1. SVD and PCA:

Question A:

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

The Σ^2 ~~is~~ in SVD is the Σ in PCA, so the columns of V are the PCA of X.

Question B:

(I) $$\|A-B\|_F = \text{trace}((A-B)^T(A-B))$$
$$= \text{trace}(V^T(A-B)^T UU^T(A-B)V)$$
$$= \|U^T(A-B)V\|_F$$
$$= \|\Sigma - U^TBV\|_F$$
$$= \sum_{i,j}(\Sigma - U^TBV)_{ij}^2$$

the ranks of U^T and V are longer than B, so rank $[U^TBV] = k$

$$\Sigma = \begin{pmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_l \end{pmatrix}$$

To minimize $\sum_{i,j}(\Sigma - U^TBV)_{ij}^2$, it is apparent that U^TBV ~~so~~should be

$$\begin{pmatrix} s_1 & & & & \\ & s_2 & & & \\ & & \ddots & & \\ & & & s_k & \\ & & & & 0 \end{pmatrix}$$

$$U^T B V = \Sigma_{1:k} = I_k \Sigma I_k$$

$$B = (U I_k) \Sigma (I_k V^T)$$

$$= U_{1:k} \Sigma V_{1:k}^T$$

$$= A_k$$

so, $\|A - A_k\|_F = \min_{\mathrm{rank}(B) \leq k} \|A - B\|_F$

(ii) We first prove $\|Y\|_* \geq \min_{Y=AB^T} \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$, $Y = U\Sigma V^T$

let $A = U\sqrt{\Sigma}$, $B = V\sqrt{\Sigma}$

$$\min_{Y=AB^T} \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$$

$$\leq \frac{1}{2}(\|U\sqrt{\Sigma}\|_F^2 + \|V\sqrt{\Sigma}\|_F^2)$$

$$= \frac{1}{2}(\mathrm{trace}(\sqrt{\Sigma} U^T U \sqrt{\Sigma}) + \mathrm{trace}(\sqrt{\Sigma} V^T V \sqrt{\Sigma}))$$

$$= \frac{1}{2}(\mathrm{trace}(\Sigma) + \mathrm{trace}(\Sigma))$$

$$= \mathrm{trace}(\Sigma)$$

$$= \|Y\|_*$$

Then we prove $\|Y\|_* \leq \min_{Y=AB^T} \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$

let $A' = U^T A$, $B' = V^T B$

$$A'B'^T = U^T A B V = U^T U \Sigma V^T V = \Sigma$$

$$\|A'\|_F^2 = \mathrm{trace}(A'^T A') = \mathrm{trace}(A^T U U^T A)$$

$$= \mathrm{trace}(A^T A) = \|A\|_F^2$$

$$\|B'\|_F^2 = \|B\|_F^2$$

So, $\min\limits_{Y=AB^T} \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$

$= \min\limits_{\Sigma=A'B'^T} \frac{1}{2}(\|A'\|_F^2 + \|B'\|_F^2)$

$= \min\limits_{\Sigma=A'B'^T} \frac{1}{2}(\text{trace}(A'A'^T) + \text{trace}(B'B'^T))$

$= \min\limits_{\Sigma=A'B'^T} \frac{1}{2}\sum_i (\|A'_i\|^2 + \|B'_i\|^2)$

$\geq \min\limits_{\Sigma=A'B'^T} \frac{1}{2}\sum_i (2\|A'_i B'^T_i\|)$

$= \min\limits_{\Sigma=A'B'^T} \sum_i (\sqrt{B'_i A'^T_i A'_i B'^T_i})$

$= \min\limits_{\Sigma=A'B'^T} \sum_i (\sqrt{(A'B'^T)_{ii}(A'B'^T)_{ii}}$

$= \min\limits_{\Sigma=A'B'^T} \sum_i (\sqrt{S_i^2})$

$= \min\limits_{\Sigma=A'B'^T} \sum_i S_i$

$= \text{trace}(\Sigma)$

$= \|Y\|_*$

In Conclusion: $\|Y\|_* = \min\limits_{Y=AB^T} \frac{1}{2}(\|A\|_F^2 + \|B\|_F^2)$

2. Matrix Factorization:

Question A:

$$\partial_{u_i} = \frac{\lambda}{N} u_i - (y_{ij} - u_i^T v_j) v_j$$

$$\partial_{v_j} = \frac{\lambda}{N} v_j - (y_{ij} - u_i^T v_j) u_i$$

Question B:

$$\frac{\partial}{\partial u_i} \left(\frac{\lambda}{2} (\| U \|_F^2 + \| V \|_F^2) + \sum_{ij} (y_{ij} - u_i^T v_j)^2 \right) = 0$$

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

$$u_i = (\lambda I_k + \sum_j v_j v_j^T)^{-1} (\sum_j v_j y_{ij})$$

$$\frac{\partial}{\partial v_j} \left(\frac{\lambda}{2} (\| U \|_F^2 + \| V \|_F^2) + \sum_{ij} (y_{ij} - u_i^T v_j)^2 \right) = 0$$

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

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

Question C:

Question D:

~~with different latent factor k, E~~

Question E:

In "problem-2-plot.pdf"

3. Word2Vec:

Question A:

$$p(w_O \mid w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W} \exp({v'_{w_O}}^T v_{w_I})}$$

computing $\nabla \log p(w_O \mid w_I)$, the cost is proportional to W, which is always large ($10^5 \sim 10^7$ terms).

Question B:

when D increases to we using large D, our learning process may overfitting. And what happens to training objective?

(i) The loss of input training objective will decrease;

(ii) The loss of output training objective will increase due to overfitting.

Question C:

Yes. When using hierarchical softmax, the cost of computing $\log p(w_O \mid w_I)$ and $\nabla \log p(w_O \mid w_I)$ is proportional to $L(w_O)$, (the length of the path)
which on average is no greater than $\log W$.
compared with Question A, it does improve.

Question D:

(i) if W words are stored in binary tree, which are all leaves,
There are $W-1$ internal nodes does the tree have.

(prove): let n be the depth of the tree

$$2^n = W$$

then internal nodes: $1 + 2^1 + 2^2 + \cdots + 2^{n-1} = 2^n - 1$
$= W - 1$

(ii) The hiearchical softmax formation has one repretation v_w for each word w and one representation v'_n for every inner node n of the binary tree.

As we come up with the outcome of (i):

we have W words and $W-1$ internal nodes,

The hiearchical softmax model represent W v_W and $(W-1)$ v'_W.

In Softmax model where it assigns ~~too~~ two representations v_W and v'_W to each words. If we have W words, we will have W v_W and W v'_W.

Thus, the hierarchical softmax model does not require representing prohibitively many more vectors than the basic softmax model.

Question E:

(i) we want to verify $\Sigma_{w=1}^{W} p(w|w_I)=1$

$$\Sigma_{w=1}^{W} p(w|w_I) =$$

$$\Sigma_{w=1}^{W} \prod_{j=1}^{L(w)-1} \sigma\left([[n(w,j+1)=ch(n(w,j))]] \cdot v'^T_{n(w,j)} v_{w_I}\right)$$

which as we have $[[x]] = \begin{cases} 1 & x \text{ is true} \\ -1 & x \text{ is not true} \end{cases}$

we can divide whether $\underbrace{n(w,j+1)=ch(n(w,j))}_{A}$ is true or not, then $[[A]]$

$$= \begin{cases} 1 & A \text{ is true} \\ -1 & A \text{ is not true.} \end{cases}$$

then $\sigma\left([[A]] \cdot v'^T_{n(w,j)} v_{w_I}\right)$ can be divided to 2 conditions:

when A is true:

$$\sigma (v'^{T}_{n(w,j)} \, v_{w_I})$$

when A is not true:

$$\sigma (- v'^{T}_{n(w,j)} \, v_{w_I})$$

~~ok~~ we already know $\sigma(x) + \sigma(-x) = 1$

$$\sum_{w=1}^{W} p(w \mid w_I) = \sum_{w=1}^{W} \prod_{j=1}^{L(w)-1} \sigma \left([[n(w,j+1) = ch(n(w,j))]] \cdot v'^{T}_{n(w,j)} v_{w_I} \right)$$

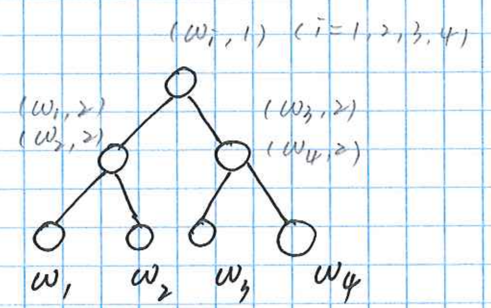
$$= \sum_{w=1}^{W} \left[\sigma (v'^{T}_{n(w_I,1)} v_{w_I}) + \sigma (- v'^{T}_{n(w_I,1)} v_{w_I}) \right]$$

$$= \sigma (v'_{n(w_I,1)}) + \sigma (- v'_{n(w_I,1)})$$

$$= 1$$

Therefore, the hierarchical softmax defines a distribution.

(ii)



set $ch(\cdot)$ = left child, we have:

$$p(w_3 \mid w_1) = \sigma(\tfrac{1}{2}) \cdot \sigma(-\tfrac{1}{3})$$

$$= 0.2199$$

Question F: if the feature vectors cannot be initialized to all zeros. we can't use the gradient descent methods as the "gradient" of initial state are all the same and cannot change to descending trend.

Question 6:

we have $P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$

we want to prove if we know $f(w_i) > f(w_j)$,

$\mathbb{E}[f(w_i)] \geq \mathbb{E}[f(w_j)]$

we know:

$$\mathbb{E}[x] = \int x P(x) dx$$

therefore:

$$\mathbb{E}[f(w_i)] = \int f(w_i) P(w_i) dw_i$$
$$= \int f(w_i) \left(1 - \sqrt{\frac{t}{f(w_i)}}\right) dw_i$$

we know, $f(w_i) > f(w_j)$

then,

$$1 - \sqrt{\frac{t}{f(w_i)}} > 1 - \sqrt{\frac{t}{f(w_j)}}$$

then,

$$f(w_i)\left(1 - \sqrt{\frac{t}{f(w_i)}}\right) > f(w_j)\left(1 - \sqrt{\frac{t}{f(w_j)}}\right)$$

therefore, we have

$\mathbb{E}(f(w_i)) > \mathbb{E}(f(w_j))$ ~~mean~~ maintains, the rank of frequencies is preserved by this subsampling policy.