1. (a) let $y = Ax$

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} A_{ik} x_k$$

only one terms depends on x_j

$$= \frac{\partial}{\partial x_j} A_{ij} x_j$$

$$= A_{ij}$$

$$\therefore \frac{\partial}{\partial x} Ax = A$$

(b)

$$\frac{\partial}{\partial x_i} w^T x = \frac{\partial}{\partial x_i} \sum_{k=1}^{n} w_k x_k$$

term which depends on x_i

ith entry of gradient

$$= \frac{\partial}{\partial x_i} w_i x_i$$

$$= w_i$$

$$\therefore \nabla_x w^T x = w$$

(c) $\underbrace{x^T}_{u} \underbrace{Ax}_{v}$

$$\nabla_x x^T A x = \nabla_x u^T v$$

$$= \left(\frac{\partial u}{\partial x}\right)^T v + \left(\frac{\partial v}{\partial x}\right)^T u$$

(dot product rule)

$$= \left(\frac{\partial}{\partial x} x\right)^T Ax + \left(\frac{\partial}{\partial x} Ax\right)^T x$$

can treat this as Ix & apply the first part

$$= I Ax + A^T x$$

$$= (A + A^T) x$$

(d) $\underbrace{a^T x}_{u} \underbrace{x^T b}_{v}$

$$\nabla_x a^T x x^T b = \nabla_x u^T v$$

$$= \left(\frac{\partial u}{\partial x}\right)^T v + \left(\frac{\partial v}{\partial x}\right)^T u$$

notice u & v are scalars x a vector

$$\therefore \left(\frac{\partial u}{\partial x}\right)^T = \nabla_x u \qquad \left(\frac{\partial v}{\partial x}\right)^T = \nabla_x v$$

$$= (\nabla_x a^T x)\, b^T x + (\nabla_x b^T x)\, a^T x$$

$$= a b^T x + b a^T x$$

$$= (ab^T + ba^T) x$$

2. (a) We compute the derivative first:

$$\frac{\partial}{\partial x} \|Ax-b\|_2 = \frac{\partial}{\partial x} \sqrt{\|Ax-b\|_2^2}$$

$$= \underbrace{\frac{1}{2\|Ax-b\|_2}}_{\frac{\partial \sqrt{\|Ax-b\|_2^2}}{\partial \|Ax-b\|_2^2}} \cdot \frac{\partial}{\partial x} \|Ax-b\|_2^2$$

(regular scalar derivative)

let $u = Ax-b$

$\frac{\partial}{\partial u} \|u\|_2^2 = \frac{\partial}{\partial u} u^T I u$

$= 2u^T$

(gradient is $2u$, so derivative is the transpose)

$$= \frac{1}{2\|Ax-b\|_2} 2(Ax-b)^T \cdot \frac{\partial}{\partial x}(Ax-b)$$

$$= \frac{2(Ax-b)^T A}{2\|Ax-b\|_2}$$

$$\therefore \nabla_x \|Ax-b\|_2 = \frac{1}{\|Ax-b\|_2} A^T(Ax-b)$$

$$\nabla_x \|x\|_2^4 = \nabla_x (x^T I x)^2$$

$$= 2\|x\|_2^2 \cdot \nabla_x (x^T I x)$$

(gradient chain rule)

$$= 4\|x\|_2^2 \, x$$

$\downarrow$

Sum to get the gradient of the entire expression

if you've taken a linear algebra class, this is just $\langle A\, xx^T \rangle$

(b) $\nabla_x \operatorname{tr}(A\, xx^T) = \nabla_x \operatorname{tr}(\underbrace{x^T A x}_{\text{this is a scalar! trace is useless}})$ ← cyclic property of trace!

$$= \nabla_x x^T A x$$

$$= (A + A^T)x$$

(c)

$$\frac{\partial}{\partial x_i} -y^T \ln x = \frac{\partial}{\partial x_i} -\sum_{k=1}^{n} y_k \ln x_k$$

$$= \frac{\partial}{\partial x_i} -y_i \ln x_i$$

$$= -\frac{y_i}{x_i}$$

$$\therefore \nabla_x -y^T \ln x = -y \odot \frac{1}{x}$$

↑ elementwise functions

(d) This one is annoying...

let $\begin{cases} f(x) = y \ln \hat{y} + (1-y)\ln(1-\hat{y}) \\ \dfrac{df}{d\hat{y}} = \dfrac{y}{\hat{y}} - \dfrac{1-y}{1-\hat{y}} \end{cases}$

will use this to apply chain rule

$\begin{cases} \sigma(u) = \dfrac{1}{1+e^{-u}} \\ \dfrac{d\sigma}{du} = \sigma(u)\left[1-\sigma(u)\right] \; (*) \end{cases}$

this is a scalar, so we don't need to worry about about order or transposes

$$\therefore \nabla_w \; y \ln \underline{g(x)} + (1-y)\ln(1-y) = \left(\frac{y}{g(x)} - \frac{1-y}{1-g(x)}\right) \nabla_w \left(\frac{1}{1+e^{-w^Tx}}\right)$$

notice this is $\sigma(w^T x)$ where σ is the sigmoid (scalar) function

$$= \left(\frac{y}{g(x)} - \frac{1-y}{1-g(x)}\right)\left[g(x)\left(1-g(x)\right)\right] \nabla_w (w^T x)$$

$$= \left(\frac{y}{g(x)} - \frac{1-y}{1-g(x)}\right) g(x)\left(1-g(x)\right) x$$

note: this notation is a bit annoying because here we are treating $g(x)$ as a function of w & not x, which is held constant...

... as you can imagine though, we write it as $g(x)$ since it represents a decision function which maps x to a probability

(*)
$$\frac{d}{du}\sigma(u) = \frac{d}{du}\frac{1}{(1+e^{-u})}$$

$$= \frac{1}{(1+e^{-u})^2}\left[\cancel{\left(\frac{d}{du}1\right)}^{\,0}(1+e^{-u}) - \underbrace{\left(\frac{d}{du}1+e^{-u}\right)}_{-e^{-u}}\cdot 1\right]$$

$$= \frac{e^{u}}{(1+e^{-u})^2}$$

$$= \frac{1}{1+e^{-u}} \; \frac{e^{-u}}{1+e^{-u}}$$

$$= \frac{1}{1+e^{-u}}\left(1 - \frac{1}{1+e^{-u}}\right)$$

$$= \sigma(u)\left[1-\sigma(u)\right] \; ✓$$

3. First we establish the following property of the trace:

$$\mathrm{tr}(X) := \sum_{i=1}^{n} X_{ii}$$

$$\therefore \mathrm{tr}(X^T Y) = \sum_{i=1}^{n} (X^T Y)_{ii}$$

assume $X \in \mathbb{R}^{m \times n}$, $Y \in \mathbb{R}^{m \times n}$, $X^T \in \mathbb{R}^{n \times m}$

$$= \sum_{i=1}^{n} \left(\sum_{j=1}^{m} (X^T)_{ij} Y_{ji} \right)$$

definition of matrix multiplication!

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ji} Y_{ji}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij}$$

(shuffle indices so this is in canonical form)

(a)

$$\frac{\partial}{\partial X_{ij}} \mathrm{tr}(A^T X) = \frac{\partial}{\partial X_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} A_{ij}$$

(i,j) entry of gradient

$$= \frac{\partial}{\partial X_{ij}} X_{ij} A_{ij}$$

$$= A_{ij}$$

$$\therefore \nabla_X \mathrm{tr}(A^T X) = A$$

(notice this says $\nabla_x \langle A, X \rangle = A$ which is as expected!)

(b)

$$\nabla_X a^T X b = \nabla_X \mathrm{tr}(a^T X b)$$

$$= \nabla_X \mathrm{tr}(\underbrace{b a^T}_{\text{"}A^T\text{"}} X)$$

$$= (b a^T)^T$$

$$= a b^T$$

(c)

$$\frac{\partial}{\partial X_{ij}} \|X\|_F^2 = \frac{\partial}{\partial X_{ij}} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2$$

$$= \frac{\partial}{\partial X_{ij}} X_{ij}^2$$

$$= 2 X_{ij}$$

$$\therefore \nabla_X \|X\|_F^2 = 2X$$

(d) This one is a bit annoying...

$$X := \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}$$

let $x_i \in \mathbb{R}^m$ denote the columns of the matrix X

$$AX = \begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ax_1 & Ax_2 & \cdots & Ax_n \\ | & | & & | \end{bmatrix}$$

$$\therefore \|AX\|_F^2 = \sum_{i=1}^{n} \|\underbrace{Ax_i}_{\text{all vectors}}\|_2^2$$

$$\nabla_{x_i} \|AX\|_F^2 = \nabla_{x_i} \sum_{k=1}^{n} \|Ax_k\|_2^2$$

will give a column of our gradient

$\frac{\partial}{\partial X_{1i}}$

$\frac{\partial}{\partial X_{2i}}$

$\vdots$

$$= \nabla_{x_i} \|Ax_i\|_2^2$$

only term which depends on x_i

$$= \underbrace{A^T}_{\left(\frac{\partial}{\partial x_i} Ax_i\right)^T} \cdot 2Ax_i$$

(gradient chain rule)

$$= 2A^TAx_i$$

so

$$\nabla_X \|AX_F\|_2^2 = \begin{bmatrix} | & | & & | \\ 2A^TAx_1 & 2A^TAx_2 & \cdots & 2A^TAx_n \\ | & | & & | \end{bmatrix}$$

$$= 2A^TA \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}$$

$$= 2A^TAX$$

4.

$w^* = \arg\min_w f(w; P\ Q\ W\ w_0)$

↑

take gradient of objective & set it equal to 0:

$$\nabla_w f(w) = \nabla_w (Xw - y)^T P (Xw - y) + (w - w_0)^T Q (w - w_0)$$

$$= \left(\frac{\partial}{\partial w} Xw - y\right)^T 2P(Xw - y) + \left(\frac{\partial}{\partial w} w - \overset{\text{constant}}{\cancel{w_0}}\right)^T 2Q(w - w_0)$$

(gradient chain rule)

$$= 2X^T P (Xw - y) + 2I^T Q(w - w_0)$$

$$= 2X^T P(Xw - y) + 2Q(w - w_0)$$

↑ want to set this equal to 0, so we collect all w terms

$$2X^T P(Xw - y) + 2Q(w - w_0) = 0$$

$$X^T P(Xw - y) + Q(w - w_0) = 0$$

$$(X^T PX + Q)w - (X^T Py + Qw_0) = 0$$

$$\underline{(X^T PX + Q)}w = (X^T Py + Qw_0)$$

notice if
$P \succ 0$ then $X^T PX \succeq 0$
$Q \succ 0$

∴ matrix positive definite & so an inverse exists

$$\boxed{w^*} = (X^T PX + Q)^{-1}(X^T Py + Qw_0)$$

↑ optimal weights!