MINIMAX SAMPLE COMPLEXITY OF FULL AND LOW-RANK BILINEAR ATTENTION SCORE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In attention models, the projection dimensions of queries and keys bound the rank of pre-softmax score matrices, yet the statistical benefit of this low-rank structure for attention score estimation remains insufficiently characterized. To this end, we show that exploiting low-rank structure reduces the sample complexity of score estimation from limited token-pair data, and quantify the gap between the full-matrix and rank-r classes in a noisy rank-1 observation model. For d-dimensional representations, we prove minimax lower bounds for squared Frobenius error with dimension factors d^2 and $2dr - r^2$ for the full-matrix and rank-r classes, respectively, under independent isotropic sub-Gaussian inputs and Gaussian noise. Correspondingly, global least-squares solutions satisfy matching-order high-probability upper bounds, with dimension factors d^2 and rd. As a convex alternative to rank-constrained estimation, nuclear norm penalized least squares attains the same rd dependence under independent sub-Gaussian noise. Together, the lower and upper bounds establish an order-d/r sample-complexity advantage for rank-r estimation at matched accuracy and confidence under Gaussian noise.

1 INTRODUCTION

Transformers (Vaswani et al., 2017) employ attention to model relationships between tokens and have been widely applied in natural language processing (Devlin et al., 2019), computer vision (Dosovitskiy et al., 2021), and speech processing (Baevski et al., 2020). In these applications, attention selectively aggregates information across token positions based on these relationships, producing context-dependent representations.

Despite their empirical success, the statistical foundations of attention-based models remain insufficiently understood. In particular, pre-softmax attention scores are determined by an interaction matrix formed from the query and key projections, whose widths limit its rank. Consequently, when relationships between tokens depend on only a few latent directions, narrow projections may capture the relevant interactions with fewer degrees of freedom (DoF). This reduction in DoF has a direct statistical implication: low-rank estimation theory links model complexity to the number of observations required for accurate estimation (Rohde and Tsybakov, 2011; Koltchinskii et al., 2011). For attention-style pairwise scores, this motivates two natural questions: How does the sample size required to achieve a given accuracy depend on the representation dimension and the number of interaction directions? How many samples can be saved by exploiting low-rank structure, holding the data distribution and noise level fixed?

In this paper, we quantify the sample-complexity benefit of low-rank structure in bilinear attention score estimation from noisy observations. For d-dimensional token representations, a full interaction matrix has d^2 DoF, whereas a rank-r matrix has $2dr - r^2$ intrinsic DoF (Absil et al., 2012). We derive finite-sample upper and lower bounds with dimension factors of order d^2 and rd for the full and rank-r classes, respectively. Together, these bounds establish an order-d/r sample-complexity advantage for rank-r estimation at matched accuracy and confidence in the non-saturated regime, subject to the stated sample size conditions.

Overall, our contributions are as follows:

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.

- We establish minimax lower bounds for the full-matrix and rank-r classes under a common noisy rank-1 observation model, with dimension factors d^2 and $2dr - r^2$, respectively, at the same Gaussian noise level and Frobenius radius.
- We prove an order-d/r sample complexity advantage of rank-r over full-matrix estimation when accuracy and confidence are held fixed. This result applies under Gaussian noise in the non-saturated regime (Theorem 5 and Section 5).
- We show that nuclear norm penalized least squares achieves the same rd dependence under sub-Gaussian noise, providing a convex alternative with the same low-rank statistical scaling (Theorem 4).
- Synthetic experiments support the theoretically predicted scaling with dimension and rank. In the two evaluated pretrained BERT attention heads, low-rank fits reach the same prediction threshold at smaller tested sample sizes than unconstrained ordinary least squares (OLS), and matched-radius comparisons show different effects of the rank constraint in the two heads.

1.1 Related Work

Learning theory for attention. Since the introduction of self-attention (Vaswani et al., 2017), the relationship between its representational capacity and learnability has been a central theoretical question. Kernel-limit analyses characterize infinite-width models (Hron et al., 2020), norm-based generalization bounds connect model complexity to out-of-sample error (Edelman et al., 2022), and convex low-rank reformulations offer another perspective on attention optimization (Sahiner et al., 2022). Further studies examine training dynamics and optimization landscapes (Jelassi et al., 2022; Tarzanagh et al., 2023; Kim and Suzuki, 2024), and task-specific analyses investigate the relationship between attention selection and prediction (Oymak et al., 2023; Deora et al., 2024; Li et al., 2024). A separate line of work studies in-context learning, in which models learn functions from examples in a prompt (Garg et al., 2022). Theoretical analyses explain this learning through algorithmic implementation and statistical inference (Akyürek et al., 2023; Bai et al., 2023; Li et al., 2023; von Oswald et al., 2023). When the focus shifts from predictive behavior to internal parameters, teacher–student models provide explicit recovery targets: Zhang et al. (2026) show that population gradient descent recovers the value matrix and positional attention weights in a position-based attention model, and characterize the convergence of parameter error and population loss over iterations. We instead focus on the score matrix representing interactions between queries and keys and derive estimation error bounds with explicit dependence on matrix rank and sample size.

Low-rank structure and statistical estimation. The query and key projection widths limit the rank of the score matrix, linking projection width to sample complexity through low-rank matrix estimation theory. Nuclear norm recovery theory shows that low-rank structure can reduce the number of measurements required under suitable measurement conditions (Recht et al., 2010; Candès and Plan, 2011). In the presence of noise, upper and lower error bounds for trace regression quantify this reduction: their dimension scaling reflects the low-rank DoF (Rohde and Tsybakov, 2011; Koltchinskii et al., 2011; Negahban and Wainwright, 2011). For nuclear norm estimation, a general framework uses restricted strong convexity and decomposable regularization to derive upper bounds on estimation error (Negahban et al., 2012). Alongside these statistical guarantees, computational approaches to low-rank matrix estimation include rank-constrained methods and factorized algorithms (Jain et al., 2010; Tu et al., 2016). Their analysis addresses nonconvex optimization landscapes (Bhojanapalli et al., 2016; Ge et al., 2017) and the implicit regularization induced by optimization (Gunasekar et al., 2017; Arora et al., 2019). Complementing these optimization analyses, we study estimation error under rank-1 measurements in the bilinear score model. We establish a lower bound on loss curvature and an upper bound on the noise term to quantify the reduction in sample complexity from low-rank structure.

Recovery and estimation under rank-1 measurements. Cai and Zhang (2015) give upper and lower bounds for rank-1 projections, including a sub-Gaussian extension of residual-constrained nuclear norm minimization with rd/n dimension scaling and a two-term upper bound. Nonconvex methods for the same measurement model also have recovery guarantees (Zhong et al., 2015; Lin and Ye, 2016). Hamidi and Bayati (2022) provide a curvature result covering our directional sub-Gaussian inputs; together with our noise bound, it yields the upper rates for comparing the two classes under a common radius and noise level at matched accuracy and confidence (Appendix D).

Link functions and estimation targets. Beyond the linear observation model, the link function determines the estimation target and the assumptions needed for recovery. When the link function is known, GL-LowPopArt gives local error guarantees in generalized low-rank trace regression and covers unbounded designs in the linear case (Lee et al., 2026). For unknown link functions, Zucker et al. (2026) establish minimax rates for estimating the composite function $\phi(x_1^\top A x_2)$. Separate work analyzes statistical properties and training dynamics after softmax normalization (Collins et al., 2024; Chen et al., 2024). We also extend the analysis to known nonlinear scalar links. The estimation error retains the rd dimension scaling for links satisfying $0 < m \le \phi' \le L < \infty$ under the stated sample size conditions (Appendix C.4).

2 Bilinear Scoring Model

We study supervised estimation of the query–key interaction matrix from observed noisy pre-softmax scores, with the fixed attention scaling absorbed into M^*. The analysis uses squared score loss and n independent input pairs.

2.1 Observation Model and Parameter Classes

We consider independent samples from the following observation model:

$$y = x_1^\top M^* x_2 + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where the inputs $x_1, x_2 \in \mathbb{R}^d$ are independent isotropic sub-Gaussian vectors (allowing dependence among coordinates in each vector), and the observation noise ϵ is independent of the input pair (x_1, x_2). Specifically, assume there is a common constant $K \ge 1$ such that both x_1 and x_2 satisfy

$$\mathbb{E}[x] = 0, \qquad \mathbb{E}[xx^\top] = I_d, \qquad \|\langle u, x\rangle\|_{\psi_2} \le K \|u\|_2 \qquad u \in \mathbb{R}^d, \tag{2}$$

where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm. For example, standard Gaussian vectors and vectors with independent Rademacher coordinates satisfy these conditions.

Known positive-definite marginal covariances can be accommodated by the rank-preserving normalization in Corollary 2. Theorem 9 separately treats structured cross-input dependence.

To compare the full-matrix and low-rank classes under the same Frobenius norm constraint, fix $B > 0$ and $1 \le r \le d$ and define

$$\mathcal{F}_B = \{M \in \mathbb{R}^{d\times d} : \|M\|_F \le B\}, \tag{3}$$

$$\mathcal{M}_r(B) = \{M \in \mathbb{R}^{d\times d} : \operatorname{rank}(M) \le r,\ \|M\|_F \le B\}, \tag{4}$$

where $\mathcal{F}_B$ is the full-matrix class with Frobenius norm at most B, and $\mathcal{M}_r(B)$ is its subclass of matrices with rank at most r.

2.2 Full and Low-Rank Parameterizations

Full-matrix parameterization. The score matrix $A = W_Q^\top W_K$ of width-d query and key projections can be any $d \times d$ matrix. The Frobenius radius constraint yields $\mathcal{F}_B$, whose minimax lower bound is Theorem 1.

Low-rank parameterization. The factorization $A = UV^\top$ with $U, V \in \mathbb{R}^{d\times r}$ represents exactly the matrices of rank at most r. This parameterization therefore yields the class $\mathcal{M}_r(B)$ under the Frobenius constraint $\|A\|_F \le B$. Moreover, U and V together contain $2dr$ scalar parameters, but different factor pairs can represent the same matrix. When A has rank exactly r, the factorization has r^2 redundant parameters, leaving $2dr - r^2$ independent DoF (Absil et al., 2012).

2.3 Prediction Risk and Estimation Error

We evaluate a candidate matrix M for matrix recovery and prediction of y on new samples. These are measured by the squared Frobenius error $\mathcal{E}(M) = \|M - M^*\|_F^2$ and the prediction risk $\mathcal{R}_{\text{pred}}(M) = \mathbb{E}[(y - x_1^\top M x_2)^2]$, respectively. For $\Delta = M - M^*$, the corresponding score error is $x_1^\top \Delta x_2 =$

$\langle x_1 x_2^\top, \Delta \rangle_F$, with vectorized design $Z = x_2 \otimes x_1 \in \mathbb{R}^{d^2}$. Since the inputs are independent and isotropic, $\mathbb{E}[ZZ^\top] = I_{d^2}$, which gives

$$\mathbb{E}[(x_1^\top \Delta x_2)^2] = \|\Delta\|_F^2, \qquad \mathcal{R}_{\text{pred}}(M) = \mathcal{E}(M) + \sigma^2. \tag{5}$$

The noise cross term vanishes by independence and zero mean. Thus excess prediction risk equals squared Frobenius error. For general independent mean-zero finite-variance noise, the floor is $\mathbb{E}[\epsilon^2]$. If both estimation errors vanish and this floor is positive, the total-risk ratio tends to 1.

3 Lower Bounds

The lower bounds in this section and the upper bounds in Section 4 jointly characterize the dependence of estimation error on dimension for the two classes under the Gaussian noise model in equation 1. The upper bounds further hold under independent sub-Gaussian noise. We derive the lower bounds using the KL divergence formula for Gaussian location models.

3.1 Full-Matrix Class

Theorem 1 (Two-regime lower bound)**.** *There is an absolute constant $c > 0$ such that, for all $d, n \geq 1$, $B > 0$, and $\sigma > 0$,*

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{F}_B} \mathbb{E}[\left\| \widehat{M} - M \right\|_F^2] \geq c \min\left\{ B^2, \frac{\sigma^2 d^2}{n} \right\}, \tag{6}$$

where $\widehat{M}$ denotes an arbitrary measurable estimator based on the n observed samples.

Proof sketch. A Varshamov–Gilbert packing and Fano bound handle large d. A Le Cam two-point argument handles the remaining dimensions (Yu, 1997). See Appendix A.1 for the complete proof.

3.2 Low-Rank Class

Theorem 2 (Two-regime low-rank lower bound)**.** *There is an absolute constant $c > 0$ such that, for every $1 \leq r \leq d$, $n \geq 1$, $B > 0$, and $\sigma > 0$,*

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{M}_r(B)} \mathbb{E}[\left\| \widehat{M} - M \right\|_F^2] \geq c \min\left\{ B^2, \frac{\sigma^2 (2dr - r^2)}{n} \right\}. \tag{7}$$

Proof sketch. We construct a set of low-rank matrices with pairwise Frobenius distances bounded below. The logarithm of the number of matrices is of order $2dr - r^2$. Controlling the KL divergence between the induced observation distributions then allows us to apply Fano's inequality and obtain the lower bound. See Appendix A.2 for the construction and complete proof.

Corollary 1 (Probability lower bounds for fixed-confidence comparison)**.** *Under the input assumptions and Gaussian noise in equation 1, there are absolute constants $a > 0$ and $\eta_0 \in (0, 1/2)$ such that, for either $(\mathcal{C}, p) = (\mathcal{F}_B, d^2)$ or $(\mathcal{M}_r(B), 2dr - r^2)$,*

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{C}} \Pr_M\left\{ \left\| \widehat{M} - M \right\|_F^2 \geq a \min\left\{ B^2, \frac{\sigma^2 p}{n} \right\} \right\} \geq \eta_0. \tag{8}$$

The proof retains the testing-error probabilities of the preceding packings (Appendix A.3).

4 Upper Bounds

We analyze the rank-constrained empirical risk minimization (ERM) estimator, which minimizes empirical risk over $\mathcal{M}_r(B)$ using squared loss:

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in \mathcal{M}_r(B)} \widehat{\mathcal{R}}_n(M), \qquad \widehat{\mathcal{R}}_n(M) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_{1,i}^\top M x_{2,i})^2. \tag{9}$$

When $M^* \in \mathcal{M}_r(B)$, rank subadditivity gives $\operatorname{rank}(\widehat{M} - M^*) \le k_r = \min\{2r, d\}$. Thus r determines the fitted class and k_r the feasible error directions. Projections of Z generally have only sub-exponential tails. We use the small-ball method for uniform curvature and conditional Bernstein bounds with sphere nets for noise control (Appendix B.2).

The sub-Gaussian condition on the observation noise is $\mathbb{E}[\exp(t\epsilon_i)] \le \exp(\sigma^2 t^2/2)$ for every $t \in \mathbb{R}$, where σ^2 is the variance proxy. Consequently, $\|\epsilon_i\|_{\psi_2} \le C\sigma$ for an absolute constant C.

Theorem 3 (High-probability upper bound for rank-constrained ERM)**.** *Assume $M^* \in \mathcal{M}_r(B)$ and let equation 2 hold. The observation noises $\epsilon_1, \dots, \epsilon_n$ are independent across samples, mean zero, independent of the inputs, and satisfy the sub-Gaussian condition above with variance proxy σ^2. Fix $\eta \in (0,1)$ and set $t_\eta = \log(8/\eta)$. There is a constant $C_K \ge 1$, depending only on K, such that if*

$$n \ge C_K(k_r d + t_\eta), \tag{10}$$

then, with probability at least $1-\eta$, every global minimizer in equation 9 satisfies

$$\left\|\widehat{M} - M^*\right\|_F^2 \le C_K \sigma^2 k_r \left\{ \frac{d + t_\eta}{n} + \frac{(d+t_\eta)^2(\log(2n) + t_\eta)}{n^2} \right\}. \tag{11}$$

If, in addition,

$$n \ge C_K \left\{ k_r d + t_\eta + (d + t_\eta)\big(\log(2(d+t_\eta)) + t_\eta\big) \right\}, \tag{12}$$

the second term is absorbed and

$$\left\|\widehat{M} - M^*\right\|_F^2 \le C_K \sigma^2 \frac{k_r(d+t_\eta)}{n}. \tag{13}$$

At fixed confidence, $n \ge C_{K,\eta} \max\{k_r d, d\log(2d)\}$ guarantees the $\sigma^2 k_r d/n$ error rate. The two-term bound also covers sample sizes satisfying the weaker condition equation 10, which ensures curvature. Condition equation 12 absorbs its second term into the first. An explicit sufficient condition for curvature is $n \ge C\{K^{20} k_r d + K^{16}\log(8/\eta)\}$ (Appendix B.2). The noise and error constants also depend on K.

Proof sketch. Let $\Delta = \widehat{M} - M^*$ and $G_n = n^{-1}\sum_i \epsilon_i x_{1,i} x_{2,i}^\top$. Empirical optimality gives $n^{-1}\sum_i (x_{1,i}^\top \Delta x_{2,i})^2 \le 2\langle G_n, \Delta\rangle$. Under equation 10, small-ball curvature and conditional Bernstein bounds with sphere nets each fail with probability at most $\eta/2$. On their joint event, the left side is at least $\kappa_K \|\Delta\|_F^2$ and $\|G_n\|_{\mathrm{op}} \le C_K \sigma b_{n,\eta}$, with $b_{n,\eta}$ defined in equation 16. Since $\|\Delta\|_* \le \sqrt{k_r}\|\Delta\|_F$, we obtain $\kappa_K\|\Delta\|_F \le 2C_K\sigma\sqrt{k_r} b_{n,\eta}$ for nonzero Δ. Squaring proves the two-term bound. Event definitions and full proofs are in Appendices B and B.2.

The exact-minimizer guarantee has a deterministic perturbation form for feasible numerical solutions. Let δ_{opt} be the empirical risk above the global minimum over $\mathcal{M}_r(B)$. On an event with restricted curvature κ and noise threshold τ, the same basic inequality gives

$$\|\widehat{M}_{\mathrm{opt}} - M^*\|_F^2 \le \frac{4\tau^2}{\kappa^2} + \frac{2\delta_{\mathrm{opt}}}{\kappa}. \tag{14}$$

Thus a certified global objective gap contributes an additive estimation term, while the first term retains the statistical rate above. Appendix C derives this perturbation bound.

4.1 A CONVEX ESTIMATOR

To obtain a convex estimator with the same rd dependence in its error bound, we replace the rank constraint with a nuclear norm penalty:

$$\widehat{M}_\lambda \in \operatorname*{arg\,min}_{M \in \mathbb{R}^{d\times d}} \left\{ \frac{1}{n}\sum_{i=1}^n (y_i - x_{1,i}^\top M x_{2,i})^2 + \lambda \|M\|_* \right\}, \tag{15}$$

where $\|M\|_*$ is the sum of the singular values of M and $\lambda > 0$ controls the penalty strength. The same theoretical rule $\lambda = c_K \sigma b_{n,\eta}$, determined by K, σ, d, n, η, applies across target ranks. The theorem below gives rank-dependent error bounds and the corresponding sample conditions,

quantifying the estimator's adaptation to low-rank structure. To specify the penalty level λ, for $\eta \in (0,1)$ let $t_\eta = \log(8/\eta)$ and define

$$b_{n,\eta} = \sqrt{\frac{d+t_\eta}{n}} + \frac{(d+t_\eta)\sqrt{\log(2n)+t_\eta}}{n}. \tag{16}$$

Theorem 4 (Convex low-rank estimation)**.** *Suppose the inputs satisfy equation 2, with independent vectors in each pair and independent samples. The noise is independent across samples and of all inputs, with* $\mathbb{E}[\exp(t\epsilon_i)] \le \exp(\sigma^2 t^2/2)$ *for every* $t \in \mathbb{R}$ *and* $\sigma > 0$*. Let* $\operatorname{rank}(M^*) \le r$*,* $1 \le r \le d$*. There are positive constants* c_K *and* C_K*, depending only on* K*, such that, for* $\lambda = c_K \sigma b_{n,\eta}$ *and* $n \ge C_K(rd + t_\eta)$*, with probability at least* $1-\eta$*, every minimizer in equation 15 satisfies*

$$\|\widehat{M}_\lambda - M^*\|_F^2 \le C_K \sigma^2 r \left\{ \frac{d+t_\eta}{n} + \frac{(d+t_\eta)^2(\log(2n)+t_\eta)}{n^2} \right\}. \tag{17}$$

Under the stronger sample size condition

$$n \ge C_K\{rd + t_\eta + (d+t_\eta)[\log(2(d+t_\eta)) + t_\eta]\}, \tag{18}$$

this bound simplifies to

$$\|\widehat{M}_\lambda - M^*\|_F^2 \le 2C_K\sigma^2 \frac{r(d+t_\eta)}{n}. \tag{19}$$

Proof sketch. The choice of λ ensures $\|G_n\|_{\mathrm{op}} \le \lambda/4$ with high probability. For $\Delta = \widehat{M}_\lambda - M^*$, the optimality of $\widehat{M}_\lambda$ and the decomposability of the nuclear norm imply $\|\Delta\|_* \le 4\sqrt{2r}\|\Delta\|_F$. The small-ball argument gives a uniform curvature lower bound $\kappa_{\mathrm{nuc}}(K) > 0$ over matrices satisfying this inequality. Combining the noise and curvature bounds gives $\kappa_{\mathrm{nuc}}(K)\|\Delta\|_F^2 \le \frac{3\lambda}{2}\sqrt{2r}\|\Delta\|_F$ with probability at least $1-\eta$. Substituting $\lambda = c_K\sigma b_{n,\eta}$ yields equation 17, which simplifies to equation 19 under equation 18. Thus nuclear norm penalized least squares attains the same rd dependence in its error bound as rank-constrained least squares. Appendix B.3 gives the full argument and specifies the constants in Theorem 8.

5 Sample-Complexity Separation

Theorem 5 (Comparison of full-matrix and low-rank estimation)**.** *Under the common Gaussian model equation 1, the lower bounds equation 6 and equation 7 match the upper bounds in their dependence on* d *and* r*. Setting* $r = d$ *in Theorem 3 and its sample size condition equation 12 gives, with probability at least* $1-\eta$*,*

$$\|\widehat{M}_{\mathrm{full}} - M^*\|_F^2 \le C_K\sigma^2 \frac{d(d+\log(8/\eta))}{n}.$$

The low-rank dimension factors satisfy $1 \le k_r d/(2dr - r^2) \le 4/3$ *and are both of order* rd*.*

To translate this comparison into sample requirements at the same accuracy and confidence level, fix $0 < \eta < \eta_0$ and a target error $0 < \varepsilon < aB^2/2$, where a, η_0 are from Corollary 1. To satisfy $\sup_{M\in\mathcal{C}} \Pr_M\{\left\|\widehat{M} - M\right\|_F^2 > \varepsilon\} \le \eta$ for $\mathcal{C} = \mathcal{F}_B$ or $\mathcal{M}_r(B)$, the probability lower bound equation 8 requires $n \gtrsim \sigma^2 d^2/\varepsilon$ for the full class and $n \gtrsim \sigma^2(2dr - r^2)/\varepsilon$ for the low-rank class. On the other hand, Theorem 3 gives sufficient conditions

$$n_{\mathrm{full}} \ge C_{K,\eta} \max\{d^2, \sigma^2 d^2/\varepsilon\},$$
$$n_r \ge C_{K,\eta} \max\{rd, d\log(2d), \sigma^2 rd/\varepsilon\},$$

where rd and $d\log(2d)$ ensure the simplified rate. The full-matrix class requires order-d/r times as many samples as the low-rank class when the noise-dependent terms dominate the other terms in these sample size conditions. The comparison concerns uniform guarantees over each parameter class, with the worst-case target chosen within that class. Equation 5 gives the same comparison for excess prediction risk.

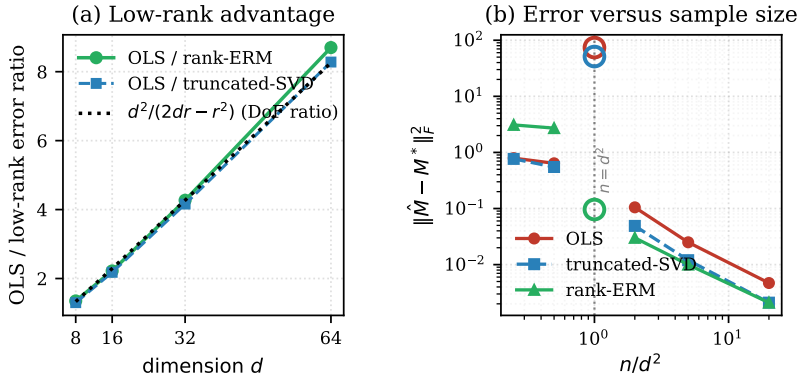


Figure 1: Synthetic scaling experiments. **(a)** Ratios of OLS error to rank-ERM and OLS+SVD-r errors versus d ($r = 4$, $n = 20d^2$). **(b)** Squared Frobenius errors versus normalized sample size n/d^2 ($d = 16, r = 4$). Filled markers denote means over 10 seeds; open diamonds at $n = d^2$ denote medians over 20 seeds.

Covariance extensions. The covariance normalization in Corollary 2 preserves rank and dimension scaling up to explicit spectral factors for independent inputs with known positive-definite marginal covariances. For structured cross-input dependence, Theorem 9 gives a population decomposition into a bilinear term and a quadratic remainder under the model with an independent linear residual. The quadratic remainder is orthogonal to bilinear scores in expectation, so the transformed interaction matrix remains a minimizer of the bilinear population squared loss. Appendix G.2 evaluates the corresponding transformation when its coefficient is estimated from auxiliary samples.

Known scalar transformations. The dimension factors in the estimation error bounds remain of order rd when the bilinear scores undergo a known scalar transformation ϕ satisfying $0 < m \leq \phi' \leq L$. The upper and lower bounds acquire factors m^{-2} and L^{-2}, respectively (Appendix C.4).

6 Numerical Experiments

This section first evaluates the theoretically predicted dimension and sample-size scaling. We then examine the roles of the rank constraint, radius, fitted rank, input distribution, and a known nonlinear link, and finally evaluate prediction of pretrained attention scores. In the synthetic experiments, data follow equation 1 with noise standard deviation $\sigma = 0.3$. We compare projected factorized rank-ERM iterates with full-matrix OLS and post hoc rank-r truncation of the OLS estimate by singular value decomposition (OLS+SVD-r). Full experimental details are provided in Appendix E.1.

6.1 Synthetic Dimension and Sample Scaling

Dimension scaling. The OLS/rank-ERM error ratio grows with d and is consistent with $d^2/(2dr - r^2)$, with the same separation for OLS+SVD-r (Figure 1(a)).

Sample scaling. Recovery error generally decreases as n grows (Figure 1(b)). Rank-ERM has the lowest error at every tested $n > 2d^2$, consistent with Section 5. At $n = d^2$, however, both OLS and OLS+SVD-r exhibit a sharp interpolation peak. We analyze this phenomenon in Appendix E.2 and find that small singular values of the design amplify noise in the OLS fit, so post hoc rank truncation retains the instability.

6.2 Estimator Comparisons and Robustness

We next compare regularization strategies, characterize fitted-rank dependence, and test the scaling pattern under non-Gaussian designs, a known nonlinear link, and real STS-B scores.

Table 1: Squared Frobenius errors for **(a)** regularized estimators at interpolation ($d = 16, r = 4, n = 256$; 25 paired runs) and **(b)** a fitted-rank sweep ($d = 16, r^* = 4, n = 5120$). Ball-constrained estimators use $B = 2$.

(a) Regularization and radius

Estimator	median	IQR
Full OLS	73.7	213.8
Ridge	0.281	0.0500
Full-ball LS	2.99	0.288
Rank-4-ball LS	0.0931	0.0162

(b) Fitted-rank sweep

r	SVD error	ERM error	DoF
1	0.5134	0.5188	31
4	0.0021	0.0021	112
8	0.0041	0.0039	192
16	0.0047	0.0046	256

Table 2: Squared Frobenius recovery error for **(a)** four input distributions ($d = 16, r = 4, n = 5120$) and **(b)** the known link $\phi(t) = t + \tanh(t)$ ($d = 16, r = 4$). In **(b)**, link-aware ERM incorporates ϕ in fitting.

(a) Input distributions

Input	OLS	ERM-r	OLS/ERM
Gaussian	0.00466	0.00210	2.22
Rademacher	0.00479	0.00191	2.51
Sparse Bernoulli	0.00495	0.00207	2.39
Correlated Gaussian	0.01656	0.00617	2.68

(b) Known nonlinear link

n	OLS	SVD-r	Link-aware ERM
512	0.4833	0.4037	**0.0119**
2560	0.3234	0.3158	**0.0017**
12800	0.3067	0.3052	**0.0003**

Regularized estimators. At interpolation, norm regularization alleviates the OLS instability, and imposing rank at the same radius lowers the median squared Frobenius recovery error by approximately 97% relative to full-ball least squares (Table 1(a)). The rank-ball advantage holds for every tested $n \geq 224$ but reverses at $n = 128$, where the less restrictive full-matrix class has lower error (Appendix Table 6). At $d = 32$ and $n = d^2$, cross-validated nuclear norm estimation also attains lower NMSE than factorized rank-ERM (0.31 versus 0.53), providing a convex low-rank alternative in this interpolation setting (Appendix Figure 4).

Fitted-rank dependence. Rank-ERM and OLS+SVD-r both attain their lowest squared Frobenius recovery error at the true rank $r^* = 4$. Errors increase for both smaller and larger fitted ranks, consistent with a bias–variance tradeoff (Table 1(b)). Complete results are in Appendix Table 7(a).

Input distributions and a known nonlinear link. Across the input distributions in Table 2(a), OLS error is 2.2–2.7 times the rank-ERM error. This error ratio is close to the 2.29 ratio between the full- and low-rank DoF. For nonlinear observations, rank-ERM using the known link improves from 0.0119 to 0.0003 as the sample size grows. OLS and truncated SVD fitted without the link remain near 0.3 (Table 2(b)). Complete results are in Appendix Tables 8(a) and 9.

Real-score scaling. On real STS-B scores, rank-ERM avoids the OLS interpolation peak (Figure 2(a)). After sample size is normalized by $2dr - r^2$, rank-ERM errors follow a common trend across nine combinations of dimension and rank (Figure 2(b)).

6.3 Pretrained Attention Scores

For head 0 of frozen BERT layers 0 and 6 (Appendix F.2), Figures 3(a)–(b) show that rank-4 ERM avoids the extreme OLS prediction errors in the limited-sample regime. In this experiment, rank-4 ERM first meets the criterion of outperforming the training-mean baseline in all eight runs at $n = 2048$ token pairs in layer 0 and $n = 1024$ in layer 6, with independently sampled training and test token pairs in each run. OLS first meets the same criterion at $n = 16384$ and $n = 8192$, respectively (Appendix Table 14(a)). Thus, the tested sample-size thresholds differ by a factor of eight in both heads. The advantage narrows as n grows, and OLS eventually attains lower median test MSE in layer 6. This crossover suggests that low rank provides an effective inductive bias when data are limited, but need not remain advantageous once OLS can be estimated from sufficiently many observations.

To characterize the low-rank structure of the pretrained attention heads, for each rank-64 interaction block we compute the fraction of squared Frobenius norm retained by the best rank-r approximation

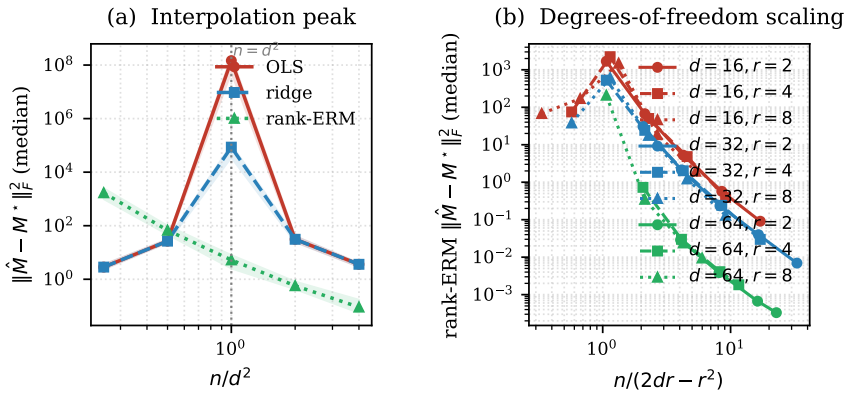


Figure 2: Real-score fits on BERT features. **(a)** Interpolation peak ($d = 16, r = 2$; IQR over 10 seeds). **(b)** Rank-ERM error across nine (d, r) settings versus samples per low-rank degree of freedom. Errors are relative to fixed reference matrices fitted on all available training pairs.

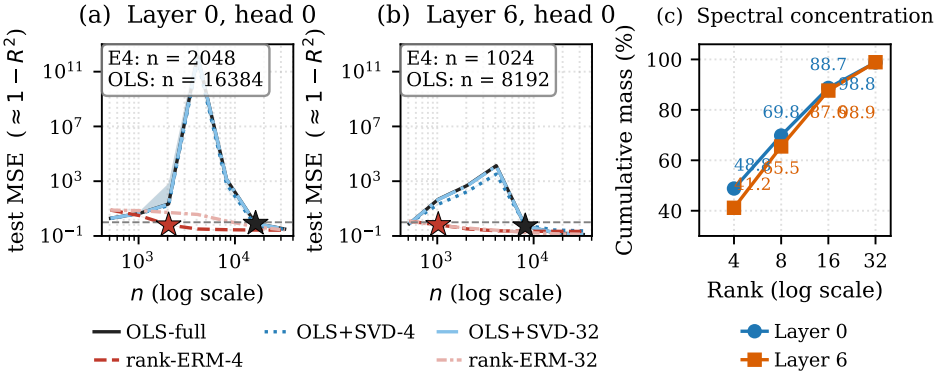


Figure 3: Pretrained BERT pre-softmax scores. **(a)–(b)** Median normalized test MSE and IQR over eight seeds; values below one beat the training-mean baseline, and upward markers exceed the displayed range. Rank-ERM uses $\|M\|_F \leq 2$; OLS is unconstrained. **(c)** Cumulative squared singular-value mass of the full-rank 64×64 interaction blocks.

for $r \in \{4, 8, 16, 32\}$ (Figure 3(c)). We find that the leading four directions retain 48.8% and 41.2% in layers 0 and 6, respectively, while rank 32 retains 98.8%–98.9%. Additionally, a matched-radius comparison shows that the predictive effect of the rank constraint differs between the two heads (Appendix Table 14(b)).

7 Conclusion

We establish dimension-matching lower and upper bounds for learning pre-softmax bilinear attention scores from noisy rank-1 measurements. The dimension factors are d^2 versus $2dr - r^2$ for the minimax lower bounds and order d^2 versus rd for the global least-squares upper bounds under the same independent isotropic sub-Gaussian inputs, Gaussian noise level, and Frobenius radius. These results imply an order-d/r sample-complexity gap at matched accuracy and confidence under the conditions in Section 5. Synthetic experiments support the predicted dimension scaling and show lower recovery error from the rank constraint at seven of eight sample sizes. On pretrained BERT scores, low-rank fits reach the prediction threshold earlier than unconstrained OLS, and matched-radius comparisons show that the effect of the rank constraint varies across attention heads.

Future work could extend this analysis to attention models with softmax normalization, value aggregation, and interactions among heads, and study the choice of query and key projection widths based on sample size and target accuracy.

DISCLOSURE OF AI USE

We used generative AI tools to improve the clarity and readability of the manuscript, to assist in proving mathematical claims, and for retrieval and discovery, including finding potentially relevant related work. The authors reviewed and verified all cited sources and take full responsibility for the final content of this work, including all mathematical claims, proofs, and text developed with AI assistance.

REFERENCES

P.-A. Absil, Luca Amodei, and Gilles Meyer. Two Newton methods on the manifold of fixed-rank matrices endowed with Riemannian quotient geometries. *arXiv preprint arXiv:1209.0068*, 2012.

A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *SIAM Journal on Information and Inference*, 3(3):224–294, 2014.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

Y. Bai, F. Chen, H. Wang, C. Xiong, and S. Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, 2023.

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, 2016.

T. T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. SemEval-2017 task 1: Semantic textual similarity: multilingual and cross-lingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14, 2017.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *IEEE Transactions on Information Theory*, 58(12):6764–6791, 2012.

S. Chen, H. Sheen, T. Wang, and Z. Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *Proceedings of the Conference on Learning Theory*, 2024.

L. Collins, A. Parulekar, A. Mokhtari, S. Sanghavi, and S. Shakkottai. In-context learning with Transformers: Softmax attention adapts to function Lipschitzness. In *Advances in Neural Information Processing Systems*, 2024.

P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024.

J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, pages 4171–4186, 2019.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

B. L. Edelman, S. Goel, S. Kakade, and C. Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the International Conference on Machine Learning*, 2022.

S. Garg, D. Tsipras, P. Liang, and G. Valiant. What can Transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022.

R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the International Conference on Machine Learning*, 2017.

C. W. J. Granger and A. P. Andersen. *An Introduction to Bilinear Time Series Models*. Vandenhoeck & Ruprecht, Göttingen, 1978.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

J. Hron, Y. Bahri, J. Sohl-Dickstein, and R. Novak. Infinite attention: NNGP and NTK for deep attention networks. In *Proceedings of the International Conference on Machine Learning*, 2020.

P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, 2010.

S. Jelassi, M. Sander, and Y. Li. Vision Transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems*, 2022.

J. Kim and T. Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the International Conference on Machine Learning*, 2024.

Y. Li, M. E. Ildiz, D. Papailiopoulos, and S. Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the International Conference on Machine Learning*, 2023.

Y. Li, Y. Huang, M. E. Ildiz, A. S. Rawat, and S. Oymak. Mechanics of next token prediction with self-attention. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2024.

M. Lin and J. Ye. A non-convex one-pass framework for generalized factorization machine and rank-one matrix sensing. In *Advances in Neural Information Processing Systems*, 2016.

S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3), 2015.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011.

S. N. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069–1097, 2011.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

S. Oymak, A. S. Rawat, M. Soltanolkotabi, and C. Thrampoulidis. On the role of attention in prompt-tuning. In *Proceedings of the International Conference on Machine Learning*, 2023.

A. Pajor. Metric entropy of the Grassmann manifold. In *Convex Geometric Analysis*, volume 34 of *Mathematical Sciences Research Institute Publications*, pages 181–188, 1998.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

A. Sahiner, T. Ergen, B. Ozturkler, J. M. Pauly, M. Mardani, and M. Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision Transformers. In *Proceedings of the International Conference on Machine Learning*, 2022.

D. A. Tarzanagh, Y. Li, X. Zhang, and S. Oymak. Max-margin token selection in attention mechanism. In *Advances in Neural Information Processing Systems*, 2023.

S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of the International Conference on Machine Learning*, 2016.

J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the International Conference on Machine Learning*, 2023.

Y. Wei, L. Zhang, B. Li, and N. He. On the benefits of weight normalization for overparameterized matrix sensing. In *International Conference on Learning Representations*, 2026.

B. Yu. Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. L. Yang, editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435. Springer, New York, 1997.

C. Zhang, Q. Zhao, Q. Gu, and Y. Cao. Transformers trained via gradient descent can provably learn a class of teacher models. In *International Conference on Learning Representations*, 2026.

K. Zhong, P. Jain, and I. S. Dhillon. Efficient matrix sensing using rank-1 Gaussian measurements. In *Algorithmic Learning Theory*, 2015.

S. Zucker, X. Wang, F. Lu, and I. Seroussi. Minimax rates for learning pairwise interactions in attention-style models. In *International Conference on Learning Representations*, 2026.

N. Hamidi and M. Bayati. On low-rank trace regression under general sampling distribution. *Journal of Machine Learning Research*, 23(321):1–49, 2022.

J. Lee, K. Jang, K.-S. Jun, M. Vojnović, and S.-Y. Yun. GL-LowPopArt: A nearly instance-wise minimax-optimal estimator for generalized low-rank trace regression. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, PMLR 300:2872–2880, 2026.

Appendix organization. Appendix A proves the minimax lower bounds for both matrix classes. Appendix B proves the rank-constrained and nuclear norm upper bounds under rank-1 measurements. Appendix C treats rank misspecification, covariance transformations, optimization error, and known scalar transformations. Appendix D compares the guarantees with existing low-rank estimation results. Appendix E gives experimental settings and synthetic results. Appendices F–H report real-feature experiments, dependence and covariance transformations, and numerical diagnostics. Appendix C.5 summarizes the scope of these extensions, and Appendix I collects the nonstandard notation.

A Minimax Lower Bounds

A.1 Full-matrix class

We use the observation model and parameter classes in Section 2: n is the number of observed pairs, d is the input dimension, B is the Frobenius radius, and σ^2 is the Gaussian noise variance. For a matrix M, let P_M denote the joint distribution of all n observations $(x_{1,i}, x_{2,i}, y_i)$ when the true matrix is M. Write $\Pr_M$ and $\mathbb{E}_M$ for probability and expectation under P_M, and $D_{\mathrm{KL}}(P\|Q)$ for the Kullback–Leibler divergence between distributions P and Q. Throughout the theoretical appendices, c, C denote positive absolute constants whose values may change between occurrences. A subscript K indicates dependence only on the input tail parameter K. For matrices, $\langle A, D\rangle = \mathrm{tr}(A^\top D)$ is the Frobenius inner product, and $\|A\|_{\mathrm{op}}$ and $\|A\|_*$ denote the largest singular value and the sum of singular values, respectively.

We prove Theorem 1 by constructing separated matrices in the Frobenius ball and bounding the KL divergence between their observation distributions. Fano's inequality then gives the risk lower bound. The low-dimensional cases are handled by Le Cam's two-point method.

Proof of Theorem 1. Fix a sufficiently large absolute integer d_0, chosen below. For $u, v \in \{1, \dots, d\}$, let $E_{uv} \in \mathbb{R}^{d\times d}$ be the matrix with 1 in row u, column v, and 0 everywhere else. **Case 1: $d \geq d_0$.** Enumerate these matrices as $E_1, \dots, E_p$ in any fixed order, where $p = d^2$ is the number of matrix coordinates. We choose the amplitude to keep the candidates in the Frobenius ball and their KL divergences small relative to the logarithm of the packing cardinality. Accordingly, set

$$a^2 = c_0 \min\left\{\frac{B^2}{p}, \frac{\sigma^2}{n}\right\},$$

where $a \geq 0$ is the entrywise amplitude and $c_0 \in (0, 1]$ is an absolute constant chosen sufficiently small below. The first term ensures $a^2 p \leq B^2$, and the second bounds na^2/σ^2, which controls the KL divergence per coordinate. Define

$$M_\omega = a \sum_{k=1}^{p} \omega_k E_k,$$

where $\omega = (\omega_1, \dots, \omega_p) \in \{0, 1\}^p$ indexes a candidate matrix by selecting which entries equal a. All other entries are zero. Let $d_H(\omega, \omega')$ count the coordinates in which two binary vectors differ. A Varshamov–Gilbert construction gives N binary vectors $\{\omega^{(i)}\}_{i=0}^{N-1} \subset \{0, 1\}^p$ with $\omega^{(0)} = 0$, $N \geq 2^{p/8}$, and $d_H(\omega^{(i)}, \omega^{(j)}) \geq p/8$ for $i \neq j$. For these hypotheses, write $P_i = P_{M_{\omega^{(i)}}}$, so P_0 is the observation distribution under the zero matrix. All hypotheses lie in $\mathcal{F}_B$, since $\|M_{\omega^{(i)}}\|_F^2 \leq a^2 p \leq c_0 B^2 \leq B^2$. Furthermore,

$$\|M_{\omega^{(i)}} - M_{\omega^{(j)}}\|_F^2 \geq \frac{c_0}{8} \min\left\{B^2, \frac{\sigma^2 d^2}{n}\right\}.$$

With $s = \frac{1}{2}\min_{i\neq j} \|M_{\omega^{(i)}} - M_{\omega^{(j)}}\|_F$, this implies

$$s^2 \geq \frac{c_0}{32} \min\left\{B^2, \frac{\sigma^2 d^2}{n}\right\}. \tag{20}$$

The covariate distribution is common to all hypotheses. By the chain rule for KL divergence, Gaussian noise, and equation 5,

$$D_{\mathrm{KL}}(P_i\|P_0) = \frac{n}{2\sigma^2}\|M_{\omega^{(i)}}\|_F^2 \leq \frac{c_0 d^2}{2}.$$

It remains to compare this KL bound with the logarithm of the packing cardinality. Fix a numerical constant $0 < \alpha < 1/8$. Because $\log N \geq (d^2/8)\log 2$, c_0 can be chosen so that

$$\frac{1}{N}\sum_{i=1}^{N-1} D_{\mathrm{KL}}(P_i\|P_0) \leq \alpha \log N, \qquad \alpha < 1/8.$$

Fano's inequality yields

$$\inf_{\widehat{M}} \sup_{M \in \mathcal{F}_B} \mathbb{E}[\left\|\widehat{M} - M\right\|_F^2] \geq s^2 \left(1 - \alpha - \frac{\log 2}{\log N}\right).$$

For fixed sufficiently large d_0, the parenthetical factor is bounded below by an absolute positive constant. Equation 20 proves the result in this case.

Case 2: $d < d_0$**.** Take $M_0 = 0$ and $M_1 = \rho E_{11}$, with $\rho \geq 0$ and $\rho^2 = c_1 \min\{B^2, \sigma^2/n\}$ for a sufficiently small absolute $c_1 > 0$. Here $P_j = P_{M_j}$ for $j = 0, 1$. Then $\|M_1\|_F \leq B$ and $D_{\mathrm{KL}}(P_1 \| P_0) = n\rho^2/(2\sigma^2) \leq c_1/2$. Writing TV for total variation distance, Le Cam's method and Pinsker's inequality give

$$\begin{aligned}\inf_{\widehat{M}} \sup_{M \in \mathcal{F}_B} \mathbb{E}[\left\|\widehat{M} - M\right\|_F^2] &\geq \frac{\rho^2}{8} \left(1 - \mathrm{TV}(P_0, P_1)\right) \\ &\geq \frac{\rho^2}{8} \left(1 - \sqrt{\frac{D_{\mathrm{KL}}(P_1 \| P_0)}{2}}\right).\end{aligned}$$

For small enough c_1, this is at least $c' \min\{B^2, \sigma^2/n\}$. Since $d < d_0$,

$$\min\left\{B^2, \frac{\sigma^2}{n}\right\} \geq \frac{1}{d_0^2} \min\left\{B^2, \frac{\sigma^2 d^2}{n}\right\}.$$

Changing the absolute constant completes the proof.

Remark 1 (Saturation and dimension)**.** The transition in equation 6 occurs at $n \asymp \sigma^2 d^2 / B^2$. This noise-dependent saturation scale is different from the algebraic requirement $n \geq d^2$ for a noiseless linear measurement map to be injective on the full matrix space.

A.2 Low-rank class

We prove Theorem 2 by first constructing a low-rank packing whose log cardinality is of order $2dr - r^2$, then applying the KL identity and Fano's inequality. In small dimensions, a rank-1 perturbation and Le Cam's two-point method give the lower bound.

Lemma 1 (Fixed-rank packing entropy)**.** *There are absolute constants $c_0, c_1, c_2 > 0$ such that, for every $1 \leq r \leq d$, $\rho > 0$, and $\varepsilon \in (0, c_0)$, the set $\mathcal{M}_r(\rho)$ contains an $\varepsilon\rho$-packing in Frobenius norm of cardinality at least $(c_1/\varepsilon)^{c_2(2dr - r^2)}$. Here an $\varepsilon\rho$-packing is a set whose distinct elements have pairwise Frobenius distance at least $\varepsilon\rho$. The packing can be chosen so that every element has Frobenius norm at least $c_0\rho$.*

Proof. Write $q = d - r$ and $p_r = 2dr - r^2 = r^2 + 2rq$. Two subclasses provide the required entropy.

First, matrices supported on the leading $r \times r$ block contain an r^2-dimensional Euclidean ball of radius $\rho/4$ centered at $(\rho/2)E_{11}$. This translated ball lies in $\mathcal{M}_r(\rho)$ and every one of its elements has norm between $\rho/4$ and $3\rho/4$. A volumetric packing therefore has cardinality at least $(c/\varepsilon)^{r^2}$.

Second, let $\mathbb{G}_{d,r}$ denote the set of r-dimensional linear subspaces of $\mathbb{R}^d$ (the Grassmann manifold), and let P_E be the orthogonal projection onto $E \in \mathbb{G}_{d,r}$. The matrices $M_E = (\rho/\sqrt{r})P_E$ have rank r and Frobenius norm ρ. Moreover, $\|M_E - M_F\|_F = (\rho/\sqrt{r}) \|P_E - P_F\|_F$. When $r \leq d/2$, the Grassmann packing bound in Frobenius projection distance (Pajor, 1998) supplies a packing of this subclass with cardinality at least $(c/\varepsilon)^{rq}$ and separation $\varepsilon\rho$. Since $q \geq r$ in this case, $p_r = r^2 + 2rq \leq 3rq$, so the Grassmann subclass alone gives the claimed exponent. When $r > d/2$, we instead use the leading-block Euclidean subclass: here $q < r$ and hence $p_r = r^2 + 2rq < 3r^2$, so its $(c/\varepsilon)^{r^2}$ packing gives the claimed exponent. This case split also avoids rescaling a packing of the orthogonal complements. When q/r is small, its diameter can be smaller than a constant fraction of ρ. Adjusting absolute constants proves the lemma.

Proof of Theorem 2. We again choose a radius that ensures feasibility and controls the KL divergence, now using a packing whose log-cardinality is of order p_r. Set

$$\rho^2 = a_0 \min\{B^2, \sigma^2 p_r/n\},$$

where $p_r = 2dr - r^2$, $\rho \geq 0$ is the packing radius, and $a_0 \in (0, 1]$ is an absolute constant chosen sufficiently small below. The two terms ensure $\rho \leq B$ and $n\rho^2/\sigma^2 \leq a_0 p_r$, respectively. Apply Lemma 1 at a fixed sufficiently small resolution $\varepsilon_0 < \min\{c_0, c_1/2\}$. The constants in the lemma then give $\log(c_1/\varepsilon_0) \geq \log 2$, so it yields matrices $M_1, \ldots, M_N \in \mathcal{M}_r(\rho) \subset \mathcal{M}_r(B)$ such that

$$\log N \geq cp_r, \qquad \|M_j - M_\ell\|_F \geq c\rho \quad (j \neq \ell).$$

Adjoin $M_0 = 0$. The annular property in Lemma 1 preserves constant-order separation from this additional point. With $P_0 = P_{M_0}$ and P_{M_j} as defined in Appendix A.1, the exact KL identity gives

$$D_{\mathrm{KL}}(P_{M_j} \| P_0) = \frac{n}{2\sigma^2} \|M_j\|_F^2 \leq \frac{n\rho^2}{2\sigma^2} \leq \frac{a_0 p_r}{2}.$$

For sufficiently small a_0, depending only on the absolute constants in Lemma 1, the average KL is at most a fixed small fraction of $\log N$. Fano's inequality then yields a lower bound of order ρ^2, proving equation 7 whenever p_r exceeds a fixed absolute constant. The remaining finite set of values is handled by the same rank-1 Le Cam construction used in Theorem 1. Since p_r is then bounded, changing the absolute constant gives the stated form.

A.3 Probability lower bounds

Proof of Corollary 1. Let $(\mathcal{C}, p)$ be either $(\mathcal{F}_B, d^2)$ or $(\mathcal{M}_r(B), 2dr - r^2)$, as in the corollary. For sufficiently large p, the constructions in Appendices A.1 and A.2 give N hypotheses $M_0, \ldots, M_{N-1}$ in $\mathcal{C}$, including $M_0 = 0$, with pairwise distances at least $2s$ for a separation radius $s > 0$, where $s^2 \geq a_1 \min\{B^2, \sigma^2 p/n\}$ and $\log N \geq c_1 p$. Taking the scale constants sufficiently small makes the average KL relative to the zero hypothesis at most $\alpha \log N$, with $\alpha \leq 1/4$. For any estimator, decode to its nearest packing point $\widehat{j}$, with a fixed tie-breaking rule. Under M_j, the event $\left\|\widehat{M} - M_j\right\|_F < s$ implies $\widehat{j} = j$. Fano's inequality yields

$$\sup_j \Pr_{M_j}\{\left\|\widehat{M} - M_j\right\|_F^2 \geq s^2\} \geq \frac{1}{N} \sum_j \Pr_{M_j}\{\widehat{j} \neq j\} \geq 1 - \alpha - \frac{\log 2}{\log N} \geq \frac{1}{2},$$

where the last inequality holds for sufficiently large p. For the remaining bounded values of p, take $M_0 = 0$ and $M_1 = \rho E_{11}$, with $\rho^2 = b \min\{B^2, \sigma^2/n\}$ for a sufficiently small absolute constant $b > 0$. Both points belong to either parameter class, and $D_{\mathrm{KL}}(P_{M_1} \| P_{M_0}) \leq b/2$. Nearest-point decoding, Le Cam's testing bound, and Pinsker's inequality give

$$\sup_{j=0,1} \Pr_{M_j}\{\left\|\widehat{M} - M_j\right\|_F^2 \geq \rho^2/4\} \geq \frac{1 - \mathrm{TV}(P_{M_0}, P_{M_1})}{2} \geq \frac{1}{4}.$$

Since p is bounded by an absolute constant in this case, $\rho^2/4$ is at least an absolute constant times $\min\{B^2, \sigma^2 p/n\}$. Combining the cases, decreasing a, taking for example $\eta_0 = 1/4$, and then taking the infimum over estimators proves the claim.

B Upper Bounds under Rank-1 Measurements

B.1 Deterministic ERM reduction

We first fix the observations and noise realizations, so the reduction below is deterministic. The population identity equation 5,

$$\mathcal{R}_{\mathrm{pred}}(M) - \mathcal{R}_{\mathrm{pred}}(M^*) = \|M - M^*\|_F^2,$$

motivates an empirical curvature lower bound that converts prediction error to Frobenius error.

B.1.1 ERROR DIRECTIONS AND CURVATURE EVENTS

Assume $M^* \in \mathcal{M}_r(B)$. We compare empirical squared loss and the noise inner product over all possible estimation error directions. Let $\mathcal{C}_{M^*} = \{t(M - M^*) : t \geq 0,\ M \in \mathcal{M}_r(B)\}$ describe these directions. Scaling allows a nonzero error to be normalized without changing its direction. For every feasible M,

$$\operatorname{rank}(M - M^*) \leq \min\{\operatorname{rank}(M) + \operatorname{rank}(M^*), d\} \leq k_r,$$

where $k_r = \min\{2r, d\}$. Thus every element of $\mathcal{C}_{M^*}$ has rank at most k_r. To express the noise term as a matrix inner product, define

$$G_n = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i x_{1,i} x_{2,i}^\top, \qquad \langle G_n, \Delta\rangle = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i x_{1,i}^\top \Delta x_{2,i}.$$

For $\kappa, \tau > 0$, define the restricted-curvature event $\mathcal{E}_{\mathrm{RC}}(\kappa)$ and noise-operator event $\mathcal{E}_{\mathrm{noise}}(\tau)$:

$$\mathcal{E}_{\mathrm{RC}}(\kappa) = \left\{ \inf_{\substack{\Delta \in \mathcal{C}_{M^*} \\ \Delta \neq 0}} \frac{n^{-1}\sum_{i=1}^{n} (x_{1,i}^\top \Delta x_{2,i})^2}{\|\Delta\|_F^2} \geq \kappa \right\}, \tag{21}$$

$$\mathcal{E}_{\mathrm{noise}}(\tau) = \left\{ \sup_{\substack{\Delta \in \mathcal{C}_{M^*} \\ \|\Delta\|_F = 1}} \langle G_n, \Delta\rangle \leq \tau \right\}. \tag{22}$$

Theorem 6 (Conditional rank-constrained error bound)**.** *Assume $M^* \in \mathcal{M}_r(B)$. On $\mathcal{E}_{\mathrm{RC}}(\kappa) \cap \mathcal{E}_{\mathrm{noise}}(\tau)$, every global minimizer in equation 9 satisfies*

$$\left\|\widehat{M} - M^*\right\|_F^2 \leq \frac{4}{\kappa^2}\tau^2. \tag{23}$$

B.1.2 SQUARED-LOSS EXPANSION AND THE BASIC INEQUALITY

Proof of Theorem 6. Write $\widehat{\Delta} = \widehat{M} - M^*$. Since $\widehat{M} \in \mathcal{M}_r(B)$, the definition of $\mathcal{C}_{M^*}$ gives $\widehat{\Delta} \in \mathcal{C}_{M^*}$. Writing $y_i = x_{1,i}^\top M^* x_{2,i} + \epsilon_i$ and $a_i = x_{1,i}^\top \widehat{\Delta} x_{2,i}$ gives

$$(y_i - x_{1,i}^\top \widehat{M} x_{2,i})^2 - (y_i - x_{1,i}^\top M^* x_{2,i})^2 = a_i^2 - 2\epsilon_i a_i.$$

Summing and using empirical optimality yields

$$\frac{1}{n}\sum_i (x_{1,i}^\top \widehat{\Delta} x_{2,i})^2 \leq 2\langle G_n, \widehat{\Delta}\rangle. \tag{24}$$

On $\mathcal{E}_{\mathrm{RC}}(\kappa) \cap \mathcal{E}_{\mathrm{noise}}(\tau)$, this inequality gives

$$\kappa \left\|\widehat{\Delta}\right\|_F^2 \leq \frac{1}{n}\sum_i (x_{1,i}^\top \widehat{\Delta} x_{2,i})^2 \leq 2\langle G_n, \widehat{\Delta}\rangle \leq 2\tau \left\|\widehat{\Delta}\right\|_F.$$

If $\widehat{\Delta} = 0$, the conclusion is immediate. Otherwise, dividing by $\kappa \left\|\widehat{\Delta}\right\|_F$ and squaring gives

$$\left\|\widehat{M} - M^*\right\|_F^2 = \left\|\widehat{\Delta}\right\|_F^2 \leq \frac{4\tau^2}{\kappa^2}.$$

B.1.3 EVENT CONTROL AND DIMENSION FACTORS

The deterministic bound becomes a high-probability guarantee once both events are controlled. The following theorem records the bounds proved in Appendix B.2, including their dependence on K.

Theorem 7 (Probabilities of the curvature and noise events)**.** *Assume $M^* \in \mathcal{M}_r(B)$ and let the inputs and observation noise satisfy the conditions of Theorem 3. For any $\eta \in (0,1)$, set $t_\eta = \log(8/\eta)$. There are constants $\kappa_K > 0$ and $C_K \geq 1$, depending only on K. Define*

$$\tau_\eta = C_K \sigma \sqrt{k_r} \left[\sqrt{\frac{d + t_\eta}{n}} + \frac{(d + t_\eta)\sqrt{\log(2n) + t_\eta}}{n} \right].$$

Then the following probability bounds hold:

$$n \geq C_K \left(k_r d + t_\eta\right) \quad \Longrightarrow \quad \Pr\{\mathcal{E}_{\mathrm{RC}}(\kappa_K)\} \geq 1 - \frac{\eta}{2}, \tag{25}$$

$$n \geq t_\eta \quad \Longrightarrow \quad \Pr\{\mathcal{E}_{\mathrm{noise}}(\tau_\eta)\} \geq 1 - \frac{\eta}{2}. \tag{26}$$

Combining Theorems 6 and 7 with a union bound proves Theorem 3. The full derivation is in Appendix B.2.

Both events are proved uniformly over the larger rank-at-most-k_r class, which contains all feasible error directions. For curvature, a fourth-moment bound supplies the small-ball probability, and an operator-norm estimate of the symmetrized rank-1 measurements supplies the Rademacher complexity (Lemma 4). For noise, G_n is the adjoint measurement operator applied to the noise. Conditional operator-norm concentration and nuclear–operator duality bound its inner product with every feasible error direction.

At fixed confidence, the leading noise threshold is $\sigma\sqrt{k_r d/n}$. The inequality $\|\Delta\|_* \leq \sqrt{k_r}\|\Delta\|_F$ produces the factor k_r after squaring, while operator-norm concentration contributes d. Under equation 12, the second noise term is absorbed, giving the error scale $\sigma^2 k_r d/n$. Since $r \leq k_r \leq 2r$ and $rd \leq 2dr - r^2 \leq 2rd$, this matches the low-rank minimax dimension scaling in the non-saturated regime. The bound holds uniformly over the entire feasible error class and establishes the optimal dimension order.

B.2 Restricted curvature and conditional noise bounds

We now prove the restricted-curvature and noise events used by the deterministic reduction. The final step substitutes both bounds into equation 23 to prove Theorem 3.

B.2.1 Curvature and noise events

Lemma 2 (Uniform small-ball curvature for rank-1 bilinear measurements)**.** *Assume $M^* \in \mathcal{M}_r(B)$ and let x_1, x_2 satisfy equation 2. There is $C_K \geq 1$ depending only on K. One may take $a_K = 1/\sqrt{2}$ and $q_K \geq cK^{-8}$, and set $\kappa_K = a_K^2 q_K/8 \geq cK^{-8}$. Then, for every Δ with $\|\Delta\|_F = 1$, $\mathbb{P}(|x_1^\top \Delta x_2| \geq a_K) \geq q_K$. Consequently, for any $\eta \in (0,1)$, with probability at least $1 - \eta$ over the sampled input pairs,*

$$\inf_{\substack{\Delta \in \mathcal{C}_{M^*} \\ \|\Delta\|_F = 1}} \frac{1}{n} \sum_{i=1}^{n} (x_{1,i}^\top \Delta x_{2,i})^2 \geq \kappa_K$$

once $n \geq C_K\big(k_r d + \log(8/\eta)\big)$, by applying the small-ball method to the containing rank-$\leq k_r$ unit-Frobenius set (Appendix B.2). The constants depend only on K. An explicit conservative curvature sample size bound is derived in that appendix.

Dependence on the input tails. For $\|\Delta\|_F = 1$, put $W = x_1^\top \Delta x_2$. Isotropy gives $\mathbb{E}[W^2] = 1$, while the directional sub-Gaussian assumption gives $\mathbb{E}[W^4] \leq CK^8$. Paley–Zygmund then yields the small-ball probability $q_K \geq cK^{-8}$. Together with the Rademacher bound $CK^2\sqrt{k_r d/n}$, this yields the conservative curvature condition stated after Theorem 3. The derivation below tracks these factors.

Lemma 3 (Rank-restricted noise bound)**.** *Let $1 \leq k \leq d$, $\eta \in (0,1)$, and $\mathcal{T}_k = \{\Delta : \mathrm{rank}(\Delta) \leq k, \|\Delta\|_F = 1\}$. Under equation 2, suppose also that the ϵ_i are independent across samples, mean zero, independent of the inputs, and satisfy the sub-Gaussian condition above with variance proxy*

σ^2. *Set* $t = \log(8/\eta)$ *and assume* $n \geq t$. *There is* $C_K \geq 1$ *depending only on* K *such that, with probability at least* $1 - \eta/2$,

$$\sup_{\Delta \in \mathcal{T}_k} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i x_{1,i}^\top \Delta x_{2,i} \right| \leq C_K \sigma \sqrt{k} \left\{ \sqrt{\frac{d+t}{n}} + \frac{(d+t)\sqrt{\log(2n)+t}}{n} \right\}. \tag{27}$$

The proof conditions on the noise vector, applies a weighted Bernstein inequality on two fixed sphere nets, and then uses $\|\Delta\|_* \leq \sqrt{k}\|\Delta\|_F$. *Details are in Appendix B.2.*

Remark 2 (Conditional interpretation of the noise bound)**.** To obtain a high-probability estimation error bound, we first bound the operator norm of the noise matrix G_n. Conditioning on $\epsilon = (\epsilon_i)_{i=1}^n$ fixes the noise weights while preserving independence and sub-exponential tails of the bilinear measurements. Weighted Bernstein and sphere nets then give an operator-norm bound depending on both $\|\epsilon\|_2$ and $\|\epsilon\|_\infty$. Substituting high-probability bounds for these norms yields equation 26. The bound on $\|\epsilon\|_\infty$ introduces the logarithmic factor in the second term of the noise threshold and the $d\log(2d)$ term in the fixed-confidence sufficient sample size condition. Step 4 of Appendix B.2 gives the derivation.

B.2.2 Completion of the rank-constrained ERM proof

Let

$$\mathcal{T}_k = \{\Delta \in \mathbb{R}^{d\times d} : \operatorname{rank}(\Delta) \leq k,\ \|\Delta\|_F = 1\}, \qquad f_\Delta(x_1, x_2) = x_1^\top \Delta x_2.$$

Here $\mathcal{T}_k$ is the class of unit-Frobenius matrices of rank at most k, and f_Δ is the score in direction Δ. Write $X = (x_1, x_2)$ and $X_i = (x_{1,i}, x_{2,i})$ for sample i. The class $\mathcal{T}_{k_r}$ contains every normalized feasible difference.

Step 1: population moments and a uniform small-ball margin. For a fixed matrix Δ, independence and isotropy give

$$\mathbb{E}[(x_1^\top \Delta x_2)^2] = \|\Delta\|_F^2,$$

while conditioning successively on x_2 and x_1, and using the directional ψ_2 assumption, gives

$$\mathbb{E}[(x_1^\top \Delta x_2)^4] \leq CK^8 \|\Delta\|_F^4.$$

The fourth-moment calculation applies to matrices of every rank and is expanded in the proof of Lemma 5. Applying Paley–Zygmund to f_Δ^2 yields constants $a_K, q_K > 0$, depending only polynomially on K, such that

$$\inf_{\|\Delta\|_F = 1} \Pr\{|f_\Delta(x_1, x_2)| \geq a_K\} \geq q_K. \tag{28}$$

Here a_K is a score-magnitude threshold and q_K is a uniform lower bound on the probability of exceeding it.

Step 2: class geometry and direct complexity control. Uniform control of the function class makes the curvature guarantee apply to data-dependent estimation errors. The following lemma describes the class through covering and increment moment bounds, and establishes its symmetrized complexity directly through an operator-norm estimate. Step 3 combines this complexity bound with the population small-ball margin.

Lemma 4 (Complexity of the rank-constrained bilinear class)**.** *Let* $1 \leq k \leq d$. *For* $0 < u < 1$, *the unit-Frobenius rank-at-most-*k *class satisfies the following covering bound, where* $N(\mathcal{T}_k, \|\cdot\|_F, u)$ *is the minimum number of Frobenius balls of radius* u *needed to cover* $\mathcal{T}_k$*:*

$$\log N(\mathcal{T}_k, \|\cdot\|_F, u) \leq (2d+1)k\log(9/u).$$

Here $\|W\|_{L_p} = (\mathbb{E}[|W|^p])^{1/p}$ *denotes a moment norm, so* p *in the next display is a moment order, not a parameter count. For every* $p \geq 2$ *and* $A = \Delta - \Delta'$ *with* $\Delta, \Delta' \in \mathbb{R}^{d\times d}$,

$$\|f_\Delta - f_{\Delta'}\|_{L_2} = \|\Delta - \Delta'\|_F, \qquad \|f_\Delta - f_{\Delta'}\|_{L_p} \leq CK^2\big(\sqrt{p}\|A\|_F + p\|A\|_{\mathrm{op}}\big). \tag{29}$$

In addition, if δ_i *are independent random signs taking* $+1$ *and* -1 *with equal probability, independent of all inputs, and* $n \geq kd$, *then*

$$\mathbb{E}\Big[\sup_{\Delta\in\mathcal{T}_k} \left| \frac{1}{\sqrt{n}} \sum_i \delta_i f_\Delta(x_{1,i}, x_{2,i}) \right|\Big] \leq CK^2\sqrt{kd}. \tag{30}$$

Here C *is an absolute constant.*

Proof. Covering and moment bounds. This part establishes the lemma's geometric and increment statements. Nuclear–operator duality then gives the direct Rademacher bound equation 30. Step 3 combines this bound with the small-ball probability to establish uniform curvature. Write each matrix in $\mathcal{T}_k$ in singular value form as $USV^\top$, where $U, V \in \mathbb{R}^{d\times k}$ have orthonormal columns and $S \in \mathbb{R}^{k\times k}$ is diagonal with nonnegative entries and $\|S\|_F = 1$. For matrices of rank below k, pad the singular values with zeros and complete the orthonormal columns. We make the factor-covering argument explicit (Candès and Plan, 2011). Set $h = u/3$. A maximal h-separated subset of any set in the unit ball of a D-dimensional normed space is an h-net of that set. The disjoint balls of radius $h/2$ centered at its points lie in the ball of radius $1 + h/2$, so volume comparison bounds its size by $(1 + 2/h)^D$. Apply this argument in the operator norm on $\mathbb{R}^{d\times k}$ to the orthonormal factors, whose operator norms equal 1, and in the Euclidean norm on $\mathbb{R}^k$ to the nonnegative unit singular-value vectors. This gives nets $\mathcal{U}_h, \mathcal{V}_h, \mathcal{D}_h$ with centers in the respective factor sets and cardinalities

$$|\mathcal{U}_h|, |\mathcal{V}_h| \le (1 + 2/h)^{dk}, \qquad |\mathcal{D}_h| \le (1 + 2/h)^k,$$

where $\mathcal{D}_h$ denotes the corresponding diagonal-matrix net, measured in Frobenius norm. Choose net points $\widetilde{U}, \widetilde{V}, \widetilde{S}$ within distance h of the three factors and put $\widetilde{\Delta} = \widetilde{U}\widetilde{S}\widetilde{V}^\top \in \mathcal{T}_k$. The factor errors propagate as

$$\begin{aligned}\|USV^\top - \widetilde{\Delta}\|_F &\le \|(U - \widetilde{U})SV^\top\|_F + \|\widetilde{U}(S - \widetilde{S})V^\top\|_F + \|\widetilde{U}\widetilde{S}(V - \widetilde{V})^\top\|_F \\ &\le \|U - \widetilde{U}\|_{\mathrm{op}}\|S\|_F + \|S - \widetilde{S}\|_F + \|\widetilde{S}\|_F\|V - \widetilde{V}\|_{\mathrm{op}} \le 3h = u.\end{aligned}$$

Thus the products of the net points form a Frobenius u-net, and

$$N(\mathcal{T}_k, \|\cdot\|_F, u) \le (1 + 6/u)^{2dk+k} \le (9/u)^{(2d+1)k}, \qquad 0 < u < 1.$$

The L_2 identity follows from isotropy. For the moment bound, let $g \sim \mathcal{N}(0, I_d)$ be independent of x_2 and put $a_p = (\mathbb{E}|g_1|^p)^{1/p} \asymp \sqrt{p}$, where g_1 is the first coordinate of g and $p \ge 2$. Conditioning first on x_2 and then on g gives

$$\begin{aligned}\big\|\|Ax_2\|_2\big\|_{L_p} &= a_p^{-1}\big(\mathbb{E}_g\mathbb{E}_{x_2}|\langle A^\top g, x_2\rangle|^p\big)^{1/p} \\ &\le \frac{CK\sqrt{p}}{a_p}\big\|\|A^\top g\|_2\big\|_{L_p} \le CK\big\|\|A^\top g\|_2\big\|_{L_p}.\end{aligned}$$

Here the equality uses the conditional Gaussian distribution of $\langle g, Ax_2\rangle$, and the first inequality follows from the directional sub-Gaussian assumption on x_2. The function $F(g) = \|A^\top g\|_2$ is $\|A\|_{\mathrm{op}}$-Lipschitz, so Gaussian concentration and its moment bound yield

$$\begin{aligned}\|F(g)\|_{L_p} &\le \mathbb{E}F(g) + \|F(g) - \mathbb{E}F(g)\|_{L_p} \\ &\le \big(\mathbb{E}\|A^\top g\|_2^2\big)^{1/2} + C\sqrt{p}\,\|A\|_{\mathrm{op}} = \|A\|_F + C\sqrt{p}\,\|A\|_{\mathrm{op}}.\end{aligned}$$

Finally, conditioning on x_2 and using independence of the two inputs,

$$\|x_1^\top Ax_2\|_{L_p} \le CK\sqrt{p}\,\big\|\|Ax_2\|_2\big\|_{L_p} \le CK^2\big(\sqrt{p}\,\|A\|_F + p\|A\|_{\mathrm{op}}\big),$$

which proves equation 29 for independent inputs that each allow dependence among their coordinates.

Direct symmetrized complexity bound. Set $H = n^{-1/2}\sum_i \delta_i x_{1,i}x_{2,i}^\top$. For fixed unit vectors u, v, the summands $\delta_i\langle u, x_{1,i}\rangle\langle v, x_{2,i}\rangle$ are centered sub-exponential with norm at most CK^2. Scalar Bernstein gives, for $s > 0$,

$$\Pr\left\{|u^\top Hv| > CK^2\left(\sqrt{s} + \frac{s}{\sqrt{n}}\right)\right\} \le 2e^{-s}.$$

Let $\mathcal{N}_1, \mathcal{N}_2$ be $1/4$-nets of the Euclidean unit sphere, each of cardinality at most 9^d. Approximating any two unit vectors by net points gives

$$\|H\|_{\mathrm{op}} \le \max_{u\in\mathcal{N}_1, v\in\mathcal{N}_2} |u^\top Hv| + \tfrac{1}{2}\|H\|_{\mathrm{op}},$$

so the operator norm is at most twice the displayed maximum. A union bound over at most 9^{2d} pairs, with $s = 2d\log 9 + t$, yields

$$\Pr\{\|H\|_{\mathrm{op}} > b(t)\} \le 2e^{-t}, \qquad t \ge 0,$$

where $b(t) = CK^2\{\sqrt{d+t} + (d+t)/\sqrt{n}\}$ and C is a sufficiently large absolute constant. To pass from this tail bound to an expectation, split the tail integral at $b(0)$ and substitute $z = b(t)$ above it:

$$\begin{aligned}\mathbb{E}\|H\|_{\mathrm{op}} &= \int_0^\infty \Pr\{\|H\|_{\mathrm{op}} > z\}\, dz \\ &\le b(0) + \int_0^\infty \Pr\{\|H\|_{\mathrm{op}} > b(t)\}\, b'(t)\, dt \\ &\le b(0) + 2CK^2 \int_0^\infty e^{-t} \left\{ \frac{1}{2\sqrt{d+t}} + \frac{1}{\sqrt{n}} \right\} dt \\ &\le b(0) + CK^2 \left(\frac{1}{\sqrt{d}} + \frac{2}{\sqrt{n}} \right) \le CK^2 \left(\sqrt{d} + \frac{d}{\sqrt{n}} \right),\end{aligned}$$

where the last inequality uses $d \ge 1$, and absolute constants are absorbed into C. Finally, nuclear–operator duality and the rank bound give

$$\sup_{\Delta \in \mathcal{T}_k} |\langle H, \Delta \rangle| \le \|H\|_{\mathrm{op}} \sup_{\Delta \in \mathcal{T}_k} \|\Delta\|_* \le \sqrt{k} \|H\|_{\mathrm{op}}.$$

Taking expectations proves equation 30 when $n \ge d$. The condition $n \ge kd$ in the lemma implies this requirement. The argument uses only directional sub-Gaussian bounds and allows dependent coordinates within each input vector.

Step 3: uniform empirical curvature. Define the expected Rademacher complexity $\mathfrak{R}_n(\mathcal{T}_k)$ and the smallest tail probability $Q_k(v)$ across the class, for $v > 0$, by

$$\mathfrak{R}_n(\mathcal{T}_k) = \mathbb{E}\Big[\sup_{\Delta \in \mathcal{T}_k} \left| \frac{1}{n} \sum_{i=1}^n \delta_i f_\Delta(X_i) \right| \Big], \qquad Q_k(v) = \inf_{\Delta \in \mathcal{T}_k} \Pr\{|f_\Delta(X)| \ge v\}.$$

To pass from the population small-ball probability to empirical curvature, let $P_n h = n^{-1} \sum_i h(X_i)$ and $Ph = \mathbb{E}[h(X)]$, where X has the same distribution as each X_i. For $u > 0$, define $h_u(s) = \min\{1, \max\{0, |s|/u - 1\}\}$ for $s \in \mathbb{R}$. Then $s^2 \ge u^2 h_u(s)$, $h_u(s) \ge \mathbb{1}\{|s| \ge 2u\}$, and h_u is $1/u$-Lipschitz with $h_u(0) = 0$. Thus

$$\begin{aligned}\inf_{\Delta \in \mathcal{T}_k} P_n f_\Delta^2 &\ge u^2 \inf_{\Delta \in \mathcal{T}_k} P_n h_u(f_\Delta) \\ &\ge u^2 \left\{ \inf_{\Delta \in \mathcal{T}_k} P h_u(f_\Delta) - D_u \right\} \ge u^2 \{ Q_k(2u) - D_u \},\end{aligned}$$

where $D_u = \sup_{\Delta \in \mathcal{T}_k} |(P_n - P) h_u(f_\Delta)|$ is the uniform empirical deviation. Symmetrization followed by contraction gives

$$\begin{aligned}\mathbb{E} D_u &\le 2\mathbb{E} \sup_{\Delta \in \mathcal{T}_k} \left| \frac{1}{n} \sum_i \delta_i h_u(f_\Delta(X_i)) \right| \\ &\le \frac{4}{u} \mathbb{E} \sup_{\Delta \in \mathcal{T}_k} \left| \frac{1}{n} \sum_i \delta_i f_\Delta(X_i) \right| = \frac{4 \mathfrak{R}_n(\mathcal{T}_k)}{u}.\end{aligned}$$

Since $0 \le h_u \le 1$, replacing one sample changes D_u by at most $1/n$. The bounded-differences inequality therefore gives $\Pr\{D_u > \mathbb{E} D_u + z\} \le e^{-2nz^2}$ for every $z > 0$. Substituting this deviation bound above proves the small-ball inequality: with probability at least $1 - 2e^{-cnz^2}$ for an absolute $c > 0$,

$$\inf_{\Delta \in \mathcal{T}_k} \frac{1}{n} \sum_{i=1}^n f_\Delta(X_i)^2 \ge u^2 \left\{ Q_k(2u) - \frac{4 \mathfrak{R}_n(\mathcal{T}_k)}{u} - z \right\}. \tag{31}$$

By equation 30, $\mathfrak{R}_n(\mathcal{T}_k) \le CK^2 \sqrt{kd/n}$. Taking $u = a_K/2$ and $z = C\sqrt{\log(4/\eta)/n}$ in equation 31, and using $Q_k(a_K) \ge q_K$, makes the expression in braces at least $q_K/2$ whenever $n \ge C_K(kd + \log(4/\eta))$. Consequently, with probability at least $1 - \eta/2$,

$$\inf_{\Delta \in \mathcal{T}_k} \frac{1}{n} \sum_{i=1}^n f_\Delta(X_i)^2 \ge \frac{a_K^2 q_K}{8}.$$

Taking $k = k_r$ proves equation 25, after absorbing the factor two into C_K.

Keeping track of K in Step 2 gives $\mathfrak{R}_n(\mathcal{T}_k) \le CK^2\sqrt{kd/n}$ for $n \ge d$. Since $a_K = 1/\sqrt{2}$ and $q_K \ge cK^{-8}$, it suffices to make both $K^2\sqrt{kd/n}$ and $\sqrt{\log(4/\eta)/n}$ at most a sufficiently small constant times K^{-8}. A sufficient condition is therefore

$$n \ge C\{K^{20}kd + K^{16}\log(4/\eta)\}.$$

In particular, one may choose a curvature sample size constant $C_{\mathrm{curv}}(K) \le CK^{20}$ in the coarser condition $n \ge C_{\mathrm{curv}}(K)\{kd + \log(4/\eta)\}$, since $K \ge 1$. This tracks the K dependence of the curvature sample size explicitly. The noise and estimation error bounds retain their respective K-dependent constants.

Remark 3 (Rank bound for the feasible-difference cone)**.** The feasible-difference cone is contained in the rank-at-most-k_r class, so the uniform bound above covers all feasible error directions.

Step 4: conditional Bernstein–net control of the noise term. Write $G_n = n^{-1}\sum_{i=1}^n \epsilon_i x_{1,i}x_{2,i}^\top$. Fix unit vectors u, v and condition on $\epsilon = (\epsilon_1, \dots, \epsilon_n)$. Independence of $x_{1,i}$ and $x_{2,i}$ implies that the variables $Y_i = \epsilon_i\langle u, x_{1,i}\rangle\langle v, x_{2,i}\rangle$ are conditionally independent and centered, with weighted sub-exponential scale $\|Y_i\|_{\psi_1|\epsilon} \le CK^2|\epsilon_i|$, where $\|\cdot\|_{\psi_1|\epsilon}$ is the sub-exponential norm under the conditional distribution given ϵ. Weighted Bernstein therefore gives, for every $s \ge 1$,

$$\Pr\left\{\left|\sum_i Y_i\right| > CK^2\big(\|\epsilon\|_2\sqrt{s} + \|\epsilon\|_\infty s\big) \,\middle|\, \epsilon\right\} \le 2e^{-s}. \tag{32}$$

Let $\mathcal{N}$ be a $1/4$-net of the unit sphere with $|\mathcal{N}| \le 9^d$. The standard bilinear net inequality gives

$$\|G_n\|_{\mathrm{op}} \le 2\max_{u,v\in\mathcal{N}} |u^\top G_n v|.$$

For $t > 0$, taking $s = Cd + t$ in equation 32 and applying a union bound over $\mathcal{N} \times \mathcal{N}$ yields, conditionally with failure probability at most $2e^{-t}$,

$$\|G_n\|_{\mathrm{op}} \le \frac{C_K}{n}\left\{\|\epsilon\|_2\sqrt{d+t} + \|\epsilon\|_\infty(d+t)\right\}. \tag{33}$$

For independent mean-zero sub-Gaussian noise, if $t \le n$, then with failure probability at most $2e^{-t}$,

$$\|\epsilon\|_2 \le C\sigma\sqrt{n}, \qquad \|\epsilon\|_\infty \le C\sigma\sqrt{\log(2n) + t}.$$

Combining these displays and using

$$\sup_{\Delta\in\mathcal{T}_k} |\langle G_n, \Delta\rangle| \le \sqrt{k}\|G_n\|_{\mathrm{op}}$$

proves equation 27. Taking $k = k_r$ proves equation 26. With $t = t_\eta = \log(8/\eta)$, the total failure probability of these tail events is at most $\eta/2$.

Step 5: completion of the ERM bound.

Proof of Theorem 3. The preceding steps establish the events summarized in Theorem 7. By a union bound, they hold simultaneously with probability at least $1 - \eta$. On their intersection, apply the deterministic bound equation 23 with $\kappa = \kappa_K$ and $\tau = \tau_\eta$, where κ_K is the curvature lower bound and τ_η is the noise threshold in equation 26. Writing $\widehat{\Delta} = \widehat{M} - M^*$ and $t = t_\eta = \log(8/\eta)$ gives

$$\begin{aligned}
\|\widehat{\Delta}\|_F^2 &\le \frac{4\tau_\eta^2}{\kappa_K^2} \\
&\le C_K\sigma^2 k_r\left[\sqrt{\frac{d+t}{n}} + \frac{(d+t)\sqrt{\log(2n)+t}}{n}\right]^2 \\
&\le C_K\sigma^2 k_r\left\{\frac{d+t}{n} + \frac{(d+t)^2(\log(2n)+t)}{n^2}\right\}.
\end{aligned}$$

Here the last inequality uses $(a+b)^2 \le 2a^2 + 2b^2$, and C_K denotes constants depending only on K that may change from line to line. This proves equation 11 under the base sample size condition. Under equation 12, the second term is absorbed by the first. To see this explicitly, put $D = d+t$ and $A = \log(2D) + t$. The sample size condition implies $n \ge C_K DA$. At the threshold $n_0 = C_K DA$,

$$\log(2n_0) + t \le C'_K\{\log(2D) + t\} = C'_K A.$$

Moreover, $s \mapsto (\log(2s)+t)/s$ is decreasing above this threshold after enlarging C_K by an absolute factor. Hence, for every $n \ge n_0$,

$$\frac{D\{\log(2n) + t\}}{n} \le C''_K,$$

where C''_K depends only on K. The left-hand side is the ratio of the second error term to the first. Absorbing this bounded ratio into C_K yields

$$\begin{aligned}\left\|\widehat{M} - M^*\right\|_F^2 &\le C_K\sigma^2\frac{k_r(d+t)}{n}\left[1 + \frac{(d+t)(\log(2n)+t)}{n}\right] \\ &\le C_K\sigma^2\frac{k_r(d+t)}{n},\end{aligned}$$

which proves equation 13. At fixed confidence, $k_r d \asymp rd \asymp 2dr - r^2$.

B.3 Nuclear norm counterpart

We combine the small-ball and operator-noise bounds with nuclear norm decomposability (Negahban et al., 2012) to obtain a convex counterpart under the same input assumptions. Write $A_i = x_{1,i}x_{2,i}^\top$ and $y_i = \langle A_i, M^*\rangle + \epsilon_i$. The input pairs are independent across samples, and within each pair the two vectors are independent and satisfy equation 2. The noises are independent across samples and independent of all inputs, with $\mathbb{E}[e^{u\epsilon_i}] \le e^{\sigma^2 u^2/2}$ for all $u \in \mathbb{R}$ and $\sigma > 0$. Consider the estimator

$$\widehat{M}_\lambda \in \operatorname*{arg\,min}_{M\in\mathbb{R}^{d\times d}}\left\{\frac{1}{n}\sum_{i=1}^n(y_i - \langle A_i, M\rangle)^2 + \lambda\|M\|_*\right\}. \tag{34}$$

This uses the same squared-loss normalization as the nuclear norm objective in equation 15. For $\lambda > 0$, the objective is continuous and coercive, so a minimizer exists.

The proof has two parts. We first establish curvature on a set of unit-Frobenius directions with bounded nuclear norm, allowing full-rank errors. We then show that choosing the penalty to dominate the noise forces the normalized estimation error into this set. The basic inequality and curvature bound together yield the error rate.

B.3.1 Curvature on the nuclear norm error set

Let $1 \le r \le d$, $\eta \in (0,1)$, and $t_\eta = \log(8/\eta)$. Define the compact set $\mathcal{S}_r$ and the empirical quadratic form Q_n by

$$\mathcal{S}_r = \{\Delta : \|\Delta\|_F = 1,\ \|\Delta\|_* \le 4\sqrt{2r}\}, \qquad Q_n(\Delta) = \frac{1}{n}\sum_i\langle A_i, \Delta\rangle^2. \tag{35}$$

The set $\mathcal{S}_r$ accommodates full-rank directions with controlled nuclear norm. Its curvature applies to the estimation error through the cone inequality established below.

Lemma 5 (Small-ball curvature on a nuclear norm set)**.** *Under the input assumptions above, there are constants $C_{\mathrm{curv}}(K) \ge 1$ and $\kappa_{\mathrm{nuc}}(K) > 0$, depending only on K, such that if $n \ge C_{\mathrm{curv}}(K)(rd + t_\eta)$, then*

$$\Pr\left\{\inf_{\Delta\in\mathcal{S}_r} Q_n(\Delta) \ge \kappa_{\mathrm{nuc}}(K)\right\} \ge 1 - \eta/2. \tag{36}$$

A sufficient sample size condition is $n \ge C\{K^{20}rd + K^{16}t_\eta\}$ for an absolute C.

Proof. For every $\|\Delta\|_F = 1$, the moment calculation in Step 1 of Appendix B.2 gives $\mathbb{E}[\langle A_i, \Delta\rangle^2] = 1$ and $\mathbb{E}[\langle A_i, \Delta\rangle^4] \le CK^8$ for every sample i. To see explicitly that this uses no rank restriction, write the rows of Δ as $b_j^\top$. Conditioning on x_2 bounds the fourth moment by $CK^4\mathbb{E}[\|\Delta x_2\|_2^4]$, and Cauchy–Schwarz gives

$$\mathbb{E}[\|\Delta x_2\|_2^4] \le \sum_{j,\ell} \sqrt{\mathbb{E}[\langle b_j, x_2\rangle^4]\mathbb{E}[\langle b_\ell, x_2\rangle^4]} \le CK^4\Big(\sum_j \|b_j\|_2^2\Big)^2.$$

Paley–Zygmund therefore yields $a = 1/\sqrt{2}$ and $q = cK^{-8}$, for a sufficiently small absolute $c > 0$, such that $\inf_{i,\Delta\in\mathcal{S}_r} \Pr\{|\langle A_i, \Delta\rangle| \ge a\} \ge q$.

Let δ_i be independent Rademacher signs independent of the data and $H = n^{-1/2}\sum_i \delta_i A_i$. The Bernstein–net argument in Step 2 of Appendix B.2 gives $\mathbb{E}[\|H\|_{\mathrm{op}}] \le CK^2(\sqrt{d} + d/\sqrt{n})$. Nuclear–operator norm duality then implies

$$\begin{aligned}\mathfrak{R}_n(\mathcal{S}_r) &:= \mathbb{E}\Big[\sup_{\Delta\in\mathcal{S}_r}\Big|\frac{1}{n}\sum_i \delta_i\langle A_i, \Delta\rangle\Big|\Big] \\ &\le \frac{4\sqrt{2r}}{\sqrt{n}}\mathbb{E}[\|H\|_{\mathrm{op}}] \le CK^2\sqrt{rd/n}, \qquad n \ge d. \end{aligned} \tag{37}$$

For $u > 0$ and $s \in \mathbb{R}$, let $h_u(s)$ equal 0 for $|s| \le u$, 1 for $|s| \ge 2u$, and $|s|/u - 1$ between these thresholds. Then $s^2 \ge u^2 h_u(s)$, h_u is $1/u$-Lipschitz, and $0 \le h_u \le 1$. Symmetrization, contraction, and bounded differences, as in equation 31, give for every $z > 0$, with probability at least $1 - 2e^{-cnz^2}$,

$$\inf_{\Delta\in\mathcal{S}_r} Q_n(\Delta) \ge u^2\left\{\inf_{\Delta\in\mathcal{S}_r}\frac{1}{n}\sum_i \Pr(|\langle A_i, \Delta\rangle| \ge 2u) - 4\mathfrak{R}_n(\mathcal{S}_r)/u - z\right\}.$$

Take $u = a/2$ and $z = C\sqrt{t_\eta/n}$. The two subtraction terms sum to at most $q/2$ once $n \ge C\{K^{20}rd + K^{16}t_\eta\}$. This condition also ensures $n \ge d$, since $K, r \ge 1$. The failure probability is at most $\eta/2$, and the lower bound is $a^2q/8$. Set $\kappa_{\mathrm{nuc}}(K) = a^2q/8$ and choose $C_{\mathrm{curv}}(K) \le CK^{20}$ sufficiently large.

B.3.2 A CONVEX TWO-TERM ERROR BOUND

Define the deterministic noise scale $b_{n,\eta}$ by

$$b_{n,\eta} = \sqrt{\frac{d + t_\eta}{n}} + \frac{(d + t_\eta)\sqrt{\log(2n) + t_\eta}}{n}. \tag{38}$$

Let $C_{\mathrm{noise}}(K)$ be a constant in the operator bound obtained from Lemma 3 with $k = 1$, so that for $n \ge t_\eta$,

$$\Pr\{\|G_n\|_{\mathrm{op}} \le C_{\mathrm{noise}}(K)\sigma b_{n,\eta}\} \ge 1 - \eta/2. \tag{39}$$

Theorem 8 (Nuclear norm upper bound)**.** *Under the input and noise assumptions of this appendix, suppose* $\mathrm{rank}(M^*) \le r$*, where* $1 \le r \le d$*. Choose*

$$\lambda = 4C_{\mathrm{noise}}(K)\sigma b_{n,\eta}. \tag{40}$$

If $n \ge C_{\mathrm{curv}}(K)(rd + t_\eta)$*, then with probability at least* $1 - \eta$*, every minimizer in equation 34 satisfies*

$$\|\widehat{M}_\lambda - M^*\|_F^2 \le C_{\mathrm{rate}}(K)\sigma^2 r\left\{\frac{d + t_\eta}{n} + \frac{(d + t_\eta)^2(\log(2n) + t_\eta)}{n^2}\right\}, \tag{41}$$

where one may take $C_{\mathrm{rate}}(K) = 144C_{\mathrm{noise}}(K)^2/\kappa_{\mathrm{nuc}}(K)^2$*. There is* $C_{\mathrm{simp}}(K) \ge C_{\mathrm{curv}}(K)$ *such that the further condition*

$$n \ge C_{\mathrm{simp}}(K)\{rd + t_\eta + (d + t_\eta)[\log(2(d + t_\eta)) + t_\eta]\} \tag{42}$$

gives, on the same event,

$$\|\widehat{M}_\lambda - M^*\|_F^2 \le 2C_{\mathrm{rate}}(K)\sigma^2\frac{r(d + t_\eta)}{n}, \tag{43}$$

where all constants depend only on K*.*

Proof. Intersect the events in Lemma 5 and equation 39. Their joint probability is at least $1-\eta$, and $\|G_n\|_{\mathrm{op}} \le \lambda/4$ there. Put $\Delta = \widehat{M}_\lambda - M^*$. The factor 4 in equation 40 ensures that the noise contribution is at most $(\lambda/2)\|\Delta\|_*$, leaving part of the penalty to control error outside the target's singular subspaces. If $M^* = 0$, optimality gives $Q_n(\Delta) + \lambda\|\Delta\|_* \le 2\langle G_n, \Delta\rangle \le (\lambda/2)\|\Delta\|_*$, so $\Delta = 0$. Otherwise let $M^* = USV^\top$ be its compact singular value decomposition, where U, V contain the left and right singular vectors corresponding to nonzero singular values and S is the diagonal matrix of those values. With $I = I_d$ denoting the identity matrix, define the two error components

$$\Delta_\perp = (I - UU^\top)\Delta(I - VV^\top), \qquad \Delta_T = \Delta - \Delta_\perp.$$

These components are Frobenius orthogonal, and $\mathrm{rank}(\Delta_T) \le 2r$. Here $\Delta_\perp$ has column and row spaces orthogonal to those of M^*, while Δ_T is the remaining component. This decomposition lets the penalty control $\Delta_\perp$ through the low-rank component Δ_T. Since M^* and $\Delta_\perp$ have orthogonal row and column spaces, their nuclear norms add. The triangle inequality therefore yields

$$\|M^*\|_* - \|M^* + \Delta\|_* \le \|\Delta_T\|_* - \|\Delta_\perp\|_*.$$

Expanding the squared loss in the optimality inequality and using duality,

$$\begin{aligned} Q_n(\Delta) &\le 2\langle G_n, \Delta\rangle + \lambda(\|M^*\|_* - \|M^* + \Delta\|_*) \\ &\le \frac{3\lambda}{2}\|\Delta_T\|_* - \frac{\lambda}{2}\|\Delta_\perp\|_*. \end{aligned} \tag{44}$$

Since $Q_n(\Delta) \ge 0$, we first obtain $\|\Delta_\perp\|_* \le 3\|\Delta_T\|_*$. Consequently,

$$\begin{aligned} \|\Delta\|_* &\le \|\Delta_T\|_* + \|\Delta_\perp\|_* \le 4\|\Delta_T\|_* \\ &\le 4\sqrt{2r}\|\Delta_T\|_F \le 4\sqrt{2r}\|\Delta\|_F, \end{aligned} \tag{45}$$

where the last two inequalities use $\mathrm{rank}(\Delta_T) \le 2r$ and $\|\Delta\|_F^2 = \|\Delta_T\|_F^2 + \|\Delta_\perp\|_F^2$, respectively. For $\Delta \ne 0$, dividing by $\|\Delta\|_F$ shows that its normalization belongs to $\mathcal{S}_r$. Lemma 5 and equation 44 then imply

$$\kappa_{\mathrm{nuc}}(K)\|\Delta\|_F^2 \le \frac{3\lambda}{2}\sqrt{2r}\|\Delta\|_F, \qquad \|\Delta\|_F^2 \le \frac{9r\lambda^2}{2\kappa_{\mathrm{nuc}}(K)^2}.$$

Substitute equation 40 and use $(a+b)^2 \le 2a^2 + 2b^2$ to obtain equation 41 with the stated $C_{\mathrm{rate}}(K)$. For the simplified rate, put $D = d + t_\eta$, $A = \log(2D) + t_\eta > 1$, and $c = C_{\mathrm{simp}}(K) \ge \max\{4, C_{\mathrm{curv}}(K)\}$. At $n_0 = cDA$, the ratio of the second term to the first satisfies

$$\frac{D(\log(2n_0) + t_\eta)}{n_0} = \frac{A + \log c + \log A}{cA} \le \frac{2 + \log c}{c} \le 1.$$

The ratio decreases for $n \ge n_0$. Condition equation 42 therefore proves equation 43.

Radius, tuning, and interpretation. Minimizing equation 34 over $\mathcal{F}_B$ with $\|M^*\|_F \le B$ preserves the error guarantee: comparison with the feasible truth gives the same basic inequality. The guarantee uses the theoretical choice in equation 40, determined by a noise-scale bound and input-distribution constants, and applies to exact minimizers. Validation-tuned and finite-step estimates are evaluated empirically below. Under the common input assumptions, the rank-constrained and nuclear norm estimators both attain error bounds with the same rd dependence, using curvature analyses tailored to their respective error sets.

C Extensions

C.1 Rank misspecification

If $\mathrm{rank}(M^*) > r$, let $M_r \in \arg\min_{\mathrm{rank}(M)\le r} \|M - M^*\|_F$. Then $\|M_r - M^*\|_F^2 = \sum_{j>r}\sigma_j^2(M^*)$, where $\sigma_j(M^*)$ are the singular values in decreasing order. If $\|M^*\|_F \le B$, then $\|M_r\|_F \le B$, so M_r is a feasible comparator for the rank-constrained estimator. Comparing the empirical loss at $\widehat{M}$ with the loss at M_r introduces the residual $R = M^* - M_r$ and the empirical process

$$\frac{1}{n}\sum_{i=1}^{n}(x_{1,i}^\top R x_{2,i})(x_{1,i}^\top(\widehat{M} - M_r)x_{2,i}).$$

This decomposition identifies the interaction term between approximation error and estimation error that is absent under correct rank specification. The rank-sweep experiment below measures the resulting approximation–estimation tradeoff.

Increasing r reduces the deterministic tail energy but enlarges the rank-at-most-k_r difference set and the effective dimension. The rank-sweep experiment examines this tradeoff between approximation and estimation (Appendix E.3).

C.2 Covariance transformations

C.2.1 Marginal covariance standardization for independent inputs

Remark 4 (Dependence on tails and conditioning)**.** The constants in the independent-input result depend on the sub-Gaussian parameter K in equation 2. Known marginal covariances allow covariance standardization. Transforming back to the original coordinates introduces the corresponding minimum-eigenvalue factors.

Corollary 2 (Estimation with known marginal covariances)**.** *Suppose $M^* \in \mathcal{M}_r(B)$ and $x_j = \Sigma_{jj}^{1/2} u_j$ for $j = 1, 2$, where u_1, u_2 are independent, isotropic, and satisfy the uniform sub-Gaussian projection assumption, and where Σ_{11}, Σ_{22} are positive definite and known. Write $\lambda_{\min}$ and $\lambda_{\max}$ for their smallest and largest eigenvalues. Apply the rank-constrained estimator to $u_j = \Sigma_{jj}^{-1/2} x_j$ to obtain $\widehat{N}$, with radius $B_N = \sqrt{\lambda_{\max}(\Sigma_{11})\lambda_{\max}(\Sigma_{22})}\, B$, so that the transformed target $N^* = \Sigma_{11}^{1/2} M^* \Sigma_{22}^{1/2}$ belongs to $\mathcal{M}_r(B_N)$. Set $\widehat{M} = \Sigma_{11}^{-1/2} \widehat{N} \Sigma_{22}^{-1/2}$. Under the noise assumptions of Theorem 3 and the stronger sample condition equation 12, with probability at least $1 - \eta$,*

$$\|\widehat{M} - M^*\|_F^2 \le \frac{C_K \sigma^2}{\lambda_{\min}(\Sigma_{11})\lambda_{\min}(\Sigma_{22})} \frac{k_r(d + \log(8/\eta))}{n}.$$

Thus the error bound in the transformed coordinates retains the same dimension and sample size order. Transforming back multiplies the error bound by the product of the two inverse minimum eigenvalues. Comparing the transformed and original radius constraints introduces the corresponding marginal condition numbers.

Remark 5 (Operator-norm and localized control)**.** The stated rate is obtained from the whole rank-at-most-k_r difference class and has dimension scaling rd, matching the minimax lower bound. Tangent localization may refine its constants and the finite-sample $-r^2$ term.

C.2.2 Linear residualization of correlated inputs

Transforming each input to have identity covariance can leave the two inputs correlated. We next subtract the part of the second input that is linearly predictable from the first and standardize the residual. The theorem explicitly requires independence of this standardized residual and the first input, which is stronger than zero cross-covariance.

Theorem 9 (Known-covariance reduction with an independent residual)**.** *Let (x_1, x_2) be a centered pair with positive-definite marginal second-moment matrices $\Sigma_{jj} = \mathbb{E}[x_j x_j^\top]$. Let $\widetilde{x}_j = \Sigma_{jj}^{-1/2} x_j$ and $C = \mathbb{E}[\widetilde{x}_1 \widetilde{x}_2^\top]$. Assume that $\widetilde{x}_1$ has uniformly sub-Gaussian one-dimensional projections, that the observation noise is independent of (x_1, x_2), mean zero, and has finite second moment, and that $\|C\|_{\mathrm{op}} < 1$. Consider the structural representation*

$$\widetilde{x}_2 = C^\top \widetilde{x}_1 + (I - C^\top C)^{1/2} \xi.$$

Equivalently, the residual input is $\xi = (I - C^\top C)^{-1/2}(\widetilde{x}_2 - C^\top \widetilde{x}_1)$. Assume ξ is centered and isotropic, has uniformly sub-Gaussian one-dimensional projections, and is independent of $\widetilde{x}_1$. This representation is automatic for a centered, nonsingular jointly Gaussian pair after linear regression. Then the bilinear model becomes

$$y = \widetilde{x}_1^\top \bar{M}^* \xi + q_C(\widetilde{x}_1) + \epsilon, \quad \bar{M}^* = \Sigma_{11}^{1/2} M^* \Sigma_{22}^{1/2} (I - C^\top C)^{1/2},$$

where $q_C(u) = u^\top \Sigma_{11}^{1/2} M^* \Sigma_{22}^{1/2} C^\top u$. The first term has independent isotropic inputs. The quadratic remainder is orthogonal in expectation to every bilinear score because $\mathbb{E}[\xi \mid \widetilde{x}_1] = 0$. The second moment of the sum of the quadratic remainder and observation noise is $\nu_C^2 = \mathbb{E}[\epsilon^2] + \mathbb{E}[q_C(\widetilde{x}_1)^2]$, which equals $\sigma^2 + \mathbb{E}[q_C(\widetilde{x}_1)^2]$ in the Gaussian model.

The decomposition follows by substituting the structural representation: terms involving the residual input form the bilinear component, and the removed linear prediction gives the quadratic component. Orthogonality follows from the zero conditional mean of the residual, so the cross term vanishes in the population squared loss, preserving the transformed interaction matrix as a minimizer of the bilinear population squared loss. The theorem therefore gives a population guarantee under structured cross-input dependence. Appendix G.2 examines residualization with estimated covariance and compares reference error and prediction before and after transforming real features.

C.3 Optimization error

The upper bound holds for every exact global minimizer of equation 9. The basic inequality is derived in Appendix B, Equation 24. To quantify how numerical suboptimality affects estimation error, consider a feasible numerical solution $\widehat{M}_{\mathrm{opt}} \in \mathcal{M}_r(B)$ and define $\delta_{\mathrm{opt}} = \widehat{\mathcal{R}}_n(\widehat{M}_{\mathrm{opt}}) - \min_{M\in\mathcal{M}_r(B)} \widehat{\mathcal{R}}_n(M) \geq 0$ and $\widehat{\Delta}_{\mathrm{opt}} = \widehat{M}_{\mathrm{opt}} - M^*$. Since M^* is feasible, the loss normalization in equation 9 gives

$$\frac{1}{n}\sum_i (x_{1,i}^\top \widehat{\Delta}_{\mathrm{opt}} x_{2,i})^2 \leq 2\langle G_n, \widehat{\Delta}_{\mathrm{opt}}\rangle + \delta_{\mathrm{opt}}.$$

On $\mathcal{E}_{\mathrm{RC}}(\kappa)\cap\mathcal{E}_{\mathrm{noise}}(\tau)$, put $u = \|\widehat{\Delta}_{\mathrm{opt}}\|_F$. Then $\kappa u^2 \leq 2\tau u + \delta_{\mathrm{opt}} \leq (\kappa/2)u^2 + 2\tau^2/\kappa + \delta_{\mathrm{opt}}$, so

$$\|\widehat{\Delta}_{\mathrm{opt}}\|_F^2 \leq \frac{4\tau^2}{\kappa^2} + \frac{2\delta_{\mathrm{opt}}}{\kappa}. \tag{46}$$

Thus a given global objective gap contributes an additive error term. Substituting the noise threshold yields a probabilistic bound under the corresponding sample size conditions. The perturbation bound quantifies how a given global objective gap affects estimation error and also applies with a valid upper bound on that gap. Loss spreads and first-order residuals measure local optimization quality rather than this global objective gap.

C.4 Known scalar transformations

C.4.1 Derivative bounded away from zero

Consider observations $y = \phi(x_1^\top M^* x_2) + \epsilon$, where ϕ is a known differentiable scalar function. Let Φ be any antiderivative of ϕ, so $\Phi' = \phi$. Assume that positive constants m, L bound the derivative as follows:

$$0 < m \leq \phi'(t) \leq L < \infty, \qquad t \in \mathbb{R}. \tag{47}$$

Define

$$\widehat{M}_\phi \in \operatorname*{arg\,min}_{M\in\mathcal{M}_r(B)} \frac{1}{n}\sum_i \{\Phi(x_{1,i}^\top M x_{2,i}) - y_i x_{1,i}^\top M x_{2,i}\}.$$

Theorem 10 (Upper bound under a known scalar transformation)**.** *Assume $M^* \in \mathcal{M}_r(B)$. On $\mathcal{E}_{\mathrm{RC}}(\kappa)\cap\mathcal{E}_{\mathrm{noise}}(\tau)$, every global minimizer satisfies*

$$\|\widehat{M}_\phi - M^*\|_F^2 \leq \frac{4\tau^2}{m^2\kappa^2}.$$

Consequently, under the input and noise assumptions of Theorem 3 and the stronger sample condition equation 12, its error is at most $C_K\sigma^2 k_r(d+\log(8/\eta))/(m^2 n)$ with probability $1-\eta$.

Proof. Let $\Delta = \widehat{M}_\phi - M^*$ and write $s_i = x_{1,i}^\top M^* x_{2,i}$ and $z_i = x_{1,i}^\top \Delta x_{2,i}$, where s_i is the true score and z_i its estimated increment. The derivative lower bound gives, for all $s, z \in \mathbb{R}$,

$$\Phi(s+z) - \Phi(s) - \phi(s)z = z\int_0^1 \{\phi(s+tz) - \phi(s)\}\,dt \geq mz^2\int_0^1 t\,dt = \frac{m}{2}z^2.$$

Since M^* is feasible and $y_i = \phi(s_i) + \epsilon_i$, empirical optimality implies

$$\begin{aligned} 0 &\geq \frac{1}{n}\sum_i \{\Phi(s_i+z_i) - \Phi(s_i) - y_i z_i\} \\ &= \frac{1}{n}\sum_i \{\Phi(s_i+z_i) - \Phi(s_i) - \phi(s_i)z_i\} - \frac{1}{n}\sum_i \epsilon_i z_i. \end{aligned}$$

Combining these inequalities with the curvature and noise events yields

$$\begin{aligned}\frac{m\kappa}{2}\|\Delta\|_F^2 &\le \frac{m}{2n}\sum_i z_i^2 \\ &\le \frac{1}{n}\sum_i \{\Phi(s_i+z_i)-\Phi(s_i)-\phi(s_i)z_i\} \\ &\le \langle G_n, \Delta\rangle \le \tau\|\Delta\|_F.\end{aligned}$$

If $\Delta = 0$ the claim is immediate. Otherwise, divide by $\|\Delta\|_F$ and square to obtain $4\tau^2/(m^2\kappa^2)$. The probabilistic bound follows by substituting the curvature and noise thresholds from Theorem 7 and absorbing the second noise term under equation 12.

Theorem 11 (Lower bounds under a known scalar transformation)**.** *Under the independent isotropic input assumptions and Gaussian observation noise $\epsilon \sim \mathcal{N}(0,\sigma^2)$, the full and rank-$r$ minimax lower bounds are respectively*

$$c\min\{B^2, \sigma^2 d^2/(L^2 n)\}, \qquad c\min\{B^2, \sigma^2(2dr-r^2)/(L^2 n)\}.$$

In this subsection, P_M denotes the joint distribution of the n observations under the transformed model with parameter M. For two parameters M, M', Gaussian noise and the upper derivative bound give

$$D_{\mathrm{KL}}(P_M\|P_{M'}) = \frac{n}{2\sigma^2}\mathbb{E}\big[\{\phi(x_1^\top M x_2)-\phi(x_1^\top M' x_2)\}^2\big] \le \frac{nL^2}{2\sigma^2}\|M-M'\|_F^2,$$

where the last inequality uses the population isometry equation 5. Applying the full and low-rank packings used in Theorems 1 and 2, with their scales adjusted by L in the noise-controlled separation term, proves the two bounds. Hence the upper and lower bounds have matching dimension dependence, with a possible factor $(L/m)^2$ from the conditioning of ϕ. Logistic sigmoid and softmax probabilities have curvature that vanishes in the tails and do not satisfy the derivative condition in Theorem 10. The logistic experiment below reports behavior in this regime.

C.5 Scope summary

Equation 46 extends the global-minimizer guarantee to feasible solutions with a certified objective gap. The covariance and known-link extensions use the structural conditions in Theorems 9 and 10. Experiments with logistic links, heavy-tailed inputs, and pretrained BERT heads complement these results with empirical behavior under their respective data-generating settings.

D Comparison with Existing Low-Rank Estimation Results

Rank-1 projection rates and estimators. Cai and Zhang (2015, Theorems 2.3 and 3.1) analyze nuclear norm minimization under constraints on the residual's ℓ_1 norm and its adjoint operator norm. For square matrices, their exact-low-rank bound is $C\sigma^2\min\{rd/n + rd^2\log(n)/n^2, 1\}$ for $n \ge Crd$. Theorem 2.3 uses independent standard Gaussian inputs and Gaussian noise. Theorem 3.1 uses independent coordinates from a symmetric, variance-one sub-Gaussian distribution and sub-Gaussian noise (with its noise parameter in place of σ). The respective failure probabilities are $11/n + 3e^{-cd}$ and $2/n + 5e^{-cd}$, after absorbing the factor $2d$ into c. Our least-squares estimators have different objectives and allow directional sub-Gaussian vectors with dependent coordinates. Their explicit η-dependent bounds retain the same two-term dimension scaling. The simplified rate requires the additional sample conditions stated in Theorems 3 and 4.

Specialization of a general curvature theorem. Theorem 3.2 of Hamidi and Bayati (2022) applies to our full input class. Here is the substitution, with its set parameter denoted by $q \ge 1$ to distinguish it from our failure probability η. Put $A = x_1 x_2^\top$ and take their auxiliary norm to be $\|\cdot\|_F$. For $L_4 = C_4 K^8$, the moment bounds below give

$$\mathbb{E}\langle A, D\rangle^2 = \|D\|_F^2, \qquad \mathbb{E}\langle A, D\rangle^4 \le L_4\|D\|_F^4.$$

Thus their covariance constants are $\gamma_{\min} = \gamma_{\max} = 1$. Set $c_{\mathrm{tr}}^2 = 2L_4$. For $\|D\|_F \le 1$,

$$\mathbb{E}[\langle A, D\rangle^2 \mathbb{1}_{\{|\langle A,D\rangle| > c_{\mathrm{tr}}\}}] \le \frac{L_4 \|D\|_F^4}{c_{\mathrm{tr}}^2} \le \tfrac{1}{2}\|D\|_F^2,$$

which verifies their Assumption 3.3. Let $R_n = n^{-1}\sum_i \xi_i A_i$, with independent Rademacher signs ξ_i. The sphere-net bound proved below yields $\mathbb{E}\|R_n\|_{\mathrm{op}} \le CK^2(\sqrt{d/n} + d/n)$. Their Theorem 3.2, with $\theta = 1$, therefore gives

$$\inf_{\|D\|_F = 1,\, \|D\|_* \le \sqrt{q}} \frac{1}{n}\sum_i \langle A_i, D\rangle^2 \ge \frac{1}{4} - 93 q c_{\mathrm{tr}}^2 (\mathbb{E}\|R_n\|_{\mathrm{op}})^2 \ge \frac{1}{8}$$

with probability at least $1 - \eta/2$ once $n \ge C_K\{qd + \log(8/\eta)\}$. Taking $q = k_r$ or $q = 32r$ covers the feasible rank differences or the nuclear norm error cone, respectively. This is a specialization of their general theorem, using the matrix complexity bound derived here. Their worked Gaussian factored-measurement example (Section 5.4) instead uses the bound in their Eq. (5.4), with a signal-size parameter, $\rho \ge \log d$, and $n \ge Cd\log^4 d$. Those particular choices are distinct from the scope of their general curvature theorem.

From curvature and noise to least-squares error. On the curvature event, empirical optimality of rank-constrained least squares gives $\kappa\|\Delta\|_F^2 \le 2\sqrt{k_r}\|G_n\|_{\mathrm{op}}\|\Delta\|_F$. For nuclear norm penalized least squares, decomposability gives $\|\Delta\|_* \le 4\sqrt{2r}\|\Delta\|_F$ when $\lambda \ge 4\|G_n\|_{\mathrm{op}}$, followed by $\|\Delta\|_F^2 \le Cr\lambda^2/\kappa^2$. This deterministic step is an instance of the low-rank regularization analysis of Negahban and Wainwright (2011) and Negahban et al. (2012, Theorem 1). For the present rank-1 measurements, conditioning on the noise weights gives, with conditional probability at least $1 - \eta/4$,

$$\|G_n\|_{\mathrm{op}} \le C_K \left\{ \frac{\|\epsilon\|_2\sqrt{d + t_\eta}}{n} + \frac{\|\epsilon\|_\infty (d + t_\eta)}{n} \right\}.$$

Concentration of the two noise norms yields $C_K \sigma b_{n,\eta}$. Combining this with either the preceding specialization or the direct small-ball proof establishes our two least-squares upper bounds.

Population quadratic loss versus empirical least squares. Koltchinskii et al. (2011, Theorem 1) give an oracle inequality for an estimator using the population quadratic term $\mathbb{E}\langle A, M\rangle^2$. In our isotropic model this term equals $\|M\|_F^2$, so it is known. Their stochastic term nevertheless differs from G_n: writing it as S_n,

$$S_n = \frac{1}{n}\sum_i (y_i A_i - \mathbb{E}[y_i A_i]) = G_n + \frac{1}{n}\sum_i (\langle A_i, M^*\rangle A_i - M^*).$$

Their bound controls $r\lambda^2$ with $\lambda \ge 2\|S_n\|_{\mathrm{op}}$. Our empirical least-squares analysis controls G_n after cancellation of the signal in the residual. This identifies the estimator and stochastic term that must be specified when comparing the two guarantees.

Bounded-class lower bounds and the full-matrix comparison. Rohde and Tsybakov (2011, Theorem 5(i)) give a probability lower bound on rank-constrained Schatten balls under an upper restricted-isometry condition. With Schatten exponent 2 and unit sampling normalization, its scale is $\min\{B^2, \sigma^2 rd/n\}$ for square matrices. Their Theorem 2 gives penalized least-squares upper bounds conditional on restricted isometry and stochastic-term control. For Gaussian rank-1 measurements, Cai and Zhang (2015, Theorem 2.4) give noise-conditional lower bounds on an unbounded rank class, on a high-probability event over the sampled inputs. Our bounded-class proof instead uses the joint law of inputs and labels and the identity

$$\mathrm{KL}(P_M, P_{M'}) = \frac{n}{2\sigma^2}\|M - M'\|_F^2.$$

Packing within the radius B gives expectation and fixed-probability lower bounds for every n, including the saturation regime. Since $rd \le 2dr - r^2 \le 2rd$, the manifold notation preserves the classical dimension order. Section 5 combines these bounded-class lower bounds with the explicit least-squares upper bounds at a common radius and confidence level. The resulting d/r sample ratio applies when the noise-dependent requirements dominate the curvature sample-size thresholds.

E SYNTHETIC EXPERIMENTS

This appendix provides the experimental settings, complete numerical results, and additional diagnostics supporting Section 6.

E.1 EXPERIMENTAL SETUP

Data generation. The core scaling and interpolation experiments use ten random seeds, $\sigma = 0.3$, and Gaussian inputs; later robustness experiments state their own input distributions and seed counts. The true interaction matrix is normalized to $\|M^*\|_F = 1$. Let r^* denote the rank of the true interaction matrix M^*. For every random seed, we draw $U^*, V^* \in \mathbb{R}^{d\times r^*}$ with independent standard Gaussian entries and set $M^* = U^*(V^*)^\top$, followed by Frobenius normalization. Training labels are generated independently as

$$y_i = x_{1,i}^\top M^* x_{2,i} + \epsilon_i, \qquad x_{1,i}, x_{2,i} \sim \mathcal{N}(0, I_d), \quad \epsilon_i \sim \mathcal{N}(0, 0.09).$$

Unconstrained estimator. Let $\mathbf{Z} \in \mathbb{R}^{n\times d^2}$ have rows $(x_{2,i} \otimes x_{1,i})^\top$, and let $y = (y_1, \dots, y_n)^\top$ be the response vector. We stack matrix columns when vectorizing, so $\widehat{m}_{\mathrm{OLS}} = \mathrm{vec}(\widehat{M}_{\mathrm{OLS}})$. The OLS coefficient is

$$\widehat{m}_{\mathrm{OLS}} = (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top y$$

whenever the inverse exists, and $\widehat{M}_{\mathrm{OLS}}$ is obtained by reshaping $\widehat{m}_{\mathrm{OLS}}$. When $n < d^2$ the normal-equation matrix $\mathbf{Z}^\top\mathbf{Z}$ is rank-deficient. In this case we use the minimum-norm least squares solution computed via a pseudoinverse based on the singular value decomposition (SVD), with singular-value thresholding at $10^{-10}\sigma_{\max}(\mathbf{Z})$, where $\sigma_{\max}(\mathbf{Z})$ is the largest singular value of the regression matrix $\mathbf{Z}$.

Low-rank post-processing estimator. Let $\widehat{\sigma}_j$ be the singular values of $\widehat{M}_{\mathrm{OLS}}$ in decreasing order, with corresponding left and right singular vectors $\widehat{u}_j, \widehat{v}_j$. Define the rank-$r$ truncation by

$$\widehat{M}_r^{\mathrm{SVD}} = \sum_{j=1}^{r} \widehat{\sigma}_j \widehat{u}_j \widehat{v}_j^\top.$$

This is the best rank-r approximation of the OLS matrix in Frobenius norm. We treat it as an SVD post-processing estimator, alongside the rank-constrained ERM that directly minimizes the empirical risk.

Projected factorized ERM estimator. The factorized estimator optimizes $U, V \in \mathbb{R}^{d\times r}$ to minimize the empirical risk $\frac{1}{n}\sum_i (y_i - x_{1,i}^\top UV^\top x_{2,i})^2$ via Adam with cosine annealing (5000 steps, learning rate 0.01). To keep the iterates in the theoretical constraint set $\mathcal{M}_r(B)$ in Section 4, we rescale U and V after each gradient step to project their product $M = UV^\top$ onto the Frobenius ball $\|M\|_F \le B = 2$. This projected gradient procedure provides the numerical constrained-ERM estimate reported below. A full-batch iteration has arithmetic cost $O(ndr)$ and uses $2dr$ trainable factor parameters.

Nuclear norm regularized estimator. This estimator solves $\min_M \frac{1}{n}\sum_i (y_i - x_{1,i}^\top M x_{2,i})^2 + \lambda\|M\|_*$ via proximal gradient descent (3000 steps) with singular-value thresholding. For the main comparisons, the regularization parameter is selected from $\lambda = c \cdot \sigma\sqrt{d/n}$ with $c \in \{0.05, 0.1, 0.5, 1, 2\}$ by best validation error. The interpolation-specific tuning grid and oracle comparison are given with the corresponding results in Section E.3. These validation-tuned, finite-step estimates are empirical implementations of equation 34; the theorem uses the choice in equation 40, determined by the noise scale and input-distribution constants, and an exact minimizer.

Metrics and uncertainty. The main metric is squared Frobenius parameter error $\|\widehat{M} - M^*\|_F^2$. Normalized mean squared error (NMSE) divides this quantity by $\|M^*\|_F^2$. For experiments with a fitted reference matrix, the same normalization uses that reference in place of M^*. When prediction risk is reported, it is evaluated on independently generated test samples and includes observation

noise. Tables report the aggregation stated in each caption: means over seeds in the regular synthetic regimes, median and interquartile range (IQR) in the heavy-tailed interpolation regime, and mean $\pm$ standard deviation for the additional Sentences Involving Compositional Knowledge relatedness task (SICK-R) case study.

E.2 Scaling and interpolation

We first report the dimension scan summarized in Section 6.1 (Table 3); the sample-size and interpolation analyses follow in separate subsections. As d increases from 8 to 64, the OLS/ERM error ratio grows from 1.35 to 8.70, closely tracking the corresponding DoF ratio from 1.33 to 8.26.

Table 3: Dimension scaling with $r = 4$ and $n = 20d^2$ fixed. Estimation error is $\|\widehat{M} - M^*\|_F^2$, and the DoF ratio is $d^2/(2dr - r^2)$.

d	OLS	SVD rank r	ERM rank r	OLS/SVD	OLS/ERM	DoF ratio
8	0.00544	0.00418	0.00404	1.30x	1.35x	1.33
16	0.00466	0.00214	0.00210	2.18x	2.22x	2.29
32	0.00475	0.00114	0.00111	4.16x	4.27x	4.27
64	0.00477	0.00058	0.00055	8.28x	8.70x	8.26

Sample-size scaling and estimator comparison. The OLS, SVD, and rank-ERM scan uses $n/d^2 \in \{0.25, 0.5, 1, 2, 5, 20\}$, while the nuclear norm comparison uses $n/d^2 \in \{2, 5, 10, 20, 50\}$ with the same baselines (Table 4). The full and low-rank dimensions are 256 and 112, respectively, giving a DoF ratio of approximately 2.29.

Table 4: Sample size comparison at $d = 16, r = 4$. Squared Frobenius errors are means over 10 seeds. The bold interpolation row ($n = 256$) reports medians [interquartile range] over 20 seeds. Ratios use the corresponding means or medians. Nuclear norm results use the validation-selected λ from the experimental setup. Dashes denote unreported entries.

n	n/d^2	OLS	SVD-r	ERM-r	Nuclear	OLS/SVD	OLS/ERM
64	0.25	0.791	0.767	3.094	—	—	0.26x
128	0.50	0.640	0.550	2.711	—	—	0.24x
256	**1.00**	**74.0 [341]**	**51.2 [232]**	**0.096 [0.020]**	—	—	**768x**
512	2.00	0.1054	0.0494	0.0302	0.0430	2.13x	3.49x
1280	5.00	0.0248	0.0115	0.0099	0.0142	2.15x	2.51x
2560	10.0	0.0101	0.0045	0.0043	0.0064	2.25x	2.37x
5120	20.0	0.0047	0.0021	0.0021	0.0031	2.18x	2.22x
12800	50.0	0.0019	0.0009	0.0009	0.0013	2.22x	2.26x

At $n = 64, 128$, pseudoinverse OLS and SVD remain bounded and outperform factorized ERM. At $n = d^2$, this ordering reverses: rank-ERM has much lower median error and a narrower spread than OLS and SVD. Section E.2 examines this transition through a spectral scan and solver diagnostics. For $n \geq 512$, rank-ERM has the lowest reported error (ties with SVD at displayed precision), while nuclear norm estimation lies between rank-ERM and OLS. SVD and ERM errors are consistent with $1/n$ decay, and the large-sample OLS/ERM ratio approaches the DoF benchmark. At smaller n, the gap combines statistical regularization and numerical optimization effects.

Instability of full-matrix OLS near interpolation. To characterize the OLS spike near interpolation, we use a dedicated 20-seed dense scan $n \in \{224, 240, 248, 252, 256, 260, 264, 272, 288, 320\}$ at $d = 16$, $r = 4$, $\sigma = 0.3$. The spike is concentrated at $n = d^2 = 256$, where the regression matrix $\mathbf{Z}$ is square. All errors are squared Frobenius. At this point, the OLS error distribution is strongly right-skewed and heavy-tailed: its mean is 589, while the seed-level errors range from approximately 14 to 5414. A mean with a normal confidence interval (CI) therefore summarizes the distribution poorly and can produce a negative lower endpoint.

Let $\mathbf{Z}$ be the vectorized regression matrix defined in the experimental setup, let $m^* = \mathrm{vec}(M^*)$, and let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. Write $\sigma_j(\mathbf{Z})$ for the singular values in decreasing order, $\mathbf{Z}^\dagger$ for the Moore–Penrose pseudoinverse, and $\kappa(\mathbf{Z}) = \sigma_{\max}(\mathbf{Z})/\sigma_{\min}(\mathbf{Z})$ for the ratio of the largest to smallest singular values. The SVD-pseudoinverse OLS estimator satisfies the exact decomposition

$$\widehat{m}_{\mathrm{OLS}} - m^* = \underbrace{(\mathbf{Z}^\dagger \mathbf{Z} - I)m^*}_{\text{bias}} + \underbrace{\mathbf{Z}^\dagger \boldsymbol{\epsilon}}_{\text{noise amplification}} .$$

At $n = d^2$ with $\mathbf{Z}$ square and invertible, the bias term vanishes ($\mathbf{Z}^\dagger\mathbf{Z} = I$) and the entire error is $\left\|\mathbf{Z}^{-1}\boldsymbol{\epsilon}\right\|_2^2$. Conditionally on $\mathbf{Z}$, the expected squared error is

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\mathbf{Z}^{-1}\boldsymbol{\epsilon}\|_2^2 \mid \mathbf{Z}\right] = \sigma^2 \operatorname{tr}\left((\mathbf{Z}^\top\mathbf{Z})^{-1}\right) = \sigma^2 \sum_{j=1}^{d^2} \sigma_j(\mathbf{Z})^{-2}. \tag{48}$$

Thus the jth singular direction contributes $\sigma^2\sigma_j(\mathbf{Z})^{-2}$ to the expected error, and the smallest singular values dominate the sum. In particular, $\sigma_{\min}(\mathbf{Z})$ drives the heavy tail of the error distribution because $1/\sigma_{\min}^2(\mathbf{Z})$ can be far larger than the remaining terms. Table 5(a) connects this mechanism to the dense scan by reporting the median and interquartile range (IQR) of the OLS error together with spectral diagnostics at representative sample sizes.

Table 5: Interpolation diagnostics at $d = 16$ over 20 seeds. **(a)** Spectrum and OLS error across the interpolation scan. **(b)** Solver, rank-truncation, and singular-value-threshold diagnostics at $n = d^2 = 256$. All errors are squared Frobenius errors.

(a) Spectral scan

n	n/d^2	$\sigma_{\min}(\mathbf{Z})$ med.	$\kappa(\mathbf{Z})$ med.	OLS med. [IQR]
224	0.88	0.99	35	0.92 [0.33]
248	0.97	0.26	132	3.33 [2.10]
252	0.98	0.17	206	5.95 [4.12]
256	1.00	0.031	1191	73.97 [340.7]
260	1.02	0.146	250	8.07 [4.92]
272	1.06	0.51	75	1.94 [0.95]
320	1.25	1.77	21	0.44 [0.12]

(b) Solver diagnostics at $n = 256$

Estimator	median	IQR
OLS (`lstsq`)	73.97	340.7
OLS + SVD rank-r	51.19	231.6
Truncated-SVD ($\mathbf{Z}$ threshold 10^{-8})	73.97	340.7
Truncated-SVD ($\mathbf{Z}$ threshold 10^{-6})	73.97	340.7

At $n = d^2$, the smallest singular value collapses and the condition number jumps, directly inflating the noise-amplification term. The measured squared bias $\|(\mathbf{Z}^\dagger\mathbf{Z} - I)m^*\|_2^2$ is negligible, indicating that nearly all of the OLS error is noise amplification, as predicted by the decomposition. An independent `torch.linalg.lstsq` implementation reproduces the same interpolation peak and the same mean OLS error of 589 at $n = 256$ as the pseudoinverse implementation.

Table 5(b) separates the effects of the linear solve, rank-r post-processing, and singular-value thresholding at the interpolation point. Rank truncation reduces the median error but retains the spike, while thresholds of 10^{-8} and 10^{-6} remove no singular directions in the tested runs and therefore leave the OLS error unchanged. SVD rank-r post-processing inherits the ill-conditioning of $\mathbf{Z}$ because it truncates the already unstable $\widehat{M}_{\mathrm{OLS}}$. By contrast, at $n = 256$ the sample size is 2.29 times the rank-r manifold dimension 112, and the tested rank-constrained estimators do not exhibit the full-matrix interpolation peak. We therefore exclude $n = d^2$ from the $1/n$ scaling fit and report it separately with robust statistics.

E.3 Matched-constraint and estimator comparisons

This subsection compares full- and low-rank estimators under a common Frobenius radius, evaluates three rank-constrained solvers under matched settings, and reports tuning and convergence diagnostics for the low-rank estimators in Section 6.2.

Matched-radius full-matrix comparison. Table 6 gives the full sample-size grid for the matched-radius comparison summarized in the main text. Under the common radius $B = 2$, the rank ball has lower median error for every tested $n \geq 224$, including a reduction from 2.99 to 0.0931 at interpolation; the ordering reverses at $n = 128$, where the less restrictive full-matrix class performs better.

Rank-constrained solver comparison. Factorized ERM, singular value projection (SVP), and iterative hard thresholding (IHT) are fitted with the same rank, radius, and iteration limit. At $\kappa(M^*) =$

Table 6: Median squared Frobenius error over 25 paired synthetic repetitions. Full-ball and rank-ball use the same radius $B = 2$. Ridge is selected on an independent validation sample.

n	OLS	Ridge	Full-ball	Rank-ball	Rank/Full
128	0.601	0.568	0.601	0.947	1.58
224	0.92	0.315	0.92	0.129	0.14
248	3.55	0.296	2.98	0.0965	0.0324
256	73.7	0.281	2.99	0.0931	0.0312
264	3.95	0.268	2.67	0.0925	0.0347
320	0.451	0.202	0.451	0.0654	0.145
512	0.104	0.089	0.104	0.0275	0.265
1280	0.0234	0.023	0.0234	0.00891	0.381

$1.7, 10, 200$, their respective median squared Frobenius errors are $(0.005865, 0.005851, 0.005878)$, $(0.006368, 0.006494, 0.006423)$, and $(0.008976, 0.008436, 0.008448)$. The largest difference between any two solvers is below 5.5×10^{-4}, so the recovery comparison on this grid is not driven by the particular rank-constrained solver.

Nuclear norm tuning at interpolation. To compare the convex and factorized estimators near interpolation, Figure 4 reports dense λ sweeps with oracle and cross-validation tuning at $n = d^2$ and $d \in \{16, 32\}$. This comparison measures the combined effects of regularization and optimization. The nuclear norm regularizer λ is selected by five-fold cross-validation over nine geometrically spaced values ($\lambda \in \{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}, 10^{-1}, 3 \cdot 10^{-1}, 1\}$), using the same tuning procedure as ridge. Each estimator is evaluated at its own validation-selected optimum.

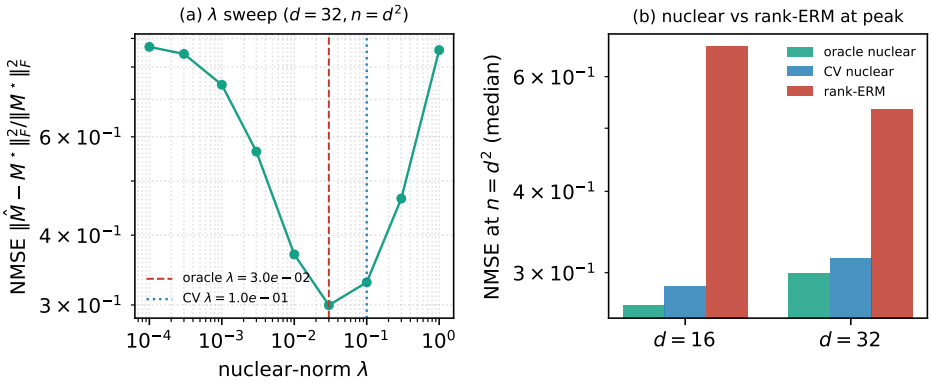


Figure 4: Nuclear norm versus factorized rank-ERM at $n = d^2$. Left: λ sweep at $d = 32$. Oracle recovery tuning and cross-validation select nearby regions. Right: median NMSE at $d \in \{16, 32\}$. Curves report oracle and cross-validation tuning for the nuclear norm estimator and the corresponding factorized outputs.

Optimization diagnostics for factorized ERM. To assess objective convergence and initialization sensitivity, Figure 5 reports training-loss trajectories and recovery-error spread. The solver uses the Adam updates, cosine annealing, learning rate 0.01, and $B = 2$ Frobenius-ball projection specified in the experimental setup. At $n/d^2 = 0.80, 1.00, 1.25$, the rank-ERM data-resampling variances are 3.0×10^{-5}, 2.5×10^{-5}, and 1.5×10^{-5}, whereas the corresponding initialization variances are 2.6×10^{-28}, 5.5×10^{-23}, and 6.0×10^{-24}. Thus, initialization variance is negligible and sampling variation accounts for essentially all observed dispersion; the closed-form baselines have zero initialization variance. Equation equation 46 relates a global objective gap to estimation error.

Rank sweep. Table 7(a) reports the complete fitted-rank grid. Both estimators exhibit a U-shaped curve with minimum at the true rank $r = r^* = 4$; underfitting leaves signal in the truncated tail, whereas overfitting adds noisy directions and increases variance. For $r > r^*$, the projected

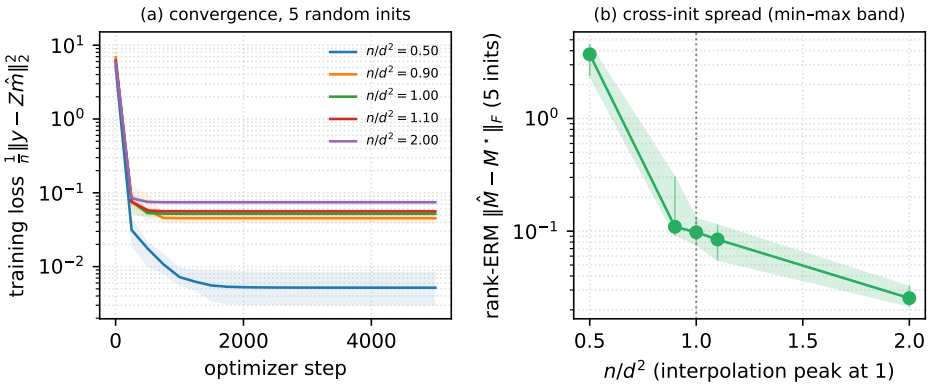


Figure 5: Optimization diagnostics for projected factorized rank-ERM at $d = 16, r = 4$: training-loss trajectories and the corresponding recovery-error spread.

Table 7: Rank specification and prediction risk. **(a)** Complete fitted-rank sweep at $d = 16$, $r^* = 4$, and $n = 5120$. **(b)** Total prediction risk on independent test samples.

(a) Fitted-rank sweep

r	SVD	ERM	DoF	ERM/SVD
1	0.5134	0.5188	31	1.01
2	0.1968	0.2002	60	1.02
3	0.0572	0.0585	87	1.02
4	0.0021	0.0021	112	1.00
5	0.0029	0.0028	135	0.97
6	0.0034	0.0032	156	0.94
8	0.0041	0.0039	192	0.95
12	0.0046	0.0045	240	0.98
16	0.0047	0.0046	256	0.99

(b) Prediction risk

n	OLS	SVD-r	ratio	σ^2
512	0.1983	0.1381	1.44	0.09
1280	0.1128	0.1005	1.12	0.09
2560	0.1020	0.0954	1.07	0.09
5120	0.0957	0.0930	1.03	0.09
12800	0.0934	0.0924	1.01	0.09

ERM achieves slightly lower error than SVD post-processing on this grid; see Appendix C.1 for the residual-process analysis. This comparison includes both the parameterization and the optimization trajectory.

Prediction risk. Table 7(b) compares total prediction risk for full-matrix OLS and rank-r SVD post-processing using independently generated test samples. The reported risks include observation noise and both approach the irreducible level $\sigma^2 = 0.09$; because prediction risk equals estimation error plus σ^2, their ratio converges to one as the estimators become consistent.

E.4 Distributional and Link-Function Robustness

Input distributions, tails, and truncation. Table 8 groups three stress tests of the input assumptions. Panel (a) compares Gaussian, Rademacher, centered sparse Bernoulli, and correlated Gaussian inputs. For the sparse design, each coordinate is $(b - p)/\sqrt{p(1-p)}$ with $b \sim \text{Bernoulli}(p)$ and $p = 0.3$. For the correlated design, x_1 and x_2 remain independent, while each has covariance $\Sigma = (1 - \rho)I + \rho\mathbf{1}\mathbf{1}^\top$ with $\rho = 0.5$; thus ρ controls within-input coordinate correlation rather than cross-input dependence. The three isotropic sub-Gaussian cases have OLS/ERM ratios close to the DoF benchmark, while correlation and non-isotropy increase absolute errors without removing the full/low-rank separation. Panel (b) compares Gaussian and bounded Rademacher marginals with a sub-exponential Laplace stress test; errors remain of the same order. Panel (c) uses Student-$t(3)$ inputs and coordinate-wise truncation at four population standard deviations; raw and truncated errors are again of the same order. These experiments extend the base comparison beyond independent Gaussian inputs without asserting that the heavy-tailed cases satisfy equation 2.

Known scalar links. Table 9 compares link-aware low-rank estimation with identity-link baselines for the known transformation $\phi(t) = t + \tanh(t)$, which satisfies $1 \le \phi'(t) \le 2$. The

Table 8: Input-robustness diagnostics. **(a)** Squared Frobenius recovery error across input distributions ($d = 16, r = 4, n = 5120$). **(b)** Rank-ERM NMSE across tail families ($d = 16, r = 4, \sigma = 0.3$; four-seed medians). **(c)** Raw and coordinate-wise truncated Student-$t(3)$ inputs (five seeds).

(a) Distribution comparison

Input	Full OLS	SVD-r	ERM-r	OLS/ERM	DoF ratio
Gaussian	0.00466	0.00214	0.00210	2.22	2.29
Rademacher	0.00479	0.00197	0.00191	2.51	2.29
Sparse ($p = 0.3$)	0.00495	0.00212	0.00207	2.39	2.29
Correlated ($\rho = 0.5$)	0.01656	0.00736	0.00617	2.68	2.29

(b) Tail families

Input marginal	$n = 128$	$n = 256$	$n = 512$
Gaussian	3.18	0.58	0.30
Rademacher	3.03	0.53	0.28
Laplace	3.88	0.55	0.32

(c) Student-$t(3)$ truncation

n/d^2	n	raw NMSE	truncated NMSE
0.5	128	2.35	2.12
1.0	256	1.73	1.38
2.0	512	1.30	1.32
5.0	1280	1.28	1.38

estimator minimizes $n^{-1}\sum_i[\Phi(s_i) - y_i s_i]$, where $s_i = x_{1,i}^\top M x_{2,i}$ and $\Phi(t) = t^2/2 + \log\cosh(t)$. Its error decreases from 0.0119 to 0.0003, whereas OLS and SVD fitted without the link remain near 0.3 at the largest sample sizes, directly matching the setting of Theorem 10. In a complementary logistic experiment, logistic- and identity-link rank-ERM NMSE values differ by at most 6% over $n/d^2 \in \{0.5, 1, 2, 5\}$.

Table 9: Recovery error under the known transformation $\phi(t) = t + \tanh(t)$ at $d = 16, r = 4$. The transformation-aware ERM incorporates the known link in fitting.

n	Id. OLS	Id. ERM	Transf. OLS	Transf. SVD	Transf. ERM
512	0.1054	0.0302	0.4833	0.4037	**0.0119**
1280	0.0248	0.0099	0.3516	0.3322	**0.0040**
2560	0.0101	0.0043	0.3234	0.3158	**0.0017**
5120	0.0047	0.0021	0.3133	0.3097	**0.0009**
12800	0.0019	0.0009	0.3067	0.3052	**0.0003**

F Real-Feature Experiments

The real-feature experiments separate three targets: recovery of a generated low-rank matrix on frozen sentence embeddings, prediction of pretrained attention scores, and agreement with a full-pool reference fitted to observed semantic-similarity ratings. This separation keeps parameter recovery, score prediction, and reference agreement as distinct evaluation questions.

For the real-score setting, let $\mathcal{P}$ denote the training pool: 5749 pairs for STS-B and 3330 pairs for the additional SICK-R case study after reserving a quarter of its training split. For a fixed feature pipeline and rank, we fit a numerical reference $\widetilde{M}_r$ on the entire pool using rank-constrained least squares:

$$L_{\mathcal{P}}(M) = \frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}(s_i - x_{1,i}^\top M x_{2,i})^2$$

under the fitting constraints. Each size-n fit is compared with this fixed reference. Table 10 summarizes the labels, references, and interpretation of matrix error for the real-feature settings.

For residualization, the map is estimated from the training pool. We apply the map to the inputs and refit the reference on the transformed pool, retaining the real-score labels. We denote transformed quantities with a bar. The before/after matrix errors are $\|\widehat{M} - \widetilde{M}_r\|_F^2$ and $\|\widehat{\overline{M}} - \widetilde{\overline{M}}_r\|_F^2$, respectively.

Table 10: Label sources and reference targets in the real-feature settings. Each Frobenius distance uses an estimate and reference expressed in the same coordinates. Held-out prediction uses the label source in the second column.

Setting	Labels	Reference matrix	Meaning of matrix error
Generated target	$x_1^\top M^* x_2 + \epsilon$	Independently generated M^*	Recovery of the generating parameter
Real scores	Observed ratings s_i	Full-pool fit $\widetilde{M}_r$	Agreement with a fixed numerical reference
Residualized real scores	The same ratings s_i	Refit $\widetilde{M}_r$ on the transformed pool	Agreement with the reference in transformed coordinates

The two references are fitted separately in their respective coordinates, so each matrix error measures agreement with that coordinate system's numerical reference. Prediction MSE uses the same observed ratings and directly assesses predictive performance within the paired comparison.

F.1 Recovery of a generated target on STS-B features

Feature construction and fixed-target experiments. We replace the synthetic Gaussian inputs with sentence embeddings from the STS-B semantic textual similarity benchmark (Cer et al., 2017) (5,749 train pairs, 1,500 dev pairs). Sentences are encoded with a frozen `bert-base-uncased` model (mean pooling, 768-dim), then reduced by principal component analysis (PCA) to $d = 64$ ($d^2 = 4096$) with normalization using the empirical covariance. We use the base masked-language-model encoder without task-specific semantic-similarity fine-tuning. This experiment retains the paired dependence and approximate isotropy of the resulting sentence features.

To separate the real feature distribution from the target-generation process, we generate a random factorized low-rank matrix M^* ($r^* = 4$, $\|M^*\|_F = 1$) independently of the embeddings and *external* to the STS-B corpus, then hold it fixed. We generate labels $y = x_1^\top M^* x_2 + \varepsilon$ using the STS-B sentence embeddings as x_1, x_2. Only the covariate pair (x_1, x_2) comes from real data, and recovery error is measured against the fixed external target. The OLS distribution is strongly right-skewed near interpolation, motivating the robust summaries in Table 11.

Table 11: Recovery of a fixed rank-4 target from frozen STS-B sentence features ($d = 64$, ten seeds). Entries are median [IQR] squared Frobenius errors.

n	n/d^2	OLS	SVD rank r	ERM rank r
256	0.06	0.944 [0.011]	0.940 [0.010]	3.44 [0.62]
512	0.12	0.895 [0.010]	0.852 [0.020]	4.64 [0.14]
1024	0.25	4×10^7	3×10^7	**0.175 [0.057]**
2048	0.50	4×10^9	2×10^9	**0.037 [0.008]**
4096	1.00	9×10^9	2×10^9	**0.015 [0.002]**
5749	1.40	0.578 [0.042]	0.076 [0.005]	**0.010 [0.000]**

At $n = 1024, 2048, 4096$, the median OLS error ranges from 4×10^7 to 9×10^9, whereas the rank-ERM error decreases from 0.175 to 0.015. At $d = 64$ and $n = 5749$, rank-ERM attains median squared Frobenius error 0.010, compared with 0.578 for OLS and 0.076 for OLS+SVD-r.

Convex baselines for recovery and prediction. To compare convex and rank-constrained baselines on real covariates, we evaluate rank-ERM, the tuned nuclear norm estimator, and OLS+SVD in the same known-target experiment (Table 12). We compare parameter recovery and held-out prediction MSE. The nuclear norm regularization parameter is selected from the grid $\lambda = c\sigma\sqrt{d/n}$, $c \in \{0.05, 0.1, 0.5, 1, 2\}$, on the validation set, and prediction error is evaluated on the test set.

At $n \geq 2048$, every estimator's held-out MSE has excess over σ^2 at most 0.07, despite differences in Frobenius error. The tuned convex and rank-constrained routes therefore give comparable prediction error on these correlated features.

Table 12: Known-target recovery and prediction on BERT features ($d = 16, r = 2$, three-seed medians). The independent target and generated test labels follow the setup in Section F.1, with noise variance 1.99.

	$\left\|\widehat{M} - M^*\right\|_F^2$			held-out MSE		
n	rank-ERM	nuclear	SVD	rank-ERM	nuclear	SVD
512	0.85	0.25	2.35	2.24	2.13	3.56
2048	0.12	0.17	0.23	1.97	2.05	2.05
5749	0.056	0.049	0.075	1.93	1.94	1.93

F.2 Prediction of pretrained BERT attention scores

For the experiment in Section 6.3, we extract the input residual stream and the query and key maps of head 0 from layers 0 and 6 of frozen `bert-base-uncased`. Each label is the actual pre-softmax score

$$y_{ij} = \frac{(h_i W_q + b_q)^\top (h_j W_k + b_k)}{\sqrt{d_h}}.$$

The label implementation agrees with a direct application of the model's query and key linear layers to within float32 rounding error on sampled token pairs. Covariates comprise 64 PCA coordinates standardized by their empirical covariance and one appended constant coordinate. The fitted prediction matrix therefore has $65^2 = 4225$ entries, whereas the spectral analysis in Figure 3(c) concerns its 64×64 nonconstant interaction block. The PCA basis is fitted on all 400 sentences before the sentence-level $80/20$ split. Labels and fitted estimators use disjoint train and test sentence subsets, with a shared unsupervised feature map. For each n, eight data seeds independently sample training and test token pairs. Rank-ERM uses ranks $r \in \{4, 8, 16, 32\}$ and radius $B = 2$.

Tables 13 and 14 provide complementary summaries of this experiment. The first reports median held-out R^2 for all repetitions, including settings with extreme OLS errors. The second combines threshold sensitivity and the matched-radius comparison: rank-4 ERM wins all eight repetitions at layer 0 and $n = 2048$, whereas full-matrix least squares wins all eight at layer 6 and $n = 1024$, showing that the effect of the rank constraint differs between the two heads.

Table 13: Median held-out R^2 over eight data seeds for pretrained BERT attention-score prediction. Sr denotes OLS followed by rank-r SVD truncation, and Er denotes rank-r ERM with $\|M\|_F \le 2$.

n	OLS	S4	S8	S16	S32	E4	E8	E16	E32
				Layer 0, head 0					
512	-0.97	-0.89	-0.91	-0.97	-0.97	-6.77	-6.06	-5.61	-6.76
1024	-3.44	-3.25	-3.11	-3.28	-3.41	-2.06	-4.98	-4.57	-6.08
2048	-20.7	-15.7	-18.0	-18.7	-20.1	0.37	-1.63	-4.19	-4.18
4096	-1.3×10^{12}	-6.5×10^{11}	-8.6×10^{11}	-1.1×10^{12}	-1.3×10^{12}	0.68	0.42	-1.05	-2.73
8192	-1.4×10^{3}	-612	-964	-1.2×10^{3}	-1.3×10^{3}	0.71	0.69	0.48	-0.23
16384	0.17	0.47	0.36	0.27	0.20	0.73	0.75	0.70	0.54
32768	0.68	0.69	0.68	0.67	0.68	0.74	0.77	0.76	0.73
				Layer 6, head 0					
512	0.22	0.16	0.20	0.21	0.22	-0.18	-0.12	-0.18	-0.19
1024	-46.5	-19.6	-27.5	-36.5	-44.5	0.34	0.21	0.37	0.41
2048	-487	-172	-262	-356	-461	0.67	0.62	0.59	0.62
4096	-1.4×10^{4}	-3.3×10^{3}	-5.5×10^{3}	-9.0×10^{3}	-1.3×10^{4}	0.76	0.77	0.75	0.74
8192	0.47	0.37	0.39	0.39	0.46	0.78	0.82	0.82	0.82
16384	0.83	0.73	0.79	0.82	0.83	0.78	0.83	0.85	0.85
32768	0.87	0.77	0.83	0.86	0.87	0.79	0.83	0.86	0.87

As a direct constraint check, we saved every fitted matrix and verified its rank and Frobenius norm. Table 15 shows that rank-ERM remains inside the radius, whereas several unconstrained OLS and truncated-OLS fits have norms above two at $n = 2048$. One repetition has $\|\widehat{M}_{\mathrm{OLS}}\|_F = 1.05 \times 10^4$ and test MSE 5.92×10^7. It is retained in all summaries. The norm audit identifies the high-norm source of these extreme errors, and the matched-radius comparison above isolates the contribution of the rank constraint.

Table 14: Attention-score prediction audits. **(a)** Smallest tested n at which the stated number of repetitions beat the training-mean baseline. **(b)** Matched-radius prediction; entries are median test MSE over eight paired repetitions, and the last column counts rank-4 wins.

(a) Threshold sensitivity

Layer	Estimator	median < 1	at least $7/8 < 1$	$8/8 < 1$
0	Full OLS	16384	16384	16384
0	Rank-4 ERM	2048	2048	2048
6	Full OLS	512	512	8192
6	Rank-4 ERM	1024	1024	1024

(b) Matched radius

Layer	n	Full-ball MSE	Rank-4-ball MSE	Rank wins
0	2048	3.877	0.2705	8/8
6	1024	2.391	3.332	0/8

Table 15: Matrix-norm diagnostics for layer 0, head 0. **(a)** Median fitted Frobenius norm and number of eight repetitions exceeding $B = 2$. **(b)** The retained extreme OLS repetition.

(a) Fitted-matrix norms

n	OLS	SVD-4	SVD-32	ERM-4	ERM-32
512	0.53 (0/8)	0.41 (0/8)	0.52 (0/8)	2.00 (0/8)	1.61 (0/8)
2048	3.14 (7/8)	2.51 (6/8)	3.11 (7/8)	1.76 (0/8)	1.66 (0/8)

(b) Extreme repetition

Quantity	Value
$\|\widehat{M}_{\mathrm{OLS}}\|_F$	1.05×10^4
Test MSE	5.92×10^7

F.3 Reference agreement for observed STS-B ratings

We use the real-score setting in Table 10: fit the fixed numerical reference $\widetilde{M}_r$ by rank-constrained least squares on all 5749 STS-B training pairs, then compare size-n fits with it. Fit residuals are retained as observation noise. We vary $d \in \{16, 32, 64\}$ and $r \in \{2, 4, 8\}$, with requested sample sizes $n_{\mathrm{req}} \in \{d^2/4, \ldots, 4d^2\}$ and actual sample sizes $n = \min\{n_{\mathrm{req}}, 5749\}$. Requests at or above the pool size use the full training pool once. Thus, for $d = 64$, the largest actual sample size is $5749 \approx 1.40d^2$. We report the median squared Frobenius error over 10 seeds per cell. Multi-start and first-order diagnostics assess both the reference and subsample solutions. The STS-B validation split is repartitioned into disjoint validation and test sets for selecting the nuclear norm regularization parameter and evaluating out-of-sample prediction error, respectively. Parameter grids and optimization diagnostics for both fits are given in this appendix.

Table 16 reports normalized reference errors by dimension and rank together with empirical condition numbers of the regression matrix, allowing us to examine their relationship near interpolation. In this table, Z denotes the regression matrix $\mathbf{Z}$. Ridge uses $\lambda = 10^{-2}/n_{\mathrm{req}}$ with mean squared loss, where n_{req} is the requested sample size and the table reports the actual sample size n. Accordingly, the $d = 64$, $n = 5749$ rows use the $n_{\mathrm{req}} = 16384$ grid value $\lambda = 10^{-2}/16384$.

Table 16: Normalized reference error ($\mathrm{NMSE} = \|\widehat{M} - \widetilde{M}_r\|_F^2 / \|\widetilde{M}_r\|_F^2$) for supervision-derived rank-r targets on real BERT features. **(a)** Full-matrix interpolation point. **(b)** Largest available sample size, with $n_{\max} = 1024$, 4096, and 5749 for $d = 16$, 32, and 64, respectively. Entries are medians over ten seeds.

(a) Interpolation: $n = d^2$

(d, r)	OLS	ridge	r-ERM	κ
$(16, 2)$	1.58×10^8	9.29×10^4	5.7	
$(16, 4)$	1.18×10^8	6.92×10^4	40	5×10^3
$(16, 8)$	1.01×10^8	5.91×10^4	1×10^3	
$(32, 2)$	7.39×10^7	2.08×10^5	0.44	
$(32, 4)$	5.01×10^7	1.41×10^5	2.6	1.15×10^4
$(32, 8)$	3.55×10^7	9.96×10^4	16	
$(64, 2)$	3.25×10^4	1.20×10^4	5.54×10^{-4}	
$(64, 4)$	1.88×10^4	7×10^3	2×10^{-3}	5×10^3
$(64, 8)$	1.13×10^4	4.2×10^3	6.9×10^{-3}	

(b) Largest sample: $n = n_{\max}$

(d, r)	OLS	ridge	r-ERM	κ
$(16, 2)$	3.9	3.9	0.098	
$(16, 4)$	2.9	2.9	0.31	2×10^1
$(16, 8)$	2.5	2.5	1.3	
$(32, 2)$	1.8	1.8	0.013	
$(32, 4)$	1.2	1.2	0.038	3×10^1
$(32, 8)$	0.86	0.86	0.12	
$(64, 2)$	3.2	3.2	2.74×10^{-4}	
$(64, 4)$	1.8	1.8	8.78×10^{-4}	1×10^2
$(64, 8)$	1.1	1.1	2.7×10^{-3}	

Residual and covariance diagnostics for these real-score experiments appear in Appendix G.1, and multi-initialization and normalized gradient diagnostics in Appendix H.1.

G Dependence and Covariance Transformations

This section measures cross-input dependence directly, varies it in controlled input models, and evaluates covariance transformations that reduce empirical second-moment dependence.

G.1 Cross-input dependence and regression-target changes

Second-moment diagnostics on real features. To quantify dependence between paired STS-B features after PCA and covariance normalization ($d = 16$), let N_{pair} denote the number of pairs used to compute the second moments and set $\widehat{\Sigma}_{jk} = N_{\text{pair}}^{-1} \sum_{i=1}^{N_{\text{pair}}} x_{j,i} x_{k,i}^\top$ for $j, k \in \{1, 2\}$. The empirical cross-covariance has $\left\|\widehat{\Sigma}_{12}\right\|_{\text{op}} = 0.97$, top five singular values $0.97, 0.93, 0.89, 0.85, 0.83$, and mean absolute entry 0.07, showing strong dependence across paired sentences. Each marginal remains nearly isotropic, with $\left\|\widehat{\Sigma}_{11} - I\right\|_{\text{op}} = \left\|\widehat{\Sigma}_{22} - I\right\|_{\text{op}} \approx 0.045$. The subsequent controlled sweep varies cross-correlation up to $|\rho| = 0.6$ while preserving isotropic marginals.

Controlled cross-input correlation. To isolate cross-input dependence from the simultaneous shifts present in real features, we draw $u_1, u_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and set $x_1 = u_1$ and $x_2 = \rho u_1 + \sqrt{1 - \rho^2}\, u_2$. Both marginals remain isotropic, while $\mathbb{E}[x_1 x_2^\top] = \rho I_d$ and $\rho = 0$ gives independent inputs. We use a general rank-r target with $\|M^*\|_F = 1$, take $d = 16, r = 2$, and scan $\rho \in \{-0.3, 0, 0.3, 0.6\}$ on the same $n/d^2 \in \{1/4, 1/2, 1, 2, 4\}$ grid as the independent experiment.

Figure 6 reports the full-matrix interpolation point and a large-sample point. At $n = d^2$, OLS NMSE increases from approximately 2×10^3 at $\rho = 0$ to approximately 10^5 at $\rho = 0.6$, whereas rank-ERM maintains low error. Across the tested correlation range, rank-ERM error varies less with ρ than OLS error. Nonzero ρ changes the risk geometry because the identity $\mathbb{E}[(x_1^\top \Delta x_2)^2] = \|\Delta\|_F^2$ no longer holds. The experiment reports the combined empirical effect on the two estimators. The main error bounds apply to the independent case $\rho = 0$.

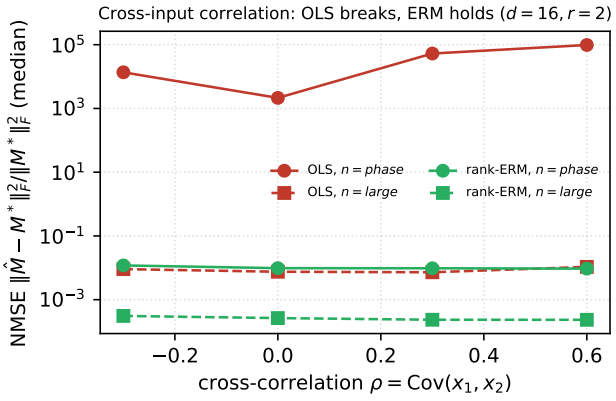


Figure 6: NMSE under controlled cross-input correlation $\mathbb{E}[x_1 x_2^\top] = \rho I_d$ for a general rank-r signal ($d = 16, r = 2$). Curves compare OLS and rank-ERM at the full-matrix interpolation point and in the large-sample regime.

Feature mixing and changes in the reference target. To track reference fitting as cross-correlation decreases, we mix the paired feature x_2 with a randomly permuted copy and refit the rank-2 reference for each mixture. As the empirical cross-correlation decreases from 0.97 to 0.10,

the squared reference norm falls from 0.92 to 0.19, while the fit residual and held-out MSE increase from 1.99 and 2.06 to 8.99 and 9.48, respectively. The sweep therefore couples lower cross-correlation with changes in both the reference target and the predictable signal.

G.2 Covariance transformations and their empirical effects

Appendix C.2.2 gives the structural decomposition under known covariance. The experiments below measure how residualization affects reference recovery and predictable signal. For real scores, reference errors use the transformed target, and fit residuals and prediction errors quantify the simultaneous change in the estimand and predictive information.

Residualization estimated from auxiliary samples. To assess residualization with estimated rather than known covariance, we use the controlled Gaussian input model $x_1 = u_1$, $x_2 = \rho u_1 + \sqrt{1-\rho^2}\, u_2$ with $\rho = 0.6$, $d = 16$, $r = 2$, $\sigma = 0.3$, and $n = 2048$. Here u_2 is the independent residual in Theorem 9. Let $\hat{\rho}$ be the estimate of ρ obtained from n' auxiliary samples. We residualize $x_2 \mapsto (x_2 - \hat{\rho} x_1)/\sqrt{1-\hat{\rho}^2}$. For $|\hat{\rho}| < 1$, this plug-in map preserves the rank of the bilinear coefficient and yields an exact bilinear–quadratic decomposition. When $\hat{\rho} \neq \rho$, the transformed inputs generally retain dependence. Thus, the experiment measures the finite-sample behavior of the estimated map, whereas Theorem 9 gives the independent-residual guarantee for known covariance. Across $n' \in \{64, 256, 1024, 4096\}$, the median squared recovery error remains between 0.021 and 0.027, indicating limited sensitivity to covariance estimation on this grid. The removed correlated component re-enters the label as a quadratic term, contributing structured noise to the bilinear fit.

Residualization and covariance standardization on STS-B. To assess how residualization affects reference error and prediction, we apply it to the STS-B features. Specifically, we residualize $x_2 \mapsto x_2 - \widehat{C}^\top x_1$ with $\widehat{C} = \widehat{\Sigma}_{11}^{-1}\widehat{\Sigma}_{12}$ estimated on the training pool, then normalize the residual using its empirical covariance. The empirical cross-covariance decreases from 0.97 to $3.8 \cdot 10^{-8}$, indicating that the transformed inputs are approximately uncorrelated in empirical second moments. Residualization lowers rank-ERM error against the respective reference fits from 1.11 to 0.55 and nuclear norm error from 0.52 to 0.17 ($n = 2048$, means over three random seeds). The fit residual increases from 1.99 to 8.85, and held-out MSE for the observed ratings increases from 3.13 to 7.80 (Table 17), showing that the transformation simultaneously changes reference recovery and the signal represented by the bilinear model.

Table 17: STS-B residualization diagnostics ($n = 2048$, means over three seeds). Parameter errors are computed against the respective reference fit before or after transformation.

Metric	raw	residualized
Rank-ERM squared error vs reference	1.11	0.55
Nuclear-norm squared error	0.52	0.17
Fit residual $\sigma^2_{\mathrm{resid}}$	1.99	8.85
Held-out MSE (observed ratings)	3.13	7.80

H Numerical Diagnostics and an Additional Dataset Case Study

The STS-B restart study reports feasibility, local stationarity, empirical loss, and reference agreement for the attained factorized solutions. The SICK-R case study then evaluates the same real-score pipeline on a second semantic-relatedness dataset. These diagnostics summarize local properties of the attained numerical solutions, while Equation 46 quantifies how a global objective gap contributes to estimation error.

H.1 Optimization diagnostics on STS-B

Multiple initializations and normalized gradient diagnostics. We restart projected factorized rank-ERM from eight independent initializations at $n = 2048$ on the real STS-B scores. All final iterates are feasible, with Frobenius norms between 1.7633 and 1.7660. We measure local stationarity by the tangent-space projected gradient, normalized by the RMS prediction score: the diagnostic ranges from $5.03 \cdot 10^{-16}$ to $3.13 \cdot 10^{-5}$, with five runs at most 10^{-8}. The attained solutions

form three groups with training MSE/reference-distance-squared pairs $(3.40, 2.23)$, $(3.10, 1.19)$, and $(3.11, 1.10)$, respectively summarizing empirical fit and agreement with the full-pool reference. For each seed in the real-feature experiments, we report the feasible initialization with the lowest empirical risk.

H.2 Additional SICK-R case study

We assess dataset dependence by applying the same frozen-BERT, PCA, covariance-normalization, and estimator pipeline to SICK-R sentence embeddings. The input and covariance diagnostics give cross-correlation $\|\widehat{\Sigma}_{12}\|_{\mathrm{op}} = 0.89$, marginal isotropy residuals $0.074/0.073$ (condition numbers 1.14, 1.14), and fit residual $\sigma^2_{\mathrm{resid}} = 1.91$.

Table 18: Additional SICK-R case study ($d = 16, r = 2$). **(a)** Squared Frobenius distance to the rank-2 full-pool reference over five seeds. **(b)** Input, covariance-transformation, and numerical-solution diagnostics.

(a) Reference error by sample size

n	rank-ERM	OLS+SVD-r	nuclear norm
512	3.02 ± 0.53	1.59 ± 0.34	1.23 ± 0.11
2048	1.81 ± 0.17	0.17 ± 0.06	0.98 ± 0.05
3330	1.88 ± 0.03	0.084 ± 0.000	0.91 ± 0.000

(b) Input and solution diagnostics

Diagnostic	Value
Cross-correlation $\|\widehat{\Sigma}_{12}\|_{\mathrm{op}}$	0.89
Marginal isotropy residual (feat. 1 / feat. 2)	0.074 / 0.073
Marginal condition numbers	1.14, 1.14
Fit residual $\sigma^2_{\mathrm{resid}}$	1.91
ERM vs truncated-OLS training loss at $n = 2048$	4.69 / 7.57
Leading singular values of subsample OLS	$0.74, 0.61, 0.55, 0.47, \ldots$

The closed-form SVD-r error decreases with sample size, whereas factorized rank-ERM plateaus over the reported grid; a run warm-started from the SVD-r solution shows the same pattern. At $n = 2048$, the ERM iterate has lower empirical loss than truncated OLS (4.69 versus 7.57) and a larger squared Frobenius distance from the reference. The subsample OLS estimate has leading singular values $0.74, 0.61, 0.55, 0.47, \ldots$. Multi-start runs likewise attain similar training losses with different reference errors. Thus, empirical loss and agreement with the full-pool reference rank the attained solutions differently in this experiment.

I Notation

Symbol	Definition	Symbol	Definition
r^*	ground-truth rank in misspecified settings	$\widehat{M}_\phi$	rank-constrained estimator for known ϕ
$\widetilde{M}_r$	numerical rank-r reference for real scores	$\widehat{M}_\lambda$	nuclear-norm penalized estimator
M_r	best rank-r approximation to M^*	$b_{n,\eta}$	noise scale in equation 38
$Z = x_2 \otimes x_1$	vectorized feature pair	$\mathcal{S}_r$	direction set in equation 35
$\mathbf{Z}$	design matrix with rows $(x_{2,i} \otimes x_{1,i})^\top$	G_n	scaled noise matrix
P_M	law of the n observations under M	κ	restricted-curvature constant
$p_r = 2dr - r^2$	dimension of the rank-r manifold	τ	restricted noise threshold
$k_r = \min\{2r, d\}$	maximum feasible-error rank	$\sigma^2_{\mathrm{resid}}$	reference-fit residual MSE
B_N	radius after marginal normalization	$\phi_{\log}$	logistic sigmoid
ϕ, Φ	known scalar link and antiderivative	$\mathcal{E}_{\mathrm{RC}}$	restricted-curvature event
m, L	derivative bounds for ϕ	$\mathcal{E}_{\mathrm{noise}}$	noise-operator event