S&DS 365 Homework 5 Solutions

Yale University, Department of Statistics

April 19, 2021

Problem 1: Gradients

a)

$$
\begin{aligned}
\mathrm{tr} AC^T &= \sum_i (AC^T)_{i,i} \\
&= \sum_i \sum_l A_{i,l} C^T_{l,i} \\
&= \sum_i \sum_l A_{i,l} C_{i,l}
\end{aligned}
$$

b)

$$
\begin{aligned}
f(v) &= \sum_i v_i^3 \\
\frac{\partial}{\partial v_i} f(v) &= 3v_i^2 \\
\nabla f(\nu) &= 3\nu^2
\end{aligned}
$$

c)

$$
\begin{aligned}
f(\beta) &= \sum_i (x_i^T \beta - y_i)^3 \\
\nabla f(\beta) &= \sum_i \nabla (x_i^T \beta - y_i)^3 \\
&= \sum_i 3(x_i^T \beta - y_i)^2 x_i
\end{aligned}
$$

let $\epsilon_i = 3(x_i^T \beta - y_i)^2$ then we have

$$
\sum_i \epsilon_i x_i = X^T \epsilon = X^T (3(X\beta - Y)^2
$$

d)

USe product rule and examples above. Recall,

$$\begin{aligned}\nabla_A \text{tr}(AB^T) &= B \\ \nabla_A \text{tr}(BA^T) &= B\end{aligned}$$

$$\begin{aligned}\nabla f(A) &= \nabla_{A_1}(A_1CA_2^T) + \nabla_{A_2}(A_1CA_2^T)|_{A_1=A_2=A} \quad \text{by applying product rule} \\ &= A_2C^T + A_1C|_{A_1=A_2=A} \\ &= AC^T + AC\end{aligned}$$

OR

$$\begin{aligned}f(A) &= \sum_{i=1}^{m}(ACA^T)_{i,i} = \sum_{i=1}^{M}\sum_{l=1}^{n}(A_{i,l}(CA^T)_{l,i}) \\ &= \sum_{i=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{n}(A_{i,l}C_{l,k}A_{i,k})\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial A_{s,t}}f(A) &= \frac{\partial}{\partial A_{s,t}}\sum_{i=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{n}(A_{i,l}C_{l,k}A_{i,k}) \\ &= \sum_{i=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{n}C_{l,k}A_{i,k}1_{i=s,l=t} + A_{i,l}C_{l,k}1_{i=s,k=t} \\ &= \sum_{k=1}^{n}C_{t,k}A_{s,k} + \sum_{l=1}^{n}C_{l,t}A_{s,l} \\ &= [AC^T]_{s,t} + [AC]_{s,t}\end{aligned}$$

implies same result $\nabla_A f(A) = AC^T + AC$.

Problem 2: Exponential families

a)

$$\begin{aligned}h(y) &= 1_{y\in\{0,1\}} \\ T(y) &= y \\ A(\theta) &= \log(1+e^{\theta})\end{aligned}$$

b)

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x^2 - 2xy + y^2)\right)$$

$$
\begin{aligned}
h(y) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{y^2}{\sigma^2}) \\
T(y) &= \frac{y}{\sigma^2} \\
A(\theta) &= \frac{\theta^2}{2\sigma^2}
\end{aligned}
$$

c)

$$A'(\theta) = \frac{e^\theta}{1+e^\theta}$$

$$E_\theta(Y) = P_\theta(Y=1) = \frac{e^\theta}{1+e^\theta}$$

d)

$$
\begin{aligned}
\nabla \log(L(\theta)) &= \sum_{i=1}^{n} y_i - \frac{e^\theta}{1+e^\theta} = 0 \\
n\frac{e^\theta}{1+e^\theta} &= \sum_i y_i \\
\frac{e^\theta}{1+e^\theta} &= \frac{1}{n}\sum_i y_i = \bar{y} \\
e^\theta &= (1-e^\theta)\bar{y} \\
e^\theta(1-\bar{y}) &= \bar{y} \\
\theta &= \ln\frac{\bar{y}}{1-\bar{y}}
\end{aligned}
$$

e)

$$
\begin{aligned}
p(y)h(y)\exp(\theta^T T(y) - A(\theta)) \\
\ln p(y) = \ln h(y) + \theta^T - A(\theta)
\end{aligned}
$$

for n iid samples we have

$$
\begin{aligned}
p(y_{[1,\cdots,n]}|\theta) &= \prod_{i=1}^{n} p(y_i) \\
l(\theta) &= \sum_{i=1}^{n} \ln p(y_i) \\
&= \sum_{i=1}^{n} \ln h(y_i) + \theta^T T(y_i) - A(\theta) \\
\nabla l(\theta) &= (\sum_{i=1}^{n} T(y_i)) - n\nabla A(\theta)
\end{aligned}
$$

at the MLE we have $\nabla l(\hat{\theta}) = 0$ so we have

$$0 = (\sum_{i=1}^{n} T(y_i)) - n\nabla A(\hat{\theta})$$
$$\nabla A(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} T(y_i)$$

Problem 3: eneralized Linear Models and Gradient Descent

a)

$$p(y_{[1,\cdots,n]}|\theta, x_{[1,\cdots,n]}) = \prod_{i=1}^{n} p(y_i|x_i,\theta)$$
$$-l(\theta) = -\sum_{i=1}^{n} \ln p(y_i|x_i,\theta)$$
$$= \sum_{i=1}^{n} -\ln h(y_i) - y_i\langle x_i,\theta\rangle + A(\langle x_i,\theta\rangle)$$

b)

Take gradient

$$\nabla - l(\theta) = \sum_{i=1}^{n} -y_i x_i + A'(\langle x_i,\theta\rangle x_i$$
$$= \sum_{i} x_i(A'(\langle x_i,\theta\rangle) - y_i)$$

c)

$$A'(s) = \frac{e^s}{1+e^s}$$

$$A'(\langle\theta, x\rangle) - y_i = \frac{e^{\langle x_i,\theta\rangle}}{1+e^{\langle x_i,\theta\rangle}} - y_i$$
$$= P[y_i = 1|x_i,\theta] - y_i$$

here we are comparing the probability that $y_i = 1$ to the actual outcome of y_i. So if y_i is likely to be 1 and the probability is high, this difference will be small. However if y_i is likely to be one and is zero, the difference will be large.

d)

$$\theta_k = \theta_{k-1} - \eta_k x_{J_k}(x_{J_k}^T\theta_{k-1} - y_{J_k})$$

e)

Still for $A(s) = \frac{s^2}{2}$, $A'(s) = s$.

$$
\begin{aligned}
&|x_{J_k}^T \theta_k - y_{J_k}| \\
&= |x_{J_K}^T (\theta_{k-1} - \eta_k x_{J_k} (x_{J_k}^T \theta_{k-1} - y_{J_k})) - y_{J_k}| \\
&= |x_{J_k}^T \theta_{k_{-1}} - \frac{1}{10} (x_{J_k}^T \theta_{k-1} - y_{J_k}) - y_{J_k}| \\
&= \frac{9}{10} |x_{J_k}^T \theta_{k-1} - y_{J_k}| \\
&\leq |x_{J_k}^T \theta_{k-1} - y_{J_k}|
\end{aligned}
$$

strict if $x_{J_k}^T \theta_{k-1} \neq y_{J_k}$.