Problem 1 Gradient

1.1 Trace Gradient

Let $A, C \in R^{m \times n}$, Show that trace $(AC^T) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{(ij)} C_{(ij)}$

Proof:

$$\begin{aligned} tr(AC^T) &= \sum_{i=1}^{m} (AC^T)_{ii} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} A_{(ij)} [C^T]_{(ji)} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} A_{(ij)} C_{(ij)} \end{aligned}$$

1.2 Cubic

Compute gradient of $f(v) = \sum_i v^3_{(i)}$ $\quad v \in R^p \quad f: \mathbb{R}^p \to R$

Since f is a function of a vector, its gradient is also a vector.
The i^{th} coordinate of the gradient can be written as:

$$\begin{aligned} [\nabla f(v)]_{(i)} &= \frac{\partial f(v)}{\partial v_{(i)}} \\ &= \frac{\partial \left(\sum_{k=1}^{p} v^3_{(k)} \right)}{\partial v_{(i)}} \\ &= \sum_{k=1}^{p} \frac{\partial v^3_{(k)}}{\partial v_{(i)}} \quad \text{Linearity of derivative} \\ &= \sum_{k=1}^{p} \left(\mathbb{1}(i=k)\, 3 v^2_{(k)} \right) \\ &= 3 v^2_{(i)} \end{aligned}$$

Since the i^{th} coordinate of the gradient is $3v^2_{(i)}$, the gradient is vector $3v^2$

1.3 $f(\beta) = \Sigma_i (x_i^T\beta - y_i)^3$ $\quad f: R^p \to R$

Show that gradient of $f(\beta) = 3X^T(X\beta - y)^2$

where $X \in \mathbb{R}^{n\times p}$ with i^{th} row x_i^T, $y_i \in \mathbb{R}$ $\quad y \in \mathbb{R}^n$

Proof:

Since f is a function of a vector, its gradient is also a vector

Set $v = g(\beta) = X\beta - y$ $\quad g: \mathbb{R}^p \to \mathbb{R}^n$

$v_{(i)} = x_i^T\beta - y_i$

then $f(v) = \Sigma_i v_{(i)}^3$

The ith coordinate of the gradient can be written as:

$$[\nabla f(\beta)]_{(i)} = \frac{\partial f(\beta)}{\partial \beta_{(i)}}$$

$$= \sum_{j=1}^{n} \frac{\partial f}{\partial v_{(j)}} \frac{\partial v_{(j)}}{\partial \beta_{(i)}} \qquad \text{Chain rule}$$

$$= \sum_{j=1}^{n} \left[\left[\nabla f(v) \big|_{v=g(\beta)}\right]_{(j)} \frac{\partial (x_j^T\beta - y_j)}{\partial \beta_{(i)}} \right]$$

$$= \sum_{j=1}^{n} \left[\left[3v^2 \big| v = g(\beta)\right]_{(j)} [x_j]_{(i)} \right]$$

$$= \sum_{j=1}^{n} \left[\left[3(X\beta - y)^2\right]_{(j)} [x_j]_{(i)} \right]$$

$$= 3 \sum_{j=1}^{n} \left[[x_j]_{(i)} \left[(X\beta - y)^2\right]_{(j)} \right]$$

Since $X^T\epsilon = \Sigma_{i=1}^n x_i \epsilon_i$, in this case $\epsilon = (X\beta - y)^2 \in \mathbb{R}^{n\times 1}$

then $[\nabla f(\beta)]_{(i)} = 3\left[X^T(X\beta - y)^2\right]_{(i)}$

Thus $\nabla f(\beta) = 3X^T(X\beta - y)^2$

1.4 quadratic trace $f(A) = trace(ACA^T)$ $A \in R^{m\times n}$, $C \in \mathbb{R}^{n\times n}$

Answer: $\nabla f(A) = AC + AC^T$ $f: R^{m\times n} \to \mathbb{R}$

Proof:

Since f is a function of a matrix, its gradient is also a matrix

$f(A) = tr(ACA^T) = tr(A^TAC)$ since $tr(ABC) = tr(CAB)$ for ABC is square

set $W = g(A) = A^TA$ matrix $W \in \mathbb{R}^{n\times n}$ $g: \mathbb{R}^{m\times n} \to \mathbb{R}^{n\times n}$

$\Rightarrow f(W) = tr(WC)$

The i,j coordinate of gradient of f is:

$$[\nabla f(A)]_{(ij)} = \frac{\partial f}{\partial A_{(ij)}}$$

$$= \sum_{\substack{k=1\\l=1}}^{n} \left[\frac{\partial f}{\partial W_{(kl)}} \frac{\partial W_{(kl)}}{\partial A_{(ij)}}\right]$$ chain rule

$$= \sum_{\substack{k=1\\l=1}}^{n} \left[\left[\nabla f(W)\big|_{W=g(A)}\right] \frac{\partial [A^TA]_{(kl)}}{\partial A_{(ij)}}\right]$$

$$= \sum_{\substack{k=1\\l=1}}^{n} \left[[C^T]_{(kl)} \frac{\partial \sum_{p=1}^{m}\left[[A^T]_{(kp)} A_{(pl)}\right]}{\partial A_{(ij)}}\right]$$

Since from eg.3 we know $[\nabla_A tr(AC^T)]_{(ij)} = C_{(ij)}$

$$= \sum_{\substack{k=1\\l=1}}^{n} \left[C_{(lk)} \sum_{p=1}^{m} \frac{\partial A_{(pk)} A_{(pl)}}{\partial A_{(ij)}}\right]$$ linearity of derivative

$$= \sum_{\substack{k=1\\l=1}}^{n} \sum_{p=1}^{m} C_{(lk)} \left[\mathbb{1}(i=p, j=k) A_{(pl)} + \mathbb{1}(i=p, j=l) A_{(pk)}\right]$$

product rule

$$= \sum_{l=1}^{n} C_{(lj)} A_{(il)} + \sum_{k=1}^{n} C_{(jk)} A_{(ik)}$$

$$= \sum_{l=1}^{n} A_{(il)} C_{(lj)} + \sum_{k=1}^{n} A_{(ik)} [C^T]_{(kj)}$$

$$= [AC]_{(ij)} + [AC^T]_{(ij)}$$

$$= [AC + AC^T]_{(ij)}$$

Since the i,j coordinate of gradient is $[AC+AC^T]_{(ij)}$, the gradient is $AC + ACT$

2.4 Bernoulli MLE

$$\log L(\theta) = \log \prod^{n} \left[\mathbb{1}(y_i \in \{0,1\})\exp\left[y_i\theta - \log(1+\exp(\theta))\right]\right]$$

Problem 2 Exponential Families

2.1 Bernoulli

$$\begin{cases} h(y) = \mathbb{1}(y \in \{0,1\}) \\ T(y) = y \\ A(\theta) = \log(1+\exp(\theta)) \end{cases}$$

2.2 Gaussian

$$Y|\theta \sim N(\mu, \sigma^2)$$

$$\begin{aligned} P(y;\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y^2}{2\sigma^2} + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y^2}{2\sigma^2}\right] \exp\left[y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] \end{aligned}$$

take $\theta = \mu$

$$\begin{cases} h(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{y^2}{2\sigma^2}\right] \\ T(y) = \frac{y}{\sigma^2} \\ A(\theta) = \frac{\mu^2}{2\sigma^2} = \frac{\theta^2}{2\sigma^2} \end{cases}$$

2.3 Bernoulli Gradient

$$\begin{aligned} A'(\theta) &= \frac{\partial \log(1+\exp(\theta))}{\partial\theta} \\ &= \frac{\exp(\theta)}{1+\exp(\theta)} \end{aligned}$$

$$\begin{aligned} E_\theta y &= 1 \times P(y=1;\theta) + 0 \times P(y=0;\theta) \\ &= P(y=1;\theta) \\ &= \exp\left[\theta - \log(1+\exp(\theta))\right] \\ &= \frac{\exp(\theta)}{\exp\left[\log(1+\exp(\theta))\right]} = \frac{\exp(\theta)}{1+\exp(\theta)} = A'(\theta) \end{aligned}$$

2.4 Bernoulli MLE

$$\log L(\theta) = \log \prod_{i=1}^{n} \left[\mathbb{1}(y_i \in \{0,1\}) \exp\left[y_i\theta - \log(1+\exp(\theta))\right]\right]$$

$$= \sum_{i=1}^{n} \log \left[\mathbb{1}(y_i \in \{0,1\}) \exp\left[y_i\theta - \log(1+\exp(\theta))\right]\right]$$

$$= \sum_{i=1}^{n} \log \left[\exp\left[y_i\theta - \log(1+\exp(\theta))\right]\right]$$

$$= \sum_{i=1}^{n} \left[y_i\theta - \log[1+\exp(\theta)]\right]$$

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \left[y_i - \frac{\exp(\theta)}{1+\exp(\theta)}\right]$$

set $\frac{\partial \log L(\theta)}{\partial \theta} = 0$

$$\sum_{i=1}^{n} \left[y_i - \frac{\exp(\hat{\theta})}{1+\exp(\hat{\theta})}\right] = 0$$

$$n\left(\frac{e^{\hat{\theta}}}{1+e^{\hat{\theta}}}\right) = \sum_{i=1}^{n} y_i$$

$$\Rightarrow \quad \hat{\theta} = -\log\left[\frac{n}{\sum_{i=1}^{n} y_i} - 1\right]$$

2.5 Exp Family Gradient

$$P(y;\theta) = h(y)\exp\left[\langle\theta, T(y)\rangle - A(\theta)\right]$$

$$L(\theta) = \prod_{i=1}^{n} P(y_i;\theta)$$

$$= \prod_{i=1}^{n} \left[h(y_i) \exp\left[\langle\theta, T(y_i)\rangle - A(\theta)\right]\right]$$

$$\log L(\theta) = \sum_{i=1}^{n} \log\left[h(y_i) \exp\left[\langle\theta, T(y_i)\rangle - A(\theta)\right]\right]$$

$$= \sum_{i=1}^{n} \left[\log h(y_i) + \left[\langle\theta, T(y_i)\rangle - A(\theta)\right]\right]$$

$$= \sum_{i=1}^{n} \log h(y_i) + \sum_{i=1}^{n} \left[\langle\theta, T(y_i)\rangle - A(\theta)\right]$$

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} [\langle \theta, T(y_i) \rangle - A(\theta)]}{\partial \theta}$$

$$= \sum_{i=1}^{n} \frac{\partial [\langle \theta, T(y_i) \rangle - A(\theta)]}{\partial \theta} \quad \text{linearity of derivative}$$

$$= \sum_{i=1}^{n} [T(y_i) - \nabla A(\theta)]$$

Since from Problem 2 eq.0 we know $\nabla_v \langle v, w \rangle = w \quad v, w \in R^s$

Set $\dfrac{\partial \log L(\theta)}{\partial \theta} = 0$

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

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

Problem 3 GLM and SGD

3.1 NLL of GLM

$$L(\theta) = \prod_{i=1}^{n} P(y_i ; x_i, \theta)$$

$$NLL(\theta) = -\log L(\theta) = -\log\left[\prod_{i=1}^{n} P(y_i ; x_i, \theta)\right]$$

$$= -\sum_{i=1}^{n} \log[P(y_i ; x_i, \theta)]$$

$$= -\sum_{i=1}^{n} \log\left[h(y_i) \exp\left[y_i \langle x_i, \theta \rangle - A(\langle x_i, \theta \rangle)\right]\right]$$

$$= -\sum_{i=1}^{n} \left[\log(h(y_i)) + y_i \langle x_i, \theta \rangle - A(\langle x_i, \theta \rangle)\right]$$

$$= \sum_{i=1}^{n} \left[A(\langle x_i, \theta \rangle) - y_i \langle x_i, \theta \rangle - \log(h(y_i))\right]$$

3.2 Gradient of NLL of GLM

Since $NLL(\theta)$ is a function of a vector $\in \mathbb{R}^s$,
the gradient of $NLL(\theta)$ is also a vector $\in \mathbb{R}^s$

set $NLL(\theta) = f(\langle x, \theta \rangle) = f(x^T\theta) = f(\theta^T x)$

from problem 1 eq.1 we know $\nabla_\theta NLL(\theta) = \nabla_\theta f(\theta^T x) = f'(\theta^T x)x$

then
$$\nabla_\theta NLL(\theta) = \sum_{i=1}^{n} \left[A'(\theta^T x_i) x_i - y_i x_i\right]$$
$$= \sum_{i=1}^{n} x_i (A'(\theta^T x_i) - y_i)$$
$$= \sum_{i=1}^{n} x_i (A'(\langle x_i, \theta \rangle) - y_i)$$

3.3 Error for logistic Regression

$$A'(t) = \frac{\partial A(t)}{\partial t} = \frac{\partial \log(1+\exp(t))}{\partial t} = \frac{\exp(t)}{1+\exp(t)}$$

$$\Rightarrow A'(\langle x_i, \theta \rangle) - y_i = \frac{\exp(\langle x_i, \theta \rangle)}{1+\exp(\langle x_i, \theta \rangle)} - y_i$$

3.4 SGD update for linear regression

$$A(s) = \frac{s^2}{2} \Rightarrow A'(s) = s$$

$$\theta_k = \theta_{k-1} - \eta_k x_{J_k} (\langle x_{J_k}, \theta_{k-1} \rangle - y_{J_k})$$

3.5 SGD improvement on random sample

$$\langle x_{J_k}, \theta_k \rangle = \left\langle x_{J_k}, \theta_{k-1} - \eta_k x_{J_k} \left[A'(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}\right] \right\rangle$$

$$= \langle x_{J_k}, \theta_{k-1} \rangle - \langle x_{J_k}, \eta_k x_{J_k} \left[A'(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}\right] \rangle$$

since additivity of inner product

$$= \langle x_{J_k}, \theta_{k-1} \rangle - \left[A'(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}\right] \eta_k \langle x_{J_k}, x_{J_k} \rangle$$

since linearity of inner product

plug in $\eta_k = \frac{1}{10\|x_{J_k}\|_2^2}$

$$= \langle x_{J_k}, \theta_{k-1} \rangle - \left[A'(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}\right] \frac{1}{10\|x_{J_k}\|_2^2} \|x_{J_k}\|_2^2$$

$$= \langle x_{J_k}, \theta_{k-1} \rangle - \frac{1}{10} \left[A'(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}\right]$$

left side $= \left| A'(\langle x_{J_k}, \theta_k \rangle) - y_{J_k} \right|$

$$= \left| \langle x_{J_k}, \theta_k \rangle - y_{J_k} \right|$$

$$= \left| \left[\langle x_{J_k}, \theta_{k-1} \rangle - \frac{1}{10} \left[A'(\langle x_{J_k}, \theta_{k-1} \rangle) - y_{J_k}\right] \right] - y_{J_k} \right|$$

$$= \left| \left[\langle x_{J_k}, \theta_{k-1} \rangle - y_{J_k} \right] - \frac{1}{10} \left[\langle x_{J_k}, \theta_{k-1} \rangle - y_{J_k} \right] \right|$$

$$= \left| \frac{9}{10} \left[\langle x_{J_k}, \theta_{k-1} \rangle - y_{J_k} \right] \right|$$

$$= \frac{9}{10} \left| \langle x_{J_k}, \theta_{k-1} \rangle - y_{J_k} \right| < \left| \langle x_{J_k}, \theta_{k-1} \rangle - y_{J_k} \right| = \text{right side}$$