Eli Pinkus
CS 155 Set 5
1 Late hour

Problem 1

Part A

For an $n \times n$ matrix $X = U\Sigma V^T$
We have singular values given by the diagonal matrix Σ.
We have:

$$XX^T = U\Sigma V^T (U\Sigma V^T)^T = U\Sigma V^T V \Sigma^T U^T = U\Sigma^2 U^T$$

So we have the matrix diagonalization for XX^T

$\Rightarrow$ diagonal entries of Σ^2 are the eigenvalues of XX^T and we know that the diagonal entries are the squares of the singular values of X. So we have found the PCA solution is $U\Lambda U^T$ where $\Lambda = \Sigma^2$ so we conclude that the columns of U are the principle components of X since the diagonal entries of Λ are the singular values squared.

Part B

Intuitively, the eigenvalues of XX^T are positive because they represent the variance upon a given dimension and variance is inherently a positive quantity

Mathematically, we showed in part a that the eigenvalues of XX^T are the squares of the singular values of X which must be positive.

Part C

We have matrices A, B

$$\mathrm{Tr}(AB) = \sum_{i=1}^{N} (AB)_{ii}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} B_{ji}$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{N} B_{ji} A_{ij}$$

$$= \sum_{j=1}^{N} (BA)_{jj} = \mathrm{Tr}(BA)$$

It is clear that this generalizes to any number of matrices
Because for square matrices E, F, G:

Eli Pinkus
CS 155 Set 5
1 Late hour

$$\mathrm{Tr}(EFG)$$

Let $A = EF, B = G$, then from above we know that

$$\mathrm{Tr}(EFG) = \mathrm{Tr}(AB) = \mathrm{Tr}(BA) = \mathrm{Tr}(GEF)$$

And the same can be done for any number of matrices and any permutation.

Part D

We know it would take N^2 values to store the entire $N \times N$ matrix

If we have the SVD of X

$$X = U\Sigma V^T$$

If we consider only the first k singular values, We have $U_{1:k}$ which is $N \times k$ and potentially dense, a truncated Σ which is a diagonal $k \times k$ matrix and thus only has k non-zero entries, and finally a truncated V^T which is size $k \times N$ so altogether we need

$$2Nk + k$$

Values.
We want to know when $2Nk + k < N^2$

$$k(2N + 1) < N^2$$

$$k < \frac{N^2}{2N + 1}$$

Part E

Sub i

Let

$$U = \begin{bmatrix} — u_1 — \\ — u_2 — \\ — u_3 — \\ \vdots \\ — u_D — \end{bmatrix}$$

Where each row u_i is of dimension D making U size $N \times N$

Let

Eli Pinkus
CS 155 Set 5
1 Late hour

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_N \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Which is of size $D \times N$

So we have:

$$U\Sigma = \begin{bmatrix} u_{11}\sigma_1 & u_{12}\sigma_2 & \cdots & u_{1N}\sigma_N \\ u_{21}\sigma_1 & & & \vdots \\ \vdots & & & \vdots \\ u_{D1}\sigma_1 & \cdots & \cdots & u_{DN}\sigma_N \end{bmatrix}$$

Let

$$U' = \begin{bmatrix} — u'_1 — \\ — u'_2 — \\ — u'_3 — \\ \vdots \\ — u'_D — \end{bmatrix}$$

Where each row $u'_i = \text{first } N \text{ elements of } u_i$
And

$$\Sigma' = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \sigma_N \end{bmatrix}$$

So we have:

$$U'\Sigma' = \begin{bmatrix} u'_{11}\sigma_1 & u'_{12}\sigma_2 & \cdots & u'_{1N}\sigma_N \\ u'_{21}\sigma_1 & & & \vdots \\ \vdots & & & \vdots \\ u'_{D1}\sigma_1 & \cdots & \cdots & u'_{DN}\sigma_N \end{bmatrix}$$

But since $u'_i = \text{first } N \text{ elements of } u_i$ we can resubstitute and get

$$U'\Sigma' = \begin{bmatrix} u'_{11}\sigma_1 & u'_{12}\sigma_2 & \cdots & u'_{1N}\sigma_N \\ u'_{21}\sigma_1 & & & \vdots \\ \vdots & & & \vdots \\ u'_{D1}\sigma_1 & \cdots & \cdots & u'_{DN}\sigma_N \end{bmatrix} = \begin{bmatrix} u_{11}\sigma_1 & u_{12}\sigma_2 & \cdots & u_{1N}\sigma_N \\ u_{21}\sigma_1 & & & \vdots \\ \vdots & & & \vdots \\ u_{D1}\sigma_1 & \cdots & \cdots & u_{DN}\sigma_N \end{bmatrix} = U\Sigma$$

As desired.

Eli Pinkus
CS 155 Set 5
1 Late hour

Sub ii

In class we said U is orthogonal if $UU^T = U^T U = I$

If U' is $D \times N$ where $D > N$, we have

$$\text{Shape}(U'U'^T) = D \times D$$

And

$$\text{Shape}(U'^T U') = N \times N$$

Thus $UU^T \neq U^T U$

So, we can't fit our definition.

Sub iii

If for some matrix $U \in \mathbb{R}^{D \times N}$
We know that the columns are orthonormal, we know that for any two columns u_i, u_j, $u_i^T u_j = 0$ when $i \neq j$
We also know that if $i = j$ then $u_i^T u_j = 1$,

Since we know how matrix multiplication works we know that some entry of the matrix

$$(U^T U)_{ij} = u_i \cdot u_j = u_i^T u_j$$

So putting it altogether we know that all non-diagonal entries of $U^T U$ are 0 and the diagonals are 1
Which is the definition of the identity matrix which in this case is $I_{D \times D}$
We note for later that $\sum_{i,j} \left(u_{ij}\right)^2 = D$ if $U = I_{D \times D}$
Because the sum of the entrants squared of the identity matrix $I_{D \times D}$ is just D

We can also show that it is false that for some $U \in \mathbb{R}^{D \times N}$ that $UU^T = I_{N \times N}$
Assume for the sake of contradiction that $UU^T = I_{N \times N}$

Aside: let u_i denote the i^{th} **ROW** in U
(I apologize I don't know a more convenient notation)
So by that definition we have an entry of the resulting matrix

$$(UU^T)_{ij} = u_i \cdot u_j = u_i^T u_j$$

We therefore know that $u_i^T u_j = 0$ when $i \neq j$ and $u_i^T u_j = 1$ when $i = j$ since U is assumed to be $I_{N \times N}$.

Eli Pinkus
CS 155 Set 5
1 Late hour

Thus we must have $\sum_{i,j}(u_{ij})^2 = N$, but we know $\sum_{i,j}(u_{ij})^2 = D$ and $N \neq D$, so we have a contradiction. Thus $UU^T \neq I_{N\times N}$.

Part F

i)

We know that Σ is a diagonal square matrix
So say

$$\Sigma = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & \sigma_D \end{bmatrix}$$

We can see that

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & 1/\sigma_D \end{bmatrix}$$

We know that all $\sigma_i \neq 0$ since we are told Σ is invertible
We also know that all $\sigma_i > 0$ from part B.

Z^+ is defined in the slides as follows:

$$\Sigma^+ = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_D \end{bmatrix} \qquad \sigma^+ = \begin{cases} 1/\sigma, & \sigma \neq 0 \\ 0, & \sigma = 0 \end{cases}$$

Thus, $\Sigma^{-1} = \Sigma^+ \Rightarrow V\Sigma^+U^T = V\Sigma^{-1}U^T$ as desired.

ii)

We have $X = U\Sigma V^T$
We know $\Sigma^T = \Sigma$

$$(X^TX) = (U\Sigma V^T)^T(U\Sigma V) = V\Sigma^TU^TU\Sigma V = V\Sigma^2V^T$$

$$\Rightarrow (X^TX)^{-1} = (V\Sigma^2V^T)^{-1} = V(\Sigma^{-1})^2V^T$$

$$(X^TX)^{-1}X^T = V(\Sigma^{-1})^2V^TV\Sigma^TU^T = V(\Sigma^{-1})^2\Sigma^TU^T = V(\Sigma^{-1})^2\Sigma U^T$$

$$= V\Sigma^{-1}U^T$$

As desired.

Eli Pinkus
CS 155 Set 5
1 Late hour

iii)
The pseudoinverse of the form $(X^T X)^{-1} X^T$ is the most complex and prone to numerical issues because it involves that inversion of a potentially large matrix $X^T X$ which is computationally difficult and prone to numerical errors. The other method only needs the inversion of a diagonal matrix which is computationally easy.

Eli Pinkus
CS 155 Set 5
1 Late hour

Question 2

We have the regularized square error:

$$L = \frac{\lambda}{2}\left(||U||_F^2 + ||V||_F^2\right) + \frac{1}{2}\sum_{i,j}\left(y_{ij} - u_i^T v_j\right)^2$$

Taking derivatives and applying chain rule gives:

$$\nabla_{u_i} L = \frac{1}{2}\lambda\frac{\partial(u_i^T u_i)}{\partial u_i} - \sum_j v_j\left(y_{ij} - u_i^T v_j\right)$$

$$= \lambda u_i - \sum_j v_j\left(y_{ij} - u_i^T v_j\right)$$

The derivation for $\nabla_{v_i} L$ is symmetric so we have:

$$\nabla_{v_j} L = \lambda v_j - \sum_i u_i\left(y_{ij} - u_i^T v_j\right)$$

Part B

We set gradient to 0 and look to solve for u_i, v_j assuming the other is constant:

$$0 = \lambda u_i - \sum_j v_j\left(y_{ij} - u_i^T v_j\right)$$

$$= \lambda u_i - \sum_j v_j y_{ij} + \sum_j v_j u_i^T v_j$$

$$= u_i\left(\lambda I_K + \sum_j v_j v_j^T\right) - \sum_j v_j y_{ij} = 0$$

$$\Rightarrow u_i = \left(\lambda I_K + \sum_j v_j v_j^T\right)^{-1}\left(\sum_j y_{ij} v_j\right)$$

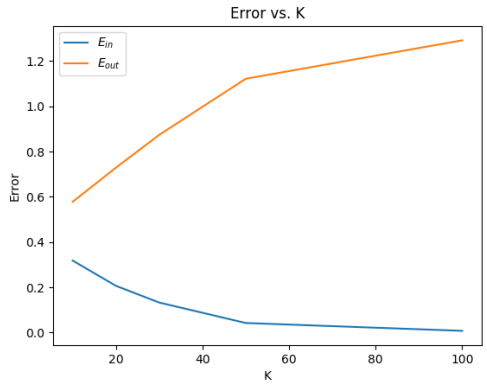
Similarly

$$v_j = \left(\lambda I_K + \sum_i u_i u_i^T\right)^{-1}\left(\sum_i y_{ij} u_i\right)$$

Eli Pinkus
CS 155 Set 5
1 Late hour

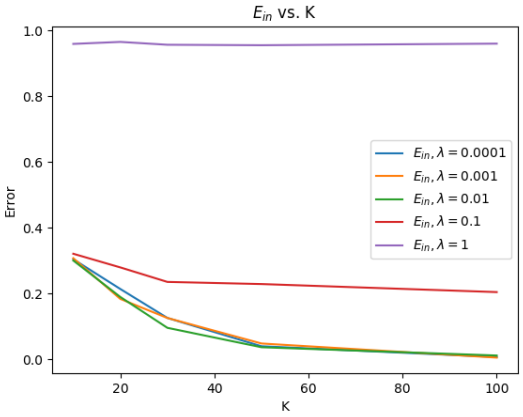
Part D

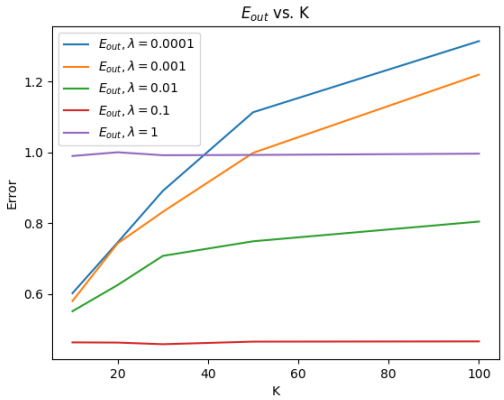
We can see that training error decreases with K which makes sense intuitively because with higher K we are essentially compressing the data less which makes it easier to predict in sample. Testing error generally increases with K which indicates that the model is overfitting with higher K which also makes sense because we are increasing the complexity of the model class and looking to predict a lot of information given relatively sparse information.



Part E

We can see that training error decreases with K for each λ, however we have best training error decreasing with increasing λ which makes sense because a lower regularization strength means that the model is more free to fit the training data. Testing error increases with K generally and there is also best performance for $\lambda = 0.1$ across all K which indicates that it is the regularization strength that results in the best generalization behavior.





Eli Pinkus
CS 155 Set 5
1 Late hour

Problem 3

Part A

We would like to calculate:

$$\log\big(p(w_O|w_I)\big) = \log\left(\frac{\exp\big(v'^T_{w_O} v_{w_I}\big)}{\sum_{w=1}^{W} \exp\big(v'^T_{w} v_{w_1}\big)}\right)$$

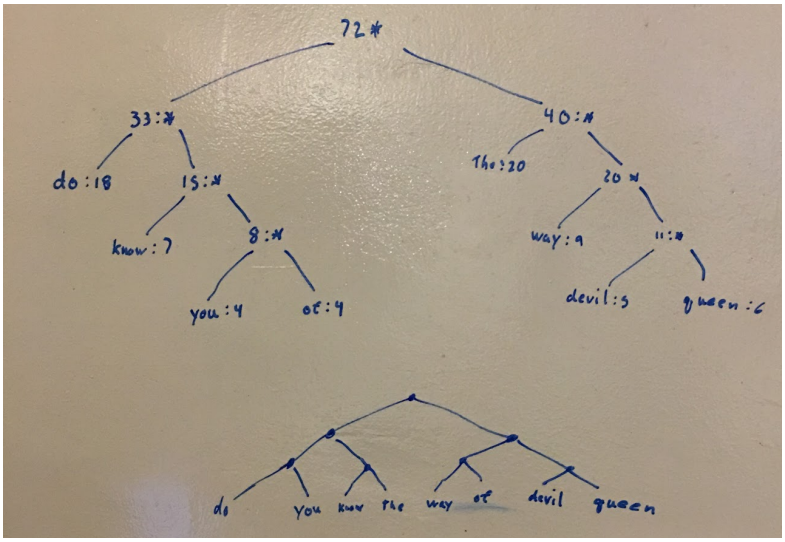
$$= \log\big(\exp\big(v'^T_{W_O} v_{w_I}\big)\big) - \log\left(\sum_{w=1}^{W} \exp\big(v'^T_{w} v_{w_1}\big)\right)$$

$$= v'^T_{W_O} v_{w_I} - \log\left(\sum_{w=1}^{W} \exp\big(v'^T_{w} v_{w_1}\big)\right)$$

We must compute a gradient with respect to each term in the summation so the computation scales linearly with W.

The numerator calculation scales with D and the denominator scales with DW so calculating both scales with $DW + D$

Part B



Part C

As one increases D, the magnitude of the training objective will increase since the embedding vector will become bigger which for one thing will increase the size of the summation of the log probability. Also higher dimensionality can result in overfitting. Additionally, we pay an increased computational cost

Eli Pinkus
CS 155 Set 5
1 Late hour

Part D/E

```
Weights shape layer 1: (308, 10)
Weights shape layer 2: (10, 308)


Pair(fox, goat), Similarity: 0.9869466
Pair(goat, fox), Similarity: 0.9869466
Pair(would, samiam), Similarity: 0.9864516
Pair(samiam, would), Similarity: 0.9864516
Pair(with, eat), Similarity: 0.98387635
Pair(eat, with), Similarity: 0.98387635
Pair(or, anywhere), Similarity: 0.98260736
Pair(anywhere, or), Similarity: 0.98260736
Pair(could, train), Similarity: 0.9825326
Pair(train, could), Similarity: 0.9825326
Pair(eggs, ham), Similarity: 0.98221785
Pair(ham, eggs), Similarity: 0.98221785
Pair(mouse, anywhere), Similarity: 0.9816431
Pair(green, ham), Similarity: 0.98051345
Pair(them, mouse), Similarity: 0.97886395
Pair(not, with), Similarity: 0.97852635
Pair(car, not), Similarity: 0.9766632
Pair(boat, with), Similarity: 0.9765761
Pair(do, eggs), Similarity: 0.9744789
Pair(tree, goat), Similarity: 0.9711042
Pair(box, with), Similarity: 0.96849716
Pair(be, not), Similarity: 0.967821
Pair(dark, tree), Similarity: 0.9654477
Pair(rain, tree), Similarity: 0.96290296
Pair(that, mouse), Similarity: 0.9579255
Pair(there, here), Similarity: 0.9575806
Pair(here, there), Similarity: 0.9575806
Pair(four, five), Similarity: 0.95323676
Pair(five, four), Similarity: 0.95323676
Pair(house, eat), Similarity: 0.9517888
```