RANDOMLY MASKED LINEAR AUTOENCODERS: A UNIFIED CLOSED-FORM FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomly masked linear autoencoders (LAEs) reconstruct data from partially observed inputs. Existing formulations, however, do not jointly provide emphasis weighting, independent diagonal control, and an explicit whole-matrix solution. We introduce a unified masked objective that combines emphasis weighting with separately tunable diagonal and full-matrix regularization. Conditioning on whether each target is retained or dropped exposes a shared quadratic structure, yielding a whole-matrix closed form through one shared inverse and diagonal rescaling. Experiments on nine recommendation datasets show that our unified model outperforms existing LAE baselines on most evaluation metrics and achieves competitive accuracy against more complex deep learning methods. Our solver also matches DEQL's column-wise solution numerically at the free-diagonal endpoint while reducing runtime and peak incremental memory. We additionally derive two closed-form low-rank approximations with $O(nk)$ factorized storage. Our code is available at `https://anonymous.4open.science/r/ICLR_2027_Unified_Recommendation-E5F2`.

1 INTRODUCTION

Linear autoencoders (LAEs) reconstruct data through learned linear mappings, providing a tractable framework for studying dimensionality reduction and low-dimensional representations (Jin et al., 2021; Steck & Garcia, 2021), denoising (Vincent et al., 2008; Chen et al., 2012), and predictive dependencies among features. Given a data matrix $\boldsymbol{R} \in \mathbb{R}^{m\times n}$, an LAE produces $\widehat{\boldsymbol{S}} = \widetilde{\boldsymbol{R}}\boldsymbol{W}$, where $\widetilde{\boldsymbol{R}}$ is the original or corrupted input and $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ is a reconstruction matrix, optionally factorized through a low-dimensional latent space. In recommendation, $\boldsymbol{R} \in \{0,1\}^{m\times n}$ is the user–item interaction matrix, whose columns correspond to items; LAEs learn item-to-item coefficients and compute recommendation scores as $\widehat{\boldsymbol{S}} = \boldsymbol{R}\boldsymbol{W}$. SLIM and EASE achieve competitive recommendation performance (Ning & Karypis, 2011; Steck, 2019). Quadratic objectives such as EASE also admit closed-form solutions without iterative training (Steck, 2019).

Random masking makes prediction across features an explicit training task: withheld features must be reconstructed from those that remain available. In recommendation, Steck (2020) formalized this approach through emphasized denoising, and Guo et al. (2026) subsequently extended its analysis through the Decoupled Expected Quadratic Loss (DEQL) framework. By withholding interactions during training, random masking reflects the information available under random-holdout evaluation, although the reconstruction loss is not itself a ranking metric. More broadly, prediction from partial inputs is central to masked image modeling and masked language modeling (He et al., 2022; Devlin et al., 2019). LAEs offer a mathematically accessible setting for analyzing how masking and reconstruction weights shape this learning process.

Specifically, let $\Delta \in \{0,1\}^{m\times n}$ have independent entries δ_{ui} with $\Pr(\delta_{ui} = 0) = p$, where $0 < p < 1$. The corrupted input is $\widetilde{\boldsymbol{R}} = \Delta \odot \boldsymbol{R}$, where $\odot$ denotes the Hadamard product, and the reconstruction target is $\boldsymbol{R}$. Define an emphasis matrix $\boldsymbol{A}(\Delta)$ with $A_{u,i} = a$ when $\delta_{ui} = 0$ and $A_{u,i} = b$ otherwise, where $a > 0$ and $b \geq 0$. The expected weighted reconstruction loss is

$$\mathcal{L}_{\mathrm{EDLAE}}(\boldsymbol{W}) = \mathbb{E}_{\Delta}\Big[\big\|\boldsymbol{A}(\Delta) \odot \big(\boldsymbol{R} - (\Delta \odot \boldsymbol{R})\boldsymbol{W}\big)\big\|_F^2\Big]. \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The effective loss weights are a^2 for dropped targets and b^2 for retained targets. Crucially, the target's mask state determines both its reconstruction weight and whether its own input value is available for self-reconstruction.

The mask Δ plays a crucial role in determining the behavior of the reconstruction in equation 1. When both $\boldsymbol{A}(\Delta)$ and Δ are removed, the minimizer of equation 1 is trivially identity, which does not perform a meaningful reconstruction. Taking $a > b$ means placing more emphasis on reconstructing masked entries (those with $\delta_{ui} = 0$); taking $a < b$ means placing more emphasis on reconstructing retained entries (those with $\delta_{ui} = 1$); and taking $a = b$ puts equal emphasis on both. DLAE takes $a = b > 0$, EDLAE takes $a \geq b \geq 0$ (Steck, 2020), and DEQL takes $b > 0$, including $b > a$ (Guo et al., 2026).

The expectation in equation 1 couples emphasis weighting to the availability of self-reconstruction, without isolating the diagonal contribution. Existing formulations treat general emphasis and diagonal control separately. DEQL derives column-wise solutions under general emphasis, but does not provide a whole-matrix expression along a continuously penalized diagonal path. We introduce an independently tunable diagonal penalty alongside full-matrix L_2 regularization and derive a unified whole-matrix closed form. The resulting expression makes the roles of emphasis and diagonal regularization explicit.

Our approach is that we first decompose equation 1 into equation 9 (Lemma 3.2), which includes three terms: the retained-state residual, the dropped-state residual, and their shared mask-induced variance. In this way, we convert equation 1 into a more interpretable expression. Moreover, this expression facilitates penalizing the diagonal: we explicitly take out a diagonal matrix $\boldsymbol{D} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))$ in equation 9. The masked conditional objective equation 9 ensures that we can add a quadratic penalty on $\boldsymbol{D}$ and still obtain a whole-matrix closed-form solution (Theorem 4.1).

Our contributions are:

- **A unified masked objective with independent diagonal control.** We combine emphasis weighting, full-matrix regularization, and a separately tunable diagonal penalty in one objective, connecting the free-diagonal setting at $\lambda_D = 0$ to the zero-diagonal limit as $\lambda_D \to \infty$.
- **An explicit whole-matrix solution and efficient solver.** We derive a closed-form solution with both diagonal and full-matrix regularization. For $b > 0$ and $\lambda_D = 0$, it recovers the corresponding DEQL solution under matching regularization. A shared inverse supports evaluation across the diagonal-penalty path.
- **Empirical gains in accuracy and efficiency.** Experiments on nine recommendation datasets show that our unified model outperforms existing LAE baselines on most evaluation metrics and achieves competitive accuracy against more complex deep learning methods. Our solver also reduces runtime and peak incremental memory relative to DEQL's column-wise implementation while preserving numerical agreement at the free-diagonal endpoint.

Finally, to obtain compact models, we further use the full-rank optimum as a teacher for low-rank approximation. We construct two closed-form approximations of the full-rank teacher, $\boldsymbol{W}^*_{L_2}$-approx and $\boldsymbol{W}^*_{L_2}$-SVD, under general emphasis and diagonal regularization. Each stores its rank-k factors in $O(nk)$ space (Appendix D).

2 PRELIMINARIES

Linear Autoencoder Recommenders. Consider an implicit-feedback dataset with m users and n items, represented by a binary interaction matrix $\boldsymbol{R} \in \{0,1\}^{m \times n}$, where $R_{u,i} = 1$ indicates that user u has interacted with item i and $R_{u,i} = 0$ denotes an unobserved interaction. A linear autoencoder (LAE) learns an item-to-item coefficient matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ and predicts the complete score matrix as $\widehat{\boldsymbol{S}} = \boldsymbol{RW}$; at inference time, the unobserved items for each user are ranked by these scores.

DEQL (Guo et al., 2026). DEQL decomposes an LAE objective into expected quadratic losses over target columns, allowing a distinct distribution for each target. It subsumes EDLAE by writing

equation 1 as a sum of expected losses over target columns

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \sum_{i=1}^{n} \mathbb{E}_{\Delta\sim\mathcal{B}} \left[\|\boldsymbol{A}_i \boldsymbol{R}_{:,i} - \boldsymbol{A}_i(\Delta \odot \boldsymbol{R})\boldsymbol{W}_{:,i}\|_2^2 \right] \tag{2}$$

where $\boldsymbol{A}_i = \mathrm{Diag}(\boldsymbol{A}(\Delta)_{:,i})$ and $\mathcal{B}$ denotes the multivariate Bernoulli distribution of Δ. For the EDLAE objective with $b > 0$, DEQL obtains each target column from

$$\begin{aligned} \boldsymbol{W}^*_{:,i} &= \left(\boldsymbol{H}^{(i)}\right)^{-1} \boldsymbol{v}^{(i)}, \\ \boldsymbol{H}^{(i)} &= \mathbb{E}_\Delta\left[(\Delta \odot \boldsymbol{R})^\top \boldsymbol{A}_i^2 (\Delta \odot \boldsymbol{R})\right], \\ \boldsymbol{v}^{(i)} &= \mathbb{E}_\Delta\left[(\Delta \odot \boldsymbol{R})^\top \boldsymbol{A}_i^2 \boldsymbol{R}_{:,i}\right]. \end{aligned} \tag{3}$$

At $b = 0$, all minimizers agree off the diagonal and have arbitrary diagonal entries; for every $b > 0$, the minimizer is unique.

DEQL decomposes $\boldsymbol{H}^{(i)}$ as $\boldsymbol{H}^{(i)} = \boldsymbol{H}_0 + \boldsymbol{E}_1^{(i)} + \boldsymbol{E}_2^{(i)}$ where $\boldsymbol{H}_0$ is positive definite and shared across targets and $\boldsymbol{E}_1^{(i)}$ and $\boldsymbol{E}_2^{(i)}$ are two rank-1 perturbation matrices supported only on column i and row i, respectively. This decomposition allows an efficient computation for the inverses ${\boldsymbol{H}^{(i)}}^{-1}$ for all i using Miller's matrix inverse theorem, giving a total cost of $O(n^3)$ while still assembling the solution column by column.

3 Masked Conditional Objective and Closed-Form Solutions

Starting from the dropout-induced loss in equation 1, we condition on whether each target interaction is retained or dropped, integrate out the mask randomness exactly, and obtain the deterministic masked conditional objective (Section 3.1). We then solve this objective under the free diagonal and zero diagonal conventions. With a free diagonal, one inverse is shared across target items and each diagonal coefficient enters through a rank-one correction, yielding a whole matrix closed form equal to DEQL's column-wise solution (Section 3.2). With a zero diagonal, the solution is independent of the emphasis ratio and agrees, up to a positive factor, with full emphasis EDLAE (Section 3.3).

3.1 Masked Conditional Objective

Let $q = 1 - p$, and let $\mathcal{L}_{\mathrm{EDLAE}}(\boldsymbol{W})$ denote the objective minimized in equation 1. In this objective, the target mask δ_{ui} determines both the target weight $A_{u,i}$ and whether the self-reconstruction term $R_{u,i}W_{i,i}$ enters the prediction. Without conditioning, the weighted target is therefore random and correlated with the prediction. Conditioning on δ_{ui} fixes both and gives the two conditional mean predictions

$$\widehat{r}_{u,i}^{(1)} = R_{u,i}W_{i,i} + q\sum_{k\neq i} R_{u,k}W_{k,i}, \qquad \delta_{ui} = 1, \tag{4}$$

$$\widehat{r}_{u,i}^{(0)} = q\sum_{k\neq i} R_{u,k}W_{k,i}, \qquad \delta_{ui} = 0. \tag{5}$$

The remaining input masks are independent of δ_{ui}, so both states share the conditional variance

$$v_{u,i}(\boldsymbol{W}) = pq\sum_{k\neq i} R_{u,k}^2 W_{k,i}^2. \tag{6}$$

Within each state the target weight is fixed: $A_{u,i} = b$ if $\delta_{ui} = 1$ and $A_{u,i} = a$ if $\delta_{ui} = 0$. Each conditional loss is therefore the weighted squared bias plus the weighted variance. The two states occur with probabilities q and p, so the law of total expectation gives, for $a > 0$ and the emphasis ratio $\alpha = b/a$,

$$\begin{aligned} &\mathbb{E}_\Delta\left[A_{u,i}^2\left(R_{u,i} - [(\Delta \odot \boldsymbol{R})\boldsymbol{W}]_{u,i}\right)^2\right] \\ &\quad = a^2\Big\{q\alpha^2\left[(R_{u,i} - \widehat{r}_{u,i}^{(1)})^2 + v_{u,i}(\boldsymbol{W})\right] + p\left[(R_{u,i} - \widehat{r}_{u,i}^{(0)})^2 + v_{u,i}(\boldsymbol{W})\right]\Big\}. \end{aligned} \tag{7}$$

The ratio α sets the relative emphasis of the two states: $\alpha < 1$ emphasizes the dropped state, $\alpha = 1$ weights them equally, and $\alpha > 1$ emphasizes the retained state. As in DEQL, $\alpha > 1$ is allowed.

Summing equation 7 over all user–item pairs and writing the result in matrix form uses

$$s = q\alpha^2 + p, \qquad \boldsymbol{G} = \boldsymbol{R}^\top \boldsymbol{R}, \qquad \boldsymbol{N} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{G})), \tag{8}$$

together with a diagonal matrix $\boldsymbol{D}$ that collects the self-reconstruction coefficients.

Definition 3.1 (Masked conditional objective)**.** For $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ and a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n\times n}$, let

$$\begin{aligned}\widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{D}) = q\alpha^2 \left\|\boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W} - p\boldsymbol{R}\boldsymbol{D}\right\|_F^2 + p \left\|\boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W} + q\boldsymbol{R}\boldsymbol{D}\right\|_F^2 \\ + pqs \left\|\boldsymbol{N}^{1/2}(\boldsymbol{W} - \boldsymbol{D})\right\|_F^2 .\end{aligned} \tag{9}$$

The masked conditional objective is $\mathcal{L}(\boldsymbol{W}) = \widetilde{\mathcal{L}}\big(\boldsymbol{W}, \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))\big)$.

The three terms in equation 9 are the retained-state residual, the dropped-state residual, and their shared mask-induced variance. Because $q\boldsymbol{W} + p\boldsymbol{D} = \boldsymbol{D} + q(\boldsymbol{W} - \boldsymbol{D})$ and $q\boldsymbol{W} - q\boldsymbol{D} = q(\boldsymbol{W} - \boldsymbol{D})$, the two residuals equal $\boldsymbol{R} - \boldsymbol{R}\boldsymbol{D} - q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D})$ and $\boldsymbol{R} - q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D})$. The self-reconstruction coefficients therefore enter only the retained state, and its weight $q\alpha^2$ controls how strongly they are fitted.

Lemma 3.2 (Conditional decomposition)**.** *Let $a > 0$, $b \geq 0$, and $\alpha = b/a$. For every $\boldsymbol{W} \in \mathbb{R}^{n\times n}$,*

$$\mathcal{L}_{\mathrm{EDLAE}}(\boldsymbol{W}) = a^2\, \mathcal{L}(\boldsymbol{W}). \tag{10}$$

The proof in Appendix B.1 sums the per-pair losses over (u, i) and converts the result to matrix form. Since a^2 does not depend on $\boldsymbol{W}$, the minimizers of $\mathcal{L}_{\mathrm{EDLAE}}$ depend on (a, b) only through α.

3.2 Free Diagonal: A Whole-Matrix Form of DEQL

Assumption 3.3. $0 < p < 1$, and every item has at least one interaction, so that $\boldsymbol{N} \succ 0$.

With the diagonal left free, we minimize the masked conditional objective over $\boldsymbol{W} \in \mathbb{R}^{n\times n}$. Let

$$\boldsymbol{K} = q\boldsymbol{G} + p\boldsymbol{N}, \qquad \boldsymbol{P} = \boldsymbol{K}^{-1}, \qquad c = \frac{pq(\alpha^2 - 1)}{s}, \tag{11}$$

where $\boldsymbol{K}$ is positive definite under Assumption 3.3. The solution follows in three steps.

Step 1: Eliminating the diagonal constraint. We treat $\boldsymbol{D}$ as a separate diagonal variable in $\widetilde{\mathcal{L}}$ and enforce $\boldsymbol{D} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))$ with a diagonal multiplier. The multiplier appears only on the diagonal of the stationarity equation for $\boldsymbol{W}$. Adding the stationarity equation for $\boldsymbol{D}$ cancels it and fixes that diagonal; combined with the off-diagonal part, this leaves the linear matrix equation (Lemma B.1)

$$\boldsymbol{G} = \boldsymbol{K}\boldsymbol{W} + c(\boldsymbol{G} - \boldsymbol{N})\boldsymbol{D}. \tag{12}$$

The diagonal enters only through the coupling term $c(\boldsymbol{G} - \boldsymbol{N})\boldsymbol{D}$, which vanishes at $\alpha = 1$.

Step 2: Column-wise rank-one structure. Let $\boldsymbol{e}^{(i)}$ be the i-th coordinate vector. The i-th column of $\boldsymbol{D}$ is $W_{i,i}\boldsymbol{e}^{(i)}$, so the i-th column of equation 12 reads

$$\boldsymbol{G}_{:,i} = \big(\boldsymbol{K} + c(\boldsymbol{G} - \boldsymbol{N})_{:,i}\boldsymbol{e}^{(i)\top}\big)\boldsymbol{W}_{:,i}. \tag{13}$$

All n column systems share $\boldsymbol{K}$ and differ only by a rank-one term.

Step 3: Sherman–Morrison. A rank-one change of a shared matrix is inverted by the Sherman–Morrison formula. This gives the tool below, which holds for any coupling matrix $\boldsymbol{M}$ and is reused in Section 4.

Lemma 3.4 (Rank-one column lemma)**.** *Let $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ be invertible, $\boldsymbol{P} = \boldsymbol{K}^{-1}$, and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$. The equation $\boldsymbol{G} = \boldsymbol{K}\boldsymbol{W} + \boldsymbol{M}\,\mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))$ has a unique solution if and only if $1 + (\boldsymbol{P}\boldsymbol{M})_{i,i} \neq 0$ for every i. Its columns are then*

$$\boldsymbol{W}_{:,i} = (\boldsymbol{P}\boldsymbol{G})_{:,i} - (\boldsymbol{P}\boldsymbol{M})_{:,i}\,\frac{(\boldsymbol{P}\boldsymbol{G})_{i,i}}{1 + (\boldsymbol{P}\boldsymbol{M})_{i,i}}, \qquad i = 1, \ldots, n. \tag{14}$$

The solvability condition follows from the matrix determinant lemma and the column formula from Sherman–Morrison (Appendix B.3). Applying Lemma 3.4 with $\boldsymbol{M} = c(\boldsymbol{G} - \boldsymbol{N})$ and stacking the columns gives the free-diagonal solution.

Proposition 3.5 (Free-diagonal solution)**.** *Under Assumption 3.3 and $\alpha > 0$, the masked conditional objective has a unique minimizer $\boldsymbol{W}^*$. It is the unique solution of equation 12 and equals*

$$\boldsymbol{W}^* = \boldsymbol{P}\boldsymbol{G} - c\,\boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N})\,\mathrm{Diag}\big(\mathrm{diag}(\boldsymbol{P}\boldsymbol{G})\big)\big[\boldsymbol{I} + c\,\mathrm{Diag}\big(\mathrm{diag}(\boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N}))\big)\big]^{-1}. \tag{15}$$

Moreover, $\boldsymbol{W}^*$ equals the column-wise DEQL solution.

Only $\boldsymbol{P}$ requires an inversion: $\boldsymbol{P}\boldsymbol{G}$ and $\boldsymbol{P}(\boldsymbol{G}-\boldsymbol{N})$ follow from it in $O(n^2)$ operations (Appendix B.4). Hence $\boldsymbol{W}^*$ costs one $n \times n$ inversion, whereas the column-wise algorithm of DEQL also forms two dense $n \times n$ products (Table 2).

3.3 Zero Diagonal: Recovering EDLAE

We now impose the zero-diagonal constraint $\mathrm{diag}(\boldsymbol{W}) = \boldsymbol{0}$ used by EASE and EDLAE. The solution follows in two steps.

Step 1: Removing the emphasis ratio. With $\boldsymbol{D} = \boldsymbol{0}$, the retained- and dropped-state residuals in equation 9 coincide, and the objective becomes

$$\mathcal{L}(\boldsymbol{W}) = \widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{0}) = s\Big(\|\boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W}\|_F^2 + pq\|\boldsymbol{N}^{1/2}\boldsymbol{W}\|_F^2\Big). \tag{16}$$

The emphasis ratio enters only through the positive factor s, so it does not affect the minimizer.

Step 2: Enforcing the zero diagonal. As in EASE, we enforce the constraint with a diagonal multiplier. The stationarity condition of equation 16 then reads

$$\boldsymbol{K}\boldsymbol{W} = \boldsymbol{G} - \boldsymbol{\Gamma} \tag{17}$$

for a diagonal matrix $\boldsymbol{\Gamma}$, so that $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{G} - \boldsymbol{P}\boldsymbol{\Gamma}$. The constraint fixes the diagonal of $\boldsymbol{\Gamma}$ entrywise as $\mathrm{diag}(\boldsymbol{\Gamma}) = \mathrm{diag}(\boldsymbol{P}\boldsymbol{G}) \oslash \mathrm{diag}(\boldsymbol{P})$, where $\oslash$ denotes element-wise division.

Proposition 3.6 (Zero-diagonal solution)**.** *Under Assumption 3.3 and $\alpha \geq 0$, the masked conditional objective restricted to $\mathrm{diag}(\boldsymbol{W}) = \boldsymbol{0}$ has the unique minimizer*

$$\boldsymbol{W}^*_{\mathrm{diag}=0} = \boldsymbol{P}\boldsymbol{G} - \boldsymbol{P}\,\mathrm{Diag}\big(\mathrm{diag}(\boldsymbol{P}\boldsymbol{G}) \oslash \mathrm{diag}(\boldsymbol{P})\big). \tag{18}$$

It does not depend on α, and $q\boldsymbol{W}^*_{\mathrm{diag}=0}$ is the full emphasis EDLAE solution of Steck (2020). At $\alpha = 0$, the minimizers of the masked conditional objective without this constraint are exactly $\boldsymbol{W}^*_{\mathrm{diag}=0} + \boldsymbol{D}$ for diagonal $\boldsymbol{D}$.

The last statement is the non-uniqueness that DEQL establishes at $b = 0$; full emphasis EDLAE selects its zero-diagonal member.

4 Unified Objective with Diagonal Control

Proposition 3.5 leaves the diagonal free, as in DEQL and DLAE, whereas Proposition 3.6 fixes it to zero, as in EASE and EDLAE. Adding a diagonal-only penalty to the masked conditional objective yields the unified objective: zero weight gives the free-diagonal solution, the infinite weight limit gives the zero-diagonal solution, and every intermediate weight changes only the coupling matrix in Lemma 3.4, preserving the closed-form structure (Section 4.1). Section 4.2 adds full-matrix L_2 regularization to obtain the model used in our experiments.

4.1 Interpolating with a Diagonal Penalty

We add a penalty on the diagonal of the masked conditional objective:

$$\begin{aligned} \widetilde{\mathcal{J}}(\boldsymbol{W}, \boldsymbol{D}) &= \widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{D}) + \lambda_D\|\boldsymbol{D}\|_F^2, \\ \mathcal{J}(\boldsymbol{W}) &= \widetilde{\mathcal{J}}\big(\boldsymbol{W}, \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))\big), \qquad \lambda_D \geq 0. \end{aligned} \tag{19}$$

The diagonal-multiplier elimination in Section 3.2 still applies (Lemma B.1). The stationarity equation retains the form in Lemma 3.4: the shared matrix $\boldsymbol{K}$ is unchanged, while the coupling matrix becomes

$$\boldsymbol{M}_{\lambda_D} = c(\boldsymbol{G} - \boldsymbol{N}) + \frac{\lambda_D}{q\alpha^2}\boldsymbol{I}. \tag{20}$$

The penalty thus shifts only the diagonal of the coupling matrix.

Theorem 4.1 (Unified closed form)**.** *Under Assumption 3.3, $\alpha > 0$, and $\lambda_D \geq 0$, the unified objective equation 19 has a unique minimizer*

$$\boldsymbol{W}^*_{\lambda_D} = \boldsymbol{P}\boldsymbol{G} - \boldsymbol{P}\boldsymbol{M}_{\lambda_D}\operatorname{Diag}\big(\operatorname{diag}(\boldsymbol{P}\boldsymbol{G})\big) \cdot \big[\boldsymbol{I} + \operatorname{Diag}\big(\operatorname{diag}(\boldsymbol{P}\boldsymbol{M}_{\lambda_D})\big)\big]^{-1}. \tag{21}$$

At $\lambda_D = 0$ it equals equation 15. As $\lambda_D \to \infty$, it converges to $\boldsymbol{W}^*_{\mathrm{diag}=0}$ in equation 18.

The proof is in Appendix B.6.

As λ_D grows, each column correction approaches the corresponding column of $\boldsymbol{P}$ normalized by $P_{i,i}$. Because $\boldsymbol{P}$ is independent of λ_D and $\boldsymbol{P}\boldsymbol{M}_{\lambda_D}$ follows from $\boldsymbol{P}$ and $\boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N})$ in $O(n^2)$ operations, the entire penalty path requires only one inversion.

4.2 Adding Full-Matrix Regularization

The diagonal penalty controls self-reconstruction, while full-matrix regularization shrinks every coefficient, as in DEQL(L2) and EASE.

$$\begin{aligned} \widetilde{\mathcal{J}}_{L_2}(\boldsymbol{W}, \boldsymbol{D}) &= \widetilde{\mathcal{J}}(\boldsymbol{W}, \boldsymbol{D}) + \lambda_W \|\boldsymbol{W}\|_F^2, \\ \mathcal{J}_{L_2}(\boldsymbol{W}) &= \widetilde{\mathcal{J}}_{L_2}\big(\boldsymbol{W}, \operatorname{Diag}(\operatorname{diag}(\boldsymbol{W}))\big), \qquad \lambda_W \geq 0. \end{aligned} \tag{22}$$

The full-matrix penalty shifts the shared system matrix and increases the effective diagonal penalty; the stationarity equation still has the form in Lemma 3.4 (Lemma B.1). Define

$$\begin{aligned} \widetilde{\lambda}_W &= \frac{\lambda_W}{qs}, & \rho &= \frac{\lambda_D + \lambda_W}{q\alpha^2}, \\ \boldsymbol{K}_F &= \boldsymbol{K} + \widetilde{\lambda}_W \boldsymbol{I}, & \boldsymbol{P}_F &= \boldsymbol{K}_F^{-1}, \\ \boldsymbol{M}_F &= c(\boldsymbol{G} - \boldsymbol{N}) + (\rho - \widetilde{\lambda}_W)\boldsymbol{I}. \end{aligned} \tag{23}$$

Proposition 4.2 (Full-matrix regularization)**.** *Under Assumption 3.3, $\alpha > 0$, and $\lambda_D, \lambda_W \geq 0$, the objective equation 22 has a unique minimizer*

$$\boldsymbol{W}^*_{L_2} = \boldsymbol{P}_F\boldsymbol{G} - \boldsymbol{P}_F\boldsymbol{M}_F\operatorname{Diag}\big(\operatorname{diag}(\boldsymbol{P}_F\boldsymbol{G})\big) \cdot \big[\boldsymbol{I} + \operatorname{Diag}\big(\operatorname{diag}(\boldsymbol{P}_F\boldsymbol{M}_F)\big)\big]^{-1}. \tag{24}$$

At $\lambda_W = 0$ it equals equation 21. At $\lambda_D = 0$ it equals the DEQL(L2) solution with regularization weight λ_W. As $\lambda_D \to \infty$, it converges to $\arg\min_{\operatorname{diag}(\boldsymbol{W})=\boldsymbol{0}}\big\{\mathcal{L}(\boldsymbol{W}) + \lambda_W\|\boldsymbol{W}\|_F^2\big\}$.

The proof is in Appendix B.7.

For fixed p and α, the regularization weights enter the solution only through $(\widetilde{\lambda}_W, \rho)$; fixing p and $\widetilde{\lambda}_W$ lets all α and λ_D share $\boldsymbol{P}_F$, with each additional setting requiring $O(n^2)$ operations. Table 1 summarizes the existing linear autoencoders recovered as special or limiting cases.

5 Experiments

In this section, we design and conduct experiments to answer the following research questions:

- **RQ1:** How fast and accurate is the closed-form solution on the whole matrix compared to the column-wise solution?
- **RQ2:** How does the unified objective's closed-form solution compare with existing linear autoencoders on real-world datasets?

Table 1: Special and limiting cases of the unified solution $\boldsymbol{W}^*_{L_2}$.

Model	Setting	Relation	Shown in
EASE (Steck, 2019)	$\lambda_D \to \infty, p \to 0, \lambda = \widetilde{\lambda}_W$	$=$	App. B.7
DLAE (Steck, 2020)	$\alpha = 1, \lambda_D = \lambda_W = 0$	$\simeq$	Prop. 3.5, App. B.8
DLAE (L_2) (Steck, 2020)	$\alpha = 1, \lambda_D = 0, \lambda_W > 0$	$\simeq$	Prop. 4.2, App. B.8
EDLAE (full emphasis) (Steck, 2020)	$\lambda_D \to \infty$ ($\alpha > 0$) or $\mathrm{diag}(\boldsymbol{W}) = \boldsymbol{0}$ (any α), $\lambda_W = 0$	$\simeq$	Prop. 3.6, Thm. 4.1
EDLAE (L_2) (Steck, 2020)	$\lambda_D \to \infty, \lambda_W > 0$	$\simeq$	App. B.8
DEQL (Guo et al., 2026)	$\alpha = b/a, \lambda_D = \lambda_W = 0$	$=$	Prop. 3.5
DEQL (L_2) (Guo et al., 2026)	$\lambda_D = 0, \lambda_W = \lambda$	$=$	Prop. 4.2
DEQL (L_2, zero diagonal) (Guo et al., 2026)	$\lambda_D \to \infty, \lambda_W = \lambda$	$=$	App. B.8

"$=$" denotes equality and "$\simeq$" denotes equality up to a positive factor.

5.1 Experimental Setup

Datasets. We use the same nine datasets and training, validation, and test splits as DEQL (Guo et al., 2026). Under strong generalization, Games, Beauty, Gowalla-1, ML-20M, Netflix, and MSD are split by user, with disjoint users across the three sets (Steck, 2019; Ni et al., 2019; Seol et al., 2024). Under weak generalization, Amazon-Book, Yelp2018, and Gowalla are split by interaction: users may appear in multiple sets, but their interactions do not overlap (He et al., 2020). Following prior recommender-system studies, we report results under the two protocols separately. Dataset details could be found in Appendix Table 8.

Baselines and Metrics. For the full-rank evaluation, we compare our method with DLAE and EDLAE (Steck, 2020), EASE (Steck, 2019), RLAE and RDLAE (Moon et al., 2023), and DEQL(L_2) (Guo et al., 2026). Under weak generalization, we also compare with PinSage, LightGCN, DGCF, SGL-ED, SimpleX, and SSM (MF and GNN); these results are taken from DEQL (Guo et al., 2026) on the same data splits, whereas all LAE-based results come from our own runs. We tune the hyperparameters of each baseline model on the range of its original paper. All methods are evaluated on the same data partitions using Recall@20 (R@20) and NDCG@20 (N@20).

Our Methods. We evaluate three full-rank instances: Uni-Sol (free diagonal, $\lambda_D = \lambda_W = 0$; Proposition 3.5), Uni-$L_2$, ($\lambda_D \to \infty$) (zero-diagonal limit; Proposition 4.2), and Uni-L_2 (diagonal and full-matrix regularization; Proposition 4.2).

Implementation Details. All experiments are conducted on a server running Linux, equipped with two Intel Xeon Platinum 8268 CPUs at 2.90 GHz (48 physical cores in total), four NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory each, and 502 GiB of system memory. We tune the emphasis ratio over $\alpha = b/a \in \{0.05, 0.10, 0.20, \ldots, 2.00, 3.00, 4.00\}$, both regularization coefficients over $\lambda_W, \lambda_D \in \{5, 10, 20, \ldots, 100, 200, 500\}$, and the dropout probability over $p \in \{0.05, 0.10, 0.20, \ldots, 0.80, 0.90\}$. All hyperparameters are selected exclusively on the validation set. After model selection, the chosen configuration is evaluated on the test set exactly once.

5.2 Experimental Results

RQ1: Computational Efficiency and Accuracy. Table 2 reports the numerical agreement between the column-wise DEQL solver and our whole-matrix closed-form solver, Uni-Sol, together with their computational efficiency. We evaluate equivalence and efficiency in separate experiments. For the equivalence experiment, let $\Delta\boldsymbol{W} = \boldsymbol{W}_{\text{Uni-Sol}} - \boldsymbol{W}_{\text{DEQL}}$ and $\Delta\widehat{\boldsymbol{S}} = \widehat{\boldsymbol{S}}_{\text{Uni-Sol}} - \widehat{\boldsymbol{S}}_{\text{DEQL}}$. We measure the relative differences in the weights and prediction scores by $\|\Delta\boldsymbol{W}\|_F / \|\boldsymbol{W}_{\text{Uni-Sol}}\|_F$ and $\|\Delta\widehat{\boldsymbol{S}}\|_F / \|\widehat{\boldsymbol{S}}_{\text{Uni-Sol}}\|_F$, respectively. Both solvers use the same hyperparameters and double-precision arithmetic. Appendix Table 5 additionally reports the Top-20 set agreement between the two solvers and their absolute Peak RSS.

Table 2: Numerical agreement and computational efficiency of the column-wise DEQL solver and the whole-matrix Uni-Sol solver under strong and weak generalization settings.

Setting	Dataset	#Items (n)	Numerical agreement		Runtime			Peak ΔRSS		
			Rel. weight difference	Rel. score difference	DEQL (s)	Uni-Sol (s)	Speedup (×)	DEQL (GiB)	Uni-Sol (GiB)	Red. (%)
Strong	Games	1,001	1.22e-15	9.17e-16	0.176	0.067	2.63	0.09	0.06	31.15
	Beauty	16,175	2.40e-15	1.51e-15	43.984	28.113	1.56	17.85	9.80	45.13
	Gowalla-1	28,969	3.04e-15	2.03e-15	172.376	103.693	1.66	57.21	31.40	45.11
	ML-20M	20,108	8.65e-15	5.98e-15	71.820	43.744	1.64	27.55	15.14	45.07
	Netflix	17,769	3.01e-14	3.56e-14	53.524	35.538	1.51	21.57	11.80	45.29
	MSD	41,140	4.37e-15	4.03e-15	414.655	225.952	1.84	115.25	63.22	45.14
Weak	Amazon-Book	91,599	5.34e-15	1.56e-15	4075.617	2327.268	1.75	319.30	310.37	2.80
	Yelp2018	38,048	3.80e-15	1.30e-15	337.768	196.047	1.72	98.58	54.04	45.18
	Gowalla	40,981	4.55e-15	1.26e-15	401.372	258.889	1.55	114.33	62.74	45.12

Table 3: Performance comparison under strong generalization. Boldface and underlining mark the best and second-best results within each block.

Model	Games		Beauty		Gowalla-1		ML-20M		Netflix		MSD	
	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
Full-rank methods												
DLAE	0.2566	0.1585	0.1322	0.0869	0.2271	0.2042	0.3924	0.3409	0.3617	0.3394	0.3291	0.3211
EASE	0.2725	0.1639	0.1304	0.0864	0.2211	0.1979	0.3905	0.3390	0.3618	0.3382	0.3341	0.3263
EDLAE	0.2546	0.1603	0.1328	0.0870	0.2268	0.2012	0.3934	**0.3428**	0.3655	0.3423	**0.3343**	**0.3268**
RLAE	0.2732	0.1734	0.1301	0.0864	0.2156	0.1948	0.3832	0.3327	0.3559	0.3330	0.3330	0.3254
RDLAE	0.2546	0.1578	**0.1361**	0.0864	0.2259	0.2018	0.3883	0.3388	0.3617	0.3393	0.3333	0.3259
DEQL(L_2)	0.2715	0.1632	0.1316	0.0868	**0.2285**	**0.2054**	0.3928	0.3410	0.3657	0.3427	0.3339	0.3264
Uni-Sol/DEQL(plain)	0.2435	0.1497	0.1087	0.0662	0.2141	0.1902	0.3838	0.3354	0.3598	0.3373	0.3329	0.3256
Uni-$L_2, (\lambda_D \to \infty)$	0.2546	0.1603	0.1322	0.0871	0.2266	0.2024	0.3931	0.3424	0.3653	0.3423	0.3342	0.3264
Uni-L_2	**0.2891**	**0.1784**	0.1359	**0.0884**	0.2276	0.2046	**0.3936**	**0.3428**	**0.3659**	**0.3428**	0.3341	0.3265

We measure runtime and memory in a separate solver-only benchmark. Runtime excludes data loading, Gram-matrix construction, score computation, and ranking, while Peak ΔRSS is the maximum increase in resident set size from its value immediately before the solver starts. Uni-Sol is faster on all nine datasets and also reduces Peak ΔRSS on every dataset.

RQ2: Full-Rank Recommendation Performance. Tables 3 and 4 report strong and weak generalization, respectively. Uni-L_2 enlarges DEQL(L_2) through λ_D and recovers DLAE, EDLAE, and EASE through parameter choices or limiting cases. The results support three observations:

1. Uni-L_2 improves over DEQL(L_2) on most strong generalization datasets and matches or improves it throughout the weak generalization comparison, with weak-generalization R@20 gains of 0.6%–2.2%.
2. No fixed diagonal treatment performs best across all datasets. With λ_W retained at both endpoints, DEQL(L_2) and Uni-$L_2, (\lambda_D \to \infty)$ each have higher R@20 on three strong generalization datasets; DEQL(L_2) leads on two weak-generalization datasets and Uni-$L_2, (\lambda_D \to \infty)$ on one.
3. With an intermediate λ_D, Uni-L_2 matches or exceeds the better endpoint in 15 of 18 dataset–metric comparisons, including one tie, consistent with the validation optima between the endpoint plateaus in Figure 1.

The comparison with deep and graph-based models is dataset dependent. Uni-L_2 improves over the strongest deep baseline by 24.6% on Amazon-Book, but trails the best deep or graph model on Yelp2018 and Gowalla. This pattern may reflect differences in the relative value of item–item co-occurrence and graph structure across datasets.

The gains over strong LAE baselines are generally modest. Appendix Table 6 reports the mean and standard deviation over five random splits, and Uni-L_2 attains the highest mean in 6 of the 12 dataset–metric combinations, including ties. These gains retain a single closed-form solve; RQ1 shows that the unified solver also improves computational efficiency.

Table 4: Performance comparison under weak generalization, with deep-learning results from DEQL (Guo et al., 2026). Boldface and underlining mark the best and second-best results within each block.

Model	Amazon-Book		Yelp2018		Gowalla	
	R@20	N@20	R@20	N@20	R@20	N@20
Deep-learning methods						
PinSage	0.0282	0.0219	0.0471	0.0393	0.1380	0.1196
LightGCN	0.0411	0.0315	0.0649	0.0530	0.1830	0.1554
DGCF	0.0422	0.0324	0.0654	0.0534	0.1842	0.1561
SGL-ED	0.0478	0.0379	0.0675	0.0555	–	–
SimpleX	0.0583	**0.0468**	0.0701	0.0575	**0.1872**	0.1557
SSM (MF)	0.0473	0.0367	0.0509	0.0404	0.1231	0.0878
SSM (GNN)	**0.0590**	0.0459	**0.0737**	**0.0609**	0.1869	**0.1571**
LAE-based methods						
DLAE	0.0718	0.0580	0.0648	0.0544	0.1768	0.1474
EASE	0.0692	0.0557	0.0567	0.0487	0.1689	0.1404
EDLAE	0.0684	0.0545	0.0648	0.0544	0.1786	0.1494
RLAE	0.0722	0.0582	0.0567	0.0488	0.1660	0.1372
RDLAE	0.0698	0.0557	0.0648	0.0544	0.1768	0.1466
DEQL(L_2)	0.0727	0.0582	0.0676	**0.0572**	0.1779	0.1469
Uni-Sol/DEQL(plain)	0.0631	0.0509	0.0647	0.0543	0.1802	0.1507
Uni-L_2, ($\lambda_D \to \infty$)	0.0706	0.0563	0.0668	0.0561	0.1806	0.1506
Uni-L_2	**0.0735**	**0.0595**	**0.0680**	**0.0572**	**0.1819**	**0.1516**

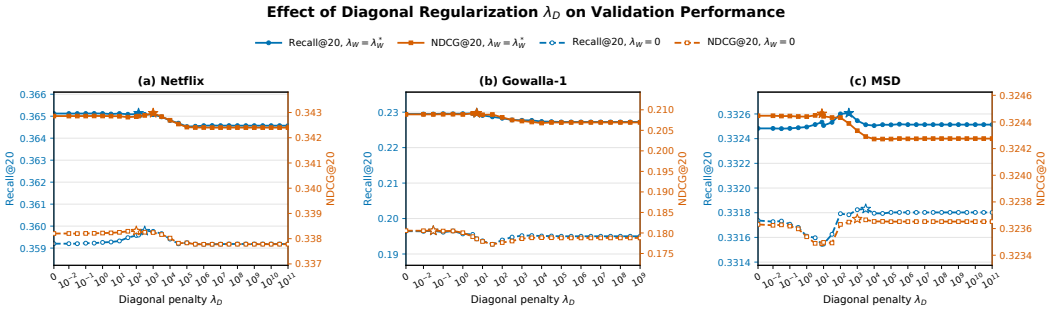


Figure 1: Validation R@20 and N@20 versus λ_D on Netflix, Gowalla-1, and MSD. Solid curves use the selected λ_W; dashed curves set $\lambda_W = 0$. Stars mark selected λ_D values.

Effect of Diagonal and Full-Matrix Regularization. All sweeps use validation performance with p and α fixed at their main-experiment values: Figure 1 scans λ_D with $\lambda_W = \lambda_W^*$ or $\lambda_W = 0$ (all six datasets appear in Appendix Figure 2), and Appendix Figure 3 scans $\widetilde{\lambda}_W$. The low- and high-λ_D plateaus represent the free- and zero-diagonal limits in Theorem 4.1 and Proposition 4.2. The best λ_D typically lies near the transition between the plateaus and offers little improvement over the better endpoint; Games is the most sensitive dataset. Most of the gain comes from λ_W: the λ_W^* curves lie above those with $\lambda_W = 0$ in Figure 1, and the λ_W scan shows a clear drop at zero relative to the selected value except on MSD, where the gap is small. The preferred λ_D can shift with λ_W, but near the selected λ_W the $\lambda_D = 0$ curves remain close to those that reselect λ_D at each point, showing that diagonal regularization fine-tunes the gains in observation 3.

6 Conclusion

In this work, we introduce a unified analytical framework for understanding how masking, emphasis weighting, and diagonal regularization jointly govern self-reconstruction and prediction across features. The explicit matrix solution makes their interactions and limiting behavior directly analyzable. Beyond recommendation, this framework offers a starting point for studying masked learning in images and language. Future work could extend the analysis to structured masking of image patches and token representations, examining how masking patterns shape contextual prediction and which information survives in low-rank representations. Understanding when reconstruction preserves information useful for downstream tasks could provide theoretical insights and testable hypotheses for nonlinear masked models.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this paper and to retrieve and discover related literature. We reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL https://icml.cc/2012/papers/416.pdf.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on Top-N recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, pp. 39–46. Association for Computing Machinery, 2010. doi: 10.1145/1864708.1864721.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109, 2019. doi: 10.1145/3298689.3347058. URL https://doi.org/10.1145/3298689.3347058.

Ruixin Guo, Xinyu Li, Hao Zhou, Yang Zhou, and Ruoming Jin. Generalizing linear autoencoder recommenders with decoupled expected quadratic loss. In *International Conference on Learning Representations*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 173–182, 2017. doi: 10.1145/3038912.3052569. URL https://doi.org/10.1145/3038912.3052569.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Seoyoung Hong, Jeongwhan Choi, Yeon-Chang Lee, Srijan Kumar, and Noseong Park. SVD-AE: simple autoencoders for collaborative filtering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/227. URL https://doi.org/10.24963/ijcai.2024/227.

Ruoming Jin, Dong Li, Jing Gao, Zhi Liu, Li Chen, and Yang Zhou. Towards a better understanding of linear models for recommendation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 776–785, 2021.

Dong Li, Zhenming Liu, Ruoming Jin, Hao Zhou, Zhi Liu, Jing Gao, and Bin Ren. On the regularization landscape for the linear recommendation models, 2026. URL https://arxiv.org/abs/2609.11876.

Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, pp. 689–698, 2018. doi: 10.1145/3178876.3186150.

Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 1243–1252, 2021. doi: 10.1145/3459637.3482297. URL https://doi.org/10.1145/3459637.3482297.

Jaewan Moon, Hye-young Kim, and Jongwuk Lee. It's enough: Relaxing diagonal constraints in linear autoencoders for recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 1639–1648, 2023.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Xia Ning and George Karypis. SLIM: Sparse linear methods for Top-N recommender systems. In *2011 IEEE 11th international conference on data mining*, pp. 497–506. IEEE, 2011.

Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. AutoRec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pp. 111–112, 2015. doi: 10.1145/2740908.2742726. URL https://doi.org/10.1145/2740908.2742726.

Jinseok Seol, Minseok Gang, Sang-goo Lee, and Jaehui Park. Proxy-based item representation for attribute and context-aware recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 616–625, 2024.

Ilya Shenbin and Sergey Nikolenko. ImplicitSLIM and how it improves embedding-based collaborative filtering. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6vF0ZJGor4.

Martin Spišák, Radek Bartyzal, Antonín Hoskovec, Ladislav Peška, and Miroslav Tůma. Scalable approximate NonSymmetric autoencoder for collaborative filtering. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 763–770, 2023. doi: 10.1145/3604915.3608827. URL https://doi.org/10.1145/3604915.3608827.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL https://jmlr.org/papers/v15/srivastava14a.html.

Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The world wide web conference*, pp. 3251–3257, 2019.

Harald Steck. Autoencoders that don't overfit towards the identity. *Advances in Neural Information Processing Systems*, 33:19598–19608, 2020.

Harald Steck and Dario Garcia Garcia. On the regularization of autoencoders, 2021. URL https://arxiv.org/abs/2110.11402.

Vojtěch Vančura, Rodrigo Alves, Petr Kasalický, and Pavel Kordík. Scalable linear shallow autoencoder for collaborative filtering. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pp. 604–609, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3551482. URL https://doi.org/10.1145/3523227.3551482.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL https://jmlr.org/papers/v11/vincent10a.html.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26, pp. 351–359, 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/38db3aed920cf82ab059bfccbd02be6a-Abstract.html.

Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1001–1010, 2020. doi: 10.1145/3397271.3401137. URL https://doi.org/10.1145/3397271.3401137.

Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 726–735, 2021. doi: 10.1145/3404835.3462862. URL https://doi.org/10.1145/3404835.3462862.

Jiancan Wu, Xiang Wang, Xingyu Gao, Jiawei Chen, Hongcheng Fu, and Tianyu Qiu. On the effectiveness of sampled softmax loss for item recommendation. *ACM Transactions on Information Systems*, 42(4):98:1–98:26, 2024. doi: 10.1145/3637061. URL https://doi.org/10.1145/3637061.

Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. Collaborative denoising auto-encoders for Top-N recommender systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 153–162, 2016. doi: 10.1145/2835776.2835837. URL https://doi.org/10.1145/2835776.2835837.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 974–983, 2018. doi: 10.1145/3219819.3219890. URL https://doi.org/10.1145/3219819.3219890.

A Related Work

Denoising, Emphasis, and Marginalization. Denoising autoencoders reconstruct clean data from corrupted inputs, and emphasized denoising assigns different reconstruction weights to corrupted and uncorrupted target dimensions (Vincent et al., 2008; 2010). For linear denoisers, marginalizing the corruption yields a deterministic problem with a closed-form solution (Chen et al., 2012). Dropout is a stochastic regularizer (Srivastava et al., 2014), with a data-adaptive interpretation in generalized linear models (Wager et al., 2013). We also integrate out the mask exactly, but first condition on whether the target interaction is retained or dropped. The resulting variance contains only off-diagonal coefficients, while self-reconstruction appears only in the retained-state residual, where its emphasis weight acts (Section 3.1).

Neural and Graph-Based Recommendation. AutoRec, CDAE, and Mult-VAE are neural autoencoder recommenders (Sedhain et al., 2015; Wu et al., 2016; Liang et al., 2018). NCF, PinSage, DGCF, and LightGCN use neural interaction or graph representations (He et al., 2017; Ying et al., 2018; Wang et al., 2020; He et al., 2020). SGL-ED, SimpleX, and SSM pursue self-supervised or ranking objectives (Wu et al., 2021; Mao et al., 2021; Wu et al., 2024). Both CDAE and SGL-ED sample masks during training: the former corrupts input interactions, and the latter applies edge dropout to graph views (Wu et al., 2016; 2021). Our LAE instead integrates out the interaction mask to obtain a deterministic objective and a closed-form solution. These neural and graph-based methods do not expose a single item–item coefficient matrix with a separately controllable diagonal. Re-evaluations of neural recommendation have also shown the importance of strong non-neural baselines (Ferrari Dacrema et al., 2019).

Linear Item–Item Recommendation. SLIM fits sparse item–item regressions with a zero-diagonal constraint, while EASE retains that constraint and obtains a closed-form ridge solution (Ning & Karypis, 2011; Steck, 2019). In our objective, the zero-diagonal solution is the $\lambda_D \to \infty$ endpoint (Theorem 4.1). With full-matrix ridge regularization, its $p \to 0$ limit is standard EASE with ridge weight $\widetilde{\lambda}_W$ (Section 4.2; Appendix B.7); without ridge, the limit is unregularized EASE. The hard constraint is thus one end of a continuous penalty path. SLIM's sparsity and nonnegativity constraints remain outside this framework. Masking removes an item's own input when its target interaction is dropped, limiting self-reconstruction without forbidding it for retained targets.

Emphasis and Diagonal Treatment in LAEs. DLAE applies dropout to input interactions and induces data-dependent regularization (Steck, 2020). At $\alpha = 1$, the diagonal coupling in our free-diagonal solution vanishes, and scaling that solution by q gives DLAE (Appendix B.8). EDLAE further emphasizes dropped targets. Under full emphasis, $b = 0$, its zero-diagonal solution and ours differ by a positive factor, while its unconstrained minimizers have an arbitrary diagonal (Proposition 3.6). The same work expands the expected loss for $a \geq b$ into a deterministic matrix quadratic, where a, b, and p fix the fraction subtracted from the diagonal, but gives a closed-form solution for emphasized reconstruction only when $b = 0$. That objective and our conditional decomposition have the same minimizers (Lemma 3.2). Separating the diagonal yields a whole-matrix closed form for any $\alpha > 0$, including $\alpha > 1$, and permits a penalty λ_D independent of emphasis. RLAE and RDLAE use a tunable upper bound on diagonal coefficients to interpolate between zero-diagonal EDLAE and unconstrained DLAE, with closed-form solutions but without emphasis (Moon et al., 2023). Their zero-diagonal endpoint agrees with ours up to a positive factor, whereas their unconstrained DLAE endpoint matches our free-diagonal solution only at $\alpha = 1$; for general α, our free-diagonal solution equals DEQL (Proposition 3.5). Our quadratic penalty varies the diagonal within the emphasized objective (Section 4). These objectives leave emphasis and an independent, continuously tunable diagonal penalty uncombined.

General Expected Quadratic Objectives. DEQL expresses emphasized denoising through target-specific expected quadratic losses, allowing an arbitrary distribution for each target item, with EDLAE as a special case (Guo et al., 2026). For the dropout objective, it derives a unique column solution when $b > 0$, including $b > a$, and uses a shared inverse to assemble the columns in $O(n^3)$. It also studies full-matrix L_2 regularization and hard zero-diagonal constraints. At $\lambda_D = 0$, the solution in Proposition 4.2 equals DEQL(L_2), while its $\lambda_D \to \infty$ endpoint equals zero-diagonal DEQL(L_2) (Table 1). Proposition 3.5 gives the same minimizer as a whole-matrix formula: target-

specific corrections become one matrix correction computed from diagonal quantities after a shared inverse (Section 3.2). We compare these solver structures in RQ1 (Table 2). DEQL covers arbitrary target distributions. Our dropout-specific objective adds independent diagonal control through λ_D (Section 4).

Low-Rank and Scalable Linear Recommendation. PureSVD and low-rank EDLAE construct low-rank solutions directly (Cremonesi et al., 2010; Steck, 2020), while ELSA, SANSA, SVD-AE, and ImplicitSLIM pursue compact or sparse item transformations (Vančura et al., 2022; Spišák et al., 2023; Hong et al., 2024; Shenbin & Nikolenko, 2024). Our approximations instead project a full-rank teacher without low-rank training, and therefore require the full-rank solve first (Appendix E).

The spectral connection between regression-based recommenders and matrix factorization studied by Jin et al. (2021) informs the full-rank-teacher approximation of Steck & Garcia (2021), which uses a zero-diagonal proxy for EDLAE. Li et al. (2026) study weighted teacher approximation and direct truncated SVD. When $\alpha \neq 1$ or $\lambda_D > 0$, the objective treats diagonal coefficients separately, and their values in a rank-constrained matrix vary with its factors, so the exact problem does not reduce to truncated SVD. DEQL's exact low-rank closed form likewise applies to its dropout objective only when $a = b$ (Guo et al., 2026). Our two approximation routes follow Li et al., but target the full-rank minimizer of $\mathcal{J}_{L_2}$, which includes emphasis and a diagonal penalty. Relative to Steck's zero-diagonal proxy, $\boldsymbol{W}^*_{L_2}$-approx uses the teacher's diagonal; $\boldsymbol{W}^*_{L_2}$-SVD truncates the teacher directly (Appendix D).

B Proofs

B.1 Conditional Decomposition and Matrix-Form Objective

Proof of Lemma 3.2. Expanding the Frobenius norm in equation 1 entrywise gives

$$\mathcal{L}_{\mathrm{EDLAE}}(\boldsymbol{W}) = \sum_{u=1}^{m}\sum_{i=1}^{n} \mathbb{E}_{\Delta\sim\mathcal{B}}\left[A_{u,i}^2\Big(R_{u,i} - \sum_{k=1}^{n}\delta_{uk}R_{u,k}W_{k,i}\Big)^2\right]. \tag{25}$$

Throughout, $\boldsymbol{R}$ and $\boldsymbol{W}$ are fixed, and all randomness comes from Δ. Let $q = 1 - p$. For a fixed user–item pair (u, i), define the weighted target and prediction as

$$Y_{ui} = A_{u,i}R_{u,i}, \qquad \widehat{Y}_{ui} = A_{u,i}\sum_{k=1}^{n}\delta_{uk}R_{u,k}W_{k,i}. \tag{26}$$

The target Y_{ui} is random before conditioning because $A_{u,i} = b$ when $\delta_{ui} = 1$ and $A_{u,i} = a$ when $\delta_{ui} = 0$. Its value is fixed within either conditional state.

For $d \in \{0, 1\}$, write $\mu_{ui}^{(d)} = \mathbb{E}_\Delta[\widehat{Y}_{ui} \mid \delta_{ui} = d]$. Adding and subtracting this conditional mean gives

$$\begin{aligned}
&\mathbb{E}_\Delta\Big[(Y_{ui} - \widehat{Y}_{ui})^2 \mid \delta_{ui} = d\Big] \\
&= \mathbb{E}_\Delta\Big[\big(Y_{ui} - \mu_{ui}^{(d)} + \mu_{ui}^{(d)} - \widehat{Y}_{ui}\big)^2 \mid \delta_{ui} = d\Big] \\
&= (Y_{ui} - \mu_{ui}^{(d)})^2 + 2(Y_{ui} - \mu_{ui}^{(d)})\mathbb{E}_\Delta[\mu_{ui}^{(d)} - \widehat{Y}_{ui} \mid \delta_{ui} = d] \\
&\quad + \mathbb{E}_\Delta\Big[(\widehat{Y}_{ui} - \mu_{ui}^{(d)})^2 \mid \delta_{ui} = d\Big] \\
&= (Y_{ui} - \mu_{ui}^{(d)})^2 + \mathrm{Var}(\widehat{Y}_{ui} \mid \delta_{ui} = d).
\end{aligned} \tag{27}$$

The cross term vanishes because $\mathbb{E}_\Delta[\widehat{Y}_{ui} \mid \delta_{ui} = d] = \mu_{ui}^{(d)}$. We next evaluate the conditional mean and variance in equation 27.

Since the mask entries are independent Bernoulli variables, for every $k \neq i$ and either $d \in \{0, 1\}$,

$$\begin{aligned}
&\mathbb{E}_\Delta[\delta_{uk} \mid \delta_{ui} = d] = q, && \mathbb{E}_\Delta[\delta_{uk}^2 \mid \delta_{ui} = d] = q, \\
&\mathrm{Var}(\delta_{uk} \mid \delta_{ui} = d) = q - q^2 = pq, && \mathrm{Cov}(\delta_{uk}, \delta_{u\ell} \mid \delta_{ui} = d) = 0 \quad (k \neq \ell,\ k, \ell \neq i).
\end{aligned} \tag{28}$$

Here $\delta_{uk}^2 = \delta_{uk}$, and conditioning on the independent target mask does not change the distribution of the other input masks.

When the target is retained, $\delta_{ui} = 1$ and $Y_{ui} = bR_{u,i}$. Separating the self-reconstruction term from the remaining inputs yields

$$\begin{aligned}\widehat{Y}_{ui}\Big|_{\delta_{ui}=1} &= bR_{u,i}W_{i,i} + b\sum_{k\neq i}\delta_{uk}R_{u,k}W_{k,i},\\ \mu_{ui}^{(1)} &= bR_{u,i}W_{i,i} + bq\sum_{k\neq i}R_{u,k}W_{k,i} = b\widehat{r}_{u,i}^{(1)}.\end{aligned} \tag{29}$$

When the target is dropped, $\delta_{ui} = 0$ and $Y_{ui} = aR_{u,i}$. The self-reconstruction term disappears, so

$$\begin{aligned}\widehat{Y}_{ui}\Big|_{\delta_{ui}=0} &= a\sum_{k\neq i}\delta_{uk}R_{u,k}W_{k,i},\\ \mu_{ui}^{(0)} &= aq\sum_{k\neq i}R_{u,k}W_{k,i} = a\widehat{r}_{u,i}^{(0)}.\end{aligned} \tag{30}$$

These are precisely the two unweighted conditional estimators in equation 4 and equation 5.

The random part of either conditional prediction is the sum over $k \neq i$. Expanding its variance before using independence gives

$$\begin{aligned}&\mathrm{Var}\left(\sum_{k\neq i}\delta_{uk}R_{u,k}W_{k,i}\,\middle|\,\delta_{ui}=d\right)\\ &= \sum_{k\neq i}R_{u,k}^2W_{k,i}^2\,\mathrm{Var}(\delta_{uk}\mid\delta_{ui}=d)\\ &\quad + 2\sum_{\substack{k<\ell\\ k,\ell\neq i}}R_{u,k}W_{k,i}R_{u,\ell}W_{\ell,i}\,\mathrm{Cov}(\delta_{uk},\delta_{u\ell}\mid\delta_{ui}=d)\\ &= pq\sum_{k\neq i}R_{u,k}^2W_{k,i}^2 = v_{u,i}(\boldsymbol{W}).\end{aligned} \tag{31}$$

The self-reconstruction term in the retained state is constant, so it does not contribute to the variance. After accounting for the fixed emphasis weight in each state, we obtain

$$\mathrm{Var}(\widehat{Y}_{ui}\mid\delta_{ui}=1) = b^2v_{u,i}(\boldsymbol{W}), \qquad \mathrm{Var}(\widehat{Y}_{ui}\mid\delta_{ui}=0) = a^2v_{u,i}(\boldsymbol{W}). \tag{32}$$

Substituting equation 29, equation 30, and equation 32 into equation 27 gives the two conditional losses:

$$\begin{aligned}\mathbb{E}_\Delta\Big[(Y_{ui}-\widehat{Y}_{ui})^2\mid\delta_{ui}=1\Big] &= b^2\left[(R_{u,i}-\widehat{r}_{u,i}^{(1)})^2 + v_{u,i}(\boldsymbol{W})\right],\\ \mathbb{E}_\Delta\Big[(Y_{ui}-\widehat{Y}_{ui})^2\mid\delta_{ui}=0\Big] &= a^2\left[(R_{u,i}-\widehat{r}_{u,i}^{(0)})^2 + v_{u,i}(\boldsymbol{W})\right].\end{aligned} \tag{33}$$

The law of total expectation weights these losses by $\Pr(\delta_{ui}=1) = q$ and $\Pr(\delta_{ui}=0) = p$. Thus,

$$\begin{aligned}&\mathbb{E}_\Delta[(Y_{ui}-\widehat{Y}_{ui})^2]\\ &= q\mathbb{E}_\Delta[(Y_{ui}-\widehat{Y}_{ui})^2\mid\delta_{ui}=1] + p\mathbb{E}_\Delta[(Y_{ui}-\widehat{Y}_{ui})^2\mid\delta_{ui}=0]\\ &= qb^2\left[(R_{u,i}-\widehat{r}_{u,i}^{(1)})^2 + v_{u,i}(\boldsymbol{W})\right] + pa^2\left[(R_{u,i}-\widehat{r}_{u,i}^{(0)})^2 + v_{u,i}(\boldsymbol{W})\right].\end{aligned} \tag{34}$$

Substituting $b = a\alpha$, with $\alpha = b/a$ as in Section 3.1, and factoring out a^2 gives equation 7. Summing over u and i yields

$$\begin{aligned}\mathcal{L}_{\mathrm{EDLAE}}(\boldsymbol{W}) = a^2\sum_{u=1}^{m}\sum_{i=1}^{n}\Big\{&q\alpha^2\big[(R_{u,i}-\widehat{r}_{u,i}^{(1)})^2 + v_{u,i}(\boldsymbol{W})\big]\\ &+ p\big[(R_{u,i}-\widehat{r}_{u,i}^{(0)})^2 + v_{u,i}(\boldsymbol{W})\big]\Big\}.\end{aligned} \tag{35}$$

The remaining steps show that the double sum of braced terms in equation 35 equals $\mathcal{L}(\boldsymbol{W}) = \widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{D})$, where $\boldsymbol{D} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))$.

To convert the two squared residuals into matrix form, first note that

$$[\boldsymbol{RD}]_{u,i} = R_{u,i}W_{i,i}, \qquad [\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D})]_{u,i} = \sum_{k \neq i} R_{u,k}W_{k,i}. \tag{36}$$

Consequently, the retained-target estimator can be written as

$$\begin{aligned} \widehat{r}^{(1)}_{u,i} &= [\boldsymbol{RD} + q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D})]_{u,i} \\ &= [q\boldsymbol{RW} + p\boldsymbol{RD}]_{u,i}, \end{aligned} \tag{37}$$

whereas the dropped-target estimator is

$$\begin{aligned} \widehat{r}^{(0)}_{u,i} &= [q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D})]_{u,i} \\ &= [q\boldsymbol{RW} - q\boldsymbol{RD}]_{u,i}. \end{aligned} \tag{38}$$

Subtracting these estimates from $R_{u,i}$ yields

$$\begin{aligned} R_{u,i} - \widehat{r}^{(1)}_{u,i} &= [\boldsymbol{R} - q\boldsymbol{RW} - p\boldsymbol{RD}]_{u,i}, \\ R_{u,i} - \widehat{r}^{(0)}_{u,i} &= [\boldsymbol{R} - q\boldsymbol{RW} + q\boldsymbol{RD}]_{u,i}. \end{aligned} \tag{39}$$

By the definition of the Frobenius norm, summing the squares of these entries over u, i gives the reconstruction part of the normalized loss:

$$\begin{aligned} & q\alpha^2 \sum_{u,i} (R_{u,i} - \widehat{r}^{(1)}_{u,i})^2 + p \sum_{u,i} (R_{u,i} - \widehat{r}^{(0)}_{u,i})^2 \\ &= q\alpha^2 \left\| \boldsymbol{R} - \boldsymbol{RD} - q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D}) \right\|_F^2 + p \left\| \boldsymbol{R} - q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D}) \right\|_F^2 \\ &= q\alpha^2 \left\| \boldsymbol{R} - q\boldsymbol{RW} - p\boldsymbol{RD} \right\|_F^2 + p \left\| \boldsymbol{R} - q\boldsymbol{RW} + q\boldsymbol{RD} \right\|_F^2 . \end{aligned} \tag{40}$$

It remains to collect the variance terms. In the normalized loss, their combined coefficient is $q\alpha^2 + p = s$. Using equation 6, then exchanging the finite sums, we have

$$\begin{aligned} \sum_{u=1}^{m} \sum_{i=1}^{n} (q\alpha^2 + p) v_{u,i}(\boldsymbol{W}) &= pqs \sum_{u=1}^{m} \sum_{i=1}^{n} \sum_{k \neq i} R_{u,k}^2 W_{k,i}^2 \\ &= pqs \sum_{i=1}^{n} \sum_{k \neq i} \left(\sum_{u=1}^{m} R_{u,k}^2 \right) W_{k,i}^2 \\ &= pqs \sum_{i=1}^{n} \sum_{k \neq i} G_{k,k} W_{k,i}^2. \end{aligned} \tag{41}$$

Because $\boldsymbol{G} = \boldsymbol{R}^\top \boldsymbol{R}$ and $\boldsymbol{N} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{G}))$, we have $N_{k,k} = G_{k,k}$, while $[\boldsymbol{W} - \boldsymbol{D}]_{k,i} = 0$ for $k = i$ and equals $W_{k,i}$ otherwise. Therefore,

$$\begin{aligned} \left\| \boldsymbol{N}^{1/2} (\boldsymbol{W} - \boldsymbol{D}) \right\|_F^2 &= \sum_{i=1}^{n} \sum_{k=1}^{n} N_{k,k} [\boldsymbol{W} - \boldsymbol{D}]_{k,i}^2 \\ &= \sum_{i=1}^{n} \sum_{k \neq i} G_{k,k} W_{k,i}^2. \end{aligned} \tag{42}$$

Combining equation 39, equation 41, and equation 42 gives

$$\begin{aligned} \widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{D}) = {} & q\alpha^2 \left\| \boldsymbol{R} - q\boldsymbol{RW} - p\boldsymbol{RD} \right\|_F^2 \\ & + p \left\| \boldsymbol{R} - q\boldsymbol{RW} + q\boldsymbol{RD} \right\|_F^2 \\ & + pqs \left\| \boldsymbol{N}^{1/2} (\boldsymbol{W} - \boldsymbol{D}) \right\|_F^2, \qquad \boldsymbol{D} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W})). \end{aligned} \tag{43}$$

This recovers equation 9 and gives $\mathcal{L}_{\mathrm{EDLAE}}(\boldsymbol{W}) = a^2 \mathcal{L}(\boldsymbol{W})$, as stated in equation 10. The equality holds for every $\boldsymbol{W}$: here $\boldsymbol{D}$ is its diagonal part, not an independent parameter. Because $a^2 > 0$ is independent of $\boldsymbol{W}$, the minimizers depend on (a, b) only through α. □

B.2 Stationarity and Uniqueness

Recall $q = 1 - p$, $s = q\alpha^2 + p$, and $c = pq(\alpha^2 - 1)/s$. The matrices $\boldsymbol{G}$ and $\boldsymbol{N}$ are defined in equation 8.

Lemma B.1. *Let Assumption 3.3 hold, $\alpha > 0$, and $\lambda_D, \lambda_W \geq 0$. For the objective $\mathcal{J}_{L_2}$ in equation 22, define $\widetilde{\lambda}_W = \lambda_W/(qs)$, $\rho = (\lambda_D + \lambda_W)/(q\alpha^2)$,*

$$\boldsymbol{K}_F = q\boldsymbol{G} + p\boldsymbol{N} + \widetilde{\lambda}_W \boldsymbol{I}, \qquad \boldsymbol{M}_F = c(\boldsymbol{G} - \boldsymbol{N}) + (\rho - \widetilde{\lambda}_W)\boldsymbol{I}. \tag{44}$$

These are the matrices in equation 23. Then $\boldsymbol{K}_F \succ 0$, $\mathcal{J}_{L_2}$ is strictly convex, and $\mathcal{J}_{L_2}$ has a unique minimizer. A matrix $\boldsymbol{W}$ minimizes $\mathcal{J}_{L_2}$ if and only if

$$\boldsymbol{G} = \boldsymbol{K}_F \boldsymbol{W} + \boldsymbol{M}_F \operatorname{Diag}(\operatorname{diag}(\boldsymbol{W})). \tag{45}$$

For $\lambda_D = \lambda_W = 0$, $\boldsymbol{K}_F = \boldsymbol{K}$ and $\boldsymbol{M}_F = c(\boldsymbol{G} - \boldsymbol{N})$.

Proof. Under Assumption 3.3, write $\Pi_{\mathrm{d}}(\boldsymbol{X}) = \operatorname{Diag}(\operatorname{diag}(\boldsymbol{X}))$ for the diagonal projection. We temporarily treat $\boldsymbol{W}$ and diagonal $\boldsymbol{D}$ as independent variables in $\widetilde{\mathcal{J}}_{L_2}(\boldsymbol{W}, \boldsymbol{D})$ and enforce $\boldsymbol{D} = \Pi_{\mathrm{d}}(\boldsymbol{W})$ with a diagonal multiplier $\boldsymbol{\Lambda}$.

Define the retained- and dropped-state residual matrices

$$\boldsymbol{E}_1 = \boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W} - p\boldsymbol{R}\boldsymbol{D}, \qquad \boldsymbol{E}_0 = \boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W} + q\boldsymbol{R}\boldsymbol{D}. \tag{46}$$

The constrained Lagrangian is

$$\begin{aligned}\mathcal{J}^{\mathrm{Lag}}(\boldsymbol{W}, \boldsymbol{D}, \boldsymbol{\Lambda}) = {} & q\alpha^2 \|\boldsymbol{E}_1\|_F^2 + p\|\boldsymbol{E}_0\|_F^2 + pqs\|\boldsymbol{N}^{1/2}(\boldsymbol{W} - \boldsymbol{D})\|_F^2 \\ & + \lambda_D \|\boldsymbol{D}\|_F^2 + \lambda_W \|\boldsymbol{W}\|_F^2 + 2\operatorname{tr}\left[\boldsymbol{\Lambda}^\top (\Pi_{\mathrm{d}}(\boldsymbol{W}) - \boldsymbol{D})\right].\end{aligned} \tag{47}$$

Stationarity with respect to $\boldsymbol{\Lambda}$ recovers the required constraint:

$$\nabla_{\boldsymbol{\Lambda}} \mathcal{J}^{\mathrm{Lag}} = 2(\Pi_{\mathrm{d}}(\boldsymbol{W}) - \boldsymbol{D}) = \boldsymbol{0}. \tag{48}$$

We first differentiate the four loss terms while holding the other variable fixed. Since $\mathrm{d}\boldsymbol{E}_1 = -q\boldsymbol{R}\,\mathrm{d}\boldsymbol{W} - p\boldsymbol{R}\,\mathrm{d}\boldsymbol{D}$ and $\mathrm{d}\boldsymbol{E}_0 = -q\boldsymbol{R}\,\mathrm{d}\boldsymbol{W} + q\boldsymbol{R}\,\mathrm{d}\boldsymbol{D}$, the identity $\mathrm{d}\|\boldsymbol{E}\|_F^2 = 2\operatorname{tr}(\boldsymbol{E}^\top \mathrm{d}\boldsymbol{E})$ gives

$$\begin{aligned}\nabla_{\boldsymbol{W}}\left(q\alpha^2\|\boldsymbol{E}_1\|_F^2\right) &= -2q^2\alpha^2 \boldsymbol{R}^\top \boldsymbol{E}_1 \\ &= 2q^2\alpha^2(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} + p\boldsymbol{G}\boldsymbol{D}),\end{aligned} \tag{49}$$

$$\begin{aligned}\nabla_{\boldsymbol{W}}\left(p\|\boldsymbol{E}_0\|_F^2\right) &= -2pq\boldsymbol{R}^\top \boldsymbol{E}_0 \\ &= 2pq(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} - q\boldsymbol{G}\boldsymbol{D}).\end{aligned} \tag{50}$$

Because variations of $\boldsymbol{D}$ are diagonal, its gradients are the diagonal projections of the corresponding unconstrained derivatives:

$$\begin{aligned}\nabla_{\boldsymbol{D}}\left(q\alpha^2\|\boldsymbol{E}_1\|_F^2\right) &= -2pq\alpha^2 \Pi_{\mathrm{d}}(\boldsymbol{R}^\top \boldsymbol{E}_1) \\ &= 2pq\alpha^2 \Pi_{\mathrm{d}}(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} + p\boldsymbol{G}\boldsymbol{D}),\end{aligned} \tag{51}$$

$$\begin{aligned}\nabla_{\boldsymbol{D}}\left(p\|\boldsymbol{E}_0\|_F^2\right) &= 2pq\Pi_{\mathrm{d}}(\boldsymbol{R}^\top \boldsymbol{E}_0) \\ &= 2pq\Pi_{\mathrm{d}}(\boldsymbol{G} - q\boldsymbol{G}\boldsymbol{W} + q\boldsymbol{G}\boldsymbol{D}).\end{aligned} \tag{52}$$

The variance and diagonal-penalty terms contribute

$$\begin{aligned}\nabla_{\boldsymbol{W}}\left(pqs\|\boldsymbol{N}^{1/2}(\boldsymbol{W} - \boldsymbol{D})\|_F^2\right) &= 2pqs\boldsymbol{N}(\boldsymbol{W} - \boldsymbol{D}), \\ \nabla_{\boldsymbol{D}}\left(pqs\|\boldsymbol{N}^{1/2}(\boldsymbol{W} - \boldsymbol{D})\|_F^2\right) &= -2pqs\Pi_{\mathrm{d}}\big(\boldsymbol{N}(\boldsymbol{W} - \boldsymbol{D})\big), \\ \nabla_{\boldsymbol{D}}\left(\lambda_D\|\boldsymbol{D}\|_F^2\right) &= 2\lambda_D\boldsymbol{D}, \\ \nabla_{\boldsymbol{W}}\left(\lambda_W\|\boldsymbol{W}\|_F^2\right) &= 2\lambda_W\boldsymbol{W}.\end{aligned} \tag{53}$$

Finally, the multiplier term contributes $2\boldsymbol{\Lambda}$ to the $\boldsymbol{W}$-gradient and $-2\boldsymbol{\Lambda}$ to the $\boldsymbol{D}$-gradient.

Adding these terms produces the two stationarity equations. For $\boldsymbol{W}$,

$$\begin{aligned}\mathbf{0} = {} & 2q^2\alpha^2(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} + p\boldsymbol{G}\boldsymbol{D}) \\ & + 2pq(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} - q\boldsymbol{G}\boldsymbol{D}) \\ & + 2pqs\boldsymbol{N}(\boldsymbol{W} - \boldsymbol{D}) + 2\lambda_W\boldsymbol{W} + 2\boldsymbol{\Lambda}. \end{aligned} \tag{54}$$

For the diagonal variable $\boldsymbol{D}$,

$$\begin{aligned}\mathbf{0} = {} & 2pq\alpha^2\Pi_{\mathrm{d}}(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} + p\boldsymbol{G}\boldsymbol{D}) \\ & + 2pq\Pi_{\mathrm{d}}(\boldsymbol{G} - q\boldsymbol{G}\boldsymbol{W} + q\boldsymbol{G}\boldsymbol{D}) \\ & - 2pqs\Pi_{\mathrm{d}}\big(\boldsymbol{N}(\boldsymbol{W} - \boldsymbol{D})\big) + 2\lambda_D\boldsymbol{D} - 2\boldsymbol{\Lambda}. \end{aligned} \tag{55}$$

Take the diagonal projection of equation 54 and add equation 55. The multiplier terms cancel. The dropped-state and variance terms also cancel pairwise. The two retained-state terms combine because $p + q = 1$, leaving

$$q\alpha^2\Pi_{\mathrm{d}}(-\boldsymbol{G} + q\boldsymbol{G}\boldsymbol{W} + p\boldsymbol{G}\boldsymbol{D}) + (\lambda_D + \lambda_W)\boldsymbol{D} = \mathbf{0}.$$

Since $\Pi_{\mathrm{d}}(\boldsymbol{G}) = \boldsymbol{N}$ and $\Pi_{\mathrm{d}}(\boldsymbol{G}\boldsymbol{D}) = \boldsymbol{N}\boldsymbol{D}$, dividing by $q\alpha^2$ gives

$$q\Pi_{\mathrm{d}}(\boldsymbol{G}\boldsymbol{W}) = \boldsymbol{N} - p\boldsymbol{N}\boldsymbol{D} - \rho\boldsymbol{D}. \tag{56}$$

To recover the full matrix equation, first collect the coefficients in equation 54. They satisfy $q^2\alpha^2 + pq = qs$ and $pq^2(\alpha^2 - 1) = qsc$, where $c = pq(\alpha^2 - 1)/s$. Dividing by $2qs$ then yields

$$-\boldsymbol{G} + \boldsymbol{K}_F\boldsymbol{W} + c\boldsymbol{G}\boldsymbol{D} - p\boldsymbol{N}\boldsymbol{D} + \frac{\boldsymbol{\Lambda}}{qs} = \mathbf{0}. \tag{57}$$

The last two terms involving $\boldsymbol{N}\boldsymbol{D}$ and $\boldsymbol{\Lambda}$ are diagonal. Hence $\boldsymbol{G} - \boldsymbol{K}_F\boldsymbol{W} - c\boldsymbol{G}\boldsymbol{D}$ is diagonal. Its diagonal part follows from equation 56:

$$\begin{aligned}&\Pi_{\mathrm{d}}\big(\boldsymbol{G} - \boldsymbol{K}_F\boldsymbol{W} - c\boldsymbol{G}\boldsymbol{D}\big) \\ &\quad = \boldsymbol{N} - q\Pi_{\mathrm{d}}(\boldsymbol{G}\boldsymbol{W}) - (p + c)\boldsymbol{N}\boldsymbol{D} - \widetilde{\lambda}_W\boldsymbol{D} \\ &\quad = \Big[(\rho - \widetilde{\lambda}_W)\boldsymbol{I} - c\boldsymbol{N}\Big]\,\boldsymbol{D}. \end{aligned} \tag{58}$$

Combining its zero off-diagonal part with equation 58 gives $\boldsymbol{G} = \boldsymbol{K}_F\boldsymbol{W} + \boldsymbol{M}_F\boldsymbol{D}$, which is equation 45.

Conversely, taking the diagonal projection of equation 45 recovers equation 56, since $\Pi_{\mathrm{d}}((\boldsymbol{G} - \boldsymbol{N})\boldsymbol{D}) = \mathbf{0}$. Its off-diagonal part is the off-diagonal part of equation 54. To account for the dependence of $\boldsymbol{D}$ on $\boldsymbol{W}$, let

$$\mathcal{J}_{L_2}(\boldsymbol{W}) = \widetilde{\mathcal{J}}_{L_2}\big(\boldsymbol{W}, \Pi_{\mathrm{d}}(\boldsymbol{W})\big) = \mathcal{L}(\boldsymbol{W}) + \lambda_D\|\Pi_{\mathrm{d}}(\boldsymbol{W})\|_F^2 + \lambda_W\|\boldsymbol{W}\|_F^2. \tag{59}$$

This is equation 22 restricted to the diagonal constraint. Each Frobenius norm separates over output columns, and $\boldsymbol{D}_{:,i} = \boldsymbol{e}^{(i)}W_{i,i}$ depends only on $\boldsymbol{W}_{:,i}$. In the following partial derivatives, $\boldsymbol{D} = \Pi_{\mathrm{d}}(\boldsymbol{W})$. For $k \neq i$, the derivative with respect to $W_{k,i}$ is $[\nabla_{\boldsymbol{W}}\widetilde{\mathcal{J}}_{L_2}]_{k,i}$, whereas at $k = i$ it is

$$\frac{\partial\mathcal{J}_{L_2}(\boldsymbol{W})}{\partial W_{i,i}} = [\nabla_{\boldsymbol{W}}\widetilde{\mathcal{J}}_{L_2}]_{i,i} + [\nabla_{\boldsymbol{D}}\widetilde{\mathcal{J}}_{L_2}]_{i,i}. \tag{60}$$

Thus equation 45 is equivalent to $\nabla\mathcal{J}_{L_2}(\boldsymbol{W}) = \mathbf{0}$.

For every nonzero $\boldsymbol{x} \in \mathbb{R}^n$, the shared matrix satisfies

$$\boldsymbol{x}^\top\boldsymbol{K}_F\boldsymbol{x} = q\|\boldsymbol{R}\boldsymbol{x}\|_2^2 + p\sum_{i=1}^{n} G_{i,i}x_i^2 + \widetilde{\lambda}_W\|\boldsymbol{x}\|_2^2 > 0. \tag{61}$$

For a nonzero perturbation $\boldsymbol{V}$, let $\boldsymbol{D}_V = \Pi_{\mathrm{d}}(\boldsymbol{V})$ and $\widetilde{\boldsymbol{V}} = \boldsymbol{V} - \boldsymbol{D}_V$. The quadratic form of $\mathcal{J}_{L_2}(\boldsymbol{W})$ is

$$\begin{aligned}\mathcal{Q}_F(\boldsymbol{V}) = {} & q\alpha^2\|\boldsymbol{R}(\boldsymbol{D}_V + q\widetilde{\boldsymbol{V}})\|_F^2 + pq^2\|\boldsymbol{R}\widetilde{\boldsymbol{V}}\|_F^2 \\ & + pqs\|\boldsymbol{N}^{1/2}\widetilde{\boldsymbol{V}}\|_F^2 + \lambda_D\|\boldsymbol{D}_V\|_F^2 + \lambda_W\|\boldsymbol{V}\|_F^2. \end{aligned} \tag{62}$$

If $\widetilde{\boldsymbol{V}} \neq \mathbf{0}$, the variance term is strictly positive because $\boldsymbol{N}$ has positive diagonal. If $\widetilde{\boldsymbol{V}} = \mathbf{0}$, then $\boldsymbol{V} = \boldsymbol{D}_V \neq \mathbf{0}$ and the retained-state term $q\alpha^2\|\boldsymbol{R}\boldsymbol{D}_V\|_F^2$ is strictly positive. Therefore $\mathcal{Q}_F(\boldsymbol{V}) > 0$ for every nonzero $\boldsymbol{V}$. The reduced objective is coercive and strictly convex. It has a unique minimizer, which by the stationary-point equivalence is the unique solution of equation 45. □

B.3 Proof of Lemma 3.4

Proof of Lemma 3.4. Let $\boldsymbol{D} = \mathrm{Diag}(\mathrm{diag}(\boldsymbol{W}))$ and denote the i-th coordinate vector by $\boldsymbol{e}^{(i)}$. Since $\boldsymbol{D}_{:,i} = W_{i,i}\boldsymbol{e}^{(i)}$, the matrix equation separates into the column systems

$$\begin{aligned}\boldsymbol{G}_{:,i} &= \boldsymbol{K}\boldsymbol{W}_{:,i} + \boldsymbol{M}_{:,i}W_{i,i} \\ &= \left(\boldsymbol{K} + \boldsymbol{M}_{:,i}\boldsymbol{e}^{(i)\top}\right)\boldsymbol{W}_{:,i}. \end{aligned} \tag{63}$$

The whole equation has a unique solution if and only if every column system has a unique solution. By the matrix determinant lemma,

$$\begin{aligned}\det\left(\boldsymbol{K} + \boldsymbol{M}_{:,i}\boldsymbol{e}^{(i)\top}\right) &= \det(\boldsymbol{K})\left[1 + \boldsymbol{e}^{(i)\top}\boldsymbol{P}\boldsymbol{M}_{:,i}\right] \\ &= \det(\boldsymbol{K})\left[1 + (\boldsymbol{P}\boldsymbol{M})_{i,i}\right]. \end{aligned} \tag{64}$$

Since $\boldsymbol{K}$ is invertible, the condition is $1 + (\boldsymbol{P}\boldsymbol{M})_{i,i} \neq 0$ for every i. Applying Sherman–Morrison then gives

$$\begin{aligned}\boldsymbol{W}_{:,i} &= \left[\boldsymbol{P} - \frac{\boldsymbol{P}\boldsymbol{M}_{:,i}\boldsymbol{e}^{(i)\top}\boldsymbol{P}}{1 + \boldsymbol{e}^{(i)\top}\boldsymbol{P}\boldsymbol{M}_{:,i}}\right]\boldsymbol{G}_{:,i} \\ &= (\boldsymbol{P}\boldsymbol{G})_{:,i} - (\boldsymbol{P}\boldsymbol{M})_{:,i}\frac{(\boldsymbol{P}\boldsymbol{G})_{i,i}}{1 + (\boldsymbol{P}\boldsymbol{M})_{i,i}}. \end{aligned} \tag{65}$$

This is equation 14. In particular,

$$W_{i,i} = \frac{(\boldsymbol{P}\boldsymbol{G})_{i,i}}{1 + (\boldsymbol{P}\boldsymbol{M})_{i,i}}. \tag{66}$$

Stacking equation 65 over $i = 1, \ldots, n$ gives the matrix form

$$\boldsymbol{W} = \boldsymbol{P}\boldsymbol{G} - \boldsymbol{P}\boldsymbol{M}\,\mathrm{Diag}\big(\mathrm{diag}(\boldsymbol{P}\boldsymbol{G})\big)\big[\boldsymbol{I} + \mathrm{Diag}\big(\mathrm{diag}(\boldsymbol{P}\boldsymbol{M})\big)\big]^{-1}. \tag{67}$$

□

B.4 Proof of Proposition 3.5

Proof of Proposition 3.5. Set $\lambda_D = \lambda_W = 0$ in Lemma B.1. Then $\widetilde{\lambda}_W = \rho = 0$, $\boldsymbol{K}_F = \boldsymbol{K}$, and $\boldsymbol{M}_F = c(\boldsymbol{G} - \boldsymbol{N})$. Thus the unique minimizer $\boldsymbol{W}^*$ exists and is the unique solution of equation 12, whose columns are equation 13. Necessity in Lemma 3.4 gives $1 + c[\boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N})]_{i,i} \neq 0$ for every i. Setting $\boldsymbol{M} = c(\boldsymbol{G} - \boldsymbol{N})$ in equation 67 then gives equation 15.

For $b > 0$, DEQL's column-wise solution minimizes equation 2 (Guo et al., 2026). With $a = 1$ and $b = \alpha$, Lemma 3.2 identifies this objective with $\mathcal{L}(\boldsymbol{W})$. Its minimizer is unique, so the DEQL solution equals $\boldsymbol{W}^*$.

At $\alpha = 1$, $c = 0$ and $\boldsymbol{W}^* = \boldsymbol{P}\boldsymbol{G}$. Since $q\boldsymbol{P} = (\boldsymbol{G} + \frac{p}{q}\boldsymbol{N})^{-1}$, $q\boldsymbol{W}^* = (\boldsymbol{G} + \frac{p}{q}\boldsymbol{N})^{-1}\boldsymbol{G}$, the DLAE closed form of Steck (2020). Finally, $\boldsymbol{P}\boldsymbol{K} = \boldsymbol{I}$ gives $q\boldsymbol{P}\boldsymbol{G} + p\boldsymbol{P}\boldsymbol{N} = \boldsymbol{I}$, hence

$$\boldsymbol{P}\boldsymbol{G} = \tfrac{1}{q}\big(\boldsymbol{I} - p\boldsymbol{P}\boldsymbol{N}\big), \qquad \boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N}) = \tfrac{1}{q}\big(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{N}\big). \tag{68}$$

Right multiplication by diagonal $\boldsymbol{N}$ rescales columns, so both matrices and $\boldsymbol{W}^*$ are obtained from $\boldsymbol{P}$ in $O(n^2)$ operations. □

B.5 Proof of Proposition 3.6

Proof of Proposition 3.6. Setting $\boldsymbol{D} = \boldsymbol{0}$ in equation 9 makes both residuals $\boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W}$. Their coefficients add to $q\alpha^2 + p = s$, giving equation 16. Since $s \geq p > 0$, the minimizer is independent of α. The Hessian of the objective inside parentheses, for each output column, is $2q(q\boldsymbol{G} + p\boldsymbol{N}) = 2q\boldsymbol{K} \succ 0$. Thus the objective is strictly convex on the zero-diagonal subspace.

Divide equation 16 by s and add the diagonal multiplier term $2\,\mathrm{tr}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{W})$. Setting the gradient to zero yields $2q(\boldsymbol{K}\boldsymbol{W} - \boldsymbol{G}) + 2\widehat{\boldsymbol{\Gamma}} = \boldsymbol{0}$. With $\boldsymbol{\Gamma} = \widehat{\boldsymbol{\Gamma}}/q$, this is equation 17, and hence $\boldsymbol{W} = \boldsymbol{P}(\boldsymbol{G} - \boldsymbol{\Gamma})$.

Because $\boldsymbol{P} \succ 0$, every $P_{i,i} > 0$. The constraint $W_{i,i} = 0$ therefore gives $(\boldsymbol{\Gamma})_{i,i} = (\boldsymbol{PG})_{i,i}/P_{i,i}$, yielding equation 18. Strict convexity makes this stationary point the unique constrained minimizer.

By equation 68, $\boldsymbol{PG} = (\boldsymbol{I} - p\boldsymbol{PN})/q$. Since $\boldsymbol{N}$ is diagonal, its contribution to the two terms of equation 18 cancels. Consequently,

$$\boldsymbol{W}^*_{\text{diag}=0} = \frac{1}{q}\left[\boldsymbol{I} - \boldsymbol{P}\big[\text{Diag}(\text{diag}(\boldsymbol{P}))\big]^{-1}\right]. \tag{69}$$

To compare with EDLAE, define $\boldsymbol{C} = (\boldsymbol{G} + \frac{p}{q}\boldsymbol{N})^{-1} = q\boldsymbol{P}$. With $\boldsymbol{B} = q\boldsymbol{W}$, dividing equation 16 by s gives

$$\frac{\widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{0})}{s} = \|\boldsymbol{R} - \boldsymbol{RB}\|_F^2 + \frac{p}{q}\|\boldsymbol{N}^{1/2}\boldsymbol{B}\|_F^2, \qquad \text{diag}(\boldsymbol{B}) = \boldsymbol{0}. \tag{70}$$

This is the full-emphasis EDLAE objective of Steck (2020), with variable $\boldsymbol{B}$ and regularizer $(p/q)\boldsymbol{N}$. Then equation 69 gives Steck's closed form,

$$q\boldsymbol{W}^*_{\text{diag}=0} = \boldsymbol{I} - \boldsymbol{C}\big[\text{Diag}(\text{diag}(\boldsymbol{C}))\big]^{-1}. \tag{71}$$

The item-dependent ridge $(p/q)\boldsymbol{N}$ is exactly EDLAE's dropout-induced regularizer. This also equals the Steck solution in Eq. (4) of DEQL (Guo et al., 2026),

$$\boldsymbol{W}^*_{\text{diag}=0} = \frac{1}{1-p}\big[\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{I} \odot \boldsymbol{C})^{-1}\big]. \tag{72}$$

Both formulations use unscaled corrupted input.

When $\alpha = 0$, the masked conditional objective becomes

$$\mathcal{L}(\boldsymbol{W}) = \widetilde{\mathcal{L}}(\boldsymbol{W}, \boldsymbol{D}) = p\|\boldsymbol{R} - q\boldsymbol{R}(\boldsymbol{W} - \boldsymbol{D})\|_F^2 + p^2q\|\boldsymbol{N}^{1/2}(\boldsymbol{W} - \boldsymbol{D})\|_F^2. \tag{73}$$

Here $\boldsymbol{D} = \text{Diag}(\text{diag}(\boldsymbol{W}))$. It depends only on $\boldsymbol{W} - \boldsymbol{D}$, whose diagonal is zero. The map $\boldsymbol{W} \mapsto (\boldsymbol{W} - \boldsymbol{D}, \boldsymbol{D})$ is a bijection between matrices and pairs comprising a zero-diagonal matrix and a diagonal matrix. Hence $\boldsymbol{W}$ minimizes the objective if and only if $\boldsymbol{W} - \boldsymbol{D} = \boldsymbol{W}^*_{\text{diag}=0}$; $\boldsymbol{D}$ is arbitrary.

Finally, as $p \downarrow 0$, $q \to 1$. If $\boldsymbol{G}$ is invertible, then $\boldsymbol{C} \to \boldsymbol{G}^{-1}$, so equation 71 reduces to the unregularized EASE solution (Steck, 2019),

$$\boldsymbol{W}^*_{\text{diag}=0} = \boldsymbol{I} - \boldsymbol{G}^{-1}\big[\text{Diag}(\text{diag}(\boldsymbol{G}^{-1}))\big]^{-1}. \tag{74}$$

□

B.6 Proof of Theorem 4.1

Proof of Theorem 4.1. Set $\lambda_W = 0$ in Lemma B.1. Then $\widetilde{\lambda}_W = 0$, $\rho = \lambda_D/(q\alpha^2)$, $\boldsymbol{K}_F = \boldsymbol{K}$, and $\boldsymbol{M}_F = \boldsymbol{M}_{\lambda_D}$. The lemma gives a unique minimizer, characterized by $\boldsymbol{G} = \boldsymbol{KW} + \boldsymbol{M}_{\lambda_D}\boldsymbol{D}$. The necessary condition in Lemma 3.4 ensures that every column denominator is nonzero. Substituting $\boldsymbol{M} = \boldsymbol{M}_{\lambda_D}$ into equation 67 gives equation 21.

At $\lambda_D = 0$, equation 20 gives $\boldsymbol{M}_0 = c(\boldsymbol{G} - \boldsymbol{N})$, so equation 21 becomes equation 15.

For the zero-diagonal limit, set $t_D = \lambda_D/(q\alpha^2)$ in equation 20. For each column,

$$\begin{aligned}(\boldsymbol{PM}_{\lambda_D})_{:,i} &= c[\boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N})]_{:,i} + t_D\boldsymbol{P}_{:,i},\\ 1 + (\boldsymbol{PM}_{\lambda_D})_{i,i} &= 1 + c[\boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N})]_{i,i} + t_D P_{i,i}.\end{aligned} \tag{75}$$

Since $P_{i,i} > 0$, their ratio converges to $\boldsymbol{P}_{:,i}/P_{i,i}$ as $t_D \to \infty$. Therefore

$$\begin{aligned}\big(\boldsymbol{W}^*_{\text{diag}=0}\big)_{:,i} &= (\boldsymbol{PG})_{:,i} - \boldsymbol{P}_{:,i}\frac{(\boldsymbol{PG})_{i,i}}{P_{i,i}},\\ \boldsymbol{W}^*_{\text{diag}=0} &= \boldsymbol{PG} - \boldsymbol{P}\,\text{Diag}\big(\text{diag}(\boldsymbol{PG})\big)\big[\text{Diag}(\text{diag}(\boldsymbol{P}))\big]^{-1},\end{aligned} \tag{76}$$

which is equation 18.

For the rate, write $\boldsymbol{H} = \boldsymbol{P}(\boldsymbol{G} - \boldsymbol{N})$. Then

$$\begin{aligned} &\frac{(\boldsymbol{P}\boldsymbol{M}_{\lambda_D})_{:,i}}{1 + (\boldsymbol{P}\boldsymbol{M}_{\lambda_D})_{i,i}} - \frac{\boldsymbol{P}_{:,i}}{P_{i,i}} \\ &\qquad = \frac{c\boldsymbol{H}_{:,i}P_{i,i} - \boldsymbol{P}_{:,i}(1 + cH_{i,i})}{P_{i,i}(1 + cH_{i,i} + t_D P_{i,i})}. \end{aligned} \tag{77}$$

The numerator is independent of λ_D, while the denominator grows linearly in λ_D because $P_{i,i} > 0$. Multiplying the difference by the fixed coefficient $(\boldsymbol{P}\boldsymbol{G})_{i,i}$ gives an $O(1/\lambda_D)$ error in each column of the minimizer. For fixed n, its Frobenius-norm error is also $O(1/\lambda_D)$. □

B.7 Proof of Proposition 4.2

Proof of Proposition 4.2. Lemma B.1 gives a unique minimizer for the general nonnegative penalties. Since $\boldsymbol{K}_F \succ 0$, take $\boldsymbol{K} = \boldsymbol{K}_F$, $\boldsymbol{P} = \boldsymbol{P}_F$, and $\boldsymbol{M} = \boldsymbol{M}_F$ in Lemma 3.4. Its matrix form equation 67 gives equation 24.

At $\lambda_W = 0$, we have $\widetilde{\lambda}_W = 0$ and $\rho = \lambda_D/(q\alpha^2)$. Thus $\boldsymbol{K}_F = \boldsymbol{K}$ and $\boldsymbol{M}_F = \boldsymbol{M}_{\lambda_D}$, reducing equation 24 to equation 21.

DEQL(L2) minimizes the objective in equation 2 plus $\lambda\|\boldsymbol{W}\|_F^2$ (Guo et al., 2026). By Lemma 3.2 with $a = 1$ and $b = \alpha$, this is $\mathcal{J}_{L_2}$ with $\lambda_D = 0$ and $\lambda_W = \lambda$. Both minimizers are unique, so they are equal.

The identity $\boldsymbol{P}_F\boldsymbol{K}_F = \boldsymbol{I}$ gives

$$\begin{aligned} \boldsymbol{P}_F\boldsymbol{G} &= \tfrac{1}{q}\big(\boldsymbol{I} - p\boldsymbol{P}_F\boldsymbol{N} - \widetilde{\lambda}_W\boldsymbol{P}_F\big), \\ \boldsymbol{P}_F(\boldsymbol{G} - \boldsymbol{N}) &= \tfrac{1}{q}\big(\boldsymbol{I} - \boldsymbol{P}_F\boldsymbol{N} - \widetilde{\lambda}_W\boldsymbol{P}_F\big), \\ \boldsymbol{P}_F\boldsymbol{M}_F &= \tfrac{c}{q}\big(\boldsymbol{I} - \boldsymbol{P}_F\boldsymbol{N} - \widetilde{\lambda}_W\boldsymbol{P}_F\big) + (\rho - \widetilde{\lambda}_W)\boldsymbol{P}_F. \end{aligned} \tag{78}$$

By the definitions in equation 23, the feasible condition $\lambda_D \geq 0$ is equivalent to $\rho \geq (s/\alpha^2)\widetilde{\lambda}_W$.

For fixed λ_W, taking $\lambda_D \to \infty$ sends $\rho \to \infty$ while leaving $\boldsymbol{P}_F$ fixed. The column limit in Appendix B.6 therefore applies with $(\boldsymbol{K}, \boldsymbol{P}, \boldsymbol{M}_{\lambda_D})$ replaced by $(\boldsymbol{K}_F, \boldsymbol{P}_F, \boldsymbol{M}_F)$, giving

$$\begin{aligned} \lim_{\lambda_D \to \infty} \boldsymbol{W}_{L_2}^* &= \boldsymbol{P}_F\boldsymbol{G} - \boldsymbol{P}_F \operatorname{Diag}\big(\operatorname{diag}(\boldsymbol{P}_F\boldsymbol{G}) \oslash \operatorname{diag}(\boldsymbol{P}_F)\big) \\ &= \frac{1}{q}\left[\boldsymbol{I} - \boldsymbol{P}_F\big[\operatorname{Diag}(\operatorname{diag}(\boldsymbol{P}_F))\big]^{-1}\right]. \end{aligned} \tag{79}$$

In the second line, the $\boldsymbol{N}$ and $\widetilde{\lambda}_W$ terms cancel by equation 78.

In the limit $p \to 0$, we have $q = 1$, $s = \alpha^2$, and $\widetilde{\lambda}_W = \lambda_W/\alpha^2$. Writing $\lambda = \widetilde{\lambda}_W > 0$, the zero-diagonal endpoint minimizes the standard EASE objective $\|\boldsymbol{R} - \boldsymbol{R}\boldsymbol{W}\|_F^2 + \lambda\|\boldsymbol{W}\|_F^2$ under $\operatorname{diag}(\boldsymbol{W}) = \boldsymbol{0}$. Its solution is

$$\boldsymbol{P}_{\mathrm{E}} = (\boldsymbol{G} + \lambda\boldsymbol{I})^{-1}, \qquad \boldsymbol{W}_{\mathrm{E}}^* = \boldsymbol{I} - \boldsymbol{P}_{\mathrm{E}}\big[\operatorname{Diag}(\operatorname{diag}(\boldsymbol{P}_{\mathrm{E}}))\big]^{-1}, \tag{80}$$

which is the standard EASE closed form (Steck, 2019). □

B.8 Special Cases

We derive the cases in Table 1 that do not follow directly from the propositions.

Equal Emphasis ($\alpha = 1$). At $\alpha = 1$, we have $s = 1$ and $c = 0$. Setting $\lambda_D = 0$ gives $\rho = \widetilde{\lambda}_W$, hence $\boldsymbol{M}_F = \boldsymbol{0}$ and $\boldsymbol{W}_{L_2}^* = \boldsymbol{P}_F\boldsymbol{G}$. Moreover,

$$q\boldsymbol{P}_F = \big(\boldsymbol{G} + (p/q)\boldsymbol{N} + (\lambda_W/q^2)\boldsymbol{I}\big)^{-1}.$$

Thus $q\boldsymbol{W}_{L_2}^*$ is the DLAE solution with L_2 weight λ_W/q^2; $\lambda_W = 0$ gives unregularized DLAE.

Zero-Diagonal Limit with Ridge. As $\lambda_D \to \infty$, the limit solves the constrained problem

$$\min_{\mathrm{diag}(\boldsymbol{W})=\boldsymbol{0}} \widetilde{\mathcal{J}}_{L_2}(\boldsymbol{W}, \boldsymbol{0}).$$

After division by s, its objective and unconstrained gradient are

$$\frac{1}{s}\widetilde{\mathcal{J}}_{L_2}(\boldsymbol{W}, \boldsymbol{0}) = \|\boldsymbol{R} - q\boldsymbol{R}\boldsymbol{W}\|_F^2 + pq\|\boldsymbol{N}^{1/2}\boldsymbol{W}\|_F^2 + \frac{\lambda_W}{s}\|\boldsymbol{W}\|_F^2,$$
$$\nabla_{\boldsymbol{W}}\left(\frac{1}{s}\widetilde{\mathcal{J}}_{L_2}(\boldsymbol{W}, \boldsymbol{0})\right) = 2q(\boldsymbol{K}_F\boldsymbol{W} - \boldsymbol{G}).$$

The constrained solution is given in equation 79. For fixed $\widetilde{\lambda}_W$, it does not depend on α. Lemma 3.2 and uniqueness show that this endpoint equals the zero-diagonal DEQL(L_2) solution for $(a, b) = (1, \alpha)$ and $\lambda_W = \lambda$. With $\boldsymbol{B} = q\boldsymbol{W}$, the same objective becomes

$$\|\boldsymbol{R} - \boldsymbol{R}\boldsymbol{B}\|_F^2 + \frac{p}{q}\|\boldsymbol{N}^{1/2}\boldsymbol{B}\|_F^2 + \frac{\widetilde{\lambda}_W}{q}\|\boldsymbol{B}\|_F^2,$$

the EDLAE objective with ridge weight $\widetilde{\lambda}_W/q$. Writing $\boldsymbol{C}_F = q\boldsymbol{P}_F$, the corresponding coefficient matrix satisfies

$$q\lim_{\lambda_D\to\infty} \boldsymbol{W}^*_{L_2} = \boldsymbol{I} - \boldsymbol{C}_F\left[\mathrm{Diag}(\mathrm{diag}(\boldsymbol{C}_F))\right]^{-1},$$

as follows from equation 79.

C Supplemental Experiments

The following subsections provide additional results for RQ1 and RQ2, followed by dataset statistics.

C.1 Solver Equivalence and Memory

Top-20 set agreement is the proportion of users for whom DEQL and Uni-Sol return identical Top-20 item sets. Unlike the Peak ΔRSS reported in the main text, absolute Peak RSS includes memory already occupied when the solver begins. Table 5 shows complete Top-20 agreement except on Beauty and Gowalla-1, where scores tie at the ranking boundary. Absolute Peak RSS does not decrease on Amazon-Book; the table note describes the bounded-lifetime DEQL implementation used for that measurement.

Table 5: Absolute peak resident set size (Peak RSS) and Top-20 set agreement under strong and weak generalization settings. Top-20 set agreement is the percentage of users for whom DEQL and Uni-Sol produce identical Top-20 recommendation sets.

Setting	Dataset	Top-20 set agreement (%)	Absolute Peak RSS		
			DEQL (GiB)	Uni-Sol (GiB)	Red. (%)
Strong	Games	100.00	1.16	1.12	2.76
	Beauty	93.33†	19.90	12.08	39.32
	Gowalla-1	99.97†	61.39	37.03	39.68
	ML-20M	100.00	30.12	18.94	37.11
	Netflix	100.00	23.80	15.24	35.98
	MSD	100.00	122.60	76.92	37.26
Weak	Amazon-Book	100.00	351.62*	360.60	-2.55
	Yelp2018	100.00	105.03	65.01	38.11
	Gowalla	100.00	121.64	74.12	39.07

* The original DEQL implementation is estimated to reach a maximum RSS of 593.88 GiB on Amazon-Book, exceeding the machine's 502 GiB memory; 351.62 GiB is measured using the mathematically equivalent bounded-lifetime implementation.
† Beauty and Gowalla-1 fall below 100% because scores are tied at the Top-20 boundary up to differences on the order of 10^{-17}.

C.2 Significance Across Five Random Splits

For the six strong generalization datasets, we fix each method's configuration selected on the original validation set and refit it on five new random splits, evaluating each split once on its test set. The new test users do not include users from the original validation set. Table 6 reports the mean and standard deviation across splits. Table 7 reports one-sided paired t-tests with four degrees of freedom at the 0.05 significance level, using raw p-values. The null that Uni-L_2 is no better than RDLAE is rejected in 9 of 12 dataset–metric comparisons. The corresponding null against Uni-L_2, $(\lambda_D \to \infty)$ is rejected in 7 of 12 comparisons. Against EDLAE, it is rejected in 4 of 12 comparisons, all on Gowalla-1 and Netflix. Against DEQL(L_2), it is rejected in 3 of 12 comparisons: ML-20M N@20 and Netflix R@20 and N@20. The remaining nine comparisons do not establish an advantage over DEQL(L_2).

Table 6: Mean recommendation performance over five random splits under the strong generalization setting, reported as mean $\pm$ standard deviation. Boldface marks the highest mean for each dataset and metric, including ties.

	Games		Beauty		Gowalla-1	
Model	R@20	N@20	R@20	N@20	R@20	N@20
EDLAE	0.2892 ± 0.0412	0.1858 ± 0.0228	0.1542 ± 0.0113	0.1041 ± 0.0105	0.2256 ± 0.0033	0.2061 ± 0.0034
RDLAE	0.2879 ± 0.0407	0.1854 ± 0.0232	0.1506 ± 0.0113	0.1042 ± 0.0101	0.2252 ± 0.0031	0.2059 ± 0.0033
DEQL(L_2)	$\mathbf{0.2985} \pm 0.0291$	0.1858 ± 0.0201	0.1540 ± 0.0109	0.1038 ± 0.0103	$\mathbf{0.2277} \pm 0.0038$	$\mathbf{0.2091} \pm 0.0034$
Uni-L_2, $(\lambda_D \to \infty)$	0.2892 ± 0.0412	0.1858 ± 0.0228	$\mathbf{0.1569} \pm 0.0106$	$\mathbf{0.1065} \pm 0.0091$	0.2260 ± 0.0032	0.2070 ± 0.0034
Uni-L_2	0.2960 ± 0.0267	$\mathbf{0.1892} \pm 0.0242$	0.1542 ± 0.0094	0.1051 ± 0.0095	0.2276 ± 0.0034	$\mathbf{0.2091} \pm 0.0036$

	ML-20M		Netflix		MSD	
Model	R@20	N@20	R@20	N@20	R@20	N@20
EDLAE	0.3958 ± 0.0026	0.3466 ± 0.0024	0.3651 ± 0.0015	0.3428 ± 0.0010	$\mathbf{0.3333} \pm 0.0005$	$\mathbf{0.3253} \pm 0.0007$
RDLAE	0.3911 ± 0.0024	0.3423 ± 0.0023	0.3619 ± 0.0014	0.3402 ± 0.0010	0.3328 ± 0.0004	0.3248 ± 0.0006
DEQL(L_2)	$\mathbf{0.3961} \pm 0.0025$	0.3451 ± 0.0024	0.3654 ± 0.0015	0.3433 ± 0.0011	0.3329 ± 0.0005	0.3250 ± 0.0006
Uni-L_2, $(\lambda_D \to \infty)$	0.3956 ± 0.0027	0.3463 ± 0.0025	0.3653 ± 0.0015	0.3430 ± 0.0011	0.3331 ± 0.0005	0.3249 ± 0.0006
Uni-L_2	$\mathbf{0.3961} \pm 0.0024$	$\mathbf{0.3467} \pm 0.0024$	$\mathbf{0.3656} \pm 0.0015$	$\mathbf{0.3434} \pm 0.0011$	0.3330 ± 0.0005	0.3250 ± 0.0006

Table 7: One-sided paired t-tests over five strong generalization splits. Acc and Rej denote failure to reject and rejection of the stated null at the 0.05 level; p-values appear beneath the decisions.

	Games		Beauty		Gowalla-1		ML-20M		Netflix		MSD	
Null hypothesis	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
Uni-$L_2 \le$ EDLAE	Acc 0.2396	Acc 0.0916	Acc 0.5022	Acc 0.2505	Rej 0.0055	Rej 0.0005	Acc 0.0839	Acc 0.1363	Rej 0.0021	Rej 0.0016	Acc 0.9998	Acc 0.9998
Uni-$L_2 \le$ RDLAE	Acc 0.1991	Rej 0.0304	Acc 0.0786	Acc 0.1837	Rej 0.0038	Rej 0.0004	Rej < 0.0001	Rej < 0.0001	Rej < 0.0001	Rej < 0.0001	Rej 0.0360	Rej 0.0034
Uni-$L_2 \le$ DEQL(L_2)	Acc 0.6684	Acc 0.2477	Acc 0.4597	Acc 0.1505	Acc 0.6940	Acc 0.4061	Acc 0.5058	Rej < 0.0001	Rej 0.0244	Rej 0.0063	Acc 0.0569	Acc 0.6128
Uni-$L_2 \le$ Uni-L_2, $(\lambda_D \to \infty)$	Acc 0.2396	Acc 0.0916	Acc 0.9958	Acc 0.9896	Rej 0.0015	Rej 0.0007	Rej 0.0224	Rej 0.0188	Rej 0.0144	Rej 0.0004	Acc 0.9544	Rej 0.0005

C.3 Regularization Parameter Sweeps

The complete λ_D and $\widetilde{\lambda}_W$ sweeps on six strong generalization datasets appear in Figures 2 and 3, extending Figure 1 and the regularization analysis in RQ2. The Games curves vary non-monotonically with λ_D. On MSD, performance changes little at smaller $\widetilde{\lambda}_W$ values but declines when it becomes too large; even after normalization by $\mathrm{mean}(\mathrm{diag}(\boldsymbol{G}))$, the validation-optimal scales differ by orders of magnitude across datasets.

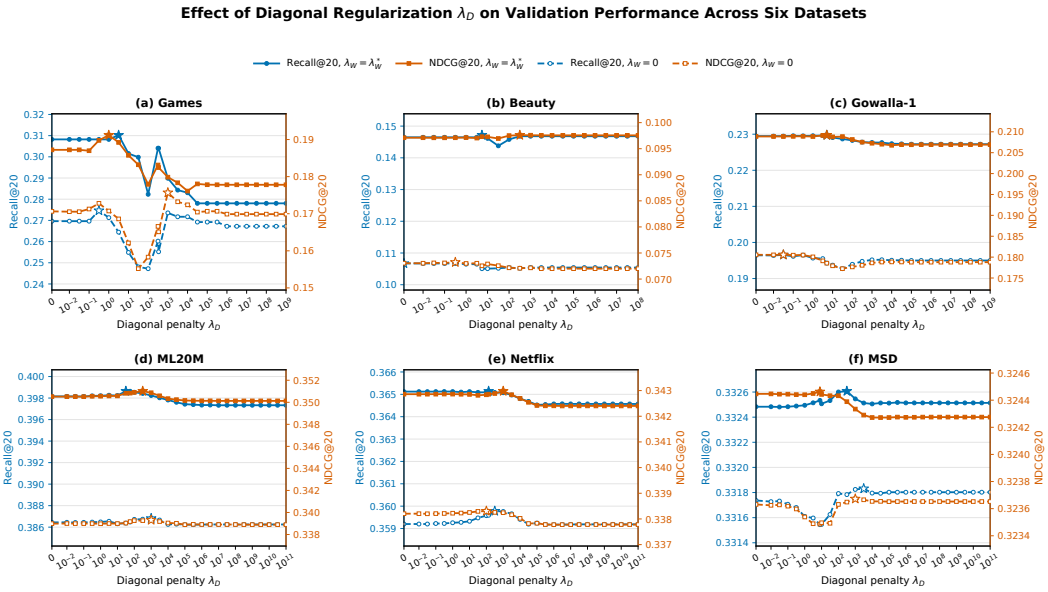


Figure 2: Validation R@20 and N@20 versus λ_D across six strong generalization datasets. Solid curves use the selected λ_W; dashed curves set $\lambda_W = 0$. Stars mark selected λ_D values.

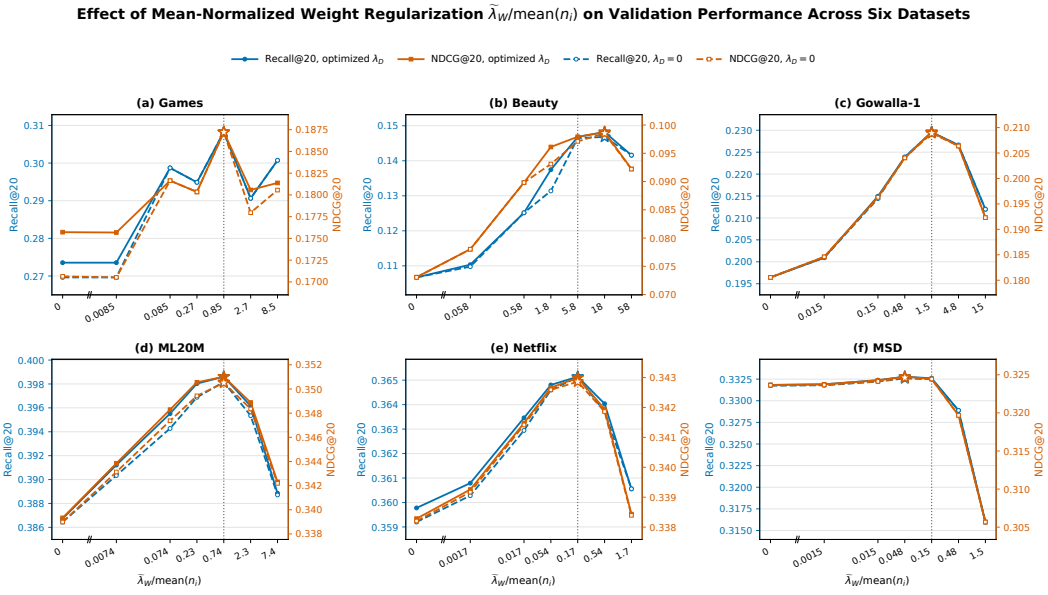


Figure 3: Validation R@20 and N@20 versus $\widetilde{\lambda}_W / \operatorname{mean}(\operatorname{diag}(\boldsymbol{G}))$ across six strong generalization datasets, with $\widetilde{\lambda}_W$ defined in equation 23. The scan uses $\widetilde{\lambda}_W^* \times \{0, 10^{-2}, 10^{-1}, 10^{-1/2}, 1, 10^{1/2}, 10\}$; the vertical dotted line marks $\widetilde{\lambda}_W^*$ and stars mark the validation-best points. Solid curves reselect λ_D at each scan point; dashed curves fix $\lambda_D = 0$.

C.4 Dataset Statistics

Table 8 reports the size and density of each user–item interaction matrix. We group the datasets by evaluation protocol: strong generalization splits users, whereas weak generalization splits interactions. Amazon-Book has the largest item catalog and lowest density, while Games has the highest density.

Table 8: Dataset statistics under strong and weak generalization.

Setting	Dataset	Users	Items	Interactions	Density (%)
Strong	Games	896	1,001	15,270	1.70
	Beauty	4,394	16,175	73,382	0.10
	Gowalla-1	13,681	28,969	677,336	0.17
	ML-20M	136,677	20,108	9,990,030	0.36
	Netflix	463,435	17,769	56,880,037	0.69
	MSD	571,355	41,140	33,633,450	0.14
Weak	Amazon-Book	52,643	91,599	2,984,108	0.06
	Yelp2018	31,668	38,048	1,561,406	0.13
	Gowalla	29,858	40,981	1,027,370	0.08

Density $= \frac{|\mathcal{R}|}{|\mathcal{U}|\times|\mathcal{I}|} \times 100\%$, where $\mathcal{R}$ is the set of observed interactions and $\mathcal{U}$ and $\mathcal{I}$ are the user and item sets.

D Low-Rank Approximation

Starting from the full-rank solution in Proposition 4.2, this appendix presents two closed-form low-rank approximations and their experimental results.

D.1 Teacher-Based Approximations

The dense $n \times n$ solution is costly to store and apply, whereas $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$ with $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times k}$ and $k \ll n$ requires only $O(nk)$ storage. Consider the exact rank-constrained problem

$$\min_{\mathrm{rank}(\boldsymbol{W})\le k} \mathcal{J}_{L_2}(\boldsymbol{W}). \tag{81}$$

Its diagonal changes with the factors, so this problem does not reduce to a closed-form truncated SVD; low-rank EDLAE instead uses ADMM (Steck, 2020). Following the full-rank-teacher approach of Jin et al. (2021), Steck & Garcia (2021), and Li et al. (2026), we use $\boldsymbol{W}^*_{L_2}$ to construct two closed-form approximations.

Weighted approximation. Around the teacher, the excess objective is a quadratic form in its approximation error. The diagonal error enters separately, so the exact rank-constrained problem cannot be solved by a truncated SVD. We replace the unknown rank-k diagonal with the teacher diagonal, whereas Steck & Garcia (2021) use zero as the diagonal proxy for EDLAE. This yields the stacked design matrix and its metric, in which λ_D acts only through the teacher:

$$\overline{\boldsymbol{R}} = \begin{bmatrix} q\sqrt{s}\,\boldsymbol{R} \\ \sqrt{pqs}\,\boldsymbol{N}^{1/2} \\ \sqrt{\lambda_W}\,\boldsymbol{I} \end{bmatrix}, \qquad \overline{\boldsymbol{R}}^\top\overline{\boldsymbol{R}} = qs\boldsymbol{K}_F. \tag{82}$$

Proposition D.1 (Low-rank approximation)**.** *Under Assumption 3.3, $\alpha > 0$, $\lambda_D, \lambda_W \ge 0$, and $1 \le k \le n$, let $\boldsymbol{W}^*_{L_2}$ be the minimizer in Proposition 4.2 and let $\overline{\boldsymbol{R}}$ be defined by equation 82.*

(a) For any $\boldsymbol{W}$ with $\mathrm{diag}(\boldsymbol{W}) = \mathrm{diag}(\boldsymbol{W}^*_{L_2})$,

$$\mathcal{J}_{L_2}(\boldsymbol{W}) - \mathcal{J}_{L_2}(\boldsymbol{W}^*_{L_2}) = \|\overline{\boldsymbol{R}}(\boldsymbol{W}^*_{L_2} - \boldsymbol{W})\|_F^2.$$

(b) For $\boldsymbol{T} \in \{\overline{\boldsymbol{R}}, \boldsymbol{I}\}$, a minimizer of

$$\min_{\mathrm{rank}(\boldsymbol{W})\le k} \|\boldsymbol{T}(\boldsymbol{W}^*_{L_2} - \boldsymbol{W})\|_F^2$$

is

$$\widehat{\boldsymbol{W}}_k = \boldsymbol{W}^*_{L_2}\boldsymbol{Q}_k\boldsymbol{Q}_k^\top, \tag{83}$$

where $\boldsymbol{Q}_k$ contains k leading right singular vectors of $\boldsymbol{T}\boldsymbol{W}^*_{L_2}$.

The proof is in Appendix D.2.

Truncated SVD. Taking $\boldsymbol{T} = \overline{\boldsymbol{R}}$ gives $\boldsymbol{W}^*_{L_2}$-approx, while $\boldsymbol{T} = \boldsymbol{I}$ gives $\boldsymbol{W}^*_{L_2}$-SVD, the rank-k truncated SVD of $\boldsymbol{W}^*_{L_2}$. The latter minimizes unweighted Frobenius error rather than the objective-induced metric. Neither approximation constrains the diagonal of $\widehat{\boldsymbol{W}}_k$; setting $\boldsymbol{U} = \boldsymbol{W}^*_{L_2}\boldsymbol{Q}_k$ and $\boldsymbol{V} = \boldsymbol{Q}_k$ stores $2nk$ entries and evaluates predictions as $(\boldsymbol{R}\boldsymbol{U})\boldsymbol{V}^\top$. With $O(k)$ matrix-free products, a top-k computation adds $O(n^2k)$ work beyond the full-rank solve (Appendix D.2).

D.2 Proof of Proposition D.1

Proof of Proposition D.1. Let $\boldsymbol{V} = \boldsymbol{W} - \boldsymbol{W}^*_{L_2}$. The reduced objective is quadratic in $\boldsymbol{W}$, with quadratic form $\mathcal{Q}_F$ given in equation 62. Lemma B.1 shows that $\boldsymbol{W}^*_{L_2}$ is its stationary point, so the linear term in the expansion vanishes:

$$\mathcal{J}_{L_2}(\boldsymbol{W}) - \mathcal{J}_{L_2}(\boldsymbol{W}^*_{L_2}) = \mathcal{Q}_F(\boldsymbol{V}). \tag{84}$$

For part (a), equality of the diagonals gives $\boldsymbol{D}_V = \Pi_{\mathrm{d}}(\boldsymbol{V}) = \boldsymbol{0}$. As $s = q\alpha^2 + p$, the first two terms of $\mathcal{Q}_F$ combine:

$$\mathcal{Q}_F(\boldsymbol{V}) = q^2 s\|\boldsymbol{R}\boldsymbol{V}\|_F^2 + pqs\|\boldsymbol{N}^{1/2}\boldsymbol{V}\|_F^2 + \lambda_W\|\boldsymbol{V}\|_F^2 = \|\overline{\boldsymbol{R}}\boldsymbol{V}\|_F^2.$$

The stacked metric satisfies

$$\overline{\boldsymbol{R}}^\top\overline{\boldsymbol{R}} = q^2 s\boldsymbol{G} + pqs\boldsymbol{N} + \lambda_W\boldsymbol{I} = qs\boldsymbol{K}_F,$$

using the definition of $\boldsymbol{K}_F$ in equation 23. Together with equation 84, this proves (a).

For a general rank-k candidate, the diagonal perturbation $\boldsymbol{D}_V = \Pi_{\mathrm{d}}(\boldsymbol{V})$ is not fixed. Writing the candidate as $\boldsymbol{U}\boldsymbol{Z}^\top$, it equals $\Pi_{\mathrm{d}}(\boldsymbol{U}\boldsymbol{Z}^\top - \boldsymbol{W}^*_{L_2})$ and depends on both factors. Its terms in $\mathcal{Q}_F$ carry different weights from the off-diagonal terms, so the exact rank-constrained objective cannot in general be written as a single shared weighted Frobenius error $\|\boldsymbol{T}\boldsymbol{V}\|_F^2$. The surrogate in (b) omits this dependence: it equals the excess objective when the diagonals match by (a), and at $k = n$ its minimizer is the full-rank teacher.

For part (b), Lemma B.1 gives $\boldsymbol{K}_F \succ 0$. Hence $\overline{\boldsymbol{R}}^\top\overline{\boldsymbol{R}} = qs\boldsymbol{K}_F \succ 0$; both choices of $\boldsymbol{T}$ have full column rank, and $\mathrm{rank}(\boldsymbol{T}\boldsymbol{W}) = \mathrm{rank}(\boldsymbol{W})$. Let $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{W}^*_{L_2}$. By the Eckart–Young–Mirsky theorem, a best rank-k approximation of $\boldsymbol{Y}$ is $\boldsymbol{Y}\boldsymbol{Q}_k\boldsymbol{Q}_k^\top$, where $\boldsymbol{Q}_k$ contains leading right singular vectors. The feasible matrix $\boldsymbol{W}^*_{L_2}\boldsymbol{Q}_k\boldsymbol{Q}_k^\top$ has rank at most k and maps under $\boldsymbol{T}$ to that best approximation, so it attains the minimum in (b). For $\boldsymbol{T} = \boldsymbol{I}$, writing a singular value decomposition of the teacher with left singular vectors $\boldsymbol{H}_k$ gives the ordinary truncated SVD $\boldsymbol{H}_k\boldsymbol{\Sigma}_k\boldsymbol{Q}_k^\top = \boldsymbol{W}^*_{L_2}\boldsymbol{Q}_k\boldsymbol{Q}_k^\top$. □

Computation. For the weighted choice, the columns of $\boldsymbol{Q}_k$ are leading eigenvectors of $(\boldsymbol{W}^*_{L_2})^\top\boldsymbol{K}_F\boldsymbol{W}^*_{L_2}$; for the unweighted choice, they come from $(\boldsymbol{W}^*_{L_2})^\top\boldsymbol{W}^*_{L_2}$. Alternatively, with the Cholesky factorization $\boldsymbol{K}_F = \boldsymbol{L}\boldsymbol{L}^\top$, the weighted vectors are the leading right singular vectors of $\boldsymbol{L}^\top\boldsymbol{W}^*_{L_2}$, so $\overline{\boldsymbol{R}}$ need not be formed. A matrix-free partial decomposition uses only products with $\boldsymbol{W}^*_{L_2}$ and $\boldsymbol{K}_F$; with $O(k)$ such products, it adds $O(n^2k)$ arithmetic beyond the full-rank solve. The approximation $\widehat{\boldsymbol{W}}_k$ has no diagonal constraint, and its own diagonal is used when evaluating the original objective.

D.3 Experimental Setup

For the low-rank evaluation, we compare against PureSVD (Cremonesi et al., 2010), EDLAE-approx (Steck & Garcia, 2021), ELSA (Vančura et al., 2022), and SVD-AE (Hong et al., 2024). We also evaluate $\boldsymbol{W}^*_{L_2}$-approx and $\boldsymbol{W}^*_{L_2}$-SVD, the weighted and unweighted rank-k approximations of the Uni-L_2 solution given by Proposition D.1. Under strong generalization, low-rank methods use $k = 1024$, except Games ($k = 256$); under weak generalization, they use $k = 1024$.

D.4 Results

Table 9 compares our two training-free projections of $\boldsymbol{W}^*_{L_2}$ with existing low-rank models. Within the strong generalization low-rank block, one of our methods is best or second-best on Games,

Table 9: Low-rank performance under strong and weak generalization. Strong generalization uses $k = 1024$ except Games ($k = 256$); weak generalization uses $k = 1024$. Boldface and underlining mark the best and second-best low-rank results; the full-rank reference rows are excluded from ranking.

Model	Games		Beauty		Gowalla-1		ML-20M		Netflix		MSD	
	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
PureSVD	0.1585	0.0899	0.0838	0.0557	0.1834	0.1649	0.0367	0.0366	0.0729	0.0838	0.1977	0.1981
SVD-AE	0.2316	0.1449	0.0967	0.0595	0.2048	0.1815	0.3726	0.3209	0.3249	0.3005	0.2418	0.2309
ELSA	0.2849	0.1640	0.1281	0.0749	0.2104	0.1849	**0.3947**	0.3388	**0.3607**	0.3358	**0.3006**	**0.2885**
EDLAE-approx	0.2526	0.1636	0.1336	**0.0868**	0.2080	0.1868	0.3892	0.3392	**0.3607**	**0.3382**	0.2654	0.2584
$\boldsymbol{W}^*_{L_2}$-SVD	**0.2870**	**0.1788**	0.1367	0.0852	**0.2133**	**0.1928**	0.3918	0.3413	0.3575	0.3349	0.2575	0.2461
$\boldsymbol{W}^*_{L_2}$-approx	0.2776	0.1729	**0.1382**	0.0860	0.2097	0.1904	0.3925	**0.3417**	0.3555	0.3323	0.2583	0.2505
Uni-L_2 (full rank)	0.2891	0.1784	0.1359	0.0884	0.2276	0.2046	0.3936	0.3428	0.3659	0.3428	0.3341	0.3265

Model	Amazon-Book		Yelp2018		Gowalla	
	R@20	N@20	R@20	N@20	R@20	N@20
PureSVD	0.0499	0.0417	0.0478	0.0403	0.1162	0.0913
SVD-AE	0.0509	0.0404	0.0672	0.0565	**0.1800**	**0.1494**
ELSA	**0.0667**	**0.0531**	**0.0717**	**0.0593**	0.1693	0.1317
EDLAE-approx	0.0504	0.0414	0.0651	0.0545	0.1636	0.1374
$\boldsymbol{W}^*_{L_2}$-SVD	0.0515	0.0418	0.0661	0.0556	0.1630	0.1373
$\boldsymbol{W}^*_{L_2}$-approx	0.0514	0.0426	0.0605	0.0503	0.1598	0.1342
Uni-L_2 (full rank)	0.0735	0.0595	0.0680	0.0572	0.1819	0.1516

Beauty, Gowalla-1, and ML-20M, whereas ELSA or EDLAE-approx leads on Netflix and MSD. Under weak generalization, ELSA leads on Amazon-Book and Yelp2018 and SVD-AE leads on Gowalla; ELSA's direct optimization of rank-k factors may explain its advantage over teacher-based projections. $\boldsymbol{W}^*_{L_2}$-approx uses the teacher diagonal rather than the zero-diagonal proxy of EDLAE-approx; the better of our two approximations exceeds EDLAE-approx in R@20 on six of nine datasets. $\boldsymbol{W}^*_{L_2}$-SVD matches or exceeds $\boldsymbol{W}^*_{L_2}$-approx in R@20 on six of nine datasets, with mostly modest absolute differences, so a closer fit in the objective-induced metric need not improve top-20 ranking.

D.5 Rank Sensitivity

Figure 4 shows that accuracy generally rises with k before saturating. The rank needed to approach the full-rank result varies by dataset and broadly grows with catalog size: ML-20M reaches its plateau at relatively small k, whereas MSD and Amazon-Book require much larger ranks. Netflix and Gowalla-1 approach full-rank performance at moderate ranks, whereas Yelp2018 does so only at the largest ranks shown. On Games and Beauty, the best low-rank test R@20 exceeds the full-rank result, suggesting that truncation may provide additional regularization for smaller catalogs. At $k = 1024$, the $2nk$ parameters of $\boldsymbol{W}^*_{L_2}$-approx amount to about 2.2%, 5.0%, and 5.4% of a dense item–item matrix on Amazon-Book, Gowalla, and Yelp2018, respectively, while retaining 69.9%, 87.8%, and 89.0% of the corresponding full-rank R@20; these differences are consistent with larger catalogs needing higher ranks to approach full-rank performance.

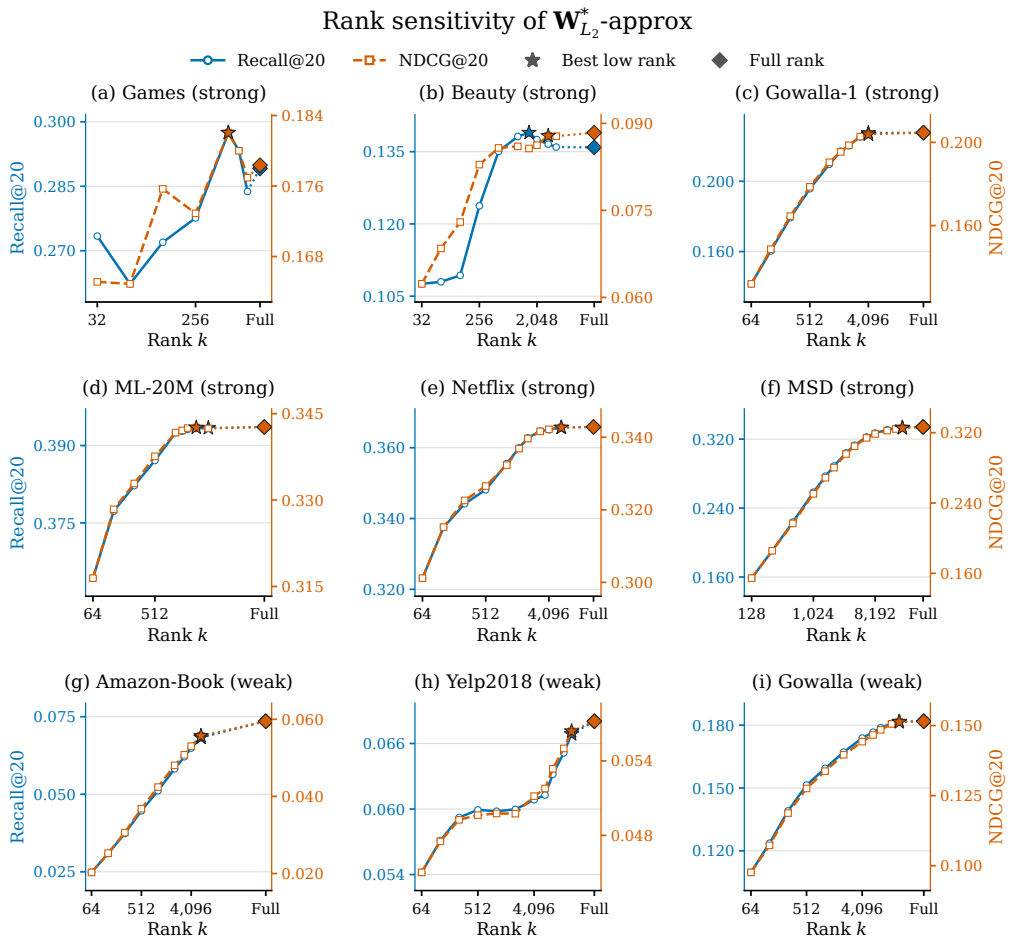


Figure 4: Rank sensitivity of $\boldsymbol{W}^*_{L_2}$-approx across nine datasets. Curves show R@20 and N@20; stars mark the test-best rank for each metric and diamonds the full-rank solution. The $k = 1024$ setting in Table 9 is a point on each curve except Games, whose rank is capped.

E LIMITATIONS

Full-Rank Cost. The full-rank solution requires $O(n^3)$ time and $O(n^2)$ storage, which limits its use for very large item catalogs. Factorized storage is available only after the full-rank teacher has been computed.

Masking Assumption. The objective assumes that interactions are masked independently with a common deletion probability. Nonuniform masking, such as item-popularity-dependent deletion, is a direction for future work.

Low-Rank Optimality. The weighted low-rank approximation uses the full-rank teacher's diagonal as a proxy for the unknown low-rank diagonal (Appendix D). Finding the exact rank-constrained optimum of the unified objective remains open.