Properties of trace:

(1) We will first show the cyclic property of Trace.

$$Tr(AB) = Tr(BA)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times n}$.

Let $A = \begin{bmatrix} - a_1^T - \\ - a_2^T - \\ \vdots \\ - a_n^T - \end{bmatrix}$, $B = \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_n \\ | & | & & | \end{bmatrix}$

Then,

$$AB = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_n \\ \vdots & & & \vdots \\ a_n^T b_1 & a_n^T b_2 & \cdots & a_n^T b_n \end{bmatrix}$$

So, $tr(AB) = \sum_{i=1}^{n} a_i^T b_i$

Now,

$$BA = b_1 a_1^T + b_2 a_2^T + \cdots + b_n a_n^T$$

Since trace is a linear operator, so

$$tr(BA) = \sum_{i=1}^{n} tr(b_i a_i^T)$$

Now,

$$b_i a_i^T = \begin{bmatrix} b_{1i} \\ b_{2i} \\ \vdots \\ b_{ni} \end{bmatrix} \begin{bmatrix} a_{i1} & a_{i2} & \cdots & a_{in} \end{bmatrix}$$

So,

$$tr(b_i a_i^T) = \sum_{m=1}^{n} a_{im} b_{mi}$$

$$\Rightarrow tr(b_i a_i^T) = a_i^T b_i$$

$$\therefore tr(BA) = \sum_{i=1}^{n} a_i^T b_i = tr(AB)$$

② By definition of trace, it's trivial that

$$tr(A) = tr(A^T)$$

③ Let's show,

$$tr(A^TB) = \sum_i \sum_j a_{ij} b_{ij}$$

Let $A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}$, $B = \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_n \\ | & | & & | \end{bmatrix}$

So,

$$A^TB = \begin{bmatrix} -a_1^T- \\ -a_2^T- \\ \vdots \\ -a_n^T- \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_n \\ | & | & & | \end{bmatrix}$$

Now,

$$tr(A^TB) = \sum_{i=1}^{n} a_i^T b_i$$

where $a_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{bmatrix}$, $b_i = \begin{bmatrix} b_{1i} \\ \vdots \\ b_{ni} \end{bmatrix}$

from above we can see that we are multiplying A and B elementwise and then summing up to get the trace of A^TB.

So, $tr(A^TB) = \sum_{i=1}^{n} \sum_{j=1} a_{ij} b_{ij}$

chain rule of derivatives:

Let $y = f(x)$ for $f: \mathbb{R}^m \Rightarrow \mathbb{R}^n$ and $z = g(y)$ for $g: \mathbb{R}^n \Rightarrow \mathbb{R}^p$. Then we have,

$$\nabla_x z = \nabla_x y \, \nabla_y z$$

Let's first do a dimensionality check on the chain rule. We know that using denominator layout

$\nabla_x z \in \mathbb{R}^{m \times p}$, $\nabla_x y \in \mathbb{R}^{m \times n}$, $\nabla_y z \in \mathbb{R}^{n \times p}$

Hence,

$$\nabla_{x} y \, \nabla_{y} z \in \mathbb{R}^{n \times p}$$

Therefore the dimensions in the chain rule is consistent. Now let's do an intuition check. from the derivative of a vector with respect to a vector we know,

$$\Delta z \approx (\nabla_{x} z)^{T} \Delta x$$

The chain rule is intuitive, since

$$\Delta y \approx (\nabla_{x} y)^{T} \Delta x$$

$$\Delta z \approx (\nabla_{y} z)^{T} \Delta y$$

Composing we get,

$$\Delta z \approx (\nabla_{y} z)^{T} (\nabla_{x} y)^{T} \Delta x$$

Hence,

$$(\nabla_{x} z)^{T} = (\nabla_{y} z)^{T} (\nabla_{x} y)^{T}$$

Practice problem on derivatives:

① Let

$$A = \begin{bmatrix} -a_1^T- \\ -a_2^T- \\ \vdots \\ -a_n^T- \end{bmatrix}, \quad B = \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_n \\ | & | & & | \end{bmatrix}$$

Then,

$$Tr(AB) = a_1^T b_1 + a_2^T b_2 + \cdots + a_n^T b_n$$

Now, using the definition

$$\nabla_A Tr(AB)$$

$$= \frac{\partial Tr(AB)}{\partial a_{ij}}$$

$$= \frac{\partial}{\partial a_{ij}} \left[\sum_{m=1}^{n} a_{1m} b_{m1} + \cdots + \sum_{m=1}^{n} a_{nm} b_{mn} \right]$$

$$= b_{ji}$$

$$\therefore \quad \nabla_A Tr(AB) = B^T$$

(2) $\nabla_A (x^T A x)$

Since $Tr(x^T A x) = x^T A x$, so

$= \nabla_A [Tr(x^T A x)]$

Using the cyclic property, we have

$= \nabla_A [Tr(A x x^T)]$

Using the result from (1),

$= x x^T$

$\therefore \nabla_A x^T A x = x x^T$

(3) We want to compute

$$\nabla_z (x-z)^T \Sigma^{-1} (x-z)$$

Let $y = f(z) = x - z$

$$r = g(y) = y^T \Sigma^{-1} y$$

Then using chain rule of derivatives,

$$\nabla_z r = \nabla_z y \; \nabla_y r$$

Now,

$$\nabla_z y = -I \in \mathbb{R}^{n \times n}$$

From lecture we know,

$$\nabla_y (y^T \Sigma^{-1} y) = \left[(\Sigma^{-1} + \{\Sigma^{-1}\}^T) y \right]$$

Since Σ^{-1} is symmetric, so

$$\nabla_y (y^T \Sigma^{-1} y) = 2 \Sigma^{-1} y$$

$$\therefore \nabla_z (x-z)^T \Sigma^{-1} (x-z) = -2\Sigma^{-1}(x-z)$$

Regularized least squares:

In RLS, we want to solve the following optimization problem

$$\arg\min_{\theta} \frac{1}{2} \sum_{i=1}^{N} \left(y^{(i)} - \theta^T x^{(i)}\right)^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

From lecture we know that

$$\frac{1}{2} \sum_{i=1}^{N} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

$$= \frac{1}{2} (Y - X\theta)^T (Y - X\theta)$$

where, $Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times 1}$

$$X = \begin{bmatrix} x^{(1)T} \\ \vdots \\ x^{(N)T} \end{bmatrix} \in \mathbb{R}^{N \times 2}$$

Then the optimization problem becomes

$$\arg\min_{\theta} \frac{1}{2}(Y-X\theta)^T(Y-X\theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

Since the cost function is convex, so we can solve for θ^* by setting the derivative w.r.t θ to zero.

$$\mathcal{L}(\theta) = \frac{1}{2}\left[Y^TY - Y^TX\theta - \theta^TX^TY + \theta^TX^TX\theta\right] + \frac{\lambda}{2}\theta^T\theta$$

$$= \frac{1}{2}\left[Y^TY - 2Y^TX\theta + \theta^TX^TX\theta\right] + \frac{\lambda}{2}\theta^T\theta$$

Now,

$$\nabla_\theta \mathcal{L} = -\nabla_\theta[Y^TX\theta] + \frac{1}{2}\nabla_\theta[\theta^TX^TX\theta] + \frac{\lambda}{2}\nabla_\theta[\theta^T\theta]$$

$$= -\nabla_\theta[Y^TX\theta] + \frac{1}{2}\nabla_\theta[\theta^T(X^TX+\lambda I)\theta]$$

$$= -X^TY + (X^TX+\lambda I)\theta$$

Setting $\nabla_\theta \mathcal{L} = 0$ and solving for θ we get

$$\theta^* = (X^T X + \lambda I)^{-1} X^T Y$$

1.5.4

$$f = \|A+\lambda B\|_f^2$$

Recall, for a matrix X the frobenius norm is defined as follows

$$\|X\|_f^2 = tr(X^T X)$$

using the above definition,

$$f = tr\left[(A+\lambda B)^T (A+\lambda B)\right]$$

$$= tr\left[A^TA\right] + \lambda Tr\left[A^TB\right] + \lambda Tr\left[B^TA\right] + \lambda^2 Tr\left[B^TB\right]$$

Since B is an orthogonal matrix, so

$$B^TB = I$$

Hence,

$$f = tr\left[A^TA\right] + \lambda Tr\left[A^TB\right] + \lambda Tr\left[B^TA\right] + \lambda^2 n$$

Dropping terms that has no B dependence, so

$$f = \lambda \, \mathrm{Tr}[A^T B] + \lambda \, \mathrm{Tr}[B^T A]$$

Since, $\mathrm{Tr}[A^T B] = \mathrm{Tr}[B^T A]$

So,

$$f = 2\lambda \, \mathrm{Tr}[A^T B]$$

Hence,

$$\nabla_B f = 2\lambda \nabla_B \mathrm{Tr}[A^T B]$$

$$= 2\lambda A \quad \text{(Using (103) from matrix cookbook)}$$

Word translation with linear models:

There are several ways to convert words into their vector representations Some of the popular word embedding models are:

(i) Word 2 Vec

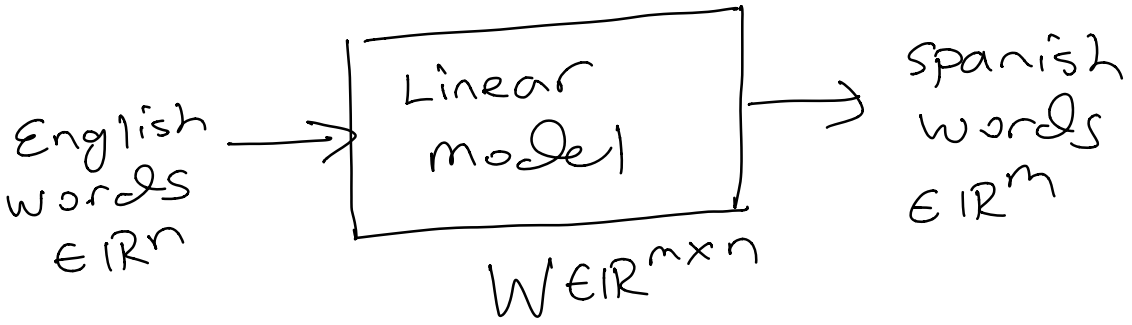
(ii) Glove

(iii) BERT

⋮

We won't get into the details of word embedding models in this discussion, but by the end of this quarter you will have all the tools necessary to

understand them.



a) $$\mathcal{L} = \frac{1}{2} \sum_{i=1}^{k} \| b^{(i)} - W a^{(i)} \|_2^2$$

The loss function is measuring the L_2 distance between the Spanish words in $\mathbb{R}^m$ and the linearly mapped english words in $\mathbb{R}^m$.

The goal of the optimization is to learn a linear mapping W such that the mapped english words are as close as possible to their spanish counterparts in the euclidean norm (L_2) sense.

If there exists a linear mapping between english and spanish words, then the linear transformation W will map the english word 'dog' to the spanish word 'perrito'.

b) $$L(W) = \frac{1}{2} \sum_{i=1}^{k} \|b^{(i)} - Wa^{(i)}\|_2^2$$

Recall that for a matrix $X \in \mathbb{R}^{m \times n}$

$$\|X\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij})^2$$

$$= \sum_{j=1}^{n} \|X(:,j)\|_2^2$$

So if we define the following matrices

$$B = \begin{bmatrix} | & | & & | \\ b^{(1)} & b^{(2)} & \cdots & b^{(k)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times k}$$

$$A = \begin{bmatrix} | & | & & | \\ a^{(1)} & a^{(2)} & \cdots & a^{(k)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times k}$$

Then,

$$L(W) = \frac{1}{2} \|B - WA\|_F^2$$

Using the trace definition of frobenius norm,

$$L(W) = \frac{1}{2} Tr\left[(B-WA)^T (B-WA) \right]$$

$$= \frac{1}{2} Tr\left[(B^T - A^T W^T)(B-WA) \right]$$

$$= \frac{1}{2} Tr\left[B^TB - B^TWA - A^TW^TB + A^TW^TWA \right]$$

Dropping the terms that has no W dependence and observing

$$Tr[B^TWA] = Tr[(B^TWA)^T]$$

$$= Tr[A^TW^TB]$$

we have

$$L(W) = -Tr\left[A^T W^T B\right] + \frac{1}{2} Tr\left[A^T W^T W A\right]$$

Now by cyclic property of trace

$$Tr\left[A^T W^T W A\right]$$

$$= Tr\left[W^T W A A^T\right]$$

Hence,

$$L(W) = -Tr\left[A^T W^T B\right] + \frac{1}{2} Tr\left[W^T W A A^T\right]$$

Now,

$$\nabla_W Tr\left[A^T W^T B\right] = BA^T \quad \text{((102) in matrix cookbook)}$$

$$\nabla_W Tr\left[W^T W A A^T\right] = 2WAA^T \quad \text{((113) in matrix cookbook)}$$

So,

$$\nabla_W L(W) = -BA^T + WAA^T$$