ON THE REGULARIZATION LANDSCAPE FOR THE LINEAR RECOMMENDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, a wide range of recommendation algorithms inspired by deep learning techniques have emerged as the performance leaders on several standard recommendation benchmarks. While these algorithms were built on different DL techniques (e.g., dropouts, autoencoder), they have similar performance and even similar cost functions. This paper studies whether the models' comparable performance are sheer coincidence, or they can be unified under a single framework. We find that all linear performance leaders effectively add only a nuclear-norm based regularizer, or a Frobenius-norm based regularizer. The former ones possess a (surprisnig) rigid structure that limits the models' predictive power but their solutions are low rank and have closed form. The latter ones are more expressive and more efficient for recommendation but their solutions are either full-rank or require executing hard-to-tune numeric procedures such as ADMM. Along this line of finding, we further propose two low-rank, closed-form solutions, derived from carefully generalizing Frobenius-norm based regularizers. The new solutions get the best of both nuclear-norm and Frobenius-norm world.

1 INTRODUCTION

Research progress on algorithms for recommendation has escalated in recent years, partially fueled by the adoption of deep learning techniques. However, recent studies have found that many new deep learning recommendation models have shown sub-par performance against the simpler linear recommendation models (Dacrema et al., 2019a; Rendle et al., 2019). Although some studies are available to analyze linear vs non-linear models (Dacrema et al., 2019a), it remains puzzling why these seemingly different techniques all result in models with similar performance or even similar cost functions. This motivates us to ask a fundamental question:

What is the barebones engine that drives the performance improvement for the recent recommendation algorithms?

Specifically, do different recommendation techniques offer different "magic", but coincidentally have similar performance, or can they be unified under a single framework, which has a potential to produce *one single algorithm* that gets *the best of all existing works*? In the latest study, Jin et al. (2021) examines the relationship between the widely used matrix factorization (MF), such as ALS (Warlop et al., 2017), and the linear autoencoders (LAE) which encompasses the recent performance leaders, such as EASE (Steck, 2019) and EDLAE (Steck, 2020). They consider two basic regularization forms (See Eq (1) and (6)) and found that the optimal (closed-form) solutions of both models recover the direction of principal components, while shrinking the corresponding singular values differently. They suggests this difference may enable LAE to utilize a larger number of latent dimensions to improve recommendation accuracy, and use this to highlight the similarity as well as difference between LAE and MF.

In this paper, we go much beyond the two basic models studied in Jin et al. (2021) to analyze a large number of recent performance leaders of (linear) recommender algorithms. We found they all can be categorized into those that implement *nuclear-norm* regularizers, and into those that implement *Frobenius-norm* regularizers. We found that the former ones possess a (surprising) rigid structure that limits the models' predictive power, and the latter ones tend to be more expressive and more effective for recommendation. In many cases, both regularizers recover the direction of principal components, while shrinking the corresponding singular values differently. Interestingly, we observe that it is not

matrix factorization or LAE that determines the shrinkage structure (as Jin et al. (2021) suggested), but instead it is the forms of regularization. Thus, this paper provides a more complete and accurate characterization on how a linear recommendation model performs under different regularizations.

To better understand the regularizations that can be transformed into the (weighted) nuclear-norm regularizer $\|W\|_*$, we first show that Variational Linear AutoEncoders (VLAE) solves a weighted nuclear-norm regularization problem, in which the weights possess a specific combinatorial structure. We also show that this technique cannot be generalized to tackle arbitrary weight sequences. Second, it has been known that using dropout techniques is equivalent to adding a squared nuclear-norm regularizer $\|W\|_*^2$, and that the solution structure is strikingly similar to the regularizer that uses $\|W\|_*$. we generalize the result to show that the solution structures for $\|W\|_*^p$ are highly similar for all $p \geq 1$. But when $p = 1, 2$, the solution and hyper-parameters possess favorable properties so hyper-parameter search becomes easier. This also partially explains why only $p = 1, 2$ have been extensively considered. Third, all nuclear-norm–based techniques possess a salient property: their estimators keep the singular vectors of the data matrix and shrink only the singular values. But this severely limits the search space and explains why models that use only nuclear-norm–based regularizers share the same performance ceiling even when hyper-parameters are extensively searched.

The (weighted) Frobenius-norm regularizers $\|\Lambda W\|_F^2$ are implemented in EASE (Steck, 2019) and EDLAE (Steck, 2020). These models produce closed form full-rank estimators; and if the zero diagonal constraint on W is enforced, their singular vectors will no longer coincide with those of the data, and can deliver (slightly) better performance. However, no closed form solutions for the low-rank estimator is known and the current approaches rely on ADMM or stochastic factorized gradients (Steck, 2020). In this paper, we propose two new low-rank, closed-form estimators that deliver comparable results to the full rank models (EASE and full-rank EDLAE) as well as the ADMM based solutions (Steck, 2020).

The new closed-form solutions for low-rank models have profound implications at both practical and conceptual fronts: First, low rank solutions are often more scalable (the full rank W can be too large to materialize) and can better serve real-world training and deployment. Second, perhaps more excitingly, these solutions get the best of nuclear (closed form and low rank) and Frobenius worlds (strong predictive power). A simple "one-liner" formula (from each solution) concisely pack all the benefits obtained by a recent long line of research and abstract out all the computation nuance (e.g., the need to tune ADMM and deal with local optimal). We believe these closed-form solutions are also powerful tools to help us compare different regularizations analytically.

2 Background and Overview

Recommendation system background. Recommendation algorithms can be categorized into explicit ones that aim to predict unseen ratings between a user and an item and implicit ones that aim to predict actions, such as user click or add-cart (Steck, 2019; Dacrema et al., 2019b; Zhang et al., 2019). We focus on the implicit problem because it is more economically relevant. Here, let n be the number of items and m be the number of users. We are given a binary matrix $X \in \{0, 1\}^{m \times n}$ that represents the interaction between users and items so far, i.e., $X_{i,j} = 1$ iff user i has purchased or made a rating on item j. Our goal is to produce a real-valued matrix $\hat{X}$, which we evaluate against future interactions using information retrieval metrics such as Top-k or nDCG. Note that while the syntax of this problem resembles matrix completion (MC) (Candès & Tao, 2010), recommendation systems and MC have different evaluation criteria so MC's results are not directly applicable here.

2.1 Nuclear-norm based regularizations

Let $X \in \mathbf{R}^{m \times n}$ be a matrix of rank at most k with k leading singular values being $\sigma_1(X) \geq \sigma_2(X) \geq \cdots \geq \sigma_k(X)$. Let $\omega = (\omega_1, \ldots, \omega_k) \in (\mathbf{R}^+)^k$. The weighted unclear norm of X with respect to ω is defined as $\|X\|_{\omega,*} = \sum_{i=1}^k \omega_i \times \sigma_i(X)$.

We can see that this is a natural generalization of the weighted nuclear norm for low-rank matrices (Gu et al., 2014). Also, despite its name, the weighted nuclear norm is neither convex nor differentiable unless ω_i's are sorted in *descending* order (Chen et al., 2013; Iglesias et al., 2020).

Nuclear-norm based regularizers perform ℓ_1-shrinkage over the estimator's singular values, which resembles performing ℓ_1-shrinkage for coefficients in a linear model in LASSO (Tibshirani, 1996). Therefore, Nuclear-norm regularizers also promote sparsity over the solution's singular values (i.e., the solution is usually low rank). We note a large fraction of recent recommendation algorithms effectively add only a nuclear-norm regularizer to MF (see also Table 3 in Appendix):

A1. Regularized PCA (Udell et al., 2016; Zheng et al., 2018) aims to solve

$$\min_{P,Q} \|X - PQ^T\|_F^2 + \lambda\|Q\|_F^2 + \lambda\|P\|_F^2. \tag{1}$$

It has been known that eq. (1) is equivalent to solving $\min_{\hat{X}} \|X - \hat{X}\|_F^2 + 2\lambda\|\hat{X}\|_*$. To solve eq. (1), one can use factored gradient descent (Bhojanapalli et al., 2016) or directly use its closed form solution (Kunin et al., 2019), which involves computation of SVD of X.

A2. Matrix Factorization via dropouts. This approach use PQ^{T} ($P \in \mathbf{R}^{m\times k}$, $Q \in \mathbf{R}^{n\times k}$) to approximate X and uses a neural net to find P and Q. A standard dropout technique is used when we train P and Q. Cavazza et al. (2018) shows that optimization with dropout is equivalent to solving $\min_{\hat{X}} \|X - \hat{X}\|_F^2 + \lambda\|\hat{X}\|_*^2$, and the closed form solution is obtained by shrinking all singular values of $\hat{X}$ by a magnitude of μ, which depends on the data X and choise of λ.

The next two approaches use techniques from (variational) auto-encoder. We show that they also effectively add variants of nuclear-norm regularizers although this may not be clear at the first glance.

A3. Linear Regression via Denoising Linear Auto-Encoder considers the following (non-uniform) weighted ℓ_2-regularization (Bao et al., 2020):

$$\|X - XW_1W_2\|_F^2 + \|W_1\Lambda^{\frac{1}{2}}\|_F^2 + \|\Lambda^{\frac{1}{2}}W_2\|_F^2, \tag{2}$$

where Λ is a diagonal matrix. Bao et al. (2020) has shown a closed-form for eq. 2 with a specific of diagonal $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_k)$ when the weight is non-descending: $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_k$. It remains unclear if a closed-form exists for an arbitrary weight order.

A4. Variational Linear Auto-encoders. While not explicitly studied before, it is also natural to consider linear simplification of Variational Autoencoders, such as Multi-VAE (Liang et al., 2018), which has shown to exhibit strong performance for recommendation. To optimize linear VAE, we need to find the MLE for the probabolistic model

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(W\mathbf{z} + \boldsymbol{\mu}, \sigma^2 I\right) \text{ and } p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(V(\mathbf{x} - \boldsymbol{\mu}), D), \tag{3}$$

where D is a diagonal covariance matrix. Then we have the following observation.

Lemma 1. *Consider optimizing the ELBO (Evidence Lower Bound) (Kingma & Welling, 2014) for the above LVAE model (Eq.3). When the optimization is over the entire dataset and $\boldsymbol{\mu} = 0$, this optimization problem is equivalent to minimizing*

$$\mathcal{L} = \|X - XV^{\mathrm{T}}W^{\mathrm{T}}\|_F^2 + N\|\sqrt{D}W^T\|_F^2 + \sigma^2\|XV^{\mathrm{T}}\|_F^2 + g(D, \sigma) \tag{4}$$

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

Here, we set $\boldsymbol{\mu} = 0$ for simplicity and following typical practices in recommendations. The proof can be found in Appendix. Since our objective is for recommendation (not purely on recovering the low rank factors of the data), we treat the covariance matrix D as hyperparameters. Thus, the term $g(D, \sigma)$ becomes a constant, and let $A = \sigma V^T$, $B = 1/\sigma W^T$, $\Lambda = \sigma\sqrt{ND}$.

If we consider D as an optimization parameter, then the optimal solution of LVAE is equivalent to that of pPCA (Tipping & Bishop, 1999). In this case, $D = \sigma^2(\mathbf{\Sigma}^2/N)^{-1}$, where $\mathbf{\Sigma}^2 = diag(\sigma_i^2)$ are the eigen-values of the covariance matrix of X, and $\sigma^2 = \frac{1}{n-k}\sum_{j=k+1}^{n}\frac{\sigma_j^2}{N}$. Further, the closed form solutions of V (A) and W (B) are characterized. However, for recommendation, the matrix D can be considered as a hyperparameter (to be learned); in this case, the closed form solution is not studied yet. We may "clean up" eq. (4) and obtain the following optimization problem:

$$\min_{A\in\mathbf{R}^{n\times k}, B\in\mathbf{R}^{k\times m}} \|X - XAB\|_F^2 + \|XA\|_F^2 + \|\Lambda B\|_F^2, \tag{5}$$

in which decision variables are A and B, and the hyper-parameter is a diagonal matrix $\Lambda \in \mathbf{R}^{k\times k}$.

Our main results: solution structure and implications (Sec. 3). *(i)* We show that solving Eq. 5 (A4) is equivalent to solving $\|X - W\|_F^2 + \lambda\|W\|_{\omega,*}$ subject to $\text{rank}(W) \leq k$, where ω consists of Λ's diagonal values, *sorted in ascending order.* In addition, the closed form solution for W is merely shrinking the i-th singular value of X by a magnitude of ω_i. When the diagonals of Λ is already sorted (i.e., $\Lambda_{1,1} \leq \cdots \leq \Lambda_{k,k}$), the problem effectively reduces to A3. When Λ is proportional to identity, the problem reduces to A1. This result has three major implications. First, A1, A3, and A4 effectively only add a variant of nuclear-norm based regularizer, and the major benefits from these algorithms are computational. Second, our result generalizes that in A3 and solves an open problem left there, i.e., the hyper-parameters Λ do not need to have sorted diagonal values, the optimization algorithm will "automatically perform the sorting". Third, while the variational auto-encoder offers a flexibility to tailor-make the prior for each entry in the latent variable $\mathbf{z}$ (in eq. 3), the solution space is quite rigid due to the auto-sorting property: the i-th smallest entry in Λ will find its way to match with the i-th largest singular values in X. In other words, it is impossible to shrink X's singular values by an arbitrary sequence via carefully choosing Λ in eq. 5.

(ii) We characterize the optimal solution for $\|X - W\|_F^2 + \lambda\|W\|_*^p$ for any $p \geq 1$. We shall show that regardless the choice of p, the optimal solution uniformly shrinks all X's singular values by a constant magnitude μ (and to 0 if a singular value is already less than μ). The specific μ depends on λ, p, as well as the data X unless $p = 1$. It has two implications. First, μ needs to be fine-tuned to optimize test performance. Therefore, choices of p (again) only produces computational gain. Second, when $p = 1$, μ does not depend on the data so we can tune this hyper-parameter in a direct manner. When $p = 2$, λ is scale-invariant, i.e., it does not need to be rescaled when all entries in X is scaled by a constant factor. Being able to directly tune μ or having the scale invariant property helps the hyper-parameter search; when $p \neq 1, 2$, it does not offer benefit in either computation or search, which explains why we see only $p = 1, 2$ in the literature.

Finally, for all nuclear-norm-based approaches discussed above, the estimators always keep singular vectors of X and shrink its singular values. Therefore, the solution space offered by nuclear-norm based regularization is quite constrained, which limits these models predictive power.

2.2 Frobenius norm based regularizations

Most algorithms below were originally motivated by the design of (denoising) auto-encoders, it has been shown that they effectively add a Frobenius-norm regularizer. See also Table 3 in Appendix.

A5. EASE (Steck, 2019) aims to optimize $\min_W ||X - XW||_F^2 + \lambda \cdot ||W||_F^2$ subject to the constraint that $diag(W) = 0$. A closed form solution exists for this problem.

A6. DLAE (Steck, 2020) adds a weighted Frobenius-norm regularizer so the objective becomes $\min_W ||X - XW||_F^2 + ||\Lambda^{1/2} \cdot W||_F^2$, where $\Lambda = \frac{p}{1-p} diagM(diag(X^T X))$.

A7. EDLAE (Steck, 2020) integrates weighted Frobenius norm in DLAE with EASE's diagonal constraint so its objective is the same as DLAE but it requires $diag(W) = 0$. A closed form solution exists for this problem. When W is required to be low rank, an ADMM algorithm may be used.

A8. Tikhonov regularization/Low Rank Regression (LRR) (Jin et al., 2021). Let V_k be the k leading right singular vectors of X. Jin et al. (2021) finds an estimator that solves

$$W = \arg \min_{rank(W)\leq k} \|X - XW\|_F^2 + \|\mathbf{\Gamma} W\|_F^2, \tag{6}$$

where $\Gamma = \Lambda^{\frac{1}{2}} V_k^T$ and $\Lambda = diag(\lambda'_1, \cdots, \lambda'_k)$ is a hyper-parameter. Its a closed-form solution is.

$$W^* = V_k diag(\frac{\sigma_1^2}{\sigma_1^2 + \lambda'_1}, \ldots, \frac{\sigma_k^2}{\sigma_k^2 + \lambda'_k}) V_k^T \tag{7}$$

Our results. Our major goal is to design a low-rank closed form estimator whose performance is comparable to the performance leaders. We first remark that A5 and A6 produce full-rank estimators. A7 can produce either full-rank or low-rank estimator (via ADMM) and has the best performance

(among all approaches we discussed). The estimator from A8 is low-rank and has a closed-form solution but it has to keep singular vectors of X so its predictive power is also limited. Nevertheless, A8 is conceptually interesting because it uses Frobenius norm regularizers but its solution space cover the solution space offered in A4 (and thus also A1-A4).

Proposition 1. *For any regularized instances in the form of (5) with regularization parameter Λ such that $\sigma_i(X) \geq \lambda_{(k-i)}$ for all i, there is a corresponding Tikhonov regularized instance with $\mathbf{\Gamma} = \Lambda^{\frac{1}{2}} V_k^{\mathrm{T}}$ which provides the same regularization effect.*

The proof is in Appendix. A major implication of Prop. 1 is that we can focus on designing Frobenius-norm regularizers because it also gets the value from using nuclear-norm regularizers. Indeed, Sec. 4 will introduce two low-rank Frobenius-norm-based model with closed form solutions that have comparable performance to linear performance leaders.

3 NUCLEAR-NORM BASED REGULARIZATION

Rigidity of VLAE. We first analyze solution for Eq. 5 (A4). To facilitate the analysis, we also consider the following problem:

$$\min_{P,Q} \|X - PQ\|_F^2 + \|\Lambda^{\frac{1}{2}} Q\|_F^2 + \|P\,\Lambda^{\frac{1}{2}}|_F^2, \text{ or equivalently }, \min_{P,Q} \|X - PQ\|_F^2 + \|\Lambda Q\|_F^2 + \|P\|_F^2, \quad (8)$$

Note that when we let $Q' = XA^*$ and $P' = B^*$ (where A^* and B^* are an optimal solution of Equation 5), the syntax of Eq. 8 matches with that of Eq. 5. This implies that solution for Eq. 8 is a lower bound of that for Eq. 5. These two solutions coincide only when the columns in the optimal P^* in Eq. 8 are spanned by the columns of X.

We shall first find a closed-form solution (P^*, Q^*) for Eq. 8, and show that indeed that the column space of P^* is in the column space of X. Below is our major Proposition.

Proposition 2. *Let $f : \mathbf{R}^{m\times n} \to \mathbf{R}^+$ be any cost function. Let $P \in \mathbf{R}^{m\times k}$ and $Q \in \mathbf{R}^{k\times n}$. Let $\Lambda \in \mathbf{R}^{k\times k}$ be a diagonal matrix such that $\lambda_i = \Lambda_{ii} \geq 0$ ($i \in [k]$). Let also $\omega = (\lambda_{\pi(1)}, \lambda_{\pi(2)}, \ldots, \lambda_{\pi(k)})$, where π is a permutation on $[k]$ such that $\lambda_{\pi(1)} \leq \lambda_{\pi(2)} \leq \cdots \leq \lambda_{\pi(k)}$. The following two optimization problems have the same optimal values*

$$OPT1: \quad \min_{P,Q} \quad f(PQ) + \|\Lambda^{\frac{1}{2}} Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2.$$

$$OPT2: \quad \begin{array}{ll} \min_W & f(W) + 2\|W\|_{\omega,*} \\ \textit{subject to} & \mathrm{rank}(W) \leq k. \end{array}$$

In addition, if (P^*, Q^*) is an optimal solution for $OPT1$, then $W^* = P^*Q^*$ is an optimal solution for $OPT2$. If W^* is an optimal solution for $OPT2$, then there exists an optimal solution (P^*, Q^*) for $OPT1$ such that $W^* = P^*Q^*$.

We reiterate three points made earlier (Sec. 2). *(i)* Both A3 and A4 effectively add a nuclear norm regularizer. *(ii)* Diagonals of Λ do not need to be sorted in ascending order as stated in (Bao et al., 2020) because any permutation of the diagonals will be equivalent to $OPT2$. This also limits the search space and affects a model's prediction power. *(iii)* Prop. 2 "compiles" a non-differentiable objective ($OPT2$) into an equivalent differentiable one ($OPT1$), which is easier to optimize. In addition, $f(\cdot)$ in A3 & A4 is the reconstruction error, in which case closed form solutions exist.

We next explain the intuition for proving Prop. 2 (see Appendix for the full analysis). Consider $OPT1$ and let $W = PQ$. Our goal is to characterize the behaviors of P and Q with the presence of the regularizers when $W = PQ$ is known (fixed). Let the SVD of W be $U_W \Sigma_W V_W^{\mathrm{T}}$. Because two regularizers $\|\Lambda^{\frac{1}{2}} Q\|_F^2$ and $\|P\Lambda^{\frac{1}{2}}\|_F^2$ are symmetric, we could "guess" $P = U_W \Sigma_W^{\frac{1}{2}} \Omega$ and $Q = \Omega^{\mathrm{T}} \Sigma_W^{\frac{1}{2}} V_W^{\mathrm{T}}$, where Ω is a unitary matrix. Now we have

$$\|\Lambda^{\frac{1}{2}} Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2 = \|\Lambda^{\frac{1}{2}} \Omega \Sigma^{\frac{1}{2}} V_W^{\mathrm{T}}\|_F^2 + \|U_W \Sigma_W^{\frac{1}{2}} \Omega \Lambda^{\frac{1}{2}}\|_F^2 = 2\|\Lambda^{\frac{1}{2}} \Omega \Sigma_W^{\frac{1}{2}}\|_F^2.$$

Now the question of finding P and Q when W is known boils down to finding a unitary matrix Ω

that minimizes $\|\Lambda^{\frac{1}{2}}\Omega\Sigma_W^{\frac{1}{2}}\|_F^2$, where diagonal matrices Λ and Σ_W are given. Recall that $\lambda_i = \Lambda_{ii}$ and let $\sigma_i = (\Sigma_W)_{ii}$. Note that λ_i's could be unsorted and, and that σ_i's are sorted in descending order.

If we restrict Ω to be only a permutation matrix, then we aim to find a permutation $\pi \in [k]$ that minimizes $\sum_{i\leq k}\lambda_{\pi(i)}\sigma_i$. Using a rearrangement inequality Yue (2020), we can see that the minimal is achieved when $\lambda_{\pi(1)} \leq \lambda_{\pi(2)} \leq \cdots \leq \lambda_{\pi(k)}$. In this case, we indeed have $\min_{\Omega \text{ a permutation}} \|\Lambda^{\frac{1}{2}}\Omega\Sigma_W^{\frac{1}{2}}\|_F^2 = \|\Sigma_W\|_{\omega,*}$, where $\omega = (\lambda_{\pi(1)}, \ldots, \lambda_{\pi(k)})$.

Note that because $PQ = P\Omega\Omega^{\mathrm{T}}Q$ for any unitary matrix Ω, it is always beneficial to use Ω to shuffle the rows and columns of P and Q so that the largest σ_i is mapped to the smallest λ_i, etc. This "degree of freedom" from Ω also explains why ordering the values along Λ's diagonal is irrelevant.

Appendix shows that even when Ω is allowed to be any unitary matrix, the optimal one is still a permutation matrix. This conclusion can be viewed as a matrix version of re-arrangement inequality.

Prop. 2 also leads to the following Corollary.

Corollary 1. *Let $\mathbf{X} \in \mathbf{R}^{m\times n}$ ($m \geq n$) be a full rank matrix. Let Λ be a diagonal matrix with $\Lambda_{ii} \geq 0$ for all i. Let $P \in \mathbf{R}^{m\times k}$ and $Q \in \mathbf{R}^{k\times n}$. Consider the optimization problems*

$$\min_{P,Q} \|X - PQ\|_F^2 + \|\Lambda^{\frac{1}{2}}Q\|_F^2 + \|P\Lambda^{\frac{1}{2}}\|_F^2 \tag{9}$$

$$\min_{A,B} \|X - XAB\|_F^2 + \|\Lambda B\|_F^2 + \|XA\|_F^2. \tag{10}$$

Let the SVD of X be $U\Sigma V^{\mathrm{T}}$, and let $\Sigma_k \in \mathbf{R}^{k\times k}$ be a matrix comprising the k largest singular values of X, and U_k and V_k be the corresponding singular vectors. Let $\lambda_{(1)} \geq \lambda_{(2)} \geq \cdots \geq \lambda_{(k)}$ be the sorted sequence of the diagonal values from Λ. (9) has a closed-form solution:

$$\begin{aligned} P^* &= U_k diag(\sqrt{(\sigma_1 - \lambda_{(k)})^+}), \ldots, \sqrt{(\sigma_k - \lambda_{(1)})^+})\Omega, \\ Q^* &= \Omega^{\mathrm{T}} diag(\sqrt{(\sigma_1 - \lambda_{(k)})^+}), \ldots)V_k^{\mathrm{T}}, \end{aligned} \tag{11}$$

where Ω is a unitary matrix that corresponds to the permutation π such that $\lambda_{\pi(1)} \leq \cdots \leq \lambda_{\pi(k)}$. In addition, (10) has a closed-form solution: $A^* = X^\dagger P^* \Lambda^{\frac{1}{2}}$ and $B^* = \Lambda^{-\frac{1}{2}}Q^*$, where $X^\dagger$ is the pseudo-inverse of X.

Uniform solution structure for regularizing $\|W\|_*^p$. In (Cavazza et al., 2018), it was observed that when we use a neural net $X = PQ^{\mathrm{T}}$ (with learnable parameters being P and Q) to train a model and a standard dropout is used, the objective is equivalent to solving $\min_W \|X - W\|_F^2 + \lambda\|W\|_*^2$. While the regularizer $\|W\|_*^2$ deviates from the standard one $\|W\|_*$, the optimal solution here is $W = US_\mu(\Sigma)V^{\mathrm{T}}$, where $U\Sigma V^{\mathrm{T}}$ is SVD of X, and $S_\mu(\Sigma)$ is a diagonal matrix such that its (i,i)-th element is $(\Sigma_{i,i} - \mu)^+$, in which μ depends on the data X and λ. In other words, the optimal solution for regularizers $\|W\|_*^2$ and $\|W\|_*$ are strikingly similar. Thus, we are interested in how regularizers with different exponents are connected. Our main observation is that for *any* regularizer $\|W\|_*^p$ ($p \geq 1$), the optimal solution has the same structure.

Lemma 2. *Let $X \in \mathbf{R}^{m\times n}$, where $m \geq n$. Let the d leading SVDs of X be U_d and V_d respectively. Let $\sigma_1, \ldots, \sigma_n$ be the singular values of X. Consider the optimization problem:*

$$\min_W \frac{1}{2}\|X - W\|_F^2 + \lambda\|W\|_*^p. \tag{12}$$

Let $\mu_k = \frac{1}{k}\sum_{i\leq k}\sigma_i(X)$, $\eta(\mu_k)$ be the positive root of the function $z + \lambda k z^{p-1} - k\mu_k$ and d be the largest value such that $\sigma_d(X) - \lambda(\eta(\mu_d))^{p-1} \geq 0$. Let $\mu = \lambda(\eta(\mu_d))^{p-1}$. The optimal solution of W is $U_d diag(\sigma_1 - \mu, \sigma_2 - \mu, \ldots, \sigma_d - \mu)V_d^{\mathrm{T}}$.

Lemma 2 is a straightforward generalization of Prop. 12 in (Cavazza et al., 2018) so the contribution here is a conceptual one: it implies that regularizers $\|W\|_*^p$ with different p's differ in how the shrinkage variable μ is obtained. Observe that μ is an important hyper-parameter that needs to be extensively tuned against data, all regularizers $\|W\|_*^p$ provide the same learning power. As noted earlier, μ is a function of X (i.e., different X will result in different μ) unless $p = 1$. In addition, λ needs to be rescaled when X is scaled by a constant factor unless $p = 2$. This implies it could be easier to tune μ when $p = 1, 2$, and explains why only $p = 1, 2$ have been extensively considered.

4 LOW-RANK FROBENIUS NORM BASED REGULARIZATIONS

This section presents the close-form low-rank estimators with comparable performance to the state of the art algorithms.

Approximate Low rank DLAE and EDLAE. Recall that DLAE solves

$$\begin{aligned} &\min_W ||X - XW||_F^2 + ||\Lambda^{\frac{1}{2}} W||_F^2 \\ s.t. \quad &\Lambda = \frac{p}{1-p} dMat(diag(X^T X)), \end{aligned} \tag{13}$$

and EDLAE with the additional $diag(W) = 0$ constraint. Even though both the Nuclear norm based regularization as well as the full rank DLAE and EDLAE solutions all have closed-form solutions, such solution is unknown for the low-rank DLAE and EDLAE, whose existing solution is based on ADMMSteck (2020). The closed form solutions will help both better understand and compare these models, and determine the hyper-parameters, which is usually difficult for the ADMM type solutions.

For DLAE, since it can be considered a special form of Tikhonov regularization (Eq 6) which has a closed form Jin et al. (2021), its closed form solution, referred to LR-DLAE, is immediately available (See Line 6 in Table 3). However, for EDLAE, it has the zero diagonal constraint, which make the exact solution difficult to express. Here, we present an approximate low-rank closed-form solution of eq. (13) by decomposing the optimization problem into two subproblems, which is similar to (Jin et al., 2021): We first consider the full-rank closed form solution for EDLAE Steck (2020), which is:

$$W^* = I - C \cdot dMat(1 \oslash diag(C)), \text{ where } C = (X^T X + \Lambda)^{-1}$$

Then, we consider two approaches to produce low-rank matrix approximation of W^*:

(Method 1 (LR-EDLAE-1):) Selecting $\widehat{W}$ to best approximate the performance of W^* without the zero diagonal constraint:

$$\begin{aligned} \widehat{W} &= \arg \min_{rank(W) \leq k} ||\overline{X} W^* - \overline{X} W||_F^2 \\ &= \arg \min_{rank(W) \leq k} ||XW^* - XW||_F^2 + ||\Lambda^{\frac{1}{2}}(W^* - W)||_F^2 \end{aligned} \tag{14}$$

where $\overline{X} = \begin{bmatrix} X \\ \Lambda^{\frac{1}{2}} \end{bmatrix}$. Noting, in the full rank problem eq. (13), it forces the diagonal of derived matrix (W^*) to be zero. Here, we relax the zero diagonal constraint - the diagonal of low rank approximate matrix $\widehat{W}$ doesn't have to be zero, which has also been discussed in (Steck, 2020). The closed-from solution of eq. (14) is given by:

$$\boxed{\widehat{W} = W^*(Q_k Q_k^T)} \tag{15}$$

where Q_k comes from SVD:

$$\overline{Y}^* = \overline{X} W^*, \text{ and } \overline{Y}^*(k) = P_k \Sigma_k Q_k^T$$

(Method 2 (LR-EDLAE-2):) SVD approximation of W^*: The alternative solution is to simply perform SVD, and which gives a low-rank estimation of W^*.

Note that in both approaches, the hard constraint of zero-diagonal on $\widehat{W}$ is relaxed. In the next section, the experimental results show both approaches can provide comparable or better performance compared with the ADMM solution, and also very close to the full rank EDLAE solution.

5 EXPERIMENTAL RESULTS

In this section, we experimentally study different regularizations for linear recommendation models. Our goal is to validate the effectiveness of various regularizations (all can be categorized under nuclear norm and Frobenius norm) together with their simple closed-form solutions. We aim to answer three questions: **Q1.** How does the closed form solution of low rank Frobenius norm perform compared with the ADMM solutions (Section 4) and how does the weighted nuclear norm regularizer

Table 1: The performance comparison between different regularizations. For notation, please refer table 3 for more details.

Model		ML-20M			Netflix			MSD		
		Recall@20	Recall@50	nDCG@100	Recall@20	Recall@50	nDCG@100	Recall@20	Recall@50	nDCG@100
Frobinius Norm	EASE	0.391	0.521	0.420	0.362	0.445	0.393	0.333	0.428	0.389
	DLAE	0.392	0.527	0.424	0.362	0.446	0.395	0.329	0.426	0.387
	EDLAE	0.393	0.523	0.424	0.366	0.449	0.398	0.334	0.429	0.392
	EDLAE-ADMM	0.392	0.524	0.424	0.365	0.448	0.396	0.330	0.424	0.386
	LRR	0.376	0.511	0.408	0.348	0.431	0.380	0.248	0.335	0.301
	LR DLAE	0.392	0.527	0.424	0.362	0.445	0.395	0.306	0.403	0.363
	LR-EDLAE-1	0.392	0.523	0.424	0.365	0.449	0.398	0.327	0.423	0.384
	LR-EDLAE-2	0.392	0.523	0.424	0.365	0.449	0.398	0.325	0.421	0.382
Nuclear Norm	MF dropout	0.367	0.501	0.393	0.334	0.418	0.365	0.270	0.367	0.328
	Regularized PCA	0.364	0.501	0.392	0.331	0.417	0.365	0.229	0.313	0.279
	LVAE	0.348	0.474	0.378	0.325	0.405	0.357	0.205	0.254	0.286
Baseline	WMF/ALS	0.360	0.498	0.386	0.316	0.404	0.351	0.211	0.312	0.257
	SLIM	0.370	0.495	0.401	0.347	0.428	0.379	no results in (Ning & Karypis, 2011)		
	CDAE	0.391	0.523	0.418	0.343	0.428	0.376	0.188	0.283	0.237
	MULT-DAE	0.387	0.524	0.419	0.344	0.438	0.380	0.266	0.363	0.313
	MULT-VAE	0.395	0.537	0.426	0.351	0.444	0.386	0.266	0.364	0.316
# items		20108			17769			41140		
# users		136677			463435			571353		
# interactions		10mil			57mil			34mil		

for matrix factorization (Proposition 2 and Corollary 1) perform? **Q2.** What is the tradeoff between the number of factors (rank k) and the recommendation accuracy? **Q3.** How does the ordering of weights from small to large (non-descending) for adjusting the singular values (Corollary 1) affect the recommendation performance?

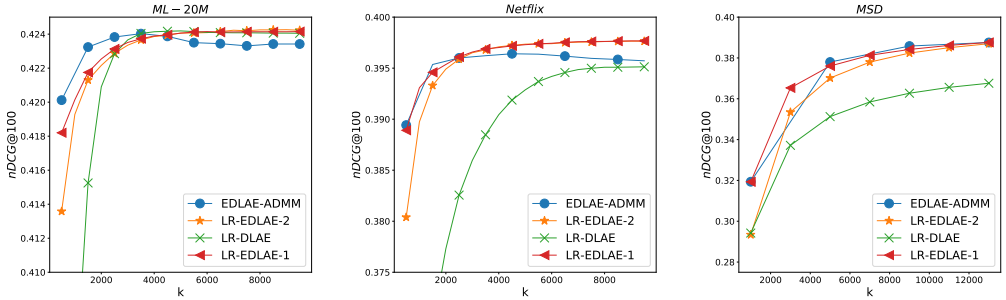
Experimental Setup: We use three commonly used datasets for recommendation studies: MovieLens 20 Million (ML-20M) (Harper & Konstan, 2015), Netflix Prize (Netflix) (Bennett et al., 2007), and the Million Song Data (MSD)(Bertin-Mahieux et al., 2011). We obtained these datasets and all benchmarks from authors of EASE (Steck, 2019), EDLAE (Steck, 2020), and Mult-VAE (Liang et al., 2018).

Similar to the latest study in EASE (Steck, 2019), and EDLAE (Steck, 2020), we consider the following state-of-the-art recommendation models: ALS (WMF) (Hu et al., 2008) for matrix factorization approaches, SLIM (Ning & Karypis, 2011), EASE (Steck, 2019), and EDLAE (Steck, 2020) for linear autoendoers, CDAE (Wu et al., 2016), Mult-DAE and Mult-VAE (Liang et al., 2018) for deep learning models. The experiment settings for these baseline are the same as (Liang et al., 2018; Steck, 2019; 2020). Also we follow their practice (Liang et al., 2018; Steck, 2019; 2020) for the *strong generalization* by splitting the users into training, validation and tests group, and report performance metrics $Recall@20$, $Recall@50$ and $nDCG@100$. Finally, we note that our code are openly available (see Appendix).

Q1: Low-Rank Frobenius Norm and (Weighted) Nuclear Norm Regularization: In this experiment, we evaluate the low rank Frobenius norm regularization and the nuclear norm regularization (eq. (1)) for the matrix factorization. Here EDLAE-ADMM, LRR, LR-DLAE, LR-EDLAE-1, LR-EDLAE-2, MF dropout and LVAE are listed in table 3 in Appendix. To determine the non-descending order of weights λ_i for the closed-form solution in (eq. (11)), we follow the practice in weighted nuclear norm regularization in (Gu et al., 2014) as well as the optimized pPCA weight (Lucas et al., 2019b). Let $\lambda_i = \frac{C}{\sigma_i}$ where C is a hyperparameter, and we perform grid-search to find the optimal one.

In Table 1, we can see that the weighted nuclear norm regularization (LVAE) based matrix factorization actually performs worse than the constant weighted version (Regularized PCA). And the latter shows very strong performance comparing against the WFM/ALS (one of the most popular implicit matrix factorization algorithm). We also observe the closed form solutions (LR-DLAE, LR-EDLAE-1 and LR-EDLAE-2) all perform very comparable with the ADMM based low rank solution and the full rank DLAE and EDLAE solutions.

Q2: nDCG vs Rank k for low-rank Frobenius norm: In this experiment, we focus on evaluating the recommendation accuracy (using nDCG) against the rank k. Specifically, we vary the rank k from around $1K$ to around $10K$, and we tune and compare four different methods, including EDLAE-ADMM, LR-DLAE, LR-EDLAE-1, and LR-EDLAE-2. We have the following observations: 1) As k increases, the recommendation accuracy also increases in general; however, most of them reaches a plateau around similar K, and for different datasets, the saturating point varies. 2) LR-DLAE performs worse than EDLAE based approaches in two out of three datasets. This partially



(a) (b) (c)

Figure 1: Low rank models $nDCG@100$ on test data for 3 datasets.

Table 2: Investigating the weight ordering of Matrix Factorization

Model	ML-20M			Netflix			MSD		
	Recall@20	Recall@50	nDCG@100	Recall@20	Recall@50	nDCG@100	Recall@20	Recall@50	nDCG@100
MF/LRR weighted	0.3806	0.5175	0.4102	0.3484	0.4320	0.3797	0.2508	0.3390	0.3037
MF sorted	0.3017	0.4507	0.3361	0.2860	0.3801	0.3265	0.2288	0.3148	0.2802

demonstrates the benefits of zero-diagonal constraint. 3) The closed form solution of EDLAE performs comparable or even slightly better than ADMM methods as K grows; but when K is relatively small, ADMM method perform slightly better. But none-the-less, for most of the reasonable choices of k when low-rank approximates full rank, the closed form solution performs comparable or better.

Q3: Impact of Weight Ordering: Finally, we study how the ordering of weights from small to large (non-descending) for adjusting the singular values (Corollary 1) affects the recommendation performance using matrix factorization (closed-form solution in Eq. 11). Our results are in Table 2. Here, we obtain the searched optimal weight parameters from weighted Tikhonov regularization (following the approach in Jin et al. (2021)), and map it back to the parameters in the closed form solution (Proposition 1). Then we sort the parameters in the non-descending order, and then report their results in the second row of Table 2. We can see that the recommendation performance becomes significant worst. This help confirm our conjecture that the strict ordering of weight on matrix factorization and other regularizations can be an inherent limitation for those approaches.

6 Conclusion and Discussion

This work provides a complete analysis on the recently proposed linear models for recommendation systems. Despite that models leverage different deep learning techniques, they achieve similar performance. We find that this is not coincident: all the models add either a nuclear-norm-based (Lemma 1) or a Frobenius-norm based regularizer. The nuclear-norm-based approach results in estimators that keep X's singular vectors and shrink its singular values in a quite rigid way (Proposition 2 and Lemma 2), which limit their prediction power. The Frobenius-norm models are more express (Proposition 1) and effective but their estimators are either full-rank or do not have closed form solutions. To get the best of both nuclear and Frobenius worlds, we propose two low-rank and closed-form estimators (Sec. 4) based on carefully generalizing Frobenius-norm based regularizers. These estimators have competitive performance against linear performance leaders, and thus concisely pack all the benefits obtained by a recent long line of research and abstract out all the computation nuance.

REFERENCES

Charu C. Aggarwal. *Recommender Systems: The Textbook*. Springer, 1st edition, 2016. ISBN 3319296574.

Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B. Grosse. Regularized linear autoencoders recover the principal components, eventually. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

James Bennett, Charles Elkan, Bing Liu, Padhraic Smyth, and Domonkos Tikk. Kdd cup and workshop 2007. 2007. doi: 10.1145/1345448.1345459. URL https://doi.org/10.1145/1345448.1345459.

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pp. 530–582. PMLR, 2016.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Jacopo Cavazza, Pietro Morerio, Benjamin Haeffele, Connor Lane, Vittorio Murino, and Rene Vidal. Dropout as a low-rank regularizer for matrix factorization. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. PMLR, 2018.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.

Evangelia Christakopoulou and George Karypis. Hoslim: Higher-order sparse linear method for top-n recommender systems. In *Advances in Knowledge Discovery and Data Mining*, 2014.

Maurizio Ferrari Dacrema, P. Cremonesi, and D. Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *RecSys'19*, 2019a.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? *RecSys'19*, 2019b.

Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.*, 2004.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2862–2869, 2014.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 2015. doi: 10.1145/2827872.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *ICDM'08*, 2008.

José Pedro Iglesias, Carl Olsson, and Marcus Valtonen Örnhag. Accurate optimization of weighted nuclear norm for non-rigid structure from motion. *arXiv preprint arXiv:2003.10281*, 2020.

Ruoming Jin, Dong Li, Jing Gao, Zhi Liu, Li Chen, and Yang Zhou. Towards a better understanding of linear recommendation models. In *KDD'21*, 2021. URL https://arxiv.org/abs/2105.12937.

Santosh Kabbur, Xia Ning, and George Karypis. Fism: Factored item similarity models for top-n recommender systems. KDD '13, 2013.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *KDD'08*, 2008.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, August 2009.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 3560–3569. PMLR, 09–15 Jun 2019.

Xiaopeng Li and James She. Collaborative variational autoencoder for recommender systems. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 305–314, 2017.

Dawen Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara. Variational autoencoders for collaborative filtering. In *WWW'18*, 2018.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the elbo! A linear VAE perspective on posterior collapse. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9403–9413, 2019b. URL https://proceedings.neurips.cc/paper/2019/hash/7e3315fe390974fcf25e44a9445bd821-Abstract.html.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pp. 1069–1097, 2011.

Xia Ning and George Karypis. Slim: Sparse linear methods for top-n recommender systems. ICDM '11, 2011.

Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems, 2019.

Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Darius Braziunas. On the effectiveness of linear models for one-class collaborative filtering. AAAI'16, 2016.

Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, pp. 528–536, 2020.

Harald Steck. Embarrassingly shallow autoencoders for sparse data. *WWW'19*, 2019.

Harald Steck. Autoencoders that don't overfit towards the identity. In *NIPS*, 2020.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Michael E. Tipping and Chris M. Bishop. Probabilistic principal component analysis. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 61(3):611–622, 1999.

Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Found. Trends Mach. Learn.*, 9(1):1–118, June 2016.

Romain Warlop, Alessandro Lazaric, and Jérémie Mary. Parallel higher order alternating least square for tensor recommender system. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. WSDM '16, 2016.

Man-Chung Yue. A matrix generalization of the hardy-littlewood-p\'olya rearrangement inequality and its applications. *arXiv preprint arXiv:2006.08144*, 2020.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 2019.

Shuai Zheng, Chris Ding, and Feiping Nie. Regularized singular value decomposition and application to recommender system, 2018.