aligned}$$

It's obvious that $D(p||q) = 0$ when $p(x) = q(x)$, but need to show that there are no other values for $p(x)$ and $q(x)$ that could produce $D(p||q) = 0$. But I'm running out of time and won't do that...

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

This proof is adapted from Theorem 2.6.4 in Cover and Thomas (2006). Let $q(x) = \frac{1}{k}$, the uniform distribution. Then:

$$\begin{aligned} D(p||q) &= \sum_{x=0}^{k-1} p(x) log \frac{p(x)}{q(x)} \\ &= \sum_{x=0}^{k-1} p(x) log\, kp(x) \\ &= \sum_{x=0}^{k-1} p(x) log\, p(x) + \sum_{x=0}^{k-1} p(x) log\, k \end{aligned}$$

The term on the left is $-H(p)$ by definition, and the term on the right is equal to $log\, k$ because $\sum_{x=0}^{k-1} p(x) = 1$. So we have:

$$D(p||q) = -H(p) + log\, k$$

From (a) we know that $D(p||q) \geq 0$, which implies that:

$$H(p) \leq log\, k$$

We know that $D(p||q) = 0$ iff $p(x) = q(x)$, which implies that equality holds only when $p(x)$ is the uniform distribution.