Nicholas Trieu
MATH 189r
Homework 1 Part 2
September 19, 2016

**4 (Murphy 2.15)** Let $\mathbb{P}_{emp}(x)$ be the empirical distribution and let $q(x|\boldsymbol{\theta})$ be some model. Show that $\arg\min_q \mathbb{KL}(\mathbb{P}_{emp}||q)$ is obtained by $q(x) = q(x;\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \mathcal{D})$ is the maximum likelihood estimate.

By the definition of the KL divergence $\mathbb{KL}$, we have

$$\begin{aligned}
\mathbb{KL}(\mathbb{P}||q) &= \int_S \mathbb{P}(x) \log \frac{\mathbb{P}_{emp}(x)}{q(x;\theta)} dx \\
&= \int_S \mathbb{P}(x) \left(\log \mathbb{P}_{emp}(x) - \log q(x;\theta)\right) dx \\
&= \int_S \mathbb{P}(x) \log \mathbb{P}_{emp}(x) dx - \int_S \mathbb{P}(x) \log q(x;\theta) dx
\end{aligned}$$

where $\mathbb{P}_{emp}(x)$ is the empirical distribution for the data $D = \{x_1, x_2, \ldots, x_n\}$.
Recall that the empirical density $\mathbb{P}_{emp}(x)$ can be expressed as

$$\mathbb{P}_{emp}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

where $\delta$ is the Dirac delta function, and thus

$$\begin{aligned}
\mathbb{KL}(\mathbb{P}||q) &= \int_S \mathbb{P}(x) \log \mathbb{P}_{emp}(x) dx - \int_S \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \log q(x;\theta) dx \\
&= \int_S \mathbb{P}(x) \log \mathbb{P}_{emp}(x) dx - \frac{1}{n} \sum_{i=1}^{n} \log q(x;\theta) \quad \text{by the sifting property.}
\end{aligned}$$

Note that only the summation term is dependent on $\theta$, so picking $\theta$ to minimize $\mathbb{KL}(\mathbb{P}||q)$ is equivalent to

$$\begin{aligned}
\arg\min_{\theta} - \sum_{i=1}^{n} \log q(x_i;\theta) &= \arg\max_{\theta} \sum_{i=1}^{n} \log q(x_i;\theta) \\
&= \arg\max_{\theta} \mathcal{L}(q, D)
\end{aligned}$$

as desired. ■

**7 (Murphy 8.3)** Gradient and Hessian of the log-likelihood for logistic regression.
(a) Let $\sigma(x) = \frac{1}{1+e^{-x}}$ be the sigmoid function. Show that

$$\sigma'(x) = \sigma(x)\left[1 - \sigma(x)\right].$$

(b) Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood for logistic regression.
(c) The Hessian can be written as $\mathbf{H} = \mathbf{X}^\top \mathbf{S} \mathbf{X}$ where $\mathbf{S} = \mathrm{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Derive this and show that $\mathbf{H} \succeq 0$ ($A \succeq 0$ means that $A$ is positive semidefinite).

(a) Taking the derivative of $\sigma(x)$, we find that

$$\begin{aligned}
\sigma'(x) &= \nabla(1+e^{-x})^{-1} \\
&= -(1+e^{-x})^{-2} \cdot -e^{-x} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}$$

as desired.

(b) Let the log likelihood for logistic regression be denoted by $l(\boldsymbol{\theta})$. Then

$$l(\boldsymbol{\theta}) = \sum_i y_i \log \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)).$$

Taking the gradient, we have:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) &= \sum_i y_i \left(1 - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)\right) \mathbf{x}_i - (1-y_i)\sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)\mathbf{x}_i \quad \text{by definition of the sigmoid function} \\
&= \sum_i \left(y_i - \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)\right) \mathbf{x}_i \\
&= X^\top(\mathbf{y} - \boldsymbol{\mu})
\end{aligned}$$

where $\boldsymbol{\mu} = \sigma(X\boldsymbol{\theta})$.

(c) The Hessian of the negative log likelihood $-l(\boldsymbol{\theta})$ is

$$\begin{aligned}
-\nabla^2 l(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \left[X^\top \boldsymbol{\mu} - X\right]^\top \\
&= \nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}^\top X \quad \text{since the } X \text{ term is independent of } \boldsymbol{\theta} \\
&= \nabla_{\boldsymbol{\theta}} \sigma(X\boldsymbol{\theta})^\top X \\
&= (\mathrm{diag}(\boldsymbol{\mu}')X)^\top X \quad \text{by the chain rule} \\
&= X^\top \mathrm{diag}(\boldsymbol{\mu}(\mathbf{1} - \boldsymbol{\mu}))X.
\end{aligned}$$

We now show that the Hessian is positive semi-definite. Note that the Hessian takes the form

$$H = \sum_i \mathbf{x}_{ii}^2(\boldsymbol{\mu}(\mathbf{1} - \boldsymbol{\mu}))$$

so since $\mathbf{x}_{ii}^2 \geq 0$, we see that the Hessian is positive semi-definite if and only if $\mathrm{diag}(\boldsymbol{\mu}(1 - \boldsymbol{\mu}))$ is positive semidefinite. Recall that the eigenvalues of a diagonal matrix are the diagonal elements, so we must show that

$$\boldsymbol{\mu}_i(1 - \boldsymbol{\mu}_i) \geq 0$$

or equivalently that

$$0 \leq \boldsymbol{\mu}_i \leq 1.$$

Recall that

$$\boldsymbol{\mu}_i = \sigma(\boldsymbol{\theta}^\top \mathbf{x}_i)$$

and that

$$0 < \sigma(\cdot) < 1$$

so our desired inequality is satisfied, and the negative log likelihood Hessian is positive semidefinite.

$\blacksquare$

**8 (Murphy 9)** Show that the multinomial distribution

$$\mathrm{Cat}(x|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

is in the exponential family and show that the generalized linear model corresponding to this distribution is the same as multinomial logistic regression.

We first put $\mathrm{Cat}(x|\boldsymbol{\mu})$ into the exponential form:

$$\begin{aligned}
\mathrm{Cat}(x|\boldsymbol{\mu}) &= \exp \log \mathrm{Cat}(x|\boldsymbol{\mu}) \\
&= \exp \left[ \sum_{k=1}^{K} x_k \log \mu_k \right] \\
&= \exp \left[ \sum_{k=1}^{K-1} x_k \log \mu_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log(1 - \sum_{k=1}^{K-1} \mu_k) \right] \\
&= \exp \left[ \sum_{k=1}^{K-1} x_k \log \left( \frac{\mu_k}{1 - \sum_{j=1}^{K-1} \mu_j} \right) + \log(1 - \sum_{k=1}^{K-1} \mu_k) \right] \\
&= \exp \left[ \sum_{k=1}^{K-1} x_k \log \left( \frac{\mu_k}{\mu_K} \right) + \log \mu_K \right]
\end{aligned}$$

where $\mu_K = 1 - \sum_{k=1}^{K-1} \mu_k$. Thus we can write the multinomial distribution in exponential family form as:

$$\begin{aligned}
\mathrm{Cat}(x|\boldsymbol{\theta}) &= \exp(\boldsymbol{\theta}^\top \phi(x) - A(\boldsymbol{\theta})) \\
\boldsymbol{\theta} &= [\log \frac{\mu_1}{\mu_K}, \ldots, \log \frac{\mu_{K-1}}{\mu_K}] \\
\phi(x) &= [\mathbb{I}(x = 1), \ldots, \mathbb{I}(x = K-1)]
\end{aligned}$$

and we derive

$$\begin{aligned}
\mu_i &= \mu_K e^{\boldsymbol{\theta}_i} && \text{from } \boldsymbol{\theta} \\
\mu_K &= 1 - \mu_K \sum_{i=1}^{K-1} e^{\boldsymbol{\theta}_i} && \text{subbing the above for } \mu_i \\
\mu_K &= \frac{1}{1 + \sum_{i=1}^{K-1} e^{\boldsymbol{\theta}_i}} && \text{solving the above equation} \\
\mu_i &= \frac{e^{\boldsymbol{\theta}_i}}{1 + \sum_{i=1}^{K-1} e^{\boldsymbol{\theta}_i}} && \text{subbing the above for } \mu_K
\end{aligned}$$

which implies that

$$A(\boldsymbol{\theta}) = -\log(\mu_K) = \log \left( 1 + \sum_{k=1}^{K-1} e^{\boldsymbol{\theta}_k} \right).$$

■