A Related Work

There have been extensive researches on recommendation (Aggarwal, 2016). Besides the basic user-based and item-based collaborative filtering (Deshpande & Karypis, 2004), the full rank linear autoencoder approaches include SLIM (Ning & Karypis, 2011), HOLISM (Christakopoulou & Karypis, 2014), EASE (Steck, 2019), DLAE (Denoising linear autoencoder) Steck (2020), whereas low-rank approaches include (Kabbur et al., 2013; Sedhain et al., 2016; Steck, 2020). All the customized recommendation has been enforcing zero diagonal constraints for generalization purpose, whereas we show an approximate closed-form solution for a two-term Tikhonov regularization without the zero diagonal constraint can be as effective as these models.

Matrix factorization has been been widely studied in practice, partially due to Netflix competition Koren et al. (2009). Methods like SVD++ Koren (2008) and implicit Alternating Least Square (ALS) method Hu et al. (2008) (also weighted matrix factorization) have been very influential. (Jin et al., 2021) shows the relationship between linear autoencoders and matrix factorization, and pointed out a potential advantage of linear autoencoders. In this work, we take a step further to reveal a deeper relationship between Tikhonov regularized linear autoencoders and a few other regularizations including matrix factorization, and show the potential limitation of the class of regularization. We also utilize the linear variational autoencoders (LVAE) to study how the deep VAE based recommendation approaches (Li & She, 2017; Liang et al., 2018; Shenbin et al., 2020) relate to linear autoencoders and matrix factorization.

Outside recommendation, there have been a few recent studies on regularization landscapes of linear (variational) autoencoders (Kunin et al., 2019; Bao et al., 2020; Lucas et al., 2019a). They do not provide the general weighted ℓ_2 regularization and thus did not find the inherent limitation on the regularization (for MF). Our LVAE inspired regularization is also never studied before.

Nuclear norm regularizers can recover low-rank matrices in the vector regression setting (Negahban & Wainwright, 2011). Its weighted generalization can be applied in the area of image processing (Gu et al., 2014). Because weighted nuclear-norm is usually not convex or differentiable, finding optimal solutions is difficult except for a few special cases (Chen et al., 2013).

B Proofs

B.1 Proof of Lemma 1

The Linear Variational AutoEncoder (LVAE) is defined in the same way as (Lucas et al., 2019b):

$$\begin{aligned} p(x \mid z) &= \mathcal{N}\left(Wz + \boldsymbol{\mu}, \sigma^2 I\right) \\ q(z \mid x) &= \mathcal{N}(V(x - \boldsymbol{\mu}), D) \end{aligned} \tag{16}$$

For simplification, we set $\mu = 0$ in following context. And the ELBO of LVAE is known as:

$$\begin{aligned} &\mathcal{L}_x = -KL(q(z|x)||p(z)) + \mathbb{E}_{q(z|x)}[\log p(x|z)] \\ &KL(q(z|x)||p(z)) = -\log|D| + x^T V^T V x + tr(D) - k \\ &\mathbb{E}_{q(z|x)}[\log p(x|z)] = -\frac{1}{2\sigma^2}\Big(tr(WDW^T) + x^T V^T W^T W V x - 2x^T W V x + x^T x\Big) - \frac{n}{2}\log 2\pi\sigma^2 \end{aligned} \tag{17}$$

Again, the (maximizing) ELBO can be written as:

$$\begin{aligned} \mathcal{L}_x &= -\frac{1}{2}\Big(-\log|D| + x^T V^T V x + tr(D) - k\Big) - \frac{n}{2}\log 2\pi\sigma^2 \\ &\quad - \frac{1}{2\sigma^2}\Big(tr(WDW^T) + x^T V^T W^T W V x - 2x^T W V x + x^T x\Big) \\ &= -\frac{1}{2}||Vx||_2^2 - \frac{1}{2\sigma^2}\Big(||W\sqrt{D}||_F^2 + ||x - WVx||_2^2\Big) + f(D,\sigma) \end{aligned} \tag{18}$$

Table 3: Investigating the closed/analytic solutions of linear models. $dMat(\cdot)$ denotes a diagonal matrix, $diag(X)$ is the vector on the diagonal of X.

Model		regularization	solution
Frobenius norm	1. EASE(full rank) (Steck, 2019)	$\min_W \|\|X - XW\|\|_F^2 + \lambda \cdot \|\|W\|\|_F^2$ $s.t. \quad diag(W) = 0$	$C = (X^T X + \lambda I)^{-1}$ $W = I - C \cdot dMat(diag(1 \oslash C))$
	2. DLAE(full rank) (Steck, 2020)	$\min_W \|\|X - XW\|\|_F^2 + \|\|\Lambda^{1/2} \cdot W\|\|_F^2$ $\Lambda = \frac{p}{1-p} dMat(diag(X^T X))$	$W = (X^T X + \Lambda)^{-1} X^T X$
	3. EDLAE(full rank) (Steck, 2020)	$\min_W \|\|X - XW\|\|_F^2 + \|\|\Lambda^{1/2} \cdot W\|\|_F^2$ $\Lambda = \frac{p}{1-p} dMat(diag(X^T X))$ $s.t. \quad diag(W) = 0$	$C = (X^T X + \Lambda)^{-1}$ $W = I - C \cdot dMat(diag(1 \oslash C))$
	4. EDLAE-ADMM (Steck, 2020)	$\min_{A,B} \|\|X - XAB^T\|\|_F^2 + \|\|\Lambda^{1/2} \cdot AB^T\|\|_F^2$ $s.t. \quad diag(W) = 0$	ADMM update A, B
	5. LRR (Jin et al., 2021)	$\min_{rank(W) \le k} \|\|X - XW\|\|_F^2 + \|\|\Gamma W\|\|_F^2$	$\overline{Y}^* = \overline{X} W^* \overset{\text{SVD}}{=} U \Sigma V$ $\widehat{W} = (X^T X + \Gamma^T \Gamma)^{-1} X^T X (V_k V_k^T)$
	6. LR-DLAE(this paper)	$\min_{rank(W) \le k} \|\|X - XW\|\|_F^2 + \|\|\Lambda^{1/2} \cdot W\|\|_F^2$ $\Lambda = \frac{p}{1-p} dMat(diag(X^T X))$	$W^* = (X^T X + \Lambda)^{-1} X^T X$ $\overline{Y}^* = \overline{X} W^* \overset{\text{SVD}}{=} U \Sigma V^T$ $\widehat{W} = W^* (V_k V_k^T)$
	7. LR-EDLAE-1(this paper)	$\min_{rank(W) \le k} \|\|X - XW\|\|_F^2 + \|\|\Lambda^{1/2} \cdot W\|\|_F^2$ $\Lambda = \frac{p}{1-p} dMat(diag(X^T X))$ $s.t. \quad diag(W) = 0$	$C = (X^T X + \Lambda)^{-1}$ $W^* = I - C \cdot dMat(diag(1 \oslash C))$ $\overline{Y}^* = \overline{X} W^* \overset{\text{SVD}}{=} U \Sigma V^T$ $\widehat{W} = W^* (V_k V_k^T)$
	8. LR-EDLAE-2(this paper)	$\min_{rank(W) \le k} \|\|X - XW\|\|_F^2 + \|\|\Lambda^{1/2} \cdot W\|\|_F^2$ $\Lambda = \frac{p}{1-p} dMat(diag(X^T X))$ s.t. $\quad diag(W) = 0$	$C = (X^T X + \Lambda)^{-1}$ $W^* = I - C \cdot dMat(diag(1 \oslash C))$ $W^* \overset{\text{SVD}}{=} U \Sigma V^T$ $\widehat{W} = U_k \Sigma_k V_k^T$
Nuclear Norm	9. Regularized PCA (Zheng et al., 2018)	$\min_{P,Q} \|\|X - PQ^T\|\|_F^2 + \lambda \cdot (\|\|P\|\|_F^2 + \|\|Q\|\|_F^2)$ $X \overset{\text{SVD}}{=} U \Sigma V^T$	$P^* = U_k$ $Q^* = V_k \Omega$ $\Omega = \sqrt{(\sigma_i - \lambda)_+}$
	10. MF dropout (Cavazza et al., 2018)	$\min_{P,Q,d} \|\|X - PQ^T\|\|_F^2 + d\frac{1-p}{p} \cdot \sum_{k=1}^{d} \|\|P_k\|\|_2^2 \cdot \|\|Q_k\|\|_2^2$ $\min_Y \|\|X - Y\|\|_F^2 + \frac{1-p}{p} \|\|Y\|\|_*^2$	$X \overset{\text{SVD}}{=} U \Sigma V^T$ $Y^* = P^* \cdot (Q^*)^T$ $= U \cdot S_\mu(\Sigma) \cdot V^T$
	11. LAE (Bao et al., 2020)	$\min_{W_1, W_2} \|X - X W_1 W_2\|_F^2 + \|W_1 \Lambda^{\frac{1}{2}}\|_F^2 + \|\Lambda^{\frac{1}{2}} W_2\|_F^2,$	$W_1^* = P(I - \Lambda S^{-2})^{\frac{1}{2}} U^T$ $W_2^* = U(I - \Lambda S^{-2})^{\frac{1}{2}} P^T$
	12. LVAE(this paper)	$\min_{P,Q} \|\|X - PQ\|\|_F^2 + \|\|\Lambda^{1/2} Q\|\|_F^2 + \|\|P \Lambda^{1/2}\|\|_F^2$ $\min_{A,B} \|\|X - XAB\|\|_F^2 + \|\|\Lambda B\|\|_F^2 + \|\|XA\|\|_F^2$ $\min_{rank(W) \le k} \|\|X - W\|\|_F^2 + 2\|\|W\|\|_{w,*}$	$X \overset{\text{SVD}}{=} U \Sigma V^T$ $P^* = U_k \cdot diag(\sqrt{\sigma_1 - \lambda_{(k)}}, \ldots, \sqrt{\sigma_1 - \lambda_{(1)}}) \cdot \Omega$ $Q^* = \Omega^T \cdot diag(\sqrt{\sigma_1 - \lambda_{(k)}}, \ldots, \sqrt{\sigma_1 - \lambda_{(1)}}) \cdot V_k^T$ $A^* = X^\dagger P^* \Lambda^{\frac{1}{2}} \quad B^* = \Lambda^{-\frac{1}{2}} Q^*$

where $f(D, \sigma) = \frac{1}{2} \log |D| - \frac{1}{2} tr(D) + \frac{k}{2} - \frac{n}{2} \log 2\pi\sigma^2$, $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^k$.

For whole data, it is equivalent to minimize:

$$\begin{aligned} \mathcal{L} &= ||X - WVX||_F^2 + N||W\sqrt{D}||_F^2 + \sigma^2 ||VX||_F^2 + g(D, \sigma) \\ &= ||X^T - X^T V^T W^T||_F^2 + N||\sqrt{D} W^T||_F^2 + \sigma^2 ||X^T V^T||_F^2 + g(D, \sigma) \end{aligned} \tag{19}$$

where $g(D, \sigma) = -\sigma^2 N\big(\log |D| - tr(D) + k - n \log 2\pi\sigma^2\big)$.

B.2 Proof of Proposition 1

Note that when $\sigma_i \le \lambda_{(k-i)}$, the new singular value shrinks to zero, and can be removed. Basically, for any $\lambda_{(1)} \ge \cdots \ge \lambda_{(k)}$, we can build the corresponding Tikhonov regularized instance by setting

$$\frac{\sigma_i^2}{\sigma_i^2 + \lambda_i'} = \frac{\sigma_i - \lambda_{(k-i)}}{\sigma_i}, \text{i.e.,} \lambda_k' = \frac{\sigma_i^3}{\sigma_i - \lambda_{(k-i)}} - \sigma_i^2. \tag{20}$$

Discussion of Proposition 1: Further, the same observation holds true for the regularization (2), and the weighted-nuclear norm regularization in when the weights are in the non-ascending order. This observation suggests a potentially limitation of the earlier regularization as they will always try to maintain the larger singular values: when a singular value is large, the shrinkage will be small.Such regularization has shown to work well in the areas such as image processing Gu et al. (2014). But it has not been studied or confirmed if it will work for the recommendation. In Section 5, we report our experimental study which shows such regularization could be too restrictive for recommendation.

B.3 Proof of Proposition 2

By slightly abusing the notation, we shall let $OPT1$ ($OPT2$) be the value of the optimal solution for $OPT1$ ($OPT2$). We need to show that $OPT1 = OPT2$. We need two directions.

$OPT2 \geq OPT1$: Let W^* be an optimal solution for $OPT2$. Let the SVD of W^* be $U^*\Sigma^*(V^*)^{\mathrm{T}}$. Recall that W^* needs to satisfy the rank constraint $\mathrm{rank}(W^*) \leq k$ so $U^* \in \mathbf{R}^{m\times k}$, $\Sigma^* \in \mathbf{R}^{k\times k}$, and $V^* \in \mathbf{R}^{n\times k}$. Let π be a permutation on $[k]$ such that $\lambda_{\pi(1)} \leq \lambda_{\pi(2)} \leq \cdots \leq \lambda_{\pi(k)}$. Let also Ω be the corresponding permutation matrix. Specifically, $\Omega \in \{0,1\}^{k\times k}$ and there is exactly one entry in each row of Ω is 1:

$$\Omega_{i,j} = \begin{cases} 1 & \text{if } j = \pi(i). \\ 0 & \text{otherwise.} \end{cases}$$

For example, consider a case in which $\lambda_1 > \lambda_2 > \cdots > \lambda_k$. Then we set $\pi = (k, k-1, \ldots, 1)$, and correspondingly,

$$\Omega = \begin{pmatrix} 0 & \ldots & 0 & 1 \\ 0 & \ldots & 1 & 0 \\ & \ldots & & \\ 1 & \ldots & 0 & 0 \end{pmatrix}.$$

Next, let $P = U^*(\Sigma^*)^{\frac{1}{2}}\Omega$ and $Q = \Omega^{\mathrm{T}}(\Sigma^*)^{\frac{1}{2}}(V^*)^{\mathrm{T}}$. We have $W^* = PQ$ and $f(W^*) = f(PQ)$. In addition,

$$\|\Lambda^{\frac{1}{2}}Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2 = 2\|\Lambda^{\frac{1}{2}}\Omega(\Sigma^*)^{\frac{1}{2}}\|_F^2 = 2\sum_{i\leq k}\lambda_{\pi(i)}\sigma_i = 2\|W^*\|_{\omega,*},$$

where σ_i is the i-th largest singular value of W^*. In other words, we have found a (P, Q) pair such that

$$f(PQ) + \|\Lambda^{\frac{1}{2}}Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2 = f(W^*) + 2\|W^*\|_{\omega,*} = OPT2,$$

which shows that $OPT1 \leq OPT2$.

$OPT2 \leq OPT1$. Let P^* and Q^* be an optimal solution for $OPT1$. Let the singular values of P^* be $\sigma_1(P^*) \geq \sigma_2(P^*) \geq \cdots \geq \sigma_k(P^*)$ and those of Q^* be $\sigma_1(Q^*) \geq \sigma_2(Q^*) \geq \cdots \geq \sigma_k(Q^*)$. Let also $\sigma_1^* \geq \cdots \geq \sigma_k^*$ be the singular values of P^*Q^*.

We shall find a lower bound of $\|\Lambda^{\frac{1}{2}}Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2$ expressed in terms of σ_i^*'s. In fact, we shall show that

$$\|\Lambda^{\frac{1}{2}}Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2 \geq 2\|P^*Q^*\|_{\omega,*}. \tag{21}$$

One can see that if Eq. 21 were true, we have

$$OPT2 \leq f(P^*Q^*) + 2\|P^*Q^*\|_{\omega,*} \leq f(P^*Q^*) + \|\Lambda^{\frac{1}{2}}Q^*\|_F^2 + \|P^*\Lambda^{\frac{1}{2}}\|_F^2 = OPT1.$$

Thus, it remains to prove Eq. 21. Let $\lambda_{(1)} \geq \lambda_{(2)} \geq \cdots \geq \lambda_{(k)}$ be a sorted sequence of λ_i's i.e., $\lambda_{(k)} = \lambda_{\pi(1)}, \lambda_{(k-1)} = \lambda_{\pi(2)}, \ldots, \lambda_{(1)} = \lambda_{\pi(k)}$.

First, we show that $\|P\Lambda^{\frac{1}{2}}\|_F^2 \geq \sum_{i=1}^k \lambda_{(k-i+1)} \times \sigma_i^2(P^*)$ and $\|\Lambda^{\frac{1}{2}}Q\|_F^2 \geq \sum_{i=1}^k \lambda_{(k-i+1)} \times \sigma_i^2(Q^*)$. We need the following Lemma (see e.g., Theorem 2 in Yue (2020)):

Lemma 3. *Let A and B be two positive definite matrices in $\mathbf{R}^{k\times k}$. Then it holds that*

$$\sum_{i=1}^{k}\sigma_i(A)\sigma_{k-i+1}(B)\le \mathrm{tr}(B^{\frac{1}{2}}AB^{\frac{1}{2}}). \tag{22}$$

Let the SVD of P^* be $U_{P^*}\Sigma_{P^*}V_{P^*}^{\mathrm{T}}$ and that of Q^* be $U_{Q^*}\Sigma_{Q^*}V_{Q^*}^{\mathrm{T}}$. We have

$$\|P^*\Lambda^{\frac{1}{2}}\|_F^2=\|U_{P^*}\Sigma_{P^*}V_{P^*}^{\mathrm{T}}\Lambda^{\frac{1}{2}}\|_F^2=\|\Sigma_{P^*}V_{P^*}^{\mathrm{T}}\Lambda^{\frac{1}{2}}\|_F^2=\mathrm{tr}(\Sigma_{P^*}V_{P^*}^{\mathrm{T}}\Lambda V_{P^*}\Sigma_{P^*}). \tag{23}$$

We now apply Lemma 3 by setting $A=V_{P^*}^{\mathrm{T}}\Lambda V_{P^*}$ and $B=\Sigma_{P^*}^2$, and obtain that

$$\|P^*\Lambda^{\frac{1}{2}}\|_F^2=\mathrm{tr}(\Sigma_{P^*}V_{P^*}^{\mathrm{T}}\Lambda V_{P^*}\Sigma_{P^*})\ge\sum_{i=1}^{k}\lambda_{(k+1-i)}\times\sigma_i^2(P^*). \tag{24}$$

We may similarly prove that $\|\Lambda^{\frac{1}{2}}Q^*\|_F^2\ge\sum_{i=1}^{k}\lambda_{(k-i+1)}\times\sigma_i^2(Q^*)$. Therefore,

$$\|\Lambda^{\frac{1}{2}}Q^*\|_F^2+\|P^*\Lambda^{\frac{1}{2}}\|_F^2\ge\sum_{i=1}^{k}\lambda_{(k+1-i)}\times(\sigma_i^2(P^*)+\sigma_i^2(Q^*)) \tag{25}$$

(25) provides a lower bound of $\|\Lambda^{\frac{1}{2}}Q^*\|_F^2+\|P^*\Lambda^{\frac{1}{2}}\|_F^2$ in terms of $\sigma_i(P^*)$ and $\sigma_i(Q^*)$. We next aim to express the lower bound in terms of σ_i^*'s (singular values of P^*Q^*) directly.

The following program gives a lower bound for $\|\Lambda^{\frac{1}{2}}Q^*\|_F^2+\|P^*\Lambda^{\frac{1}{2}}\|_F^2$:

$$\begin{aligned}\min:\quad & \|\Lambda^{\frac{1}{2}}Q^*\|_F^2+\|P^*\Lambda^{\frac{1}{2}}\|_F^2\\ \text{subject to}\quad & W=P^*Q^*\\ & \sigma_i(W)=\sigma_i^*\quad\text{for } i\le k.\end{aligned} \tag{26}$$

Write the SVD of W be $U_W\Sigma_WV_W^{\mathrm{T}}$. Also, let $\tilde{P}=U_W^{\mathrm{T}}P^*$ and $\tilde{Q}=Q^*V_W$. Noting that the columns in P^* are in the column space of W and the rows in Q^* are in the row space of W, we have *(i)* $\sigma_i(P^*)=\sigma_i(\tilde{P})$ and $\sigma_i(Q^*)=\sigma_i(\tilde{Q})$ for $i\le k$, and *(ii)* $\|\Lambda^{\frac{1}{2}}Q^*\|_F^2+\|P^*\Lambda^{\frac{1}{2}}\|_F^2=\|\Lambda^{\frac{1}{2}}\tilde{Q}\|_F^2+\|\tilde{P}\Lambda^{\frac{1}{2}}\|_F^2$.

Therefore, (26) can be equivalently written as

$$\begin{aligned}\min:\quad & \|\Lambda^{\frac{1}{2}}\tilde{Q}\|_F^2+\|\tilde{P}\Lambda^{\frac{1}{2}}\|_F^2\\ \text{subject to}\quad & \Sigma_W=\tilde{P}\tilde{Q}\\ & (\Sigma_W)_{i,i}=\sigma_i^*\quad\text{for } i\le k.\end{aligned} \tag{27}$$

Now $\tilde{P}\tilde{Q}$ is positive definite. Using a similar technique developed in (Bao et al., 2020) (Theorem 1), one can see that $\tilde{P}=\tilde{Q}^{\mathrm{T}}$. See also Lemma 4. This implies that $\tilde{P}=\Sigma_W^{\frac{1}{2}}\Omega$ for some unitary matrix Ω and $\sigma_i(P^*)=\sigma_i(\tilde{P})=\sigma_i(Q^*)=\sigma_i(\tilde{Q})=\sqrt{\sigma_i^*}$ for $i\le k$. Together with (25), we have

$$\|\Lambda^{\frac{1}{2}}Q^*\|_F^2+\|P^*\Lambda^{\frac{1}{2}}\|_F^2\ge 2\sum_{i\le k}\lambda_{(k+1-i)}\times\sigma_i^2(P^*)=2\sum_{i\le k}\lambda_{(k+1-i)}\times\sigma_i^*=2\|P^*Q^*\|_{\omega,*}.$$

B.4 Proof of Corollary 1

We first find an optimal solution for (9). Let the SVD of X be $X=U_X\Sigma_XV_X^{\mathrm{T}}$, where $U_X\in\mathbf{R}^{m\times n}$, $\Sigma_X\in\mathbf{R}^{n\times n}$, and $V_X\in\mathbf{R}^{n\times n}$. Let $\bar{U}_X$ be an arbitrary basis for the subspace that is orthogonal to X's column space so $\bar{U}_X\in\mathbf{R}^{m\times(m-n)}$ and $[U_X,\bar{U}_X]$ form a basis for $\mathbf{R}^m$. We have

$$\begin{aligned}&\|X-PQ\|_F^2+\|P\Lambda^{\frac{1}{2}}\|_F^2+\|\Lambda^{\frac{1}{2}}Q\|_F^2\\ =&\left\|\begin{pmatrix}U_X^{\mathrm{T}}\\ \bar{U}_X^{\mathrm{T}}\end{pmatrix}XV_X-\begin{pmatrix}U_X^{\mathrm{T}}\\ \bar{U}_X^{\mathrm{T}}\end{pmatrix}PQV_X\right\|_F^2+\left\|\begin{pmatrix}U_X^{\mathrm{T}}\\ \bar{U}_X^{\mathrm{T}}\end{pmatrix}P\Lambda^{\frac{1}{2}}\right\|_F^2+\|\Lambda^{\frac{1}{2}}QV_X\|_F^2.\end{aligned}$$

Let $\tilde{P} = \begin{pmatrix} U_X^{\mathrm{T}} \\ \bar{U}_X^{\mathrm{T}} \end{pmatrix} P$ and $\tilde{Q} = QV_X$. Then our objective becomes

$$\min_{\tilde{P},\tilde{Q}} \left\| \begin{pmatrix} \Sigma_X \\ 0_{(m-n)\times n} \end{pmatrix} - \tilde{P}\tilde{Q} \right\|_F^2 + \|\tilde{P}\Lambda^{\frac{1}{2}}\|_F^2 + \|\Lambda^{\frac{1}{2}}\tilde{Q}\|_F^2. \tag{28}$$

Let $\tilde{W} = \tilde{P}\tilde{Q}$ and the singular values of $\tilde{W}$ be $\sigma_1^* \geq \sigma_2^* \geq \cdots \geq \sigma_k^*$. Let also $\tilde{\Sigma} = \begin{pmatrix} \Sigma_X \\ 0_{(m-n)\times n} \end{pmatrix}$.

Recall also that σ_i is the i-th largest singular value of X. We next show that

$$\left\| \begin{pmatrix} \Sigma_X \\ 0 \end{pmatrix} - \tilde{P}\tilde{Q} \right\|_F^2 = \|\tilde{\Sigma} - \tilde{P}\tilde{Q}\|_F^2 \geq \sum_{i=1}^{k} (\sigma_i - \sigma_i^*)^2 + \sum_{i=k+1}^{n} \sigma_i^2.$$

Note first that

$$\|\tilde{\Sigma} - \tilde{W}\|_F^2 = \|\tilde{\Sigma}\|_F^2 + \|\tilde{W}\|_F^2 - 2\langle \tilde{\Sigma}, \tilde{W} \rangle. \tag{29}$$

Next, we have ((Zheng et al., 2018)):

$$|\langle \tilde{\Sigma}, \tilde{W} \rangle| = |\mathrm{tr}(\tilde{\Sigma}\tilde{W}^{\mathrm{T}})\| \leq |\mathrm{tr}(\tilde{\Sigma}\Sigma_{\tilde{W}})| = \sum_{i=1}^{k} \sigma_i \sigma_i^*.$$

Therefore, $\langle \tilde{\Sigma}, \tilde{W} \rangle$ is maximized when

$$\tilde{W}_{i,j} = \begin{cases} \sigma_i^* & \text{if } i = j \leq k \\ 0 & \text{Otherwise.} \end{cases}$$

When we plug in this optimized $\tilde{W}$ to Eq. 29, we get

$$\|\tilde{\Sigma} - \tilde{W}\|_F^2 \geq \sum_{i=1}^{k} (\sigma_i - \sigma_i^*)^2 + \sum_{i=k+1}^{n} \sigma_i^2.$$

Next, from Proposition 2, we have

$$\|\tilde{P}V^{\frac{1}{2}}\|_F^2 + \|\Lambda^{\frac{1}{2}}\tilde{Q}\|_F^2 \geq \sum_{i=1}^{k} \lambda_{(k-i+1)}\sigma_i^*.$$

Therefore, we can find a lower bound for Eq. 5 in terms of σ_i^*'s:

$$\mathcal{L}(\sigma_1^*, \ldots, \sigma_k^*) = \sum_{i=1}^{k} (\sigma_i - \sigma_i^*)^2 + 2\sum_{i=1}^{k} \lambda_{(k-i+1)}\sigma_i^* + \sum_{i=k+1}^{m} \sigma_i^2 \quad (\sigma_1^* \geq \cdots \geq \sigma_k^* \geq 0). \tag{30}$$

We next find a minimal value of $\mathcal{L}$ (by treating σ_i^*'s as decision variables). This will give us a lower bound (and is independent of σ_i^*) on our optimization problem. We then show that this lower bound can be achieved by carefully constructing $\tilde{W}$ (as well as $\tilde{P}$ and $\tilde{Q}$). This means such $\tilde{W}$ is optimal.

Specifically, we need to find an optimal solution for the following program:

$$\begin{aligned} \text{minimize}_{\sigma_1^*,\ldots,\sigma_k^*} \quad & \mathcal{L}(\sigma_1^*, \ldots, \sigma_k^*) \\ \text{subject to:} \quad & \sigma_i^* \geq 0 \\ & \sigma_1^* \leq \sigma_2^* \leq \cdots \leq \sigma_k^* \quad \text{(Ordering constraint)} \end{aligned} \tag{31}$$

We shall first find an optimal solution for

$$\begin{aligned} \text{minimize}_{\sigma_1^*,\ldots,\sigma_k^*} \quad & \mathcal{L}(\sigma_1^*,\ldots,\sigma_k^*) \\ \text{subject to:} \quad & \sigma_i^* \geq 0 \end{aligned} \tag{32}$$

Note here, the ordering constraint is removed so the optimal value for (32) should be no more than that for (31). We shall see that the optimal solution for (31) also satisfies the ordering constraint so indeed optimal solutions for (31) and (32) are the same.

The problem (32) boils down to finding

$$\min_{\sigma_i^* \geq 0} (\sigma_i - \sigma_i^*)^2 + 2\sum_{i=1}^{k} \lambda_{(k-i+1)}\sigma_i^*.$$

We note that σ_i^*'s do not interact with each other so we can optimize each σ_i^*'s independently. We get

$$\sigma_i^* = (\sigma_i - \lambda_{(k-i+1)})^+.$$

We can check that $\sigma_1^* \geq \cdots \geq \sigma_k^*$. Therefore, the optimal value for (31) is

$$\sum_{i=1}^{k} (\sigma_i - (\sigma_i - \lambda_{(k-i+1)})^+)^2 + 2\sum_{i=1}^{k} \lambda_{(k-i+1)}(\sigma_i - \lambda_{(k-i+1)})^+ + \sum_{i=k+1}^{n} \sigma_i^2.$$

This is also a lower bound for (9). One can check that when we set P and Q as

$$\begin{aligned} P^* &= U_k diag(\sqrt{(\sigma_1 - \lambda_{(k)})^+}, \ldots, \sqrt{(\sigma_k - \lambda_{(1)})^+})\Omega, \\ Q^* &= \Omega^{\mathrm{T}} diag(\sqrt{(\sigma_1 - \lambda_{(k)})^+}, \ldots, \sqrt{(\sigma_k - \lambda_{(1)})^+})V_k^{\mathrm{T}}, \end{aligned} \tag{33}$$

the lower bound is achieved so (33) gives an optimal solution. Here, U_k and V_k are leading left and right singular vectors of X.

Now we move to analyze (10). Our goal is to reduce (10) to (9). Let

$$P = XA\Lambda^{-\frac{1}{2}} \qquad Q = \Lambda^{\frac{1}{2}}B.$$

Then (10) becomes

$$\begin{aligned} \text{minimize}_{P,Q} \quad & \|X - PQ\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2 + \|\Lambda^{\frac{1}{2}}Q\|_F^2 \\ \text{subject to} \quad & P = XA\Lambda^{-\frac{1}{2}} \quad \text{(Constraint P)} \\ & Q = \Lambda^{\frac{1}{2}}B \quad \text{(Constraint Q)}. \end{aligned} \tag{34}$$

Here, X and Λ are given, whereas P, Q, A, and B are decision variables. The (Constraint P) says that each column of P needs to be in a column space of X (it is a necessary and sufficient condition for A to exist). The (Constraint Q) simply says Q and B are linearly related and does not have tangible impact to the optimization problem.

But we note that when we put aside the constraints, an optimal (P, Q) is specified by (33). The columns of the optimal P indeed is in the column space of X. So (P, Q) is also an optimal solution for (34). We may find the corresponding A and B:

$$A^* = X^\dagger P^* \Lambda^{\frac{1}{2}} \quad \text{ and } B^* = \Lambda^{-\frac{1}{2}} Q^*,$$

B.5 SYMMETRIC LEMMA

Lemma 4. *Let $\tilde{P}, \tilde{Q} \in \mathbf{R}^{k\times k}$ be full rank, Λ be a diagonal matrix, and Σ_W be a diagonal matrix so that $(\Sigma_W)_{i,i} = \sigma_i^*$, where σ_i^*'s are sorted in descending order. Consider the optimization problem:*

$$\begin{aligned} \min: \quad & \|\Lambda^{\frac{1}{2}}\tilde{Q}\|_F^2 + \|\tilde{P}\Lambda^{\frac{1}{2}}\|_F^2 \\ \textit{subject to} \quad & \Sigma_W = \tilde{P}\tilde{Q} \\ & (\Sigma_W)_{i,i} = \sigma_i^* \quad \textit{for } i \leq k. \end{aligned} \tag{35}$$

There is an optimal solution such that $\tilde{P} = \tilde{Q}^{\mathrm{T}}$

Proof. Let $\hat{P} = \tilde{P}\Lambda^{\frac{1}{2}}$ and $\hat{Q} = \Lambda^{-\frac{1}{2}}\tilde{Q}$. The program (36) is equivalent to

$$\begin{aligned} \min: \quad & \|\Lambda\hat{Q}\|_F^2 + \|\hat{P}\|_F^2 \\ \text{subject to} \quad & \Sigma_W = \hat{P}\hat{Q} \\ & (\Sigma_W)_{i,i} = \sigma_i^* \quad \text{for } i \le k. \end{aligned} \tag{36}$$

Let the SVD of $\hat{Q}$ be $U_{\hat{Q}}\Sigma_{\hat{Q}}V_{\hat{Q}}^{\mathrm{T}}$ so $\hat{Q}^{-1} = V_{\hat{Q}}\Sigma_{\hat{Q}}^{-1}U_{\hat{Q}}^{\mathrm{T}}$. We can also see that $\hat{P} = \Sigma_W\hat{Q}^{-1}$. Therefore, the objective term becomes

$$\|\Lambda U_{\hat{Q}}\Sigma_{\hat{Q}}V_{\hat{Q}}^{\mathrm{T}}\|_F^2 + \|\Sigma_W V_{\hat{Q}}\Sigma_{\hat{Q}}^{-1}U_{\hat{Q}}^{\mathrm{T}}\|_F^2 = \|\Lambda U_{\hat{Q}}\Sigma_{\hat{Q}}\|_F^2 + \|\Sigma_W V_{\hat{Q}}\Sigma_{\hat{Q}}^{-1}\|_F^2.$$

Let us consider the stationary points $U_{\hat{Q}}$ and $V_{\hat{Q}}$ when $\Sigma_{\hat{Q}}$ is fixed. We can see that they need to be permutation matrices to minimize both terms in the objective (using the rearrangement inequality again). Therefore, we can see $\hat{Q} = \Sigma_1\Sigma_{\hat{Q}}\Sigma_2$ for two permutation matrices Σ_1 and Σ_2. This implies that $\tilde{Q} = \Lambda^{\frac{1}{2}}\Sigma_1\Sigma_{\hat{Q}}\Sigma_2$, i.e., each row (column) of $\tilde{Q}$ has exactly one non-zero entry. We may similarly show that each row (column) of $\tilde{P}$ has exactly one non-zero entry. In addition, the locations of non-zero entries of $\tilde{P}$ and $\tilde{Q}^{\mathrm{T}}$ are identical because $\tilde{P}\tilde{Q}$ is a diagonal matrix. We may thus write

$$\tilde{P} = \Sigma_{(1)}\Sigma_{\tilde{P}}\Sigma_{(2)} \quad \tilde{Q} = \Sigma_{(2)}^{\mathrm{T}}\Sigma_{(\tilde{Q})}\Sigma_{(1)}^{\mathrm{T}},$$

where $(\Sigma_{\tilde{P}})_{i,i} = \sigma_i(\tilde{P})$ and $(\Sigma_{(\tilde{Q})})_{i,i} = \sigma_{\tau(i)}(\tilde{Q})$, where τ is a permutation on $[k]$. The set of (possibly unsorted) singular values for $\tilde{P}\tilde{Q}$ thus is $\sigma_i(\tilde{P})\sigma_{\tau(i)}(\tilde{Q})$. Thus, we can see that there exists a permutation $\bar{\pi}$ such that

$$\begin{aligned} \|\Lambda^{\frac{1}{2}}\tilde{Q}\|_F^2 + \|\tilde{P}\Lambda^{\frac{1}{2}}\|_F^2 &= \sum_{i\le k}\left(\sigma_i^2(P^*)\lambda_{\bar{\pi}(i)} + \sigma_{\tau(i)}^2(Q^*)\lambda_{\bar{\pi}(i)}\right) \\ &\ge \sum_{i\le k} 2\sigma_i(P^*)\sigma_{\tau(i)}^*(Q)\lambda_{\bar{\pi}(i)} \\ &\ge 2\|P^*Q^*\|_{\omega,*}. \end{aligned}$$

One can see that we can set $\tilde{P} = \tilde{Q}^{\mathrm{T}}$ to make all inequality becomes equality so there is an optimal solution such that $\tilde{P} = \tilde{Q}^{\mathrm{T}}$. □

C Experimental Details

C.1 Anonymous Code

`https://anonymous.4open.science/r/ICLR-2022-Anonymous-Demo-Code-B9FF/README.md`

C.2 DLAE hyperparameters tuning

This section presents the hyperparameter tuning process on the validation data over three (ML-20M, Netflix, MSD) datasets for the full rank DLAE formula, which was introduced by Steck (2020) yet not investigated:

$$\begin{aligned} &\min_W ||X - XW||_F^2 + ||\Lambda^{1/2} \cdot W||_F^2 \\ &\Lambda = \frac{p}{1-p} dMat(diag(X^T X)) \\ &\hat{W} = (X^T X + \Lambda)^{-1} X^T X \end{aligned}$$

In practical, l_2 regularization is also imposed:

$$\hat{W} = (X^T X + \Lambda + \lambda)^{-1} X^T X$$

The tables 4 to 6 show the results of $nDCG@100$ over three datasets respectively. And the optimal parameters are highlighted.

Table 4: ml-20m, DLAE full rank, parameter tuning on validation dataset by $nDCG@100$

		λ					
		800	900	**1000**	1100	1200	1300
p	0.1	0.42024	0.42063	0.42073	0.42102	0.42131	0.4212
	0.2	0.43132	0.43139	0.43154	0.4314	0.43147	0.43136
	0.3	0.43203	0.43211	**0.43214**	0.43206	0.43203	0.43196
	0.4	0.43001	0.43001	0.42995	0.42996	0.42984	0.42978
	0.5	0.42754	0.42745	0.42729	0.42718	0.42715	0.42704

Table 5: netflix, DLAE full rank, parameter tuning on validation dataset by $nDCG@100$

		λ						
		800	900	1000	**1100**	1200	1300	1400
p	0.2	0.3904	0.3904	0.39027	0.39024	0.3902	0.3903	0.39018
	0.25	0.39247	0.39252	0.39248	0.39249	0.3925	0.3925	0.39256
	0.3	0.39359	0.39359	0.39366	0.39358	0.39362	0.39368	0.39369
	0.35	0.39402	0.39403	0.394	**0.39405**	0.39399	0.39403	0.39397
	0.4	0.39399	0.39393	0.39395	0.39393	0.39389	0.39388	0.39387
	0.45	0.39346	0.3935	0.39343	0.39344	0.39338	0.3933	0.39329
	0.5	0.39249	0.39241	0.39247	0.39241	0.39242	0.3923	0.39224

C.3 Matrix Factorization with Dropout Hyperparameters Tuning

Cavazza et al. (2018) shows that optimization with dropout (allowing rank optimizing) is equivalent to solving a matrix approximation problem with nuclear norm:

$$\min_{P,Q,d} ||X - PQ^T||_F^2 + d\frac{1-p}{p} \cdot \sum_{k=1}^{d} ||P_k||_2^2 \cdot ||Q_k||_2^2$$

$$\min_{Y} ||X - Y||_F^2 + \frac{1-p}{p}||Y||_*^2$$

and the solution is given by:

$$\begin{aligned} X &\overset{\text{SVD}}{=} U\Sigma V^T \\ Y^* &= P^* \cdot (Q^*)^T \\ &= U \cdot S_\mu(\Sigma) \cdot V^T \\ S_\mu(\sigma) &= \max(\sigma - \mu, 0) \\ \mu &= \frac{1-p}{p + (1-p)\bar{d}} \sum_{i=1}^{\bar{d}} \sigma_i(X) \end{aligned}$$

where $\bar{d}$ denotes the largest integer such that:

$$\sigma_{\bar{d}}(X) > \frac{1-p}{p + (1-p)\bar{d}} \sum_{i=1}^{\bar{d}} \sigma_i(X)$$

Hence, there is only one parameter p to tuning. We present the tuning process on the validation set below, see tables 7 to 9. Optimal parameters as well as induced rank $\bar{d}$ are highlighted.

Table 6: msd, DLAE full rank, parameter tuning on validation dataset by $nDCG@100$

		λ					
		10	20	**30**	40	50	60
p	0.3	0.38514	0.38515	0.38517	0.38505	0.38492	0.38474
	0.4	0.38596	0.38599	**0.38602**	0.386	0.38597	0.38592
	0.5	0.38556	0.38555	0.38553	0.38557	0.38553	0.38549
	0.6	0.38382	0.3838	0.38381	0.38374	0.38373	0.38366

Table 7: ml-20m, matrix factorization with dropout, hyper parameter tuning by $nDCG@100$ on validation dataset and its induced rank .

p	0.9	0.99	0.995	**0.996**	0.997
induced rank $\tilde{d}$	10	200	385	**467**	602
nDCG@100	0.29723	0.39369	0.40045	**0.40046**	0.39925

Table 8: netflix, matrix factorization with dropout, hyper parameter tuning by $nDCG@100$ on validation dataset and its induced rank .

p	0.9	0.99	0.996	**0.997**	0.998
induced rank $\tilde{d}$	9	209	524	**653**	883
nDCG@100	0.26026	0.35462	0.36453	**0.36495**	0.36406

Table 9: msd, matrix factorization with dropout, hyper parameter tuning by $nDCG@100$ on validation dataset and its induced rank .

p	0.99	0.999	0.9995	**0.9999**	0.99995
induced rank $\tilde{d}$	249	2054	3783	**11380**	19308
nDCG@100	0.18986	0.28532	0.307	**0.32634**	0.30995

C.4 RESOURCES

Our code are mainly implemented in Numpy 1.19, Pytorch 1.7.1 on CUDA 11.0. Our experiments are performed on nodes with two sockets, each containing a 24-core Intel(R) Xeon(R) Platinum 8268 CPU @ 2.90GHz and 4 GeForce RTX 3090 24GB memory GPU.