# THE RANK COST OF FORGETTING: WHEN ONE DELETION REQUIRES A HIGH-RANK UPDATE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Removing a record from a training dataset leaves the fitted model unchanged. Machine unlearning seeks to remove its influence without repeating the full training procedure. When the model parameter is a matrix, a low-rank update is compact, but can it reproduce the change caused by retraining? We answer this question in regularized bilinear regression. A rank-one deletion gradient can produce a retraining correction that low-rank updates cannot approximate to an arbitrarily small relative error. For independent Gaussian feature pairs, when the sample size is proportional to the number of matrix entries and regularization is fixed and positive, we derive an exact asymptotic formula for the minimum relative squared prediction error, uniformly over all positive ranks. For a nonzero deletion effect, every nonzero rank growing more slowly than the matrix dimension leaves the same positive asymptotic error, even if an oracle knows the exact retrained model. Any fixed relative tolerance below this threshold requires a rank proportional to the matrix dimension. This is a relative error, so it can persist while the absolute deletion effect becomes small. We also identify Hessians that preserve rank and give computable approximation bounds to distinguish an insufficient rank from an inaccurate numerical solution. Experiments on frozen text representations show improved agreement with retraining at higher update ranks, including binary decisions after regularization is selected for validation accuracy. On one real-data instance, a verified bound excludes every rank-one correction at a stated absolute prediction tolerance, while an explicit rank-eight update meets it.

## 1 INTRODUCTION

A training record may need to be removed because it has been withdrawn or found to be unreliable. Deleting it from the stored dataset leaves the trained model unchanged. Machine unlearning asks how to remove its influence from the model itself (Cao and Yang, 2015). This matters because models can retain information about individual examples: Carlini et al. (2021) recovered training text from a language model.

We compare an update with the model obtained by training again on the remaining observations. Repeating this computation after every deletion can be expensive, motivating methods that reuse computations or limit how much training must be repeated (Ginart et al., 2019; Bourtoule et al., 2021). We study whether the allowed parameter changes can express the effect of retraining on predictions.

For a matrix parameter $W \in \mathbb{R}^{d\times d}$, an update $B = A_1 A_2^\top$, with $A_1, A_2 \in \mathbb{R}^{d\times r}$, stores $2dr$ numbers instead of $d^2$ and has rank at most $r$. This form underlies LoRA (Hu et al., 2022) and recent unlearning methods (Cha et al., 2025; Prakash et al., 2026). We compare such updates with retraining over all matrices. If both fits instead use rank-$r_0$ adapters on the same frozen base, their difference has rank at most $2r_0$, and our lower bounds do not apply to that comparator. We ask: *how much of full retraining's prediction change can a low-rank update reproduce?*

We isolate this question in regularized bilinear regression, with predictor $x^\top W z$. The entries of $W$ describe interactions between two observed feature vectors, and the quadratic training objective has a unique minimizer. One observation contributes a gradient proportional to the rank-one matrix $xz^\top$. Retraining requires applying the inverse Hessian of the retained objective to this gradient, an operation



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.

that need not preserve rank. The Hessian describes how the loss couples different parameter directions. Thus, inferring a low-rank correction from a low-rank gradient requires additional information about curvature. This inverse-Hessian dependence also appears in influence-function approximations (Koh and Liang, 2017). In our quadratic setting, the retained Hessian gives the exact correction.

We give an oracle the exact retrained model and ask for the best correction of rank at most $r$. Any error left by this oracle is unavoidable under the rank constraint, regardless of how the update is computed. We measure squared prediction error relative to the squared prediction change caused by retraining. Under the independent, centered test-feature model with identity covariance used below, this is the relative squared Frobenius error. An error of zero means exact agreement with retraining, while an error of one is obtained by leaving the original model unchanged.

To our knowledge, we give the first exact formula showing how closely the best low-rank update can reproduce the predictions obtained by retraining after deleting one observation, as the model and training sample grow. We establish this result for regularized bilinear regression with independent Gaussian feature pairs. Denote this minimum relative error by $\mathcal{E}_r$. The matrix dimension $d$ and sample size $n$ grow with $d^2/n \to \gamma > 0$, so the number of observations is proportional to the number of parameters. For fixed regularization $\lambda > 0$ and ranks $1 \leqslant r_d \leqslant d$, we obtain

$$\mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \chi(\gamma, \lambda)\tau(\rho), \qquad r_d/d \to \rho. \tag{1}$$

Here $\xrightarrow{\mathbb{P}}$ denotes convergence in probability. The error left by the best rank-one correction converges to $\chi(\gamma, \lambda) > 0$. The function $\tau(\rho)$ describes how this error decreases as the retained proportion of singular directions increases, from $\tau(0) = 1$ to $\tau(1) = 0$. In particular, for every nonzero rank growing more slowly than $d$, the error converges to the same positive value $\chi$. Reducing the error below this level requires rank proportional to $d$. Section 5 defines both factors, and Theorem 2 gives convergence uniformly over all positive ranks.

The relative error can remain positive even when the absolute deletion effect becomes small. We also identify Hessians for which low-rank corrections suffice and derive bounds from approximate numerical solutions to distinguish an insufficient rank from numerical error.

## 2 Related Work

Unlearning guarantees depend on what is compared with retraining. Guo et al. (2020) compare the distributions of randomized model outputs, while Golatkar et al. (2020) study information about forgotten observations that remains in network weights. For convex models, Neel et al. (2021) analyze gradient-based deletion methods over sequences of data updates. Sekhari et al. (2021) relate deletion guarantees to generalization and memory requirements. Our quantity is the prediction difference from exact retraining for a fixed sample. A bound on this difference alone does not establish these distributional or information guarantees. Yu et al. (2025) analyze failures of retraining equivalence for gradient ascent after staged training. Our rank constraint imposes an approximation error even when the exact retrained model is available, independently of the algorithm used to choose the correction.

Zeng and Lee (2024) study the expressive power of low-rank adaptation for matching a target network, while Rushka and Klabjan (2026) analyze rank–accuracy trade-offs under gradient flow. For unlearning, SEMU uses singular value decomposition to select update subspaces and proves an optimality result for approximating a gradient (Sendera et al., 2025). LoKU uses Fisher information to initialize adapters and reports that the rank constraint can be too restrictive for its objective (Cha et al., 2025). Our analysis determines the singular-value tail of the correction generated by deleting an observation, and hence the minimum error achievable under each rank constraint.

Obliviate motivates low-rank corrections through the support of deletion gradients in user and item embeddings (Prakash et al., 2026, Appendices H.6 and J.8). Proposition H.11 establishes a restriction on this support. Extending that restriction to the inverse-Hessian correction requires an additional argument about the Hessian. Obliviate separately considers weak parameter coupling in Assumption H.3 and retains a projection-error term in Theorem K.12. We quantify the error imposed by the rank constraint in bilinear regression. For independent Gaussian feature pairs in the proportional regime with fixed positive regularization, every nonzero sublinear rank leaves a positive asymptotic relative error, even for a rank-one deletion gradient.

The accuracy of Newton-based deletion in high dimensions is studied by Zou et al. (2025). The retained objective is quadratic here, so there is no Taylor approximation error, but restricting the correction's rank can still prevent agreement with retraining. Matrix factorization of curvature is a separate issue. K-FAC approximates blocks of a neural network's Fisher information by Kronecker products (Martens and Grosse, 2015). Our deterministic results explain why an exactly separable Hessian preserves rank, and why this property cannot be assumed for a general empirical Hessian.

The probabilistic analysis uses known limits for tensor-product sample covariance matrices (Lytova, 2017; Collins et al., 2022; Yaskov, 2016). These determine scalar resolvent quantities but do not directly give the singular values of the inverse-Hessian image after reshaping it as a matrix. For this additional step, we apply the Gaussian matrix criterion of Bandeira et al. (2023), verifying its hypotheses through a Schur complement and concentration of partial traces. The key step is to determine the singular-value tail of the reshaped inverse-Hessian image of an independent rank-one direction. We then obtain an exact error law uniformly over all positive ranks, including ranks selected from the data. A covariance eigenvalue limit, a norm calculation, or the Eckart–Young theorem alone does not determine this tail. No Haar assumption on Hessian eigenvectors is used.

## 3 Deletion and Two Approximation Errors

Each observation contains two feature vectors $x_i, z_i \in \mathbb{R}^d$ and a response $y_i \in \mathbb{R}$. Write $D := ((x_i, z_i, y_i))_{i\in[\![1,n]\!]}$ for the training sample. We predict the response using $f_W(x, z) := x^\top W z$, where $W \in \mathbb{R}^{d\times d}$ assigns a coefficient to each interaction between the two vectors. To express this prediction as a matrix inner product, define $\langle A, B\rangle := \operatorname{tr}(A^\top B)$, $\|A\|_{\mathrm{F}}^2 := \langle A, A\rangle$ and $X_i := x_i z_i^\top$. Then $f_W(x_i, z_i) = \langle W, X_i\rangle$. For observations indexed by $T \subseteq [\![1, n]\!]$, we minimize the regularized squared loss

$$L_T(W) := \frac{1}{2n}\sum_{i\in T}(\langle W, X_i\rangle - y_i)^2 + \frac{\lambda}{2}\|W\|_{\mathrm{F}}^2 . \tag{2}$$

The denominator $n$ and the regularization parameter $\lambda > 0$ stay fixed after deletion. This convention removes the observation's contribution while keeping the scale of every remaining term unchanged.

Let $i_0 \in [\![1, n]\!]$ be the index of the observation to be deleted and define $R := [\![1, n]\!] \setminus \{i_0\}$. Let $W_D, W_R$ minimize the full and retained objectives, and define the retraining correction by $\Delta := W_R - W_D$. Retraining restores a zero gradient for the retained objective. This objective is quadratic, so its Hessian maps a parameter change to the corresponding gradient change:

$$\mathcal{H}_R[A] := \lambda A + \frac{1}{n}\sum_{i\in R}\langle A, X_i\rangle X_i. \tag{3}$$

For the vector form, let $\operatorname{vec}(A)$ stack the columns of $A$. With $\phi_i := \operatorname{vec}(X_i) \in \mathbb{R}^{d^2}$, the Hessian is represented by $H_R = \lambda I_{d^2} + n^{-1}\sum_{i\in R}\phi_i\phi_i^\top \succeq \lambda I_{d^2}$. The positive regularization therefore makes the Hessian invertible.

**Proposition 1** (Exact deletion). *Let $e_{i_0} := \langle W_D, X_{i_0}\rangle - y_{i_0}$ be the residual of the deleted observation. Then*

$$\mathcal{H}_R[\Delta] = \frac{e_{i_0}}{n}X_{i_0}, \qquad \operatorname{vec}(\Delta) = \frac{e_{i_0}}{n}H_R^{-1}\phi_{i_0}. \tag{4}$$

*For a set $F$ of deleted observations, the right-hand side is $n^{-1}\sum_{i\in F} e_i X_i$, where $e_i = \langle W_D, X_i\rangle - y_i$.*

At $W_D$, removing the forgotten terms changes the gradient by minus that right-hand side. The retained objective is quadratic, so solving the displayed linear system is exactly retraining. Full details, leverage formulas, and alternative normalization conventions appear in Appendix B.

The question is now how much of the retraining change can be represented with a prescribed number of matrix directions. We restrict the correction to a matrix $B$ of rank at most $r \in [\![0, d]\!]$. Write $s_1(A) \geqslant \cdots \geqslant s_d(A) \geqslant 0$ for the singular values of a matrix $A$. For a nonzero retraining correction $\Delta$, define

$$\mathcal{E}_r := \frac{\inf_{\operatorname{rg}(B)\leqslant r}\|B - \Delta\|_{\mathrm{F}}^2}{\|\Delta\|_{\mathrm{F}}^2} = \frac{\sum_{j>r} s_j(\Delta)^2}{\sum_j s_j(\Delta)^2}. \tag{5}$$

The second equality is the Eckart–Young theorem (Eckart and Young, 1936). It states that keeping the $r$ largest singular values of $\Delta$ minimizes the squared Frobenius error. This is the oracle correction of rank at most $r$. In particular, $\mathcal{E}_0 = 1$ and $\mathcal{E}_d = 0$. If $\Delta = 0$, the zero correction already agrees with retraining and the relative error is undefined.

For independent, mean-zero test features with covariance $I$, independent of the training sample,

$$\mathbb{E}_{x,z}\left[(f_{W_D+B} - f_{W_R})^2\right] = \|B - \Delta\|_{\mathrm{F}}^2 . \tag{6}$$

Thus (5) is an exact prediction discrepancy relative to the change produced by retraining. Appendix C gives the corresponding weighted error and optimal rank approximation for independent, mean-zero test features with general invertible covariance matrices.

Fitting the retained objective answers a different question from matching predictions. Its excess loss is a squared error weighted by the retained Hessian, so truncated singular value decomposition (SVD) need not minimize it. Appendix C defines this second relative error, denoted by $\mathcal{E}_r^H$, and compares the two criteria. Convergence of a numerical factorization method alone does not establish global optimality.

## 4 When the Hessian Preserves Low Rank

If the *full* Hessian is separable, $\mathcal{H}_R[A] = PAQ$ with $P, Q \succ 0$, then

$$\mathcal{H}_R^{-1}[xz^\top] = P^{-1}xz^\top Q^{-1} \tag{7}$$

is rank one. For $k$ deletions, the correction has rank at most $k$. This control also explains why ordinary multi-output ridge regression does not exhibit the same obstruction: its curvature acts by multiplication on one side.

Adding $\lambda A$ to a separable operator need not preserve this property. Appendix D bounds the error for nearly separable Hessians and derives a condition-number bound (Gustafson, 1968). It also shows that two Hessians with identical eigenvalues can give different rank errors.

For any nonzero matrix $A$, write $\mathcal{E}_r(A)$ for (5) with $A$ in place of $\Delta$, and $\|A\|_{\mathrm{op}} := s_1(A)$.

## 5 Rank Approximation for Gaussian Features

The deterministic results show that the Hessian can preserve or increase rank. We now ask what happens for a specified random feature distribution. Here $d$ is the side length of the parameter matrix, which has $d^2$ entries. The asymptotic regime keeps the number of observations proportional to this parameter count.

**Assumption 1** (Independent Gaussian feature pairs)**.** For every $i \in [\![1, n]\!]$, the vectors $x_i, z_i$ are independent with law $\mathcal{N}(0, I_d)$, and feature pairs are independent across observations. The forgotten index is fixed independently of the realized sample. We let $d^2/n \to \gamma \in ]0, +\infty[$.

Convergence below is in probability, denoted by $\xrightarrow{\mathbb{P}}$. The notation $O_{\mathbb{P}}(a_d)$ means boundedness in probability after division by $a_d$, and $o_{\mathbb{P}}(a_d)$ means convergence to zero after division by $a_d$.

Multiplying a nonzero correction by a nonzero scalar leaves its relative rank-approximation error unchanged. We can therefore study its direction separately from its magnitude. Normalize the deleted feature vectors by setting $u := x_{i_0}/\|x_{i_0}\|$ and $v := z_{i_0}/\|z_{i_0}\|$. Their outer product $U := uv^\top$ has rank one and unit Frobenius norm. To apply the inverse Hessian in vector form, set $q := \mathrm{vec}(U)$ and define the retained feature covariance $S := n^{-1}\sum_{i\in R}\phi_i\phi_i^\top$. The exact deletion identity then gives

$$M := \mathrm{mat}[(\lambda I + S)^{-1}q], \quad \Delta = \frac{e_{i_0}\,\|x_{i_0}\|\,\|z_{i_0}\|}{n} M. \tag{8}$$

In the Gaussian analysis below, we write $\mathcal{E}_r := \mathcal{E}_r(M)$. It equals the deletion error in (5) whenever $\Delta \neq 0$. If $\Delta = 0$, no correction is needed.

We now compare the correction's component along $U$ with its total squared norm. Their limits will be denoted by $m$ and $m_2$. The scalar $\chi$ will measure the asymptotic fraction of the squared norm

orthogonal to $U$. These quantities are expressed through the mean-one Marchenko–Pastur law $\nu_\gamma$, including its zero atom when $\gamma > 1$ (Marchenko and Pastur, 1967). Define

$$m := \int \frac{\nu_\gamma(\mathrm{d}s)}{\lambda + s} = \frac{2}{\lambda + 1 - \gamma + \sqrt{(\lambda + 1 - \gamma)^2 + 4\gamma\lambda}}, \tag{9}$$

$$m_2 := \int \frac{\nu_\gamma(\mathrm{d}s)}{(\lambda + s)^2}, \qquad \chi := \gamma(1 - \lambda m)^2. \tag{10}$$

For fixed $\lambda > 0$, $0 < \chi < 1$ and $m_2 = m^2/(1 - \chi)$. Tensor covariance limits and concentration over the independent $u, v$ imply

$$\langle M, U \rangle \xrightarrow{\mathbb{P}} m, \qquad \|M\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} m_2. \tag{11}$$

Since $\|U\|_{\mathrm{F}} = 1$, the projection coefficient $a_d := \langle M, U \rangle$ satisfies

$$\|M - a_d U\|_{\mathrm{F}}^2 = \|M\|_{\mathrm{F}}^2 - a_d^2 \xrightarrow{\mathbb{P}} m_2 - m^2 = m_2\chi. \tag{12}$$

Thus the fraction of squared Frobenius norm orthogonal to the gradient converges to $\chi$. This orthogonal component need not itself have high rank. Its distribution across singular values determines the required rank. The next definition describes that distribution, and the fixed-regularization theorem connects it to the exact correction.

### 5.1 The spectral tail

The best rank-$r$ Frobenius approximation retains the $r$ largest singular values, so its squared error is the sum of squares of the discarded values. The theorem below describes the normalized squared singular values of the remaining correction block through the aspect-one Marchenko–Pastur law $\mu_1 := \mathrm{MP}_1$. In contrast, $\nu_\gamma$ describes eigenvalues of the feature covariance. The law $\mu_1$ has support $[0, 4]$, and integrating its density with the change of variable below gives the trigonometric formulas. For a retained proportion $\rho \in [0, 1]$, let $q_\rho \in [0, 4]$ satisfy $\mu_1([q_\rho, 4]) = \rho$. Define the contribution below this cutoff by

$$\tau(\rho) := \int_0^{q_\rho} t\, \mu_1(\mathrm{d}t), \qquad 0 \leqslant \rho \leqslant 1. \tag{13}$$

Since $\mu_1$ has mean one, $\tau(\rho)$ is the fraction of the total squared singular values left after retaining the largest proportion $\rho$ under this distribution. It decreases continuously from one to zero. Setting $q_\rho = 4\cos^2\theta$ with $\theta \in [0, \pi/2]$ gives

$$\rho = \frac{2\theta - \sin 2\theta}{\pi}, \quad \tau = 1 - \frac{2\theta}{\pi} + \frac{\sin 4\theta}{2\pi}. \tag{14}$$

## 6 Exact Rank Law at Fixed Regularization

The Hessian eigenvalue law alone does not determine the singular values of $M$. We must also understand how the inverse Hessian acts on the rank-one direction $q$. We establish this directly in the paired model, without replacing the Hessian eigenvectors by Haar rotations.

**Theorem 2** (Exact asymptotic error at fixed rank fraction)**.** *Under Assumption 1, fix $\lambda > 0$ and define the minimum error using $M$ in* (8)*. It equals the deletion minimum error whenever $\Delta \neq 0$. Then*

$$\sup_{1 \leqslant r \leqslant d} |\mathcal{E}_r - \chi(\gamma, \lambda)\tau(r/d)| \xrightarrow{\mathbb{P}} 0. \tag{15}$$

*Consequently, $r_d/d \to \rho$ with $1 \leqslant r_d \leqslant d$ implies $\mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \chi\tau(\rho)$. In particular, $\mathcal{E}_{r_d} \to \chi$ for every nonzero sublinear rank.*

For example, $\gamma = 1$ and $\lambda = 0.5$ give $\chi = 1/4$. Every nonzero rank growing more slowly than $d$ then leaves asymptotically one quarter of the squared prediction change caused by retraining. The comparison is with the deletion effect, not the prediction error against the response labels. The relative error can therefore converge to a positive value even when task predictions are accurate and the absolute deletion effect is small.

Uniform convergence also permits ranks selected from the data. Appendix H gives complementary bounds obtained from an explicit auxiliary matrix.

### 6.1 WHY THE CORRECTION HAS THIS SPECTRAL TAIL

Choose left and right orthonormal bases beginning with $u$ and $v$, so that $u = v = e_1$, and write $x_i = (\alpha_i, a_i)$ and $z_i = (\beta_i, b_i)$, with $a_i, b_i \in \mathbb{R}^{d-1}$. The first-row and first-column entries of $M$ outside position $(1,1)$ have squared norm tending to zero in probability. The entry $a_d := M_{11}$ converges to $m > 0$. Appendix G.2 constructs a matrix $Z_d$ with zero first row and column such that

$$\|M - a_d(U - Z_d)\|_{\mathrm{F}} = O_{\mathbb{P}}(\sqrt{\log d/d}). \tag{16}$$

Conditional on $(a_i, b_i, \beta_i)_i$, the entries of $Z_d$ are jointly Gaussian because they depend linearly on the independent $\alpha_i$. These entries need not be independent. Bounds on their conditional second moments and covariance verify the assumptions of Bandeira et al. (2023, Corollary 2.11), giving the Marchenko–Pastur law for the normalized squared singular values and an operator-norm bound.

The terms $a_d U$ and $-a_d Z_d$ have orthogonal row and column spaces. Truncated SVD first retains $a_d U$ with probability tending to one, and the remaining singular values give $\tau$. Operator-norm control makes their tail convergence uniform over ranks. The distance to the set of matrices of rank at most $r$ is 1-Lipschitz in Frobenius norm, so (16) transfers the limit to $M$. Appendix G gives the proof and intermediate formulas.

### 6.2 THE MINIMUM RANK FOR A PRESCRIBED ERROR

For a prescribed relative error, we seek the smallest rank meeting that tolerance. For $0 < \varepsilon < \chi$, define $r_\varepsilon := \min\{r : \mathcal{E}_r \leqslant \varepsilon\}$. Strict monotonicity of $\tau$ and the uniform law imply

$$r_\varepsilon/d \xrightarrow{\mathbb{P}} \tau^{-1}(\varepsilon/\chi). \tag{17}$$

For $\varepsilon > \chi$, rank one suffices with probability tending to one. At $\varepsilon = \chi$, the minimum rank fraction tends to zero, but rank-one success need not follow. A nonzero fraction of the matrix dimension is therefore necessary for every fixed tolerance strictly below $\chi$.

Appendices I and J give the joint high-regularization limit and the comparison with Haar-distributed Hessian eigenvectors. Appendix E.5 gives scalar limits of the rank law. The proof for Gaussian feature pairs uses neither an interchange of limits nor this additional eigenvector model.

## 7 CHOOSING THE RANK FROM ERROR BOUNDS

We first give a diagnostic that separates an insufficient rank from numerical error. It approximates the full correction and computes its singular values.

To determine whether a numerical correction is accurate enough, we measure how well it satisfies the deletion equation. Let $\widetilde{\Delta}$ be a computed approximation of the retraining correction. Define $G := e_{i_0} X_{i_0}/n$ and its linear-system residual $R_{\mathrm{lin}} := G - \mathcal{H}_R[\widetilde{\Delta}]$. Since $H_R \succeq \lambda I$, this residual bounds the correction error:

$$\left\|\widetilde{\Delta} - \Delta\right\|_{\mathrm{F}} \leqslant \eta, \qquad \eta := \|R_{\mathrm{lin}}\|_{\mathrm{F}}/\lambda. \tag{18}$$

We also need the norm of the computed correction and its rank-$r$ truncation error. Define $N := \left\|\widetilde{\Delta}\right\|_{\mathrm{F}}$ and $t_r^2 := \sum_{j>r} s_j(\widetilde{\Delta})^2$. If $N > \eta$, the exact correction has positive norm and the following relative bounds are well defined:

$$\underline{E}_r = \frac{(t_r - \eta)_+^2}{(N+\eta)^2} \leqslant \mathcal{E}_r, \tag{19}$$

$$\frac{\left\|\widetilde{\Delta}_r - \Delta\right\|_{\mathrm{F}}^2}{\|\Delta\|_{\mathrm{F}}^2} \leqslant \overline{E}_r = \frac{(t_r + \eta)^2}{(N-\eta)^2}, \tag{20}$$

where $\widetilde{\Delta}_r$ is truncated SVD. A lower bound above tolerance proves impossibility for every rank-$r$ update. An upper bound below tolerance proves that the constructed update meets the tolerance.

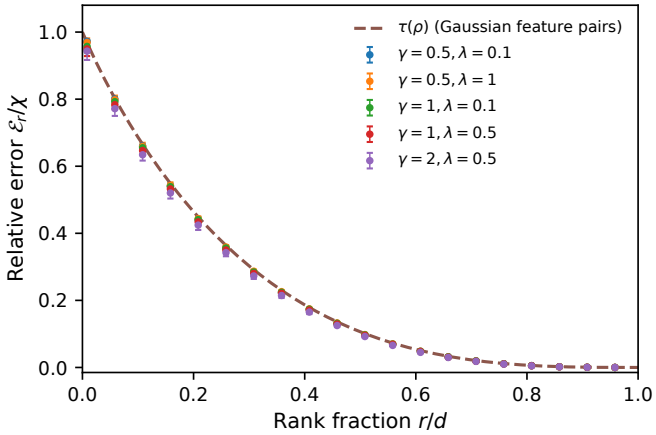


Figure 1: Gaussian features at $d = 120$. The oracle errors, divided by $\chi$, follow the predicted curve $\tau(r/d)$ for five pairs $(\gamma, \lambda)$. Points average three independent draws. Bars show one sample standard deviation. No coefficient is fitted.

Appendix K illustrates how a more accurate solve can identify the minimum sufficient rank when the initial bounds overlap the tolerance, and compares these calculations with direct retraining.

Algorithm 1 gives the procedure. Fitting a factorized correction to the retained objective seeks the Hessian-weighted minimum $\mathcal{E}_r^H$. A computed rank-$r$ correction gives an upper bound on this minimum, without establishing global optimality. Appendix C gives both formulations.

### 7.1 Computational cost

A Hessian application costs $O(nd^2)$ arithmetic for dense features and $O(nd + d^2)$ working storage without blocking, avoiding $O(d^4)$ Hessian storage. Algorithm 1 selects the rank using a full correction and its spectrum. These computations are included in its cost even though the resulting rank-$r$ update can be stored in $2dr$ entries.

Appendix M calibrates a rank-selection rule on $b$ independent datasets from the same distribution. At fixed $(d, n, \lambda)$ and for a nonadaptive deletion rule, the probability of missing any tolerance on an independent new dataset is at most $1/(b + 1)$, averaged over both calibration and new datasets. The rule selects an oracle rank. Computing the update remains a separate problem.

For an approximate original fit, the correction-error bound must also include the original gradient norm divided by $\lambda$ (Appendix B.4). Under Assumption 1, at fixed $\lambda > 0$ and with $e_{i_0} = O_{\mathbb{P}}(1)$, the squared deletion effect is $O_{\mathbb{P}}(n^{-1})$. In this regime, even a small error caused by incomplete original training can exceed the deletion effect.

## 8 Experiments

We evaluate the Gaussian law and rank selection on simulated features, then study bilinear heads on frozen text representations and a softmax head on frozen image features. In each experiment, retraining optimizes the full head matrix. The feature extractor remains fixed.

### 8.1 Gaussian features and numerical error

The baseline uses independent $x_i, z_i \sim \mathcal{N}(0, I_d)$, six dimensions from 16 to 120, five parameter pairs and three seeds per setting. We set $n = \mathrm{round}(d^2/\gamma) + 1$ and use the actual aspect ratio. At $d = 120$, we compare the mean oracle error over three seeds with $\chi\tau(r/d)$ for five parameter pairs and eight ranks. The median absolute relative discrepancy over these 40 comparisons is 4.1%, with no fitted curve coefficient. At $d = 256$ and $(\gamma, \lambda) = (1, 0.5)$, five seeds give mean $\mathcal{E}_1/\chi = 0.983$ and

mean uniform error $\max_{1\leqslant r\leqslant d}|\mathcal{E}_r - \chi\tau(r/d)| = 0.00172$. Figure 1 compares the Gaussian error curves with the theoretical law. Appendix K also displays convergence of the rank-one error.

The complete synthetic study contains 1284 correction solves with relative residual below $10^{-11}$ and 21 direct-retraining comparisons. All 252 computed error intervals in these comparisons contain the oracle and constructed-approximation errors. Across 270 further solves and five tolerances, the asymptotic rank rule misses 22 of 1350 dependent targets, whereas the numerical bounds identify the minimum sufficient rank in every case. The independently calibrated rule of Appendix M misses none of 2700 tested tolerances on 540 new draws, compared with 33 failures for the unadjusted rule, at the cost of additional rank. Appendix K reports the protocols and uncertainty summaries.

### 8.2 Heads on frozen pretrained representations

We use MiniLM on STSB, BGE on SICK, and E5 on QQP and PAWS (Sentence Transformers, n.d.a; BAAI, n.d.; Intfloat, n.d.; Sentence Transformers, n.d.b; Marelli et al., 2014; Wang et al., 2018; Zhang et al., 2019). Each predictor adds $x^\top Wz$ to an affine function of the original embedding cosine. The encoder, PCA whitening and affine baseline remain fixed when the head is retrained. The score comparisons use $d \in \{16, 32\}$, $n = 2d^2$, fixed $\lambda \in \{0.1, 1\}$ and three training subsamples. Each head has 20 single deletions and ten groups of each size 4 and 16. For the binary-decision comparisons below, regularization is selected on separate validation pairs. Evaluation pairs never select a head hyperparameter or an update candidate.

On an evaluation set $T$ separate from training, define the relative squared prediction discrepancy of a candidate $B$ by

$$\widehat{\mathcal{E}}_T(B) := \frac{\sum_{i\in T}(x_i^\top(B-\Delta)z_i)^2}{\sum_{i\in T}(x_i^\top\Delta z_i)^2}, \tag{21}$$

when the denominator is positive. This ratio compares the candidate's squared prediction discrepancy with the squared prediction change caused by retraining. It equals one for no update and can exceed one if the candidate increases the discrepancy. It generally differs from the relative squared Frobenius error. We compare truncated SVD, factors fitted to the retained objective, and factors matched to the retrained predictions on 2048 separate input pairs. The last method has oracle access to retraining. Each optimization uses three initializations and at most 500 L-BFGS-B iterations. Convergence does not establish global optimality.

An entry of $20.3\%$ means that the squared discrepancy from retraining is $0.203$ times its no-update value, not that $20.3\%$ of labels are predicted incorrectly. The fixed baseline cancels from (21).

Table 3 reports prediction agreement for the fitted candidates. We also compute the minimum relative squared Frobenius error over every single deletion in the same heads. At $d = 32$, $\lambda = 1$, its mean over deletions and heads ranges from $8.6\%$ to $12.3\%$ at rank one and from $1.9\%$ to $2.5\%$ at rank eight across the four corpora. Appendix N.6 compares these minima with the Gaussian law and examines the effect of shuffling feature pairs.

**When an update is needed.** We compare the RMS score discrepancy with a numerical tolerance of $0.1\%$ of the target-score range. Appendix N.7 specifies the protocol, and Appendix L relates relative error to requirements on predictions and noisy response laws. Figure 2 shows that 11, 12 and 43 of 60 single deletions require an update on STSB, SICK and QQP. Among them, the rank-one candidate suffices for 9, 11 and 20, respectively, and rank eight suffices for all.

We also evaluate binary decisions on QQP and PAWS, selecting each head's regularization by accuracy on 2000 separate validation pairs and keeping it fixed after deletion. On 5000 evaluation pairs per corpus, mean initial accuracy is $79.51\%$ on QQP and $56.39\%$ on PAWS, compared with $76.52\%$ and $55.76\%$ for the affine baselines. Across 60 single deletions, retraining changes 268 and 340 of 300000 decisions at threshold $0.5$, respectively. Retained-objective updates recover $74.3\%$ of these changes at rank one and $96.6\%$ at rank eight on QQP, with total disagreements falling from 135 to 30. On PAWS, recovery rises from $73.2\%$ to $90.6\%$, and disagreements fall from 212 to 64 (Appendix N.7).

**Excluding every rank-one correction.** Candidate comparisons alone cannot establish a rank obstruction. Appendix N.8 bounds the actual prediction metric below by a separable quadratic form,

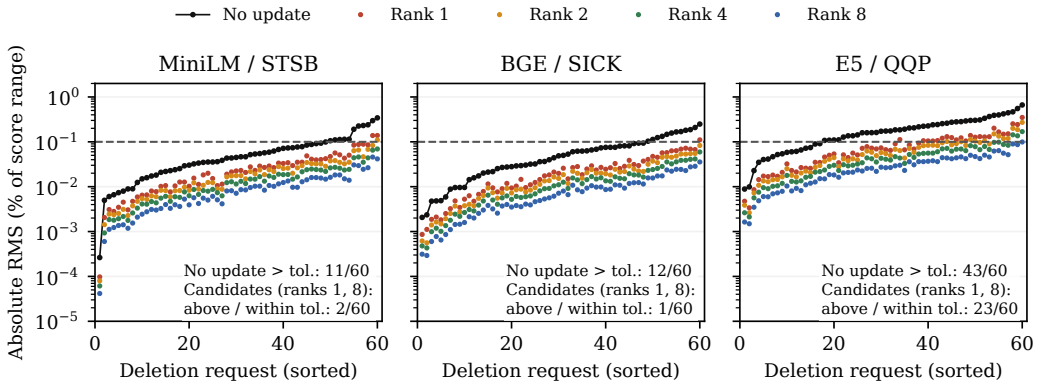


Figure 2: Every single deletion at $d = 32$, $\lambda = 1$, ordered by its no-update effect. Points give normalized RMS disagreement with retraining for no update and retained-objective candidates. The $0.1\%$ line marks the fixed numerical tolerance used for this comparison. All settings and tolerance curves appear in Appendix N.7.

then applies Eckart–Young. On one E5/QQP single deletion with $d = 16$, $\lambda = 1$, interval arithmetic verifies

$$\inf_{\mathrm{rg}(B)\leqslant 1} e_T(B) > 0.001088 > 0.001 > e_T(B_8), \qquad e_T(B_8) < 0.000760,$$

where $e_T$ is absolute RMS discrepancy and $B_8$ is an explicit rank-eight candidate. The no-update effect is $0.0103202$. Thus rank one cannot meet this tolerance, regardless of optimization. One of the 60 surveyed requests satisfies these verified inequalities at the fixed threshold. The guarantee concerns these 2000 evaluation pairs and the stored finite head instance, not future-pair error. The appendix gives the full search, proof, interval bounds and numerical-accuracy diagnostics.

**Task quality.** For the score comparisons, task RMSE is $0.858$ on STSB and $0.570$ on SICK, compared with $0.860$ and $0.569$ for the affine cosine baselines. Appendix N reports task performance for every fixed-regularization setting, together with the complete approximation results and optimization diagnostics.

**A nonquadratic training objective.** We also refit a ridge-regularized softmax head $W \in \mathbb{R}^{10\times 65}$ on fixed Fashion-MNIST features (Xiao et al., 2017). The removed image's cross-entropy term has a rank-one gradient. For each of 18 combinations of sample size and regularization, we average the minimum relative squared Frobenius error over 32 deletion requests. These means range from approximately $2.4\%$ to $7.0\%$ at rank one and are at most $0.41\%$ at rank four. Here the error measures agreement between parameter matrices. Appendix N.9 gives the protocol and results, with a separate numerical-accuracy study on 16 deletions.

## 9 Conclusion

We determine asymptotically the minimum rank needed to approximate full retraining after a single deletion. Under Assumption 1 and fixed $\lambda > 0$, every fixed relative tolerance below $\chi$ requires rank proportional to the matrix dimension whenever the deletion correction is nonzero. This requirement applies even to an oracle that knows the exact retrained model. The proof establishes the singular-value tail of the reshaped inverse-Hessian correction and yields the exact error uniformly over positive ranks.

Under Assumption 1, with fixed $\lambda > 0$ and $e_{i_0} = O_{\mathbb{P}}(1)$, the squared deletion effect is $O_{\mathbb{P}}(n^{-1})$. No update is needed when its RMS size is below tolerance. Appendix L gives the scaling, including zero residuals, and Appendix L specifies a noisy observation model. On real text pairs, one verified instance excludes every rank-one correction at a stated absolute tolerance, while an explicit rank-eight update meets it.

## AI USE STATEMENT

LLMs were exclusively used to polish the English writing of the manuscript and to refactor author-written code for readability. All concepts, methodological contributions, experimental logic, and empirical results are entirely original and produced solely by the authors.

## REPRODUCIBILITY STATEMENT

The assumptions and proofs are given in Appendices A–J and M. Appendices K and N specify the experimental protocols and report their numerical accuracy. Reproduction code, configurations, results and instructions are in `code_lora.zip` and its README files.

## REFERENCES


BAAI. bge-small-en-v1.5 model card. https://huggingface.co/BAAI/bge-small-en-v1.5.

Afonso S. Bandeira, March T. Boedihardjo, and Ramon van Handel. Matrix concentration inequalities and free probability. *Inventiones mathematicae*, 234:419–487, 2023. https://doi.org/10.1007/s00222-023-01204-6.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy*, 2021.

Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *IEEE Symposium on Security and Privacy*, pages 463–480, 2015. https://doi.org/10.1109/SP.2015.35.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *USENIX Security*, pages 2633–2650, 2021.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and parameter-efficient knowledge unlearning for LLMs. In *ICLR*, 2025.

Benoît Collins, Jianfeng Yao, and Wangjun Yuan. On spectral distribution of sample covariance matrices from large dimensional and large $k$-fold tensor products. *Electronic Journal of Probability*, 27:1–18, 2022. https://doi.org/10.1214/22-EJP825.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. https://doi.org/10.1007/BF02288367.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making AI Forget You: Data Deletion in Machine Learning. In *NeurIPS*, volume 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *CVPR*, pages 9304–9312, 2020.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *ICML*, volume 119 of *PMLR*, pages 3832–3842, 2020.

Karl Gustafson. The angle of an operator and positive operator products. *Bulletin of the American Mathematical Society*, 74(3):488–492, 1968.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Intfloat. e5-small-v2 model card. https://huggingface.co/intfloat/e5-small-v2.

Fredrik Johansson. Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66(8):1281–1292, 2017. https://arxiv.org/abs/1611.02831.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *ICML*, volume 70 of *PMLR*, pages 1885–1894, 2017.

Anna Lytova. CLT for linear eigenvalue statistics for a tensor product version of sample covariance matrices, 2017. arXiv:1602.08613v2.

Vladimir A. Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. https://doi.org/10.1070/SM1967v001n04ABEH001994.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*, 2014. Dataset: https://zenodo.org/records/2787612.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *ICML*, volume 37 of *PMLR*, pages 2408–2417, 2015.

L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. https://doi.org/10.1093/qmath/11.1.50.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning. In *ALT*, volume 132 of *PMLR*, pages 931–962, 2021.

Tushar Prakash, Brijraj Singh, Niranjan Pedanekar, and Narayan Chaturvedi. Obliviate: Efficient unlearning in recommender systems, 2026. arXiv:2607.22665. Accepted at ICML 2026.

Michael Rushka and Diego Klabjan. Rank-accuracy trade-off for LoRA: A gradient-flow analysis. arXiv:2602.10212, 2026.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *NeurIPS*, 2021.

Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Damian Rymarczyk. SEMU: Singular value decomposition for efficient machine unlearning. In *ICML*, volume 267 of *PMLR*, pages 53843–53866, 2025.

Sentence Transformers. all-MiniLM-L6-v2 model card. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.

Sentence Transformers. STSB dataset release. https://huggingface.co/datasets/sentence-transformers/stsb.

Joel A. Tropp. An introduction to matrix concentration inequalities, 2015. arXiv:1501.01571.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. https://doi.org/10.1017/9781108231596.

Cédric Villani. *Optimal Transport: Old and New*. Volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. https://doi.org/10.1007/978-3-540-71050-9.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *ACML*, volume 25 of *PMLR*, pages 475–490, 2012.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the EMNLP Workshop BlackboxNLP*, 2018. QQP release: https://huggingface.co/datasets/nyu-mll/glue.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. https://doi.org/10.2307/2276774.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017.

Pavel Yaskov. A short proof of the Marchenko–Pastur theorem. *C. R. Math.*, 354(3):319–322, 2016. https://doi.org/10.1016/j.crma.2015.12.008.

Jiatong Yu, Yinghui He, Anirudh Goyal, and Sanjeev Arora. On the impossibility of retrain equivalence in machine unlearning. arXiv:2510.16629, 2025.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *ICLR*, 2024.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of NAACL-HLT*, 2019. Dataset: https://huggingface.co/datasets/google-research-datasets/paws.

Haolin Zou, Arnab Auddy, Yongchan Kwon, Kamiar Rahnama Rad, and Arian Maleki. Certified machine unlearning under high dimensional regime. *Journal of Machine Learning Research*, 26 (308):1–58, 2025.

## APPENDIX CONTENTS


**A Contents and Notation** **15**

**B Exact Deletion, Leverage, and Normalization** **15**
B.1 One deleted observation . . . 15
B.2 Several observations . . . 16
B.3 Changing the loss normalization changes regularization . . . 16
B.4 Inexact original training . . . 17

**C Prediction Error, Training Objective, and Error Bounds** **17**
C.1 Prediction error and feature second moments . . . 17
C.2 Rank approximation under two error measures . . . 17
C.3 Bounds from an approximate solution . . . 18
C.4 Factorized objectives . . . 19

**D Separable Hessians and a Condition-Number Bound** **19**
D.1 Separable and nearly separable operators . . . 19
D.2 Universal condition-number upper bound . . . 19
D.3 Attaining the bound while keeping the eigenvalues fixed . . . 20

**E Scalar Limits for the Paired-Feature Model** **21**
E.1 The Marchenko–Pastur law and its scaling . . . 21
E.2 Independent product-sphere quadratic forms . . . 22
E.3 Algebra of the scalar factor . . . 22
E.4 A fourth-moment bound for the resolvent remainder . . . 23
E.5 Limits of the scalar rank law . . . 23

**F An Auxiliary Matrix for the Error Lower Bound** **24**
F.1 Conditional Gaussian representation . . . 24
F.2 Weighted covariance concentration . . . 24
F.3 Norms and sums of squared singular values . . . 25
F.4 First-row and first-column terms in the resolvent expansion . . . 25
F.5 Frobenius inner product of the correction and the auxiliary matrix . . . 25

**G Exact Law for Gaussian Feature Pairs at Fixed Regularization** **26**
G.1 Statement and scalar quantities . . . 26
G.2 Exact block elimination and a finite-dimensional bound . . . 27
G.3 Operator-norm concentration of partial traces . . . 28
G.4 A conditionally Gaussian matrix with controlled entry covariance . . . 29
G.5 A spectral theorem for Gaussian matrices . . . 31
G.6 Transferring the singular-value limit to the exact correction . . . 31
G.7 Consequences and assumptions . . . 32

**H Proofs of the Rank Bounds for Gaussian Feature Pairs** **33**
H.1 Sublinear ranks . . . 33
H.2 Optimized rank-fraction lower bound . . . 34
H.3 Upper bounds . . . 34
H.4 A positive rank fraction is necessary below the error lower bound . . . 34

**I Joint High-Regularization Limit** **35**
I.1 Finite-regularization correction to the scale . . . 35

**J Fixed Regularization under Orthogonal Invariance** **36**
J.1 Norms of the projection and the remainder . . . 36
J.2 Why the residual direction is uniform . . . 36
J.3 Reshaping and removal of cross blocks . . . 36
J.4 Gaussian singular values and approximation error . . . 37
J.5 Specialization to a shifted Marchenko–Pastur spectrum . . . 37
J.6 Rank quantiles and asymptotic regimes . . . 37

**K Experiments** **37**
K.1 Protocol . . . 38
K.2 Gaussian rank errors and spectra . . . 38
K.3 Direct retraining, prediction error, and numerical bounds . . . 39
K.4 Rank selection at finite dimension . . . 40
K.5 Curvature and the gradient direction . . . 41
K.6 Other feature distributions and the prediction metric . . . 42
K.7 Larger dimensions and absolute correction size . . . 42
K.8 Numerical verification of the algebra . . . 42

**L Absolute Error and Interpretation** **44**
L.1 Choosing a tolerance from an observable prediction requirement . . . 44

**M Finite-Dimensional Rank Selection by Independent Calibration** **45**
M.1 A simultaneous upper bound for all ranks . . . 46
M.2 Absolute requirements without the new correction norm . . . 46
M.3 Independent numerical validation . . . 47

**N Experiments on Frozen Representations** **47**
N.1 Models, data and preprocessing . . . 48
N.2 Heads, deletion requests and candidate updates . . . 48
N.3 Metrics, aggregation and complete results . . . 49
N.4 Task performance across datasets and regularization levels . . . 49
N.5 Numerical accuracy and optimizer convergence . . . 50
N.6 Minimum Frobenius errors on the observed features . . . 50
N.7 Absolute prediction errors and binary decisions . . . 51
N.8 A lower bound for every rank-one correction on real pairs . . . 53
N.9 A softmax head on fixed image features . . . 54

# A Contents and Notation

The proofs address three questions: what correction does retraining produce, how well can a matrix of prescribed rank approximate it, and how does this error behave as the dimension grows? Sections B–D give the deletion identities and deterministic approximation bounds. Sections E–F establish the scalar limits and properties of an auxiliary matrix. Section G proves the exact rank law for Gaussian feature pairs at fixed regularization. Section H gives complementary bounds from the auxiliary matrix. Sections I–J treat two distinct extensions: growing regularization and orthogonally invariant Hessian eigenvectors. Sections K–M describe the synthetic experiments, prediction tolerances and calibrated rank choices. Section N reports the experiments on frozen pretrained representations.

| Result | Setting | Proof |
|---|---|---|
| Exact deletion and metric identities | Arbitrary paired-feature data, fixed normalization | Sections B–C |
| Separable Hessians and an attained condition-number bound | Arbitrary positive-definite matrix-space operator | Section D |
| Exact law for Gaussian feature pairs at fixed regularization | Original independent Gaussian pairs, fixed positive ridge, uniform over ranks $1, \dots, d$ | Section G |
| Error bounds for sublinear and proportional ranks | Independent Gaussian paired features, fixed $\lambda > 0$ | Sections F, H |
| Exact limit under high regularization | Same feature distribution, joint $\lambda_d \to \infty$ | Section I |
| Exact law at fixed regularization | Independent Haar-distributed Hessian eigenvectors | Section J |
| Experiments | Protocols and numerical results | Section K |
| Prediction tolerances and calibrated ranks | Specified response model and independent calibration datasets | Sections L–M |

The main probabilistic proof separates the component in the gradient direction, the entries in the first row and column, and the remaining square block. Equation (G.13) gives this exact decomposition. Lemma S4 bounds the squared norm of the first-row and first-column terms. The singular values of the remaining block are then studied using the cited Gaussian spectral theorem.

Throughout, $\langle A, B\rangle := \operatorname{tr}(A^\top B)$ and $\|A\|_{\mathrm{F}}^2 := \langle A, A\rangle$. The singular values are ordered as $s_1(A) \geqslant \cdots \geqslant s_d(A) \geqslant 0$, so $\|A\|_{\mathrm{op}} = s_1(A)$. A norm without a subscript on a vector is Euclidean. The map $\mathrm{vec}$ stacks matrix columns, and $\mathrm{mat}$ is its inverse. A Hessian $\mathcal{H}$ acts on matrix space and is represented by the matrix $H$ after vectorization. The notation $[\![a, b]\!]$ denotes an integer interval, and $[t]_+ := \max\{t, 0\}$ denotes the positive part. We use $\xrightarrow{\mathbb{P}}$ for convergence in probability, $O_{\mathbb{P}}(a_d)$ for boundedness in probability after division by $a_d$, and $o_{\mathbb{P}}(a_d)$ for convergence to zero after that division.

# B Exact Deletion, Leverage, and Normalization

## B.1 One deleted observation

We first compute the exact change in the fitted matrix when one observation is removed. Each observation consists of $x_i, z_i \in \mathbb{R}^d$ and a response $y_i \in \mathbb{R}$. Fix $\lambda > 0$. For every $T \subseteq [\![1, n]\!]$, define the objective

$$L_T(W) := \frac{1}{2n} \sum_{i \in T} (\langle W, X_i\rangle - y_i)^2 + \frac{\lambda}{2} \|W\|_{\mathrm{F}}^2, \qquad X_i := x_i z_i^\top. \tag{B.1}$$

The denominator remains $n$ for every subset. Let $D := [\![1, n]\!]$ be the original index set, $i_0 \in D$ the deleted index, and $R := D \setminus \{i_0\}$ the retained set. Write $W_D$ and $W_R$ for the corresponding minimizers, and define the retraining correction by $\Delta := W_R - W_D$. With $\phi_i := \mathrm{vec}(X_i)$, the

retained Hessian is $H_R = \lambda I + n^{-1} \sum_{i \in R} \phi_i \phi_i^\top$. The inequality $H_R \succeq \lambda I$ ensures that the minimizer is unique.

Removing observation $i_0$ changes the gradient at $W_D$ by exactly its contribution to the original objective. Define its fitted residual by $e_{i_0} := \langle W_D, X_{i_0} \rangle - y_{i_0}$ and its gradient contribution by $G_{i_0} := e_{i_0} X_{i_0}/n$. Since $\nabla L_D(W_D) = 0$,

$$\nabla L_R(W_D) = -G_{i_0}. \tag{B.2}$$

Since the objective is quadratic,

$$0 = \nabla L_R(W_R) = \nabla L_R(W_D) + \mathcal{H}_R[W_R - W_D]. \tag{B.3}$$

Therefore $\mathcal{H}_R[\Delta] = G_{i_0}$. The quadratic objective makes this identity exact. No small-deletion or local Taylor approximation is used.

The correction can also be expressed using the original Hessian $H_D = H_R + \phi_{i_0} \phi_{i_0}^\top / n$. The scalar $h_{i_0} := \phi_{i_0}^\top H_D^{-1} \phi_{i_0}/n$ records how its inverse acts in the deleted feature direction. The rank-one inverse identity gives

$$H_R^{-1} \phi_{i_0} = \frac{H_D^{-1} \phi_{i_0}}{1 - h_{i_0}}. \tag{B.4}$$

To show that the denominator is positive, set $t := \phi_{i_0}^\top H_R^{-1} \phi_{i_0}/n \geqslant 0$. The same inverse identity yields $h_{i_0} = t/(1+t) < 1$. Hence

$$\mathrm{vec}(\Delta) = \frac{e_{i_0}}{n(1 - h_{i_0})} H_D^{-1} \phi_{i_0}. \tag{B.5}$$

### B.2 Several observations

For a deleted set $F \subset [\![1, n]\!]$, define $R := [\![1, n]\!] \setminus F$ and $\Delta_F := W_R - W_D$. Collect the deleted feature vectors and their original fitted residuals in $\Phi_F := [\phi_i]_{i \in F}$ and $e_F := (\phi_i^\top w_D - y_i)_{i \in F}$, where $w_D := \mathrm{vec}(W_D)$. The same stationarity argument gives

$$\mathrm{vec}(\Delta_F) = \frac{1}{n} H_R^{-1} \Phi_F e_F. \tag{B.6}$$

Because $H_R = H_D - \Phi_F \Phi_F^\top / n \succ 0$, the $|F| \times |F|$ matrix $I - \Phi_F^\top H_D^{-1} \Phi_F / n$ is positive definite. Multiplying the Woodbury identity by $\Phi_F$ gives

$$\mathrm{vec}(\Delta_F) = \frac{1}{n} H_D^{-1} \Phi_F \left( I - \frac{1}{n} \Phi_F^\top H_D^{-1} \Phi_F \right)^{-1} e_F. \tag{B.7}$$

The matrix gradient $n^{-1} \sum_{i \in F} e_i x_i z_i^\top$ has rank at most $|F|$. The correction applies $\mathcal{H}_R^{-1}$ to that gradient, so its rank also depends on this operator. The displayed identities provide no rank-preservation property for a general Hessian.

### B.3 Changing the loss normalization changes regularization

The fixed denominator matters because changing it also changes the relative weight of regularization. Let $\ell_i(W) := (\langle W, X_i \rangle - y_i)^2$ and suppose $|R| > 0$. If retraining uses $(2|R|)^{-1} \sum_{i \in R} \ell_i + \lambda \|W\|^2 / 2$ with the same numerical $\lambda$, multiplying the objective by $|R|/n$ leaves its minimizer unchanged and replaces the ridge term by $(|R|/n)\lambda \|W\|^2 / 2$. In this subsection, $W_R$ denotes the minimizer of this renormalized retained objective and $\Delta := W_R - W_D$, while $W_D$ remains the original minimizer. Consequently, for one deletion,

$$\left( S_R + \frac{n-1}{n} \lambda I \right) \mathrm{vec}(\Delta) = \frac{1}{n} \left( e_{i_0} \phi_{i_0} + \lambda w_D \right), \quad S_R := \frac{1}{n} \sum_{i \in R} \phi_i \phi_i^\top. \tag{B.8}$$

For $k$ deletions the additional term is $k \lambda w_D / n$. Its matrix rank need not be small. The main paper fixes normalization precisely to avoid counting this regularization change as a rank effect caused by the forgotten gradient.

### B.4 Inexact original training

An approximate original fit introduces another error, even before imposing a rank constraint. Let $\bar{W}_D$ be that fit and define its stationarity residual by $g_D := \nabla L_D(\bar{W}_D)$. Evaluate the deleted gradient at the same fit, $G_F := n^{-1}\sum_{i\in F}(\langle \bar{W}_D, X_i\rangle - y_i)X_i$. Then the exact retained minimizer satisfies

$$\mathcal{H}_R[W_R - \bar{W}_D] = G_F - g_D. \tag{B.9}$$

If the original stationarity error is ignored, the resulting correction differs by $\mathcal{H}_R^{-1}[g_D]$, with Frobenius norm at most $\|g_D\|_{\mathrm{F}}/\lambda$. This error is separate from both the low-rank constraint and the correction solve error. In small dimensions, our experiments solve the original training problem directly and compare the deletion formula with direct retraining. In larger dimensions, we compute the correction direction directly from the features alone.

To account also for an approximate correction $\widetilde{\Delta}$, define its linear-system residual by $R_{\mathrm{lin}} := G_F - \mathcal{H}_R[\widetilde{\Delta}]$. Let $\eta_{\mathrm{total}}$ denote the sum of the resulting two norm bounds. Strong convexity and the triangle inequality give

$$\left\|\widetilde{\Delta} - (W_R - \bar{W}_D)\right\|_{\mathrm{F}} \leqslant \eta_{\mathrm{total}} := \frac{\|R_{\mathrm{lin}}\|_{\mathrm{F}} + \|g_D\|_{\mathrm{F}}}{\lambda}. \tag{B.10}$$

The approximation bounds in Section C therefore remain valid with $\eta_{\mathrm{total}}$ whenever a valid bound on the original gradient norm is available. This addition is deterministic and requires no independence between the two residuals. When the deletion effect is small, even a small error caused by incomplete original training can exceed it. In that case the rank of $W_R - \bar{W}_D$ reflects both incomplete training and data removal.

## C Prediction Error, Training Objective, and Error Bounds

### C.1 Prediction error and feature second moments

We next relate matrix approximation to prediction. Fix the training sample and a proposed update $B$. The prediction difference from exact retraining is $x^\top Az$, where $A := B - \Delta$. Assume that the test vectors $x, z$ are independent of each other and of the training sample, with finite second moments. Define their second-moment matrices by $\Sigma_x := \mathbb{E}[xx^\top]$ and $\Sigma_z := \mathbb{E}[zz^\top]$. Conditioning on $z$ gives

$$\mathbb{E}(x^\top Az)^2 = \mathbb{E}_z[z^\top A^\top(\mathbb{E}xx^\top)Az] \tag{C.1}$$

$$= \mathrm{tr}(A^\top\Sigma_x A\Sigma_z) = \left\|\Sigma_x^{1/2}A\Sigma_z^{1/2}\right\|_{\mathrm{F}}^2. \tag{C.2}$$

Thus, when $\Sigma_x = \Sigma_z = I$, the expected squared prediction difference is exactly $\|B - \Delta\|_{\mathrm{F}}^2$. When both second-moment matrices are positive definite, the map $B \mapsto \Sigma_x^{1/2}B\Sigma_z^{1/2}$ is invertible and preserves rank. A prediction-optimal update of rank at most $r$ is obtained by keeping the $r$ largest singular values of $\Sigma_x^{1/2}\Delta\Sigma_z^{1/2}$ and transforming back. For $\Delta \neq 0$, the minimum relative error is the squared singular-value tail of $\Sigma_x^{1/2}\Delta\Sigma_z^{1/2}$ divided by its squared Frobenius norm. If a second-moment matrix is singular, prediction only measures the corresponding observable subspace. The inverse transformation then requires a separate treatment.

### C.2 Rank approximation under two error measures

To measure the best approximation of a matrix $A$ at a prescribed rank, define $d_r(A) := \inf_{\mathrm{rg}(B)\leqslant r}\|A - B\|_{\mathrm{F}}$, for $r \in [\![0, d]\!]$. The set of matrices of rank at most $r$ is closed, and the infimum is attained. The Eckart–Young theorem gives $d_r(A)^2 = \sum_{j>r} s_j(A)^2$ (Eckart and Young, 1936). This distance also changes continuously with its target: for any two matrices,

$$|d_r(A) - d_r(A')| \leqslant \|A - A'\|_{\mathrm{F}}. \tag{C.3}$$

Indeed, evaluate one infimum at an optimizer of the other and apply the triangle inequality in both directions.

Prediction error and retained training loss use different quadratic forms. Indeed, expansion around $W_R$ gives

$$L_R(W_D + B) - L_R(W_R) = \tfrac{1}{2} \langle B - \Delta, \mathcal{H}_R[B - \Delta] \rangle . \tag{C.4}$$

For $\Delta \neq 0$, define the relative retained-objective error by

$$\mathcal{E}_r^H := \frac{\inf_{\mathrm{rg}(B) \leqslant r} \langle B - \Delta, \mathcal{H}_R[B - \Delta] \rangle}{\langle \Delta, \mathcal{H}_R[\Delta] \rangle}. \tag{C.5}$$

The numerator is twice the smallest retained-objective excess, and the denominator is twice the excess without an update. If $\mu I \preceq H_R \preceq LI$, then for every $B$,

$$\mu \left\| B - \Delta \right\|_{\mathrm{F}}^2 \leqslant \langle B - \Delta, \mathcal{H}_R[B - \Delta] \rangle \leqslant L \left\| B - \Delta \right\|_{\mathrm{F}}^2 . \tag{C.6}$$

Taking the infimum over $\mathrm{rg}(B) \leqslant r$ proves the lower bound $\mu d_r(\Delta)^2$ for the squared Hessian error. Evaluating that error at a Frobenius-optimal truncation proves the upper bound $L d_r(\Delta)^2$. Together with $\mathcal{E}_r := d_r(\Delta)^2 / \left\| \Delta \right\|_{\mathrm{F}}^2$, the error $\mathcal{E}_r^H$ is therefore bounded by the same quadratic-form comparison. Dividing the preceding bounds by $\langle \Delta, \mathcal{H}_R[\Delta] \rangle \in [\mu \left\| \Delta \right\|_{\mathrm{F}}^2 , L \left\| \Delta \right\|_{\mathrm{F}}^2]$ proves

$$\kappa^{-1} \mathcal{E}_r \leqslant \mathcal{E}_r^H \leqslant \kappa \mathcal{E}_r, \qquad \kappa := L/\mu. \tag{C.7}$$

These bounds compare optimal values, even when their optimizers differ. For a separable positive Hessian $\mathcal{H}[A] = PAQ$ with $P, Q \succ 0$, the invertible change of coordinates above reduces weighted rank approximation to a singular value decomposition of $P^{1/2} \Delta Q^{1/2}$. A general Hessian has no such left–right factorization.

### C.3 Bounds from an approximate solution

A numerical solve provides an approximate correction and a computable residual. The residual bounds how much the singular-value approximation errors can differ from their exact values. Let $\widetilde{\Delta}$ approximate the solution of $\mathcal{H}_R[\Delta] = G$, and define $R_{\mathrm{lin}} := G - \mathcal{H}_R[\widetilde{\Delta}]$. Then

$$\widetilde{\Delta} - \Delta = -\mathcal{H}_R^{-1}[R_{\mathrm{lin}}], \qquad \left\| \widetilde{\Delta} - \Delta \right\|_{\mathrm{F}} \leqslant \eta := \left\| R_{\mathrm{lin}} \right\|_{\mathrm{F}} / \lambda. \tag{C.8}$$

Both the size of the approximate correction and its rank-approximation error are available from the solve. Denote them by $N := \left\| \widetilde{\Delta} \right\|_{\mathrm{F}}$ and $t_r := d_r(\widetilde{\Delta})$. Equation (C.3) and the reverse triangle inequality give

$$d_r(\Delta) \geqslant (t_r - \eta)_+, \qquad N - \eta \leqslant \left\| \Delta \right\|_{\mathrm{F}} \leqslant N + \eta. \tag{C.9}$$

For $\Delta \neq 0$, these inequalities prove

$$\mathcal{E}_r \geqslant \frac{(t_r - \eta)_+^2}{(N + \eta)^2}. \tag{C.10}$$

For the explicit truncated SVD $\widetilde{\Delta}_r$,

$$\left\| \widetilde{\Delta}_r - \Delta \right\|_{\mathrm{F}} \leqslant t_r + \eta. \tag{C.11}$$

If $N > \eta$, division by the lower norm bound proves the upper error bound in the main paper. If $\mu I \preceq H_R \preceq LI$ is known, division of the lower relative bound by $\kappa = L/\mu$ gives a lower bound for $\mathcal{E}_r^H$. Multiplication of the candidate's upper relative squared Frobenius-error bound by $\kappa$ bounds its squared Hessian error divided by $\langle \Delta, \mathcal{H}_R[\Delta] \rangle$. An estimated largest eigenvalue can be used here only if its estimation error is bounded. The deterministic choice $L = \lambda + n^{-1} \sum_i \left\| X_i \right\|_{\mathrm{F}}^2$ is always valid, although it may give a loose bound.

For zero tolerance, this refinement procedure does not guarantee termination in finite precision.

The Hessian application is matrix-free and requires retained features but not retained labels:

$$\mathcal{H}_R[A] = \lambda A + \frac{1}{n} \sum_{i \in R} (x_i^\top A z_i) x_i z_i^\top . \tag{C.12}$$

It can also be obtained by subtracting the forgotten rank-one operator from a stored full-data Hessian operator. Dense $d^2 \times d^2$ storage is unnecessary, although storing and processing the features still has a cost.

**Algorithm 1** Rank selection for a target prediction tolerance $\varepsilon$

**Require:** Exact original fit $W_D$, forgotten observation, $\lambda > 0$, Hessian operator $\mathcal{H}_R$, tolerance $\varepsilon \geqslant 0$.

1: Form $G := (x_{i_0}^\top W_D z_{i_0} - y_{i_0}) x_{i_0} z_{i_0}^\top / n$.
2: If $G = 0$, return the zero correction. The model already equals the exact retained fit, and the relative error is undefined.
3: Approximately solve $\mathcal{H}_R[\widetilde{\Delta}] = G$.
4: Set $N := \left\|\widetilde{\Delta}\right\|_{\mathrm{F}}$ and compute the error bound $\eta$ and the truncation errors $t_r$.
5: If $N > \eta$ and some $r \in [\![0, d]\!]$ satisfies $\overline{E}_r \leqslant \varepsilon$, return the smallest such $r$, together with $\widetilde{\Delta}_r$ and both error bounds.
6: Otherwise, repeat the solve with greater accuracy and recompute the bounds.

### C.4 Factorized objectives

To optimize the retained objective under a rank constraint, write $B = A_1 A_2^\top$ with $A_1, A_2 \in \mathbb{R}^{d \times r}$. The terms that depend on these factors define

$$Q(A_1, A_2) := -\left\langle G, A_1 A_2^\top \right\rangle + \tfrac{1}{2} \left\langle A_1 A_2^\top, \mathcal{H}_R[A_1 A_2^\top] \right\rangle. \tag{C.13}$$

Self-adjointness of $\mathcal{H}_R$ and the chain rule give

$$\nabla_{A_1} Q = (\mathcal{H}_R[A_1 A_2^\top] - G) A_2, \qquad \nabla_{A_2} Q = (\mathcal{H}_R[A_1 A_2^\top] - G)^\top A_1. \tag{C.14}$$

Since $G = \mathcal{H}_R[\Delta]$, $Q$ differs from one half of the squared Hessian norm of $A_1 A_2^\top - \Delta$ by a constant independent of $A_1, A_2$. Its minimum determines $\mathcal{E}_r^H$. A particular numerical solution need not attain this minimum. To compare with the oracle prediction error, the relevant objective is instead $\frac{1}{2} \left\|A_1 A_2^\top - \Delta\right\|_{\mathrm{F}}^2$.

For a computed correction of rank at most $r$, the retained-objective excess divided by the excess without an update bounds $\mathcal{E}_r^H$ from above. A better retained loss can coexist with a worse prediction discrepancy, since $\mathcal{E}_r$ uses a different metric.

## D Separable Hessians and a Condition-Number Bound

### D.1 Separable and nearly separable operators

We first identify a case in which the inverse Hessian preserves the rank of the deleted gradient. If $\mathcal{H}[A] = PAQ$ with $P, Q \succ 0$, then $\mathcal{H}^{-1}[xz^\top] = P^{-1} x z^\top Q^{-1}$. For nonzero $x, z$, this matrix has rank one. A sum of $k$ such images has rank at most $k$.

If the Hessian is close to an operator with this property, the corresponding correction provides a low-rank approximation. Let $\mathcal{H} = \mathcal{H}_0 + \mathcal{A}$, where $\mathcal{H}$ and $\mathcal{H}_0$ are invertible and $\mathcal{H}_0^{-1}$ preserves rank. Define $\Delta := \mathcal{H}^{-1}[G]$ and $\Delta_0 := \mathcal{H}_0^{-1}[G]$. Applying $\mathcal{H}_0^{-1}$ to $G = \mathcal{H}[\Delta]$ gives

$$\Delta_0 = \Delta + \mathcal{H}_0^{-1} \mathcal{A}[\Delta]. \tag{D.1}$$

For rank-one $G$, $\Delta_0$ is an admissible rank-one approximation, so

$$\mathcal{E}_1 \leqslant \frac{\|\Delta_0 - \Delta\|_{\mathrm{F}}^2}{\|\Delta\|_{\mathrm{F}}^2} \leqslant \left\|\mathcal{H}_0^{-1} \mathcal{A}\right\|_{\mathrm{F} \to \mathrm{F}}^2. \tag{D.2}$$

Here $\|\cdot\|_{\mathrm{F} \to \mathrm{F}}$ denotes the operator norm induced by the Frobenius norm. For an added ridge term $\mathcal{A} = \lambda I$ this yields $\mathcal{E}_1 \leqslant \lambda^2 \left\|\mathcal{H}_0^{-1}\right\|^2$ provided $\mathcal{H}_0$ is fixed and invertible. This is an approximation-error bound, not an equality of matrix ranks.

### D.2 Universal condition-number upper bound

Even without rank preservation, a well-conditioned Hessian limits how far its inverse can move a vector from its original direction. The required angle estimate is classical (Gustafson, 1968). We give its elementary derivation, then construct a matrix example that attains the resulting rank-approximation bound.

**Proposition S1** (Conditioning bounds but does not identify rank cost)**.** *For any positive-definite $H$, define its condition number by $\kappa := \lambda_{\max}(H)/\lambda_{\min}(H)$. Let $U$ have rank one and $\|U\|_{\mathrm{F}} = 1$, and let $M = \mathrm{mat}(H^{-1}\,\mathrm{vec}(U))$. Then*

$$\mathcal{E}_1(M) \leqslant b_\kappa, \qquad b_\kappa := \left(\frac{\kappa-1}{\kappa+1}\right)^2. \tag{D.3}$$

*For every $\kappa > 1$ and all sufficiently large $d$, there are two operators with exactly the same eigenvalues and condition number $\kappa$ such that one has $\mathcal{E}_1 = 0$, while the other has*

$$\mathcal{E}_r = b_\kappa \frac{d-r}{d-1}, \qquad 1 \leqslant r \leqslant d. \tag{D.4}$$

Let $H \succ 0$ and let $U$ have rank one and $\|U\|_{\mathrm{F}} = 1$. Define $q := \mathrm{vec}(U)$ and $R := H^{-1}$, and suppose $0 < a \leqslant \lambda_j(R) \leqslant b$. The best scalar multiple of $U$ is an admissible rank-one approximation to $\mathrm{mat}(Rq)$, so

$$\mathcal{E}_1 \leqslant 1 - \frac{(q^\top Rq)^2}{q^\top R^2 q}. \tag{D.5}$$

To bound this ratio, assign each eigenvalue of $R$ the weight given by the corresponding squared coordinate of $q$. The weights sum to one. Let $t$ denote the resulting scalar random variable on $[a, b]$. This auxiliary distribution is determined by $H$ and $U$ and introduces no randomness into the data. Since $(t-a)(t-b) \leqslant 0$,

$$\mathbb{E}t^2 \leqslant (a+b)\mathbb{E}t - ab. \tag{D.6}$$

Writing $s := \mathbb{E}t$, elementary maximization over $s \in [a, b]$ gives

$$\frac{\mathbb{E}t^2}{s^2} \leqslant \frac{(a+b)^2}{4ab}. \tag{D.7}$$

Substitution in (D.5) proves

$$\mathcal{E}_1 \leqslant \frac{(b-a)^2}{(b+a)^2} = \left(\frac{\kappa-1}{\kappa+1}\right)^2, \qquad \kappa = b/a. \tag{D.8}$$

This proof is deterministic and dimension-free. In particular, no positive relative lower bound on the rank-one approximation error can persist as $\kappa \to 1$.

### D.3 Attaining the bound while keeping the eigenvalues fixed

The next construction shows both that the bound can be attained and that eigenvalues alone do not determine the rank error. Fix $\kappa > 1$ and $d \geqslant 2$, set $a := 1/\kappa$, $b := 1$, and choose

$$U = e_1 e_1^\top, \qquad V = \frac{1}{\sqrt{d-1}} \sum_{j=2}^{d} e_j e_j^\top. \tag{D.9}$$

These matrices are Frobenius-orthonormal and have orthogonal row and column supports. On their two-dimensional span, define the inverse Hessian

$$R_2 := \frac{1}{a+b} \begin{pmatrix} 2ab & \sqrt{ab}(b-a) \\ \sqrt{ab}(b-a) & a^2+b^2 \end{pmatrix}. \tag{D.10}$$

The trace is $a+b$ and determinant is $ab$, so its eigenvalues are $a, b$. Set $R = aI$ on the orthogonal complement. The full $H = R^{-1}$ has condition number $\kappa$. Moreover,

$$\mathrm{mat}(Rq) = \alpha U + \beta V, \quad \alpha = \frac{2ab}{a+b}, \quad \beta = \frac{\sqrt{ab}(b-a)}{a+b}, \quad \alpha^2 + \beta^2 = ab. \tag{D.11}$$

If $d - 1 \geqslant \beta^2/\alpha^2 = (\kappa-1)^2/(4\kappa)$, the largest singular value is $\alpha$ and each remaining singular value is $\beta/\sqrt{d-1}$. Thus

$$\mathcal{E}_r = \frac{\beta^2}{\alpha^2+\beta^2}\frac{d-r}{d-1} = \left(\frac{\kappa-1}{\kappa+1}\right)^2 \frac{d-r}{d-1}. \tag{D.12}$$

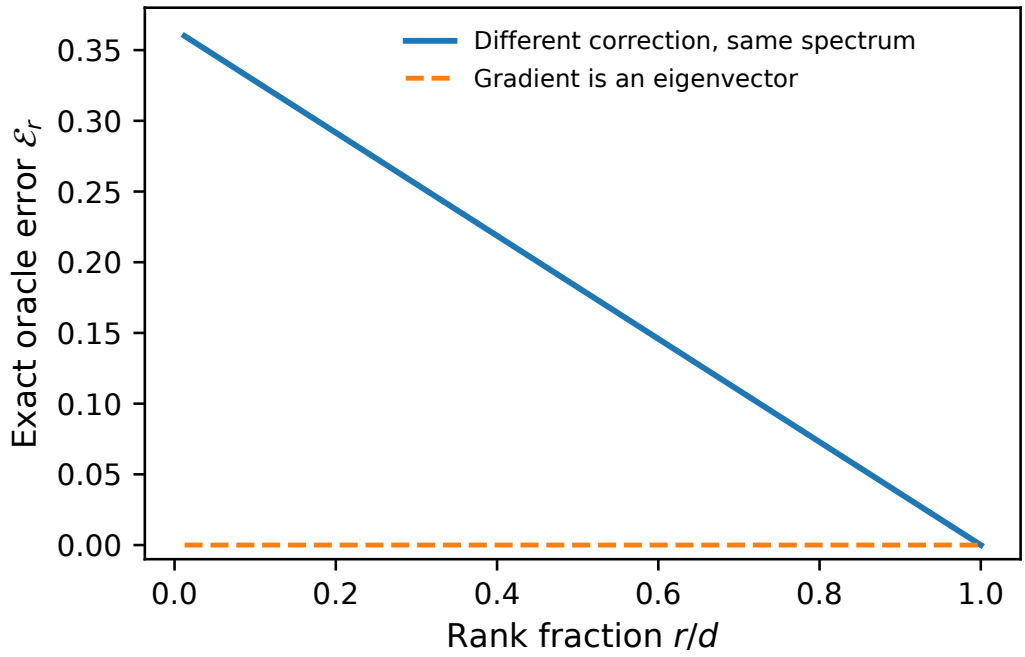


Figure 3: Two deterministic Hessians with identical eigenvalues. At $d = 80$ and $\kappa = 4$, rotating the Hessian eigenvectors changes rank cost from zero to $0.36(d - r)/(d - 1)$ without changing any Hessian eigenvalue.

An orthogonal conjugate $R'$ with $q$ as an eigenvector has the same spectrum but produces a multiple of $U$, so its rank-one error is zero. These operators can be Hessians of positive quadratic objectives. They are deterministic controls and are not asserted to be typical paired-Gaussian Hessians.

The contrast in Figure 3 is realizable by a positive quadratic objective: every positive operator is a quadratic Hessian, and an additional rank-one observation generates the specified inverse-Hessian direction whenever its residual is nonzero. It is not asserted that these deliberately constructed operators have the paired-Gaussian sampling distribution. Their purpose is to isolate which deterministic information about curvature can, and cannot, determine rank cost.

## E Scalar Limits for the Paired-Feature Model

We now study random feature pairs. All retained and forgotten feature vectors $x_i, z_i$ are independent with law $\mathcal{N}(0, I_d)$, the forgotten index is fixed independently of the sample, and $d^2/n \to \gamma \in ]0, +\infty[$. Unless stated otherwise, $\lambda > 0$ is fixed. The normalized forgotten directions are $u := x_{i_0}/\|x_{i_0}\|$ and $v := z_{i_0}/\|z_{i_0}\|$, which are well defined almost surely. Write $U := uv^\top$ and $q := \mathrm{vec}(U) = v \otimes u$. Separating these directions from the residual and feature lengths allows us to study the relative approximation error without specifying a response model.

### E.1 The Marchenko–Pastur law and its scaling

The retained covariance determines the correction through its resolvent. Define the ambient vector dimension $p := d^2$, the retained sample size $N := n - 1$, and

$$S := \frac{1}{n}\sum_{i=1}^{N}(z_i \otimes x_i)(z_i \otimes x_i)^\top. \tag{E.1}$$

The matrix $M := \mathrm{mat}[(\lambda I + S)^{-1} q]$ has the same relative rank-approximation errors as every nonzero deletion correction in this feature direction. We first determine the limits of its squared norm and its component along $U$.

The tensor-product covariance theorem of Lytova (2017) applies with tensor order two and normalized base vectors $x_i/\sqrt{d}, z_i/\sqrt{d}$. Its aspect parameter is $N/d^2$. Rescaling by $d^2/n$ yields the mean-one law with aspect $p/n \to \gamma$. Removing one observation changes $N/n$ by $1/n$ and does not affect the limit. The resulting measure is

$$\nu_\gamma := (1-\gamma^{-1})_+ \delta_0 + \frac{\sqrt{(b_\gamma - s)(s - a_\gamma)}}{2\pi\gamma s} \mathbf{1}_{\{a_\gamma \leqslant s \leqslant b_\gamma\}} \, \mathrm{d}s, \quad a_\gamma = (1-\sqrt{\gamma})^2, \; b_\gamma = (1+\sqrt{\gamma})^2. \tag{E.2}$$

The empirical eigenvalue distribution of $S$ converges weakly in probability to $\nu_\gamma$. Its first two resolvent moments will be used throughout the proof. Define $m := \int (\lambda + s)^{-1} \nu_\gamma(\mathrm{d}s)$ and $m_2 := \int (\lambda + s)^{-2} \nu_\gamma(\mathrm{d}s)$. Since both integrands are bounded and continuous on $[0, +\infty[$ at fixed $\lambda > 0$,

$$\frac{1}{p} \operatorname{tr}(\lambda I + S)^{-1} \xrightarrow{\mathbb{P}} m, \quad \frac{1}{p} \operatorname{tr}(\lambda I + S)^{-2} \xrightarrow{\mathbb{P}} m_2. \tag{E.3}$$

These trace limits concern eigenvalues and give no Haar distribution for the eigenvectors. Coordinates of a tensor feature are dependent, which is why we use a tensor-product covariance theorem here. General criteria based on quadratic forms are also discussed by Yaskov (2016). Spectral moment extensions appear in Collins et al. (2022).

### E.2 Independent product-sphere quadratic forms

The trace limits average over all directions. To use them for the independent forgotten direction $q$, we need its quadratic forms to concentrate around those averages.

**Lemma S2.** *Let $d \geqslant 2$, let $u, v$ be independent uniform vectors on the unit sphere of $\mathbb{R}^d$, set $q := v \otimes u$, and let $A$ be a symmetric $p \times p$ matrix independent of $u, v$. Then*

$$\mathbb{E}[q^\top A q \mid A] = \operatorname{tr}(A)/p, \qquad \operatorname{Var}(q^\top A q \mid A) \leqslant \frac{8 \operatorname{tr}(A^2)}{(d-1)p}. \tag{E.4}$$

*Proof.* Isotropy gives $\mathbb{E} q q^\top = I/p$, proving the mean. For $f(u, v) = q^\top A q$, the Euclidean derivative with respect to either sphere argument has norm at most $2 \|Aq\|$. Projection to the corresponding tangent space only decreases that norm. The product-sphere Poincaré inequality, with spectral gap $d - 1$ for each unit sphere, therefore gives

$$\operatorname{Var} f \leqslant \frac{1}{d-1} \mathbb{E}(\|\nabla_u f\|^2 + \|\nabla_v f\|^2) \leqslant \frac{8}{d-1} \mathbb{E} \|Aq\|^2 = \frac{8 \operatorname{tr} A^2}{(d-1)p}. \tag{E.5}$$

The Poincaré inequality follows from the spherical-harmonic decomposition: the first nonconstant eigenvalue of minus the sphere Laplacian is $d - 1$. On the product of the two independent spheres, the bound uses the sum of the squared tangential gradients. □

Apply the lemma conditionally on $S$ with $A = (\lambda I + S)^{-1}$ or $A = (\lambda I + S)^{-2}$. Their norms are at most $\lambda^{-1}$ and $\lambda^{-2}$, respectively, so the variances vanish. Combining with (E.3) proves

$$a_d = \langle M, U \rangle = q^\top (\lambda I + S)^{-1} q \xrightarrow{\mathbb{P}} m, \quad \|M\|_{\mathrm{F}}^2 = q^\top (\lambda I + S)^{-2} q \xrightarrow{\mathbb{P}} m_2. \tag{E.6}$$

The forgotten directions are independent of the retained covariance because the forgotten index was fixed before seeing the sample. Adaptive selection would require a separate argument.

### E.3 Algebra of the scalar factor

The two preceding limits show that the asymptotic fraction of squared norm outside the gradient direction is $1 - m^2/m_2$. We express it through the scalar used in the main paper, $\chi := \gamma(1 - \lambda m)^2$. The positive solution of

$$\frac{1}{m} = \lambda + \frac{1}{1 + \gamma m} \tag{E.7}$$

is the formula in the main paper. Rearranging gives

$$1 - \lambda m = \frac{m}{1+\gamma m}, \quad \chi = \frac{\gamma m^2}{(1+\gamma m)^2}. \tag{E.8}$$

Differentiating (E.7) and using $m_2 = -\partial_\lambda m$ yields

$$m_2 \left( \frac{1}{m^2} - \frac{\gamma}{(1+\gamma m)^2} \right) = 1, \qquad m_2 = \frac{m^2}{1-\chi}. \tag{E.9}$$

The variance of $(\lambda + s)^{-1}$ under $\nu_\gamma$ is positive for $\gamma > 0$, so $m_2 > m^2$ and $0 < \chi < 1$. The lower-bound proof will also use the inner product between $M$ and an auxiliary matrix. Define $\beta_* := m(1 - \lambda m)$, which determines the limit of that inner product. The identities needed below are

$$\gamma\beta_*^2 = m_2\chi(1-\chi), \quad m_2 - m^2 = m_2\chi, \quad m_2 - m^2 - \gamma\beta_*^2 = m_2\chi^2. \tag{E.10}$$

Another useful identity is

$$1 - (\lambda + 1)m = -\gamma m(1 - \lambda m) = -\gamma\beta_*. \tag{E.11}$$

### E.4 A FOURTH-MOMENT BOUND FOR THE RESOLVENT REMAINDER

When $\lambda$ grows with $d$, we will approximate the resolvent by its first two terms. Its remainder depends on $\left\|(S-I)^2 q\right\|$, so we need this quantity to be bounded in probability uniformly along the dimensional sequence. It suffices to prove

$$\sup_d \frac{1}{p} \mathbb{E} \operatorname{tr} S^4 < \infty \tag{E.12}$$

when $p/n$ stays in a compact subset of $]0, +\infty[$. This can be shown directly for our features. Let $A_i := \phi_i \phi_i^\top$. Expand $\operatorname{tr} S^4 = n^{-4} \sum_{i,j,k,\ell} \operatorname{tr}(A_i A_j A_k A_\ell)$. Gaussian radial moments give $\mathbb{E} \left\|\phi\right\|^{2s} = O(p^s)$ for fixed $s$. Also $\mathbb{E}\phi\phi^\top = I$ and $\mathbb{E}[\left\|\phi\right\|^2 \phi\phi^\top] = (d+2)^2 I$. Group the tuples $(i, j, k, \ell)$ by their number of distinct sample indices:

| Distinct samples | Number of tuples | Largest expectation order |
|---|---|---|
| 1 | $O(n)$ | $O(p^4)$ |
| 2 | $O(n^2)$ | $O(p^3)$ |
| 3 | $O(n^3)$ | $O(p^2)$ |
| 4 | $O(n^4)$ | $p$ |

For two distinct samples, one index may occur three times and the other once. This type $(3, 1)$ integrates to $\mathbb{E} \left\|\phi\right\|^6 = O(p^3)$. When both indices occur twice, adjacent occurrences give $\operatorname{tr}(\mathbb{E} A_i^2)^2 = O(p^3)$, whereas alternating occurrences give $\mathbb{E}(\phi_i^\top \phi_j)^4 = 9[d(d+2)]^2 = O(p^2)$. For three distinct samples, integrate the single-occurrence indices using isotropy to obtain $O(\mathbb{E} \left\|\phi\right\|^4) = O(p^2)$. Four distinct indices integrate to $p$. Multiplying each row's count by its expectation bound and dividing by $pn^4$ proves (E.12). The same calculations give uniform bounds for the lower moments. Since $q$ is independent of $S$ and isotropic,

$$\mathbb{E} \left\|(S-I)^2 q\right\|^2 = \frac{1}{p} \mathbb{E} \operatorname{tr}(S-I)^4 \leqslant 8 \left( \frac{1}{p} \mathbb{E} \operatorname{tr} S^4 + 1 \right) = O(1). \tag{E.13}$$

Markov's inequality proves the desired stochastic bound. This avoids assuming convergence of the top eigenvalue of a tensor covariance matrix.

### E.5 LIMITS OF THE SCALAR RANK LAW

The behavior near full rank follows from the density of $\mathrm{MP}_1$, which is $(2\pi)^{-1}\sqrt{(4-t)/t}$ on $]0, 4[$. Define its distribution function by $F(t) := \mu_1([0, t])$. Near zero, the density is $1/(\pi\sqrt{t})[1 + O(t)]$. Writing $q := q_\rho$ for the cutoff in this calculation, we have $F(q) = 1 - \rho$, hence $1 - \rho = 2\sqrt{q}/\pi + O(q^{3/2})$. Integrating $t$ times the density gives

$$\tau(\rho) = \frac{2q^{3/2}}{3\pi}[1 + O(q)] = \frac{\pi^2}{12}(1-\rho)^3[1 + o(1)]. \tag{E.14}$$

Thus, after the dimensional limit, the fraction of singular directions that may be discarded scales as $(\varepsilon/\chi)^{1/3}$ as $\varepsilon/\chi \downarrow 0$. The gradient's rank does not determine this spectral tail.

The high-regularization expansion of Section I gives $\chi \sim \gamma/\lambda^2$. For the scalar small-$\lambda$ limits, use the explicit root for $m$. If $\gamma < 1$, $m \to 1/(1-\gamma)$, hence $\chi \to \gamma$. If $\gamma > 1$, $\lambda m \to 1 - 1/\gamma$, hence $\chi \to 1/\gamma$. At $\gamma = 1$, $m \sim \lambda^{-1/2}$ and $\lambda m \to 0$, hence $\chi \to 1$. Here the dimensional limit is taken first at fixed $\lambda$, followed by $\lambda \downarrow 0$. A joint vanishing-ridge theorem for Gaussian feature pairs would require additional control. The small-$\lambda$ limits are

$$\chi \longrightarrow \begin{cases} \gamma, & 0 < \gamma < 1, \\ 1, & \gamma = 1, \\ 1/\gamma, & \gamma > 1. \end{cases} \tag{E.15}$$

The joint high-regularization result in Theorem S10 uses a separate proof, rather than an interchange of limits.

# F An Auxiliary Matrix for the Error Lower Bound

## F.1 Conditional Gaussian representation

To obtain an error lower bound, we seek a matrix whose Frobenius inner product with $M$ stays away from zero but whose inner product with matrices of sublinear rank and bounded Frobenius norm vanishes. We construct it from the part of the first resolvent expansion orthogonal to the forgotten feature directions. Let $P_u^\perp := I - uu^\top$ and $P_v^\perp := I - vv^\top$, and define

$$T := P_u^\perp \,\mathrm{mat}[(S - I)q] P_v^\perp. \tag{F.1}$$

Conditional on the forgotten directions $u, v$, choose orthonormal bases starting with these vectors. In these coordinates $u = v = e_1$, and rotational invariance leaves the retained Gaussian law unchanged. Set $k := d - 1$ and write $x_i = (\alpha_i, a_i)$ and $z_i = (\beta_i, b_i)$ with $a_i, b_i \in \mathbb{R}^k$ and all indicated Gaussian components independent. The remaining square block is

$$T_0 := \frac{1}{n} \sum_{i=1}^{N} \alpha_i \beta_i a_i b_i^\top \in \mathbb{R}^{(d-1)\times(d-1)}. \tag{F.2}$$

To describe its conditional column covariance, define $w_i := \alpha_i^2 \beta_i^2$ and $C_n := n^{-1} \sum_i w_i a_i a_i^\top$. Conditional on $(\alpha_i, \beta_i, a_i)_i$, only the $b_i$ remain random. The columns of $T_0$ are then independent centered Gaussian vectors of covariance $C_n/n$, giving the exact conditional representation

$$T_0 \overset{\mathcal{L}}{=} n^{-1/2} C_n^{1/2} G, \tag{F.3}$$

where $G$ has independent $\mathcal{N}(0,1)$ entries. This conditional Gaussian assertion is exact for $T_0$. It is *not* an assertion about the full inverse-Hessian correction $M$.

## F.2 Weighted covariance concentration

If $C_n$ is close to the identity in operator norm, the representation above compares $T_0$ with a matrix of independent Gaussian entries. We prove that closeness by first controlling each quadratic form of $C_n$, then making the bound uniform over unit vectors. The weights are independent of $a_i$, with $\mathbb{E}w_i = 1$ and $\mathbb{E}w_i^2 = 9$. The law of large numbers gives

$$\frac{1}{n} \sum_i w_i \xrightarrow{\mathbb{P}} 1, \qquad \frac{1}{n} \sum_i w_i^2 \xrightarrow{\mathbb{P}} 9. \tag{F.4}$$

A union bound for the Gaussian tails gives $\max_i w_i = O_{\mathbb{P}}(\log^2 n)$. Conditional on the weights, for a unit vector $h \in \mathbb{R}^{d-1}$, the log-moment-generating function of $\sum_i w_i[(h^\top a_i)^2 - 1]$ is bounded by a constant times $t^2 \sum_i w_i^2$ whenever $|t| \max_i w_i$ is sufficiently small. Chernoff optimization therefore yields

$$\mathbb{P}\left( \left| h^\top \left[ C_n - \frac{1}{n} \sum_i w_i I \right] h \right| > s \,\middle|\, (w_i)_i \right) \leqslant 2 \exp\left[ -c \min\left\{ \frac{n^2 s^2}{\sum_i w_i^2}, \frac{ns}{\max_i w_i} \right\} \right]. \tag{F.5}$$

A $1/4$-net of the sphere has at most $9^{d-1}$ elements, and the operator norm of a symmetric matrix is bounded by twice the largest absolute value of its quadratic form on that net. Choosing a sufficiently large constant in

$$s = C\left(\frac{\sqrt{d\sum_i w_i^2}}{n} + \frac{d\max_i w_i}{n}\right) \tag{F.6}$$

makes the union bound vanish exponentially in $d$. Because $n \asymp d^2$, both terms vanish in probability. Hence

$$\|C_n - I\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{F.7}$$

### F.3 Norms and sums of squared singular values

A Gaussian matrix of size $k \times k$ satisfies $\|G\|_{\mathrm{op}} = O_{\mathbb{P}}(\sqrt{k})$, obtainable from a pair of fixed-radius nets and Gaussian linear-form tails. Also $\|G\|_{\mathrm{F}}^2/k^2 \xrightarrow{\mathbb{P}} 1$. Equations (F.3)–(F.7) imply

$$\|T_0\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} \gamma, \qquad \|T_0\|_{\mathrm{op}} = O_{\mathbb{P}}(d^{-1/2}). \tag{F.8}$$

Moreover, on a coupling using the same $G$,

$$\left\|T_0 - G/\sqrt{n}\right\|_{\mathrm{F}} \leqslant \left\|C_n^{1/2} - I\right\|_{\mathrm{op}} \|G\|_{\mathrm{F}}/\sqrt{n} = o_{\mathbb{P}}(1). \tag{F.9}$$

For Gaussian $G$, the empirical squared singular values of $G/\sqrt{k}$ converge to $\mu_1 := \mathrm{MP}_1$, whose first moment is one. The empirical first moment also tends to one by the law of large numbers. Weak convergence together with convergence of the first moment gives convergence in Wasserstein distance of order one (Villani, 2009, Theorem 6.9). On the real line, this is convergence of quantile functions in $L^1$, and hence of their integrals.

To express the squared norm left after retaining a fraction $\rho$ of the singular values, let $q_\rho \in [0,4]$ satisfy $\mu_1([q_\rho, 4]) = \rho$ and define $\tau(\rho) := \int_0^{q_\rho} t\,\mu_1(\mathrm{d}t)$. This is the contribution of the smallest fraction $1-\rho$ to the first moment. Therefore, whenever $r_d/d \to \rho$,

$$\sum_{j>r_d} s_j(T)^2 \xrightarrow{\mathbb{P}} \gamma\tau(\rho). \tag{F.10}$$

Here $T$ is obtained by adding a zero first row and column to $T_0$. The Frobenius perturbation above converges to zero in probability, so the 1-Lipschitz property (C.3) transfers the sums of squared singular values from the Gaussian matrix to $T$. A shift of rank by one is immaterial. The endpoint $\rho = 0$ also follows from $r_d \|T\|_{\mathrm{op}}^2 = o_{\mathbb{P}}(1)$ when $r_d = o(d)$.

### F.4 First-row and first-column terms in the resolvent expansion

We must also control the entries removed when defining $T$. Let $B := \mathrm{mat}[(S-I)q] - T$, whose only nonzero entries are in the first row or column. Its $(1,1)$ entry is $n^{-1}\sum_i \alpha_i^2\beta_i^2 - 1$. Its off-diagonal first-row entries are $n^{-1}\sum_i \alpha_i^2\beta_i b_{ij}$, and the first-column expression is analogous. Direct variance calculations give

$$\mathbb{E}\,\|B\|_{\mathrm{F}}^2 = \frac{8N}{n^2} + \frac{1}{n^2} + \frac{6(d-1)N}{n^2} = O(d/n) \to 0. \tag{F.11}$$

The first-row and first-column terms of $\mathrm{mat}[(S-I)q]$ therefore vanish in Frobenius norm. The analogous statement for the inverse at fixed $\lambda$ needs a separate argument, given in Section G by Schur elimination.

### F.5 Frobenius inner product of the correction and the auxiliary matrix

Since $\|M\|_{\mathrm{F}} \leqslant 1/\lambda$ and $\|B\|_{\mathrm{F}} = o_{\mathbb{P}}(1)$,

$$\langle M, T\rangle = q^\top(\lambda I + S)^{-1}(S-I)q + o_{\mathbb{P}}(1) \tag{F.12}$$

$$= 1 - (\lambda+1)q^\top(\lambda I + S)^{-1}q + o_{\mathbb{P}}(1) \tag{F.13}$$

$$\xrightarrow{\mathbb{P}} 1 - (\lambda+1)m = -\gamma\beta_*. \tag{F.14}$$

Thus the Frobenius inner product of $M$ and $T$ converges to a nonzero value. A matrix of sublinear rank has a vanishing inner product with $T$ when its Frobenius norm stays bounded. This difference will force a positive approximation error.

# G Exact Law for Gaussian Feature Pairs at Fixed Regularization

The proof has two main steps. We first show that the terms in the first row and column, outside the $(1,1)$ entry, have vanishing squared norm. We then identify the singular-value distribution of the remaining square block. An exact Schur complement gives a matrix that is Gaussian after conditioning on part of the sample. We verify the assumptions of Bandeira et al. (2023, Corollary 2.11) for that matrix and transfer the resulting approximation errors to the exact correction.

## G.1 Statement and scalar quantities

For completeness, we restate the Gaussian setting and the scalar quantities used in this proof. Let $N := n - 1$, let $x_i, z_i \in \mathbb{R}^d$ be independent standard Gaussian vectors for $i \leqslant N$, and let the independent unit vectors $u, v$ be the normalized directions of a further Gaussian pair. Define the unit gradient direction, retained covariance, and inverse-Hessian image by

$$U := uv^\top, \quad q := \mathrm{vec}(U), \quad S := \frac{1}{n}\sum_{i=1}^{N}(z_i \otimes x_i)(z_i \otimes x_i)^\top, \quad M := \mathrm{mat}[(\lambda I + S)^{-1} q]. \tag{G.1}$$

Assume $d^2/n \to \gamma \in \,]0, +\infty[$ and fix $\lambda > 0$. All constants below may depend on $\gamma, \lambda$, which do not vary with $d$. The limits of the projection and squared norm of $M$ are determined by the first two resolvent moments. Recall their definitions and the resulting relative factor:

$$m := \int \frac{\nu_\gamma(\mathrm{d}s)}{\lambda + s}, \quad m_2 := \int \frac{\nu_\gamma(\mathrm{d}s)}{(\lambda + s)^2}, \quad \chi := \gamma(1 - \lambda m)^2, \tag{G.2}$$

where $\nu_\gamma$ is the mean-one Marchenko–Pastur law. The rank dependence is described by the integrated lower quantiles of $\mu_1 := \mathrm{MP}_1$:

$$\mu_1([q_\rho, 4]) = \rho, \qquad \tau(\rho) := \int_0^{q_\rho} t\, \mu_1(\mathrm{d}t), \quad 0 \leqslant \rho \leqslant 1. \tag{G.3}$$

The endpoint values are $\tau(0) = 1, \tau(1) = 0$. With $d_r(A) := \inf_{\mathrm{rg}\, B \leqslant r} \|A - B\|_{\mathrm{F}}$, the relative squared approximation error is $\mathcal{E}_r(M) := d_r(M)^2 / \|M\|_{\mathrm{F}}^2$.

**Theorem S3** (Uniform rank law in the original paired model)**.** *Under* (G.1)*,*

$$\sup_{1 \leqslant r \leqslant d} |\mathcal{E}_r(M) - \chi\tau(r/d)| \xrightarrow{\mathbb{P}} 0. \tag{G.4}$$

*In particular, if* $1 \leqslant r_d \leqslant d$ *and* $r_d/d \to \rho \in [0,1]$*, then* $\mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \chi\tau(\rho)$*. The conclusion holds for the exact one-deletion correction whenever its scalar residual multiplier is nonzero. Rank zero is excluded, since* $\mathcal{E}_0 = 1$*.*

The tensor covariance limit (Lytova, 2017, Theorem 1.2) and independent product-sphere concentration give

$$a_d := u^\top M v \xrightarrow{\mathbb{P}} m > 0, \qquad \|M\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} m_2. \tag{G.5}$$

These two limits require only the empirical eigenvalue distribution of $S$. Indeed, $q$ is independent of $S$, $\mathbb{E}qq^\top = I/d^2$, and for symmetric $A$ independent of $q$ the product-sphere Poincaré inequality gives

$$\mathrm{Var}(q^\top A q \mid A) \leqslant \frac{8\,\mathrm{tr}(A^2)}{(d-1)d^2}. \tag{G.6}$$

Apply this with $A = (\lambda I + S)^{-1}$ and $A = (\lambda I + S)^{-2}$, whose norms are bounded by fixed powers of $\lambda^{-1}$. Tensor covariance convergence then determines their normalized traces, since the resolvent functions are bounded and continuous. This establishes (G.5) without an eigenvector assumption.

The Gaussian matrix constructed below involves another function of a sample covariance matrix. To identify the limit of its variance, define this scalar function, its spectral mean, and its scaled mean:

$$h(s) := \frac{s}{(\lambda + s)^2}, \qquad h_* := \int h(s)\nu_\gamma(\mathrm{d}s) = m - \lambda m_2 > 0, \qquad \sigma^2 := \gamma h_*. \tag{G.7}$$

The scalar fixed-point equation $m^{-1} = \lambda + (1 + \gamma m)^{-1}$ gives

$$m_2 = \frac{m^2}{1 - \chi}, \qquad h_* = \frac{m_2}{(1 + \gamma m)^2}, \qquad m^2\sigma^2 = m_2 - m^2 = m_2\chi. \tag{G.8}$$

For example, the middle identity follows by multiplying the fixed-point equation by $m_2/m$ and substituting the differentiated equation $m_2/m^2 = 1 + \gamma m_2/(1 + \gamma m)^2$.

### G.2 Exact block elimination and a finite-dimensional bound

Conditional on $u, v$, choose orthonormal bases starting with these vectors. The retained-feature law is unchanged by these rotations, and the forgotten directions become $u = v = e_1$. Set $k := d - 1$ and write

$$x_i = (\alpha_i, a_i), \quad z_i = (\beta_i, b_i), \quad a_i, b_i \in \mathbb{R}^k, \tag{G.9}$$

with all displayed Gaussian families independent. We group the feature coordinates according to their position in the matrix: the $(1, 1)$ entry, the rest of the first row and column, and the lower-right $k \times k$ block. For the last group, define the $N \times k^2$ matrix $V$. Also collect the original $k$-dimensional feature components in $N \times k$ matrices $A_0, B_0$, and the $(1, 1)$ entries in a vector $t$:

$$V_{i,:} := \mathrm{vec}(a_i b_i^\top)^\top, \quad (A_0)_{i,:} := a_i^\top, \quad (B_0)_{i,:} := b_i^\top, \quad t_i := \alpha_i\beta_i. \tag{G.10}$$

The first-row and first-column coordinates form the $N \times 2k$ matrix $K$. Including the $(1, 1)$ coordinate gives the $N \times (1 + 2k)$ matrix $A$:

$$K := [\mathrm{diag}(\alpha)B_0,\ \mathrm{diag}(\beta)A_0], \qquad A := [t, K]. \tag{G.11}$$

In this ordering, the full feature matrix is $[A, V]$. Eliminating the coordinates associated with $V$ introduces its covariance, resolvent, and the following $N \times N$ matrix:

$$S_0 := V^\top V/n, \quad R_0 := (\lambda I + S_0)^{-1}, \quad \Omega := I - VR_0V^\top/n = (I + VV^\top/(n\lambda))^{-1}. \tag{G.12}$$

Thus $0 \preceq \Omega \preceq I$, and $\Omega$ is independent of $\alpha, \beta$.

To express the same decomposition independently of the chosen bases, define $P_u^\perp := I - uu^\top$ and $P_v^\perp := I - vv^\top$. Then

$$M = a_d U + C_d + D_d, \quad C_d := uu^\top MP_v^\perp + P_u^\perp Mvv^\top, \quad D_d := P_u^\perp MP_v^\perp. \tag{G.13}$$

Here $C_d$ contains the first-row and first-column entries outside $(1, 1)$, while $D_d$ contains the remaining square block. Let $c_d \in \mathbb{R}^{2k}$ collect the entries of $C_d$, so $\|c_d\| = \|C_d\|_{\mathrm{F}}$. The Schur complement of the $V$ coordinates gives

$$\begin{pmatrix} a_d \\ c_d \end{pmatrix} = \left(\lambda I + A^\top\Omega A/n\right)^{-1} e_1, \tag{G.14}$$

$$c_d = -a_d\left(\lambda I + K^\top\Omega K/n\right)^{-1} b_d, \qquad b_d := K^\top\Omega t/n. \tag{G.15}$$

These are finite-dimensional identities. In particular $0 < a_d \leqslant \lambda^{-1}$.

**Lemma S4** (Vanishing cross-block norm)**.** *For every $d, n$ and $\lambda > 0$ in this model,*

$$\mathbb{E}\,\|C_d\|_{\mathrm{F}}^2 \leqslant \frac{6N(d-1)}{n^2\lambda^4}. \tag{G.16}$$

*Consequently $\|C_d\|_{\mathrm{F}} = O_{\mathbb{P}}(d^{-1/2})$ in the proportional regime.*

*Proof.* Equation (G.15) bounds $\|c_d\|$ by $\lambda^{-2}\|b_d\|$. It therefore suffices to bound the second moment of $b_d$. For its $j$th coordinate in the first block,

$$(b_d)_j = \frac{1}{n}\sum_{i,\ell} b_{ij}\alpha_i\Omega_{i\ell}\alpha_\ell\beta_\ell. \tag{G.17}$$

Condition on $(a_i, b_i)_i$, so $\Omega$ is fixed. Integrating over $\beta$ first and using $\mathbb{E}[\alpha_i \alpha_{i'} \alpha_\ell^2] = \mathbf{1}_{\{i=i'\}} + 2\mathbf{1}_{\{i=i'=\ell\}}$ yields

$$\mathbb{E}_{\alpha,\beta}[(b_d)_j^2 \mid (a_i, b_i)_i] = \frac{1}{n^2}\left(\sum_i b_{ij}^2 \sum_\ell \Omega_{i\ell}^2 + 2\sum_i b_{ij}^2 \Omega_{ii}^2\right) \leqslant \frac{3}{n^2}\sum_i b_{ij}^2. \tag{G.18}$$

The last inequality uses $\Omega^2 \preceq I$. The other block has the same bound with $a_{ij}$ in place of $b_{ij}$. Sum over the $2k$ coordinates and integrate over $a, b$ to obtain $\mathbb{E}\,\|b_d\|^2 \leqslant 6Nk/n^2$. Equation (G.15), the bound on $a_d$, and $\left\|(\lambda I + K^\top \Omega K/n)^{-1}\right\| \leqslant \lambda^{-1}$ finish the proof. □

After controlling $C_d$, we can isolate the leading term of the remaining block. In the following block equations, $D_d$ denotes its nonzero $k \times k$ block, and mat reshapes a vector in $\mathbb{R}^{k^2}$ into a $k \times k$ matrix. Define

$$Z_d := \mathrm{mat}(R_0 V^\top t/n). \tag{G.19}$$

The equation for the remaining block is

$$\mathrm{vec}(D_d) = -a_d R_0 V^\top t/n - R_0 V^\top K c_d/n, \tag{G.20}$$

and singular-value calculus gives

$$\left\|R_0 V^\top/\sqrt{n}\right\|_{\mathrm{op}} \leqslant \frac{1}{2\sqrt{\lambda}}. \tag{G.21}$$

Moreover $\|A_0/\sqrt{n}\|_{\mathrm{op}} + \|B_0/\sqrt{n}\|_{\mathrm{op}} = O_{\mathbb{P}}(1)$ and $\max_i(|\alpha_i| + |\beta_i|) = O_{\mathbb{P}}(\sqrt{\log k})$. Therefore

$$\|D_d + a_d Z_d\|_{\mathrm{F}} \leqslant \frac{1}{2\sqrt{\lambda}}\left\|K/\sqrt{n}\right\|_{\mathrm{op}} \|c_d\| = O_{\mathbb{P}}\left(\sqrt{\frac{\log k}{k}}\right) = o_{\mathbb{P}}(1). \tag{G.22}$$

Thus $D_d$ is approximated in Frobenius norm by $-a_d Z_d$. The construction of $Z_d$ makes its dependence on $t_i = \alpha_i \beta_i$ linear. This will give an exact Gaussian distribution after conditioning on all variables except the $\alpha_i$.

### G.3 Operator-norm concentration of partial traces

To study the singular values of $Z_d$, we first control the two matrix second moments that enter the Gaussian spectral theorem. They are determined by partial traces of $h(S_0)$, each of which sums over one tensor coordinate while retaining the other. For a symmetric operator $T$ on $\mathbb{R}^k \otimes \mathbb{R}^k$, define the partial traces by their quadratic forms, for every $u \in \mathbb{R}^k$:

$$u^\top \mathrm{Tr}_2(T) u = \mathrm{tr}[(uu^\top \otimes I_k)T], \qquad u^\top \mathrm{Tr}_1(T) u = \mathrm{tr}[(I_k \otimes uu^\top)T]. \tag{G.23}$$

This fixes the convention for the two labels. We need the same bound for both partial traces, so interchanging their labels under another vectorization convention would leave the argument unchanged. The symbol tr is the unnormalized trace throughout.

**Lemma S5** (Partial-trace isotropy)**.** *Let $S_0$ be as in* (G.12)*. For $j = 1, 2$,*

$$\left\|\frac{1}{k}\mathrm{Tr}_j\, h(S_0) - h_* I_k\right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{G.24}$$

*More precisely, with $b_k = k^{-2}\mathbb{E}\,\mathrm{tr}\, h(S_0)$,*

$$\left\|\frac{1}{k}\mathrm{Tr}_j\, h(S_0) - b_k I_k\right\|_{\mathrm{op}} = O_{\mathbb{P}}(\sqrt{k/n}), \qquad b_k \to h_*. \tag{G.25}$$

*Proof.* We prove the statement for $\mathrm{Tr}_2$ using its defining quadratic form. Exchanging the two feature families proves it for $\mathrm{Tr}_1$. Within this proof, $u \in \mathbb{R}^k$ denotes a deterministic test unit vector. Define $P := uu^\top \otimes I_k$ and regard

$$F_u := \frac{1}{k}\,\mathrm{tr}[P h(S_0)] \tag{G.26}$$

as a smooth function of the $2Nk$ independent scalar Gaussian coordinates in $a, b$. Since $0 \preceq h(S_0) \preceq (4\lambda)^{-1} I$, its values lie in $[0, (4\lambda)^{-1}]$. With $R = R_0$ and $h(S_0) = R - \lambda R^2$, differentiation gives

$$\mathrm{d}F_u = \frac{1}{k} \operatorname{tr}(B_u \, \mathrm{d}S_0), \qquad B_u = -RPR + \lambda(R^2 PR + RPR^2). \tag{G.27}$$

The symmetric operator $B_u$ describes the derivative with respect to $S_0$. We now bound the derivative with respect to the underlying Gaussian features, which determines the Lipschitz constant needed for concentration. Define $w_i := \operatorname{vec}(a_i b_i^\top)$ and $L_{\max} := \max_i(\|a_i\|^2 + \|b_i\|^2)$. Differentiating $S_0 = n^{-1} \sum_i w_i w_i^\top$ in each feature coordinate, and bounding the norms of the tensor multiplication maps, gives

$$\|\nabla F_u\|_2^2 \leqslant \frac{4}{n^2 k^2} \sum_i (\|a_i\|^2 + \|b_i\|^2) \, \|B_u w_i\|^2 \tag{G.28}$$

$$\leqslant \frac{4 L_{\max}}{nk^2} \operatorname{tr}(B_u S_0 B_u). \tag{G.29}$$

The functional calculus inequalities $\left\| S_0^{1/2} R \right\|_{\mathrm{op}} \leqslant (2\sqrt{\lambda})^{-1}$ and $\|R\|_{\mathrm{op}} \leqslant \lambda^{-1}$, together with $\|P\|_{\mathrm{F}} = \sqrt{k}$, bound each of the three terms in $S_0^{1/2} B_u$ by $\sqrt{k}/(2\lambda^{3/2})$ in Frobenius norm. Hence

$$\operatorname{tr}(B_u S_0 B_u) \leqslant \frac{9k}{4\lambda^3}, \qquad \|\nabla F_u\|_2^2 \leqslant \frac{9 L_{\max}}{nk\lambda^3}. \tag{G.30}$$

The random feature norms in $L_{\max}$ prevent a deterministic global Lipschitz bound. Fix a sufficiently large constant $K_0$ and let $\mathcal{G}_k$ be the convex set where all $\|a_i\|, \|b_i\| \leqslant K_0 \sqrt{k}$. Gaussian norm tails give $\mathbb{P}(\mathcal{G}_k^c) \leqslant 2N \mathrm{e}^{-c_{K_0} k}$. On this set, (G.30) bounds the Lipschitz constant by $\sqrt{18} K_0/(\sqrt{n}\lambda^{3/2})$, uniformly over $u$. Extend $F_u$ to the entire Euclidean space with the same Lipschitz constant, then clip the extension to $[0, (4\lambda)^{-1}]$. Denote the result by $\widetilde{F}_u$. It agrees with $F_u$ on $\mathcal{G}_k$ and differs from it in expectation by at most $(4\lambda)^{-1} \mathbb{P}(\mathcal{G}_k^c)$.

Applying Gaussian concentration to the Lipschitz extension $\widetilde{F}_u$ (Vershynin, 2018, Chapter 5) gives constants $c, C > 0$, depending only on $\lambda, K_0$, such that for each net point $u$,

$$\mathbb{P}\big(|\widetilde{F}_u - \mathbb{E}\widetilde{F}_u| > s\big) \leqslant 2\mathrm{e}^{-cns^2}. \tag{G.31}$$

Product rotational invariance of $S_0$ implies $\mathbb{E}h(S_0) = b_k I_{k^2}$, so $\mathbb{E}F_u = b_k$ for every $u$. Take a deterministic $1/4$-net of the unit sphere with at most $9^k$ elements. The norm of a symmetric matrix is bounded by twice its largest absolute quadratic form on this net. Using the common event $\mathcal{G}_k$ once, rather than once per net point, yields

$$\mathbb{P}\Big( \big\| k^{-1} \operatorname{Tr}_2 h(S_0) - b_k I \big\|_{\mathrm{op}} > 2s \Big) \leqslant \mathbb{P}(\mathcal{G}_k^c) + 2\, 9^k \mathrm{e}^{-c' n s^2} \tag{G.32}$$

for $s$ larger than twice the negligible expectation shift. Set $s = C\sqrt{k/n}$ with $C$ sufficiently large. This proves (G.25). Finally $k^2/n \to \gamma$, $N/n \to 1$, and the tensor covariance limit applies to $S_0$. Because $h$ is bounded continuous, $k^{-2} \operatorname{tr} h(S_0) \xrightarrow{\mathbb{P}} h_*$ and boundedness also gives convergence of expectations. Thus $b_k \to h_*$. □

The lemma controls the two partial traces in operator norm. The full operator $h(S_0)$ need not be close to a scalar multiple of the identity, and the next step only uses the two proved limits.

### G.4 A conditionally Gaussian matrix with controlled entry covariance

To obtain a nondegenerate limit for the singular-value distribution, we rescale $Z_d$ by $\sqrt{k}$. Define its coefficient matrices and rescaled version by

$$Q_i := \frac{\sqrt{k}}{n} \operatorname{mat}(R_0 w_i), \qquad Y_d := \sqrt{k} Z_d = \sum_{i=1}^{N} \alpha_i \beta_i Q_i. \tag{G.33}$$

The coefficients $Q_i$ depend only on $V$ and are independent of both scalar families $\alpha, \beta$. The covariance of all entries of $Y_d$, conditional only on $V$, is determined by the identity

$$\sum_i \operatorname{vec}(Q_i)\operatorname{vec}(Q_i)^\top = \frac{k}{n} R_0 S_0 R_0 = \frac{k}{n} h(S_0). \tag{G.34}$$

Taking its two partial traces and using Lemma S5 gives, with the appropriate matching of partial-trace labels,

$$\sum_i Q_i Q_i^\top \xrightarrow{\mathbb{P}} \sigma^2 I_k, \qquad \sum_i Q_i^\top Q_i \xrightarrow{\mathbb{P}} \sigma^2 I_k \quad \text{in operator norm.} \tag{G.35}$$

Both sums are deterministically bounded above by $k^2/(4n\lambda)$ times the identity. On an event of probability tending to one, $\max_i \|a_i\| \, \|b_i\| \leqslant Ck$. Consequently,

$$\max_i \|Q_i\|_{\mathrm{op}}^2 \leqslant \max_i \|Q_i\|_{\mathrm{F}}^2 \leqslant \frac{k}{n^2\lambda^2} \max_i \|w_i\|^2 \leqslant \frac{C}{k}. \tag{G.36}$$

Conditioning on $\beta$ will replace the preceding sums by sums weighted by $\beta_i^2$. The next lemma shows that this leaves their limits unchanged.

**Lemma S6** (Gaussian reweighting of the variances)**.** *For the independent scalars $\beta_i \sim \mathcal{N}(0,1)$,*

$$\left\| \sum_i (\beta_i^2 - 1) Q_i Q_i^\top \right\|_{\mathrm{op}} + \left\| \sum_i (\beta_i^2 - 1) Q_i^\top Q_i \right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{G.37}$$

*Proof.* Condition on $V$ on the event where (G.36) holds. For $W_i := Q_i Q_i^\top$ (or $Q_i^\top Q_i$), one has $0 \preceq W_i$, $\max_i \|W_i\| \leqslant C/k$, and $\|\sum_i W_i\| \leqslant C$. The weights $\beta_i^2$ are unbounded, so truncate them before applying matrix concentration. Define $\zeta_i := \min\{\beta_i^2, 20 \log k\}$ and $\xi_i := \zeta_i - \mathbb{E}\zeta_i$. Since $N \asymp k^2$, Gaussian tails imply that all $\zeta_i = \beta_i^2$ with probability tending to one. Also $\mathbb{E}\zeta_i \to 1$, $\mathbb{E}\xi_i^2 \leqslant \mathbb{E}\beta_i^4 = 3$, and $|\xi_i| \leqslant 20 \log k + 1$. Therefore

$$\max_i \|\xi_i W_i\| \leqslant C \log k / k, \qquad \left\| \sum_i \mathbb{E}[\xi_i^2] W_i^2 \right\| \leqslant C/k, \tag{G.38}$$

where $W_i^2 \preceq (C/k) W_i$. The self-adjoint matrix Bernstein inequality (Tropp, 2015, Theorem 6.6.1), applied to the sum and its negative, gives

$$\left\| \sum_i \xi_i W_i \right\|_{\mathrm{op}} = O_{\mathbb{P}}\left( \sqrt{\frac{\log k}{k}} + \frac{\log^2 k}{k} \right) = o_{\mathbb{P}}(1). \tag{G.39}$$

The difference between the truncated mean and one contributes at most $|\mathbb{E}\zeta_i - 1| \, \|\sum_i W_i\| = o(1)$. Removing the truncation on its high-probability equality event proves the claim for each of the two sums. □

We can now condition on the feature vectors and the $\beta_i$, leaving only Gaussian coefficients $\alpha_i$ random. The corresponding sigma-algebra is

$$\mathcal{F}_d := \sigma((a_i, b_i, \beta_i)_{i=1}^N). \tag{G.40}$$

The entries of $Y_d$ are exactly jointly centered Gaussian, since only the independent Gaussian coefficients $\alpha_i$ remain random. Equations (G.35)–(G.37) give

$$\left\| \mathbb{E}_\alpha[Y_d Y_d^\top \mid \mathcal{F}_d] - \sigma^2 I_k \right\|_{\mathrm{op}} + \left\| \mathbb{E}_\alpha[Y_d^\top Y_d \mid \mathcal{F}_d] - \sigma^2 I_k \right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{G.41}$$

The spectral theorem also requires a bound on the joint covariance of all $k^2$ entries, which controls their dependence. Define that covariance operator by

$$\mathcal{K}_d := \operatorname{Cov}_\alpha(\operatorname{vec} Y_d \mid \mathcal{F}_d) = \sum_i \beta_i^2 \operatorname{vec}(Q_i)\operatorname{vec}(Q_i)^\top \tag{G.42}$$

$$\preceq \left( \max_i \beta_i^2 \right) \frac{k}{n} h(S_0), \qquad \|\mathcal{K}_d\|_{\mathrm{op}} = O_{\mathbb{P}}(\log k / k). \tag{G.43}$$

Here $h(S_0) \preceq (4\lambda)^{-1} I$ and $\max_i \beta_i^2 = O_{\mathbb{P}}(\log k)$.

### G.5 A SPECTRAL THEOREM FOR GAUSSIAN MATRICES

We use the following special case of Bandeira et al. (2023, Corollary 2.11). If $X_j$ are self-adjoint matrices of dimensions $p_j \geqslant j$ tending to infinity, with jointly Gaussian entries, and

$$\|\mathbb{E}X_j\|_{\mathrm{op}} \to 0, \quad \left\|\mathbb{E}X_j^2 - I\right\|_{\mathrm{op}} \to 0, \quad v(X_j) := \|\mathrm{Cov}(\mathrm{vec}\, X_j)\|_{\mathrm{op}}^{1/2} = o((\log p_j)^{-3/2}), \tag{G.44}$$

then their empirical eigenvalue distributions converge to the standard semicircle law and $\|X_j\|_{\mathrm{op}} \to 2$ almost surely. This theorem allows dependent entries and requires no eigenvector invariance.

The theorem concerns self-adjoint matrices. We apply it to a symmetric block matrix whose eigenvalues are the signed singular values of $Y_d/\sigma$. Here the sequence index is $k$, the matrix dimension is $2k$, and $N \asymp k^2$ is the number of Gaussian summands. Define

$$\mathcal{X}_d := \frac{1}{\sigma}\begin{pmatrix} 0 & Y_d \\ Y_d^\top & 0 \end{pmatrix}. \tag{G.45}$$

Conditional on $\mathcal{F}_d$, this matrix has Gaussian entries and zero mean. Its conditional second moment tends to $I_{2k}$ in probability by (G.41). The linear map from $\mathrm{vec}\, Y_d$ to its entries has norm $\sqrt{2}/\sigma$, so (G.43) gives

$$v_\alpha(\mathcal{X}_d \mid \mathcal{F}_d)^2 \leqslant \frac{2}{\sigma^2}\,\|\mathcal{K}_d\|_{\mathrm{op}} = O_{\mathbb{P}}(\log k/k). \tag{G.46}$$

In particular $v_\alpha(\mathcal{X}_d \mid \mathcal{F}_d)(\log(2k))^{3/2} \xrightarrow{\mathbb{P}} 0$. Thus the conditional mean, second moment, and full covariance satisfy the required bounds in probability. The remaining step is to justify using a theorem stated for deterministic Gaussian covariance sequences.

Use the subsequence characterization of convergence in probability. From any dimension subsequence, choose a further subsequence on which the errors in (G.41) and the scaled quantity in (G.46) vanish almost surely with respect to the variables in $\mathcal{F}_d$. For almost every fixed sequence of these variables, the remaining Gaussian matrices satisfy (G.44). Apply the cited theorem to the $\alpha$ variables. For each fixed positive error threshold, the resulting conditional probability of exceeding it tends to zero. These probabilities lie in $[0, 1]$, so bounded convergence gives convergence in probability under the joint law along the extracted subsequence. Since the initial subsequence was arbitrary, convergence in probability holds on the original sequence. The later transfer to $M$ requires no independence between $a_d$ and $Y_d$.

The eigenvalues of $\mathcal{X}_d$ are the positive and negative singular values of $Y_d/\sigma$. Squaring a standard semicircle variable gives the aspect-one Marchenko–Pastur law. We obtain

$$\frac{1}{k}\sum_{j=1}^{k} \delta_{k s_j(Z_d)^2/\sigma^2} \Longrightarrow \mathrm{MP}_1 \quad \text{in probability}, \qquad \sqrt{k}\,\|Z_d\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 2\sigma. \tag{G.47}$$

Weak convergence alone would not control sums of squared singular values. The operator-norm convergence also bounds the largest point in each empirical measure in probability, which gives convergence of the first moments. Together, these properties imply convergence in Wasserstein distance of order one (Villani, 2009, Theorem 6.9). On the real line, this distance is the $L^1$ distance between quantile functions. It therefore bounds the difference between their integrals over every lower quantile interval, simultaneously. We obtain

$$\|Z_d\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} \sigma^2, \qquad \sup_{0\leqslant \ell \leqslant k} \left| \sum_{j>\ell} s_j(Z_d)^2 - \sigma^2 \tau(\ell/k) \right| \xrightarrow{\mathbb{P}} 0. \tag{G.48}$$

The operator-norm limit in (G.47) concerns $Z_d$. A Frobenius perturbation that only vanishes in probability will suffice to transfer the integrated errors to $D_d$, but does not transfer the scaled largest-singular-value limit.

### G.6 TRANSFERRING THE SINGULAR-VALUE LIMIT TO THE EXACT CORRECTION

To return to the original $d \times d$ matrix, add a zero first row and column to $Z_d$ and denote the result by $\bar{Z}_d$. Keep the exact coefficient $a_d$ in the gradient direction and define the comparison matrix

$$N_d := a_d U - a_d \bar{Z}_d. \tag{G.49}$$

Equations (G.16) and (G.22) imply

$$\|M - N_d\|_{\mathrm{F}} = O_{\mathbb{P}}\left(\sqrt{\frac{\log k}{k}}\right) = o_{\mathbb{P}}(1). \tag{G.50}$$

The two terms of $N_d$ have orthogonal row and column spaces. Since $a_d \xrightarrow{\mathbb{P}} m > 0$ and $\|Z_d\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0$, the largest singular value of $N_d$ is $a_d$ with probability tending to one. On this event, a rank-$r$ truncation retains that singular value and the $r-1$ largest singular values from the other block. Thus simultaneously for every $1 \leqslant r \leqslant d$,

$$d_r(N_d)^2 = a_d^2 \sum_{j>r-1} s_j(Z_d)^2. \tag{G.51}$$

Because $|(r-1)/(d-1) - r/d| \leqslant 1/(d-1)$ and $\tau$ is uniformly continuous, (G.48) gives

$$\sup_{1\leqslant r\leqslant d} \left|d_r(N_d)^2 - m^2\sigma^2\tau(r/d)\right| \xrightarrow{\mathbb{P}} 0. \tag{G.52}$$

The distance to the closed set of matrices of rank at most $r$ is 1-Lipschitz in Frobenius norm, with a constant independent of $r$. Factoring the difference of the squared distances therefore gives

$$\sup_r |d_r(M)^2 - d_r(N_d)^2| \leqslant \|M - N_d\|_{\mathrm{F}} \left(\|M\|_{\mathrm{F}} + \|N_d\|_{\mathrm{F}}\right) = o_{\mathbb{P}}(1). \tag{G.53}$$

Divide by $\|M\|_{\mathrm{F}}^2 \to m_2$ and use (G.8). This proves Theorem S3.

The same approximation determines the size and singular-value distribution of $D_d$. Equation (G.16) already controls the omitted first-row and first-column terms. For the remaining block, combine (G.22) and (G.48) with the singular-value perturbation inequality of Mirsky (1960) to obtain

$$\|D_d\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} m^2\sigma^2 = m_2\chi, \qquad \frac{1}{k}\sum_{j=1}^{k} \delta_{k s_j(D_d)^2/\|D_d\|_{\mathrm{F}}^2} \Longrightarrow \mathrm{MP}_1 \quad \text{in probability.} \tag{G.54}$$

Indeed, that inequality bounds the Euclidean distance between the ordered singular-value vectors by the Frobenius distance between the matrices. Factoring each difference of squares and applying Cauchy–Schwarz gives

$$\sum_j |s_j(D_d)^2 - a_d^2 s_j(Z_d)^2| \leqslant \|D_d + a_d Z_d\|_{\mathrm{F}} \left(\|D_d\|_{\mathrm{F}} + a_d \|Z_d\|_{\mathrm{F}}\right) = o_{\mathbb{P}}(1). \tag{G.55}$$

After normalizing by the respective squared Frobenius norms, whose limits are positive, this bounds the Wasserstein distance of order one between the normalized squared-singular-value empirical measures. It proves (G.54). Thus both the discarded entries and the change in the remaining singular values are controlled by explicit norm bounds.

### G.7 Consequences and assumptions

**Corollary S7** (Sublinear rank and the required rank fraction)**.** *For any $1 \leqslant r_d = o(d)$, $\mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \chi$. For $0 < \varepsilon < \chi$, let $r_\varepsilon := \min\{r : \mathcal{E}_r \leqslant \varepsilon\}$. Then*

$$r_\varepsilon/d \xrightarrow{\mathbb{P}} \tau^{-1}(\varepsilon/\chi). \tag{G.56}$$

*For $\chi < \varepsilon < 1$, rank one meets the tolerance with probability tending to one. At $\varepsilon = \chi$, one can conclude $r_\varepsilon/d \to 0$ in probability, but not that rank one succeeds with probability tending to one. The uniform law also permits a rank chosen from the data: if a random $1 \leqslant \widehat{r}_d \leqslant d$ satisfies $\widehat{r}_d/d \xrightarrow{\mathbb{P}} \rho$, then $\mathcal{E}_{\widehat{r}_d} \xrightarrow{\mathbb{P}} \chi\tau(\rho)$.*

*Proof.* Uniform convergence and continuity of $\tau$ prove the sublinear and data-dependent statements. For $0 < \varepsilon < \chi$, evaluate the law at rank fractions strictly below and above $\tau^{-1}(\varepsilon/\chi)$. Strict monotonicity places the minimum sufficient rank between these two fractions with probability tending to one. For $\varepsilon > \chi$, use $\mathcal{E}_1 \to \chi$. At equality, the error at every fixed positive rank fraction converges to a value strictly below $\chi$, so the minimum fraction tends to zero in probability. This does not determine the sign of the finite-dimensional difference $\mathcal{E}_1 - \chi$. □

The lower bound $\chi(1-\chi)$ from Section H remains valid. The exact theorem identifies the sublinear-rank limit as $\chi$ and determines the limit for every rank fraction. The result for $\lambda_d \to +\infty$ requires the separate argument in Section I, since convergence for each fixed $\lambda$ does not justify an interchange of limits.

The proof applies to independent Gaussian paired features, a forgotten observation selected independently of the sample, and fixed positive ridge. It does not establish a joint vanishing-ridge limit, universality for arbitrary non-Gaussian features, an exact law for the Hessian-weighted minimum error at a given rank, or distributional guarantees for randomized removal. Numerical comparisons supplement the analytic proof by evaluating the stated Schur identities and derivatives.

# H Proofs of the Rank Bounds for Gaussian Feature Pairs

The bounds in this section use only the scalar limits and the auxiliary matrix $T$. The exact limit at fixed regularization was proved in Section G using the additional block-decomposition and Gaussian spectral arguments.

The two bounds below use the same setting as Section E: all retained and forgotten feature vectors $x_i, z_i$ are independent with law $\mathcal{N}(0, I_d)$, the forgotten index is fixed independently of the sample, and its normalized directions $u, v$ are independent of the retained features. We assume $d^2/n \to \gamma \in ]0, +\infty[$ and define $M := \mathrm{mat}[(\lambda I + S)^{-1} \operatorname{vec}(uv^\top)]$. The error is $\mathcal{E}_r := d_r(M)^2 / \|M\|_{\mathrm{F}}^2$, which equals the relative deletion error when $\Delta \neq 0$. The quantities $m, m_2, \chi$ are those of Section E. Recall that, for $\mu_1 = \mathrm{MP}_1$, the cutoff satisfies $\mu_1([q_\rho, 4]) = \rho$ and $\tau(\rho) = \int_0^{q_\rho} t\, \mu_1(\mathrm{d}t)$.

**Theorem S8** (Error bound for sublinear ranks)**.** *In the setting above, fix $\lambda > 0$. For every sequence $1 \leqslant r_d \leqslant d$ with $r_d = o(d)$,*

$$\chi(1-\chi) - o_{\mathbb{P}}(1) \leqslant \mathcal{E}_{r_d} \leqslant \chi + o_{\mathbb{P}}(1). \tag{H.1}$$

**Theorem S9** (Error bounds for proportional ranks)**.** *In the setting above, let $1 \leqslant r_d \leqslant d$, $r_d/d \to \rho$ and fix $\lambda > 0$. Put $c = \sqrt{\chi(1-\chi)}$. Then*

$$\mathcal{E}_{r_d} \geqslant \left[c\sqrt{\tau(\rho)} - \chi\sqrt{1-\tau(\rho)}\right]_+^2 - o_{\mathbb{P}}(1), \tag{H.2}$$

$$\mathcal{E}_{r_d} \leqslant \min\{(1-\rho)\chi, [c\sqrt{\tau(\rho)} + \chi]^2\} + o_{\mathbb{P}}(1). \tag{H.3}$$

## H.1 Sublinear ranks

Let $B_r$ be a best approximation to $M$ of rank at most $r$. Truncating the singular values of $M$ gives $\|B_r\|_{\mathrm{F}} \leqslant \|M\|_{\mathrm{F}}$. To bound its inner product with $T$, use the nuclear norm $\|B_r\|_* := \sum_j s_j(B_r)$. The duality between the nuclear and operator norms, followed by Cauchy–Schwarz on the at most $r$ nonzero singular values, implies

$$|\langle B_r, T\rangle| \leqslant \|B_r\|_* \|T\|_{\mathrm{op}} \leqslant \sqrt{r}\, \|M\|_{\mathrm{F}} \|T\|_{\mathrm{op}} = o_{\mathbb{P}}(1) \tag{H.4}$$

for every $r = o(d)$. The inequality holds for every such $B_r$, including one chosen from the observed matrices. It requires no independence between $B_r$ and $T$. Cauchy–Schwarz now yields

$$\|M - B_r\|_{\mathrm{F}} \geqslant \frac{|\langle M - B_r, T\rangle|}{\|T\|_{\mathrm{F}}} \geqslant \sqrt{\gamma}\beta_* - o_{\mathbb{P}}(1). \tag{H.5}$$

Squaring this norm bound, dividing by $\|M\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} m_2$, and using equation (E.10) prove the lower bound $\chi(1-\chi)$. For every $r \geqslant 1$, the candidate $(\langle M, U\rangle)U$ gives

$$\mathcal{E}_r \leqslant 1 - \frac{\langle M, U\rangle^2}{\|M\|_{\mathrm{F}}^2} \xrightarrow{\mathbb{P}} 1 - m^2/m_2 = \chi. \tag{H.6}$$

This proves Theorem S8. Rank zero is excluded: $\mathcal{E}_0 = 1$ exactly.

### H.2 Optimized rank-fraction lower bound

For proportional ranks, a comparison matrix with known singular values gives a more informative bound. Assume $r_d/d \to \rho$. For a fixed scalar $b \geqslant 0$, define $N_b := mU - bT$. The two terms have orthogonal row and column spaces. Since $\|T\|_{\mathrm{op}} \to 0$ in probability and $m > 0$, the largest singular value of $N_b$ is $m$ with probability tending to one. For $r_d \geqslant 1$,

$$d_{r_d}(N_b)^2 \xrightarrow{\mathbb{P}} b^2\gamma\tau(\rho). \tag{H.7}$$

By the limits proved above,

$$\|M - N_b\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} m_2 - m^2 - 2b\gamma\beta_* + b^2\gamma. \tag{H.8}$$

The 1-Lipschitz property of $d_r$ bounds the error for $M$ below by the error for $N_b$ minus their Frobenius distance. To optimize over $b$, use the rescaled coefficient $x$ and abbreviate the two scalar factors as follows:

$$x := \frac{b\sqrt{\gamma}}{\sqrt{m_2}}, \quad c := \sqrt{\chi(1-\chi)}, \quad t := \tau(\rho). \tag{H.9}$$

Then, for every fixed $x \geqslant 0$,

$$\sqrt{\mathcal{E}_r} \geqslant \left[x\sqrt{t} - \sqrt{(x-c)^2 + \chi^2}\right]_+ - o_{\mathbb{P}}(1). \tag{H.10}$$

For $0 < t < 1$, the expression inside the positive part is maximized at $x = c + \chi\sqrt{t/(1-t)}$, with value $c\sqrt{t} - \chi\sqrt{1-t}$. For $t = 0$, its maximum is negative, so the positive part is zero. For $t = 1$, take successively larger fixed $x$ after the dimensional limit. The supremum is then $c$. Therefore

$$\mathcal{E}_r \geqslant \left[c\sqrt{\tau(\rho)} - \chi\sqrt{1-\tau(\rho)}\right]_+^2 - o_{\mathbb{P}}(1). \tag{H.11}$$

The positive part is nonzero precisely when $\tau(\rho) > \chi$.

### H.3 Upper bounds

The same comparison also provides an upper bound. Taking $b = \beta_*$ and using (E.10) gives

$$\frac{\|M - N_{\beta_*}\|_{\mathrm{F}}^2}{m_2} \xrightarrow{\mathbb{P}} \chi^2. \tag{H.12}$$

The triangle inequality for $d_r$ proves

$$\mathcal{E}_r \leqslant \left[\sqrt{\chi(1-\chi)\tau(\rho)} + \chi\right]^2 + o_{\mathbb{P}}(1). \tag{H.13}$$

A second upper bound follows directly from ordering the squared singular values. The average of the smallest $d - r$ values cannot exceed the average of the smallest $d - 1$, so

$$\sum_{j>r} s_j(M)^2 \leqslant \frac{d-r}{d-1} \sum_{j>1} s_j(M)^2. \tag{H.14}$$

The rank-one upper bound already proved yields $\mathcal{E}_r \leqslant (1-\rho)\chi + o_{\mathbb{P}}(1)$. Taking the minimum proves the stated bounds.

### H.4 A positive rank fraction is necessary below the error lower bound

For a tolerance $\varepsilon \geqslant 0$, define the smallest sufficient oracle rank by $r_\varepsilon := \min\{r \in [\![0, d]\!] : \mathcal{E}_r \leqslant \varepsilon\}$. The lower bound is continuous as a function of $\rho$ at zero, where it equals $\chi(1-\chi) > 0$. For every $\varepsilon < \chi(1-\chi)$, choose a sufficiently small $\rho_0 > 0$ so that the lower bound at $\rho_0$ exceeds $\varepsilon$. Applying the theorem to $r_d = \lfloor \rho_0 d \rfloor$ and using monotonicity of the oracle error shows

$$\mathbb{P}(r_\varepsilon > \lfloor \rho_0 d \rfloor) \to 1. \tag{H.15}$$

This is stronger than ruling out each fixed rank separately. It does not assert that the current lower bound is optimal at larger rank fractions.

# I JOINT HIGH-REGULARIZATION LIMIT

We keep the Gaussian setting and $d^2/n \to \gamma$ from Section E, but now let $\lambda = \lambda_d \to \infty$. For ranks $1 \leqslant r_d \leqslant d$ with $r_d/d \to \rho$, we show that the auxiliary matrix $T$ gives the leading approximation error.

**Theorem S10** (Exact asymptotic error under high regularization)**.** *Under Assumption 1, if $\lambda_d \to \infty$, $1 \leqslant r_d \leqslant d$ and $r_d/d \to \rho$, then*

$$\lambda_d^2 \mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \gamma\tau(\rho). \tag{I.1}$$

To expand around the identity covariance, define $E := S - I$ and $a := \lambda + 1$. Since $E$ commutes with $(aI + E)^{-1}$, direct multiplication verifies the exact identity

$$(\lambda I + S)^{-1} = \frac{1}{a} I - \frac{1}{a^2} E + \frac{1}{a^2}(\lambda I + S)^{-1} E^2. \tag{I.2}$$

After applying it to $q$, the remainder has Euclidean/Frobenius norm at most

$$\frac{\left\|E^2 q\right\|}{a^2 \lambda}. \tag{I.3}$$

Section E proved $\left\|E^2 q\right\| = O_{\mathbb{P}}(1)$ uniformly in dimension. For every sequence $\lambda_d \to \infty$, the remainder is therefore $o_{\mathbb{P}}(\lambda_d^{-2})$. The first-order estimate for the first row and column in Section F also implies

$$M = N_d + o_{\mathbb{P}}(\lambda_d^{-2}), \qquad N_d := \frac{U}{\lambda_d + 1} - \frac{T}{(\lambda_d + 1)^2} \tag{I.4}$$

in Frobenius norm.

The terms of $N_d$ are block diagonal in bases starting with $u, v$. The singular value contributed by the rank-one term is $1/(\lambda_d + 1)$ and its other singular values are those of $T/(\lambda_d + 1)^2$. This singular value is the largest with probability tending to one because $\|T\|_{\mathrm{op}} = O_{\mathbb{P}}(d^{-1/2})$. Thus, for $r_d \geqslant 1$,

$$\lambda_d^4 d_{r_d}(N_d)^2 = \frac{\lambda_d^4}{(\lambda_d + 1)^4} \sum_{j > r_d - 1} s_j(T)^2 \xrightarrow{\mathbb{P}} \gamma\tau(\rho). \tag{I.5}$$

The distance Lipschitz bound and (I.4) transfer this limit to $M$. Also $\lambda_d M \to U$ in Frobenius norm in probability, so $\lambda_d^2 \|M\|_{\mathrm{F}}^2 \to 1$. Dividing proves

$$\lambda_d^2 \mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \gamma\tau(\rho). \tag{I.6}$$

The uniform fourth-moment bound controls the remainder along every sequence $\lambda_d \to \infty$, which establishes a joint limit in $d$ and $\lambda_d$. Rank zero remains excluded because its relative error is exactly one.

## I.1 FINITE-REGULARIZATION CORRECTION TO THE SCALE

The first two moments of $\nu_\gamma$ are 1 and $1 + \gamma$. Expanding its Stieltjes transform at infinity gives

$$m = \lambda^{-1} - \lambda^{-2} + (1 + \gamma)\lambda^{-3} + O(\lambda^{-4}). \tag{I.7}$$

Consequently,

$$\frac{\lambda^2 \chi}{\gamma} = 1 - \frac{2(1 + \gamma)}{\lambda} + O(\lambda^{-2}). \tag{I.8}$$

This explains why the leading high-regularization scale can remain noticeably inaccurate at moderate $\lambda$ even when the exact fixed-$\lambda$ asymptotic formula is accurate. It is an expansion of the scalar factor, not a uniform finite-$d$ error bound.

# J Fixed Regularization under Orthogonal Invariance

Under full orthogonal invariance, the direction orthogonal to the gradient has an explicit distribution. This yields the exact rank law directly from Gaussian singular values. The assumption belongs only to the model in this section. Section G proves the corresponding law for the original paired features without imposing it.

**Theorem S11** (Rank law under orthogonal invariance)**.** *Let $p = d^2$ and $H_d = O_d \Lambda_d O_d^\top$, where $O_d$ is Haar on $O(p)$, independent of the positive diagonal eigenvalue matrix $\Lambda_d$ and a unit rank-one query $q = \mathrm{vec}(uv^\top)$. Suppose the inverse eigenvalues are bounded by a dimension-independent constant, and*

$$\frac{1}{p}\operatorname{tr} H_d^{-1} \xrightarrow{\mathbb{P}} a > 0, \qquad \frac{1}{p}\operatorname{tr} H_d^{-2} \xrightarrow{\mathbb{P}} b > a^2. \tag{J.1}$$

*For $M_d = \mathrm{mat}(H_d^{-1} q)$ and $1 \leqslant r_d \leqslant d$, $r_d/d \to \rho \in [0,1]$,*

$$\frac{d_{r_d}(M_d)^2}{\|M_d\|_{\mathrm{F}}^2} \xrightarrow{\mathbb{P}} \left(1 - \frac{a^2}{b}\right) \tau(\rho). \tag{J.2}$$

## J.1 Norms of the projection and the remainder

Condition on the eigenvalues and $q$. The coordinates of $O_d^\top q$ form a uniform sphere vector $h \in \mathbb{R}^p$. Its fourth moments are

$$\mathbb{E} h_i^4 = \frac{3}{p(p+2)}, \qquad \mathbb{E} h_i^2 h_j^2 = \frac{1}{p(p+2)} \quad (i \neq j). \tag{J.3}$$

Thus for a diagonal matrix $A$,

$$\mathrm{Var}(h^\top A h \mid A) = \frac{2}{p(p+2)} \left[\operatorname{tr} A^2 - \frac{(\operatorname{tr} A)^2}{p}\right]. \tag{J.4}$$

The component of $H_d^{-1} q$ along $q$ and its total squared norm are quadratic forms of this type. Define

$$a_d := q^\top H_d^{-1} q, \qquad b_d := q^\top H_d^{-2} q. \tag{J.5}$$

Uniform boundedness of the inverse eigenvalues lets us apply the variance identity to $A = \Lambda_d^{-1}$ and $A = \Lambda_d^{-2}$. We obtain $a_d \xrightarrow{\mathbb{P}} a$ and $b_d \xrightarrow{\mathbb{P}} b$. The squared norm orthogonal to $q$ is therefore $\sigma_d^2 := b_d - a_d^2 \xrightarrow{\mathbb{P}} b - a^2 > 0$.

## J.2 Why the residual direction is uniform

For every orthogonal $V$ fixing $q$, the conditional law of $H_d$ given its eigenvalues is invariant under $H_d \mapsto V H_d V^\top$. Therefore $H_d^{-1} q - a_d q$, conditional on the scalars $(a_d, b_d)$ and the eigenvalues, is invariant under all rotations of $q^\perp$. Its norm is $\sigma_d$. The unique rotation-invariant probability measure on that sphere is the uniform measure. We may consequently represent

$$H_d^{-1} q = a_d q + \sigma_d \frac{(I - qq^\top) g}{\|(I - qq^\top) g\|}, \qquad g \sim \mathcal{N}(0, I_p), \tag{J.6}$$

where the normalized Gaussian direction is independent of the conditioning scalars. The argument uses every orthogonal transformation that fixes $q$, a group isomorphic to $O(p-1)$. The smaller group of separate left and right rotations, $O(d-1) \times O(d-1)$, would not imply a uniform direction on this sphere.

## J.3 Reshaping and removal of cross blocks

Conditional on $q$, apply fixed left and right rotations so that $u = v = e_1$. These rotations preserve singular values. Reshaping $g$ gives a $d \times d$ matrix $G$ of independent standard Gaussian entries. The projection $I - qq^\top$ replaces $G_{11}$ by zero. Denote the resulting matrix by $G_\perp$, its first-row and

first-column part by $G_C$, and its lower-right $(d-1)\times(d-1)$ block by $G_D$. The disjoint supports are Frobenius-orthogonal. Since

$$\|G_\perp\|_{\mathrm{F}}^2/d^2 \xrightarrow{\mathbb{P}} 1, \qquad \mathbb{E}\,\|G_C\|_{\mathrm{F}}^2 = 2(d-1), \tag{J.7}$$

we have

$$\|M_d - N_d\|_{\mathrm{F}} = O_{\mathbb{P}}(d^{-1/2}), \qquad N_d := a_d U + \frac{\sigma_d}{\|G_\perp\|_{\mathrm{F}}}\begin{pmatrix} 0 & 0 \\ 0 & G_D \end{pmatrix}. \tag{J.8}$$

The matrix $N_d$ is block diagonal. The singular value contributed by the rank-one term tends to $a > 0$, whereas the operator norm of the remaining block is $O_{\mathbb{P}}(d^{-1/2})$. Thus this singular value is the largest with probability tending to one.

### J.4 Gaussian singular values and approximation error

Write $k := d-1$. The empirical eigenvalue distribution of $G_D G_D^\top/k$ converges to $\mathrm{MP}_1$, and its mean tends to one. The integrated-quantile argument from Section F therefore gives, for $r_d \geqslant 1$,

$$d_{r_d}(N_d)^2 = \frac{\sigma_d^2}{\|G_\perp\|_{\mathrm{F}}^2}\sum_{j>r_d-1} s_j(G_D)^2 \tag{J.9}$$

$$\xrightarrow{\mathbb{P}} (b-a^2)\tau(\rho). \tag{J.10}$$

For $r_d = o(d)$, the sum of squares of the largest $r_d$ singular values of the remaining block is bounded by $r_d$ times its squared operator norm, hence $O_{\mathbb{P}}(r_d/d) = o_{\mathbb{P}}(1)$. This gives the same limit at $\rho = 0$. For $r_d/d \to 1$, the integrated-quantile limit gives zero. The $o_{\mathbb{P}}(1)$ Frobenius perturbation from $N_d$ to $M_d$ transfers the tail limit by (C.3). Since $\|M_d\|_{\mathrm{F}}^2 = b_d \to b$, division proves the theorem.

### J.5 Specialization to a shifted Marchenko–Pastur spectrum

Take eigenvalues $\lambda + s_j$, where $s_j \geqslant 0$ and their empirical distribution converges weakly in probability to $\nu_\gamma$. Fixed $\lambda > 0$ bounds all inverse eigenvalues, and the hypotheses hold with $a = m$, $b = m_2$. Therefore

$$\mathcal{E}_{r_d} \xrightarrow{\mathbb{P}} \chi(\gamma,\lambda)\tau(\rho). \tag{J.11}$$

For example, sample a paired-feature covariance, retain all its eigenvalues $s_j$, and independently Haar-rotate the eigenvectors of $\lambda I + S$. The resulting Hessian has exactly the original Hessian's spectrum but a different eigenvector distribution. The theorem applies to this rotated model. Transferring its conclusion to the unrotated sample requires a separate proof.

### J.6 Rank quantiles and asymptotic regimes

Because $\tau$ is continuous and strictly decreasing, the pointwise rank law and monotonicity of $\mathcal{E}_r$ imply that, for $0 < \varepsilon < \chi$,

$$r_\varepsilon/d \xrightarrow{\mathbb{P}} \rho_\varepsilon = \tau^{-1}(\varepsilon/\chi). \tag{J.12}$$

To prove this, evaluate the law at $\lfloor(\rho_\varepsilon - \delta)d\rfloor$ and $\lceil(\rho_\varepsilon + \delta)d\rceil$ for any sufficiently small fixed $\delta > 0$. The two error limits lie strictly on either side of $\varepsilon$, forcing $r_\varepsilon/d$ into that interval with probability tending to one. For $\varepsilon > \chi$, $\mathcal{E}_1 \to \chi < \varepsilon$ so rank one suffices with probability tending to one. No analogous conclusion at equality follows from convergence alone.

The scalar regimes of $\tau$ and $\chi$ are derived in Section E.5. The same scalar formulas occur in the Gaussian-pair theorem, whose proof does not use orthogonal invariance.

## K Experiments

The experiments evaluate the deletion identities, compare finite-dimensional approximation errors with the asymptotic law, and examine rank selection and alternative feature distributions. Installation and reproduction commands are provided in the `README.md` included in `code_lora.zip`.

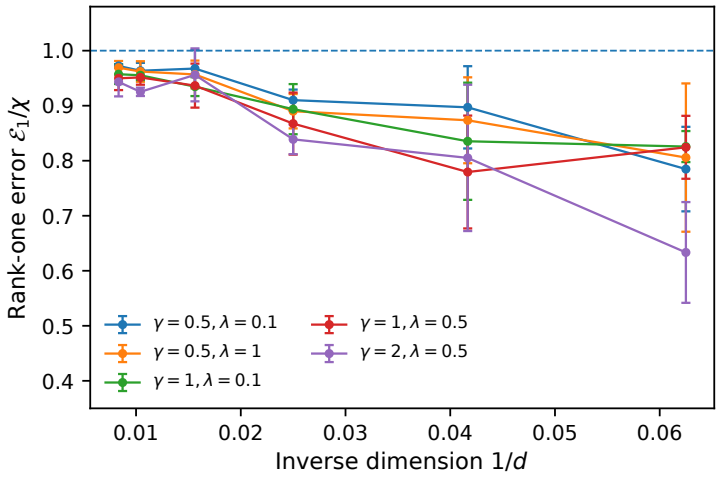


Figure 4: Gaussian features. Rank-one error divided by $\chi$ against inverse dimension. The horizontal line is the asymptotic limit, not a finite-sample lower bound.

### K.1 Protocol

The baseline grid uses independent standard Gaussian feature pairs, dimensions $d \in \{16, 24, 40, 64, 96, 120\}$, parameter pairs $(\gamma, \lambda) \in \{(0.5, 0.1), (0.5, 1), (1, 0.1), (1, 0.5), (2, 0.5)\}$, and seeds $1701, 1702, 1703$. For each configuration, $n = \mathrm{round}(d^2/\gamma) + 1$ and $N = n - 1$ retained observations are generated. The theoretical formulas use $d^2/n$. The random-number generator is NumPy's `default_rng`. The draw order is the retained $X$ array, retained $Z$ array, forgotten $x_{i_0}$, and forgotten $z_{i_0}$. Identical seeds at a fixed dimension and sample size reuse features across regularization values. Different seeds provide independent replicates within each setting.

We solve for $M = \mathcal{H}_R^{-1}[uv^\top]$, where $u = x_{i_0}/\|x_{i_0}\|$ and $v = z_{i_0}/\|z_{i_0}\|$. The scalar deleted residual cancels from relative errors. Conjugate gradients use relative tolerance $10^{-11}$, zero absolute tolerance, and at most 3000 iterations. A nonzero solver status or excessive explicitly recomputed residual raises an error. The largest relative residual over all 744 correction solves was $9.993 \times 10^{-12}$.

Computations use float64 arithmetic on CPU.

The experiments contain 90 baseline solves, four high-regularization solves, 270 rank-decision solves, 240 non-Gaussian or correlated solves, 120 curvature-interpolation solves, and 20 larger-dimensional solves. The resulting 744 solves reuse features in the regularization and curvature comparisons, so they are not 744 independent samples. Unless stated otherwise, plotted dispersion is one sample standard deviation with denominator one less than the number of seeds. The protocol was fixed before executing the grids.

### K.2 Gaussian rank errors and spectra

To summarize agreement with the asymptotic error curve, we use $d = 120$, the five baseline parameter pairs, and eight interior ranks $r = 12, 24, \ldots, 96$. Let $\bar{E}_r$ denote the mean oracle error over the three seeds. For each parameter pair and rank, we compute the absolute relative discrepancy from the value predicted by the limit. The summary is the median of the resulting 40 values

$$\left| \frac{\bar{E}_r}{\chi \tau(r/d)} - 1 \right|. \tag{K.1}$$

Its computed value is $0.0408712$, which rounds to $4.1\%$. No coefficient is fitted. The plot for the remaining block uses the prespecified seed $1701$ and divides its squared singular values by their empirical mean.

The deterministic construction at $d = 80$, $\kappa = 4$ gives $\mathcal{E}_r = 0.36(80 - r)/79$ for $r \geqslant 1$, while the operator with the gradient direction as an eigenvector gives zero error. A direct SVD verifies this

formula and the two eigenvalues of the inverse Hessian. This comparison holds the full spectrum fixed. It is separate from the random Hessian experiment below.

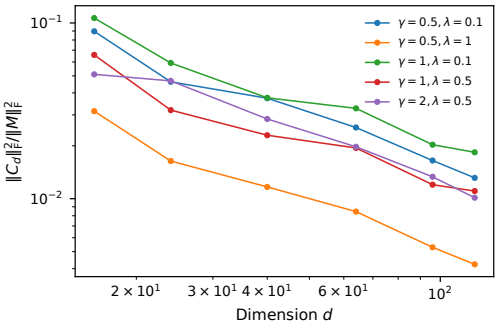


Figure 5: Squared Frobenius norm of the cross blocks divided by that of the correction, averaged over three draws. This ratio is nonzero at the tested finite dimensions.

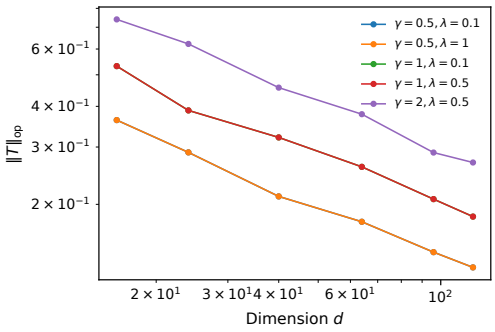


Figure 6: Operator norm of the auxiliary matrix, averaged over three draws. Settings with the same features and different regularizations coincide because this matrix does not depend on $\lambda$.

The high-regularization experiment uses $d = 40$, nominal aspect ratio $\gamma = 1$, $\lambda \in \{1, 4, 16, 64\}$, rank eight, and seed 1901. All four solves share their features. Figure 7 separates the finite-$\lambda$ scalar factor from the remaining effects of finite dimension.

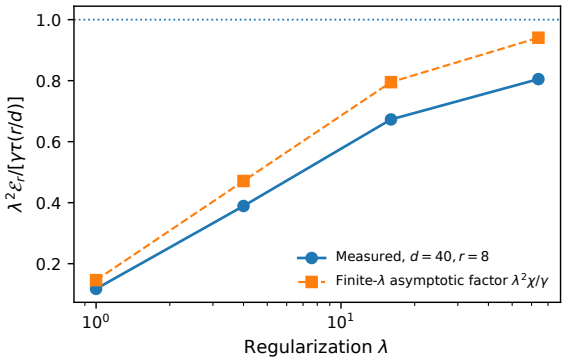


Figure 7: High-regularization scaling at $d = 40$, $r = 8$ and seed 1901. Both curves use the actual ratio $\gamma = d^2/n$. The measured ratio $\lambda^2\mathcal{E}_r/[\gamma\tau(r/d)]$ is compared with the scalar prediction $\lambda^2\chi/\gamma$. Their difference measures the remaining discrepancy at finite dimension. The horizontal line is the asymptotic limit one. The scalar prediction itself differs from one by $O(1/\lambda)$.

### K.3 Direct retraining, prediction error, and numerical bounds

We first compare the deletion formula with separately fitted original and retained models. This direct retraining experiment uses $d = 8$, $n = 128$, $\lambda = 0.2$, and seed 9129. Its draw order is all $n$ rows of $X$, all $n$ rows of $Z$, then $n$ independent Gaussian labels. The final observation is deleted and the original denominator $n$ is retained. Dense solutions for $W_D$ and $W_R$ agree with the deletion identity to relative error $5.46 \times 10^{-15}$. A matrix-free solve at tolerance $10^{-13}$ agrees with the dense correction to $4.51 \times 10^{-14}$. For the rank-two SVD approximation, 200,000 independent test pairs give mean squared prediction discrepancy $0.000938574$, compared with the exact Frobenius value $0.000934049$. The Monte Carlo standard error is $0.000004205$, and the relative discrepancy is $0.485\%$.

For example, at $d = 8$, $n = 128$ and $\lambda = 0.2$, fix the target relative squared prediction error at $0.13$. The separate CG tolerance controls the accuracy of the linear-system solve. At CG tolerance $10^{-2}$,

the rank-two error interval is $[0.1011319, 0.1533114]$, so the upper bound does not yet establish that rank two meets the target. Tightening the CG tolerance to $10^{-5}$ gives $[0.1250926, 0.1251263]$, entirely below the target. The rank-one lower bound is then $0.2028674 > 0.13$. These bounds identify two as the minimum sufficient rank at the reported numerical precision. The improvement comes from resolving numerical uncertainty about the same correction, not from changing its rank requirement. The direct-retraining comparisons above validate these calculations.

We additionally draw 20 datasets for direct retraining at $d = 16$, $n = 512$, $\lambda = 0.2$, using seeds 8001 through 8020. In each of the 21 instances, we evaluate three CG tolerances, $10^{-2}, 10^{-5}, 10^{-9}$, and four ranks, $1, 2, 4, d - 1$. All 252 lower bounds lie below their SVD oracle errors, and all upper bounds lie above the errors of the constructed truncated solves. Residuals and bounds use ordinary floating-point arithmetic, not outward-rounded interval arithmetic.

To compare the training and prediction objectives, we apply L-BFGS-B to the retained quadratic objective at rank two, using exact gradients and initialization at the SVD factorization. We use at most 3000 iterations, gradient tolerance $10^{-12}$, and relative objective tolerance $10^{-15}$. All 21 optimizer calls report success. In the $d = 8$ instance, the relative squared errors are

| Candidate | Relative Frobenius error | Relative Hessian error |
|---|---|---|
| Truncated SVD | 0.1251094 | 0.3295370 |
| Numerically optimized factors | 0.1876276 | 0.1920848 |

All 20 additional instances exhibit the same direction of change. This comparison demonstrates different objectives selecting different candidates. It does not prove a global minimum for factor optimization.

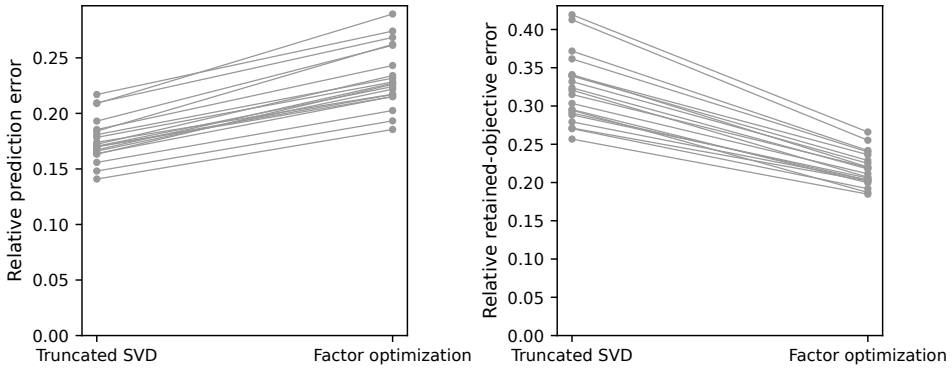


Figure 8: Paired rank-two candidates on 20 direct-retraining instances. Each line joins two candidates for the same data. In these instances, improving the retained objective increases prediction discrepancy relative to the SVD optimum.

### K.4 Rank selection at finite dimension

We use dimensions $d \in \{40, 80, 120\}$, pairs $(\gamma, \lambda) \in \{(0.5, 0.1), (1, 0.1), (1, 0.5)\}$, and seeds 2701 through 2730, for 270 solves. Each setting has 30 independent seed replicates. For each correction, the target tolerances are $\varepsilon = c\chi$ with $c \in \{0.1, 0.25, 0.5, 0.9, 1.1\}$. We compare three choices: the smallest positive integer with $\chi\tau(r/d) \leqslant \varepsilon$, the exact SVD minimum, and the smallest rank whose numerical upper bound is at most $\varepsilon$.

The asymptotic rule selects a rank whose oracle error exceeds the target in 22 of the 1350 evaluated decisions. This is a descriptive count, not a binomial sample of size 1350: tolerances within a draw and settings with shared features are dependent. Figure 9 reports all 45 setting-specific counts. No setting has more than two failures among 30 seeds. With zero failures among 30 independent seeds, the Wilson score interval at nominal level $95\%$ has upper endpoint approximately $11.4\%$ (Wilson, 1927). These samples therefore cannot establish negligible failure probability. Figure 10 gives the corresponding intervals at $d = 120$ and shows the selected rank minus the exact minimum.

In all 1350 decisions, the numerical upper bound at the selected rank is below the target and the lower bound at the preceding rank is above it. The two bounds therefore identify the same minimal rank as the exact SVD oracle. Rank selection here uses a computed full correction and its spectrum.

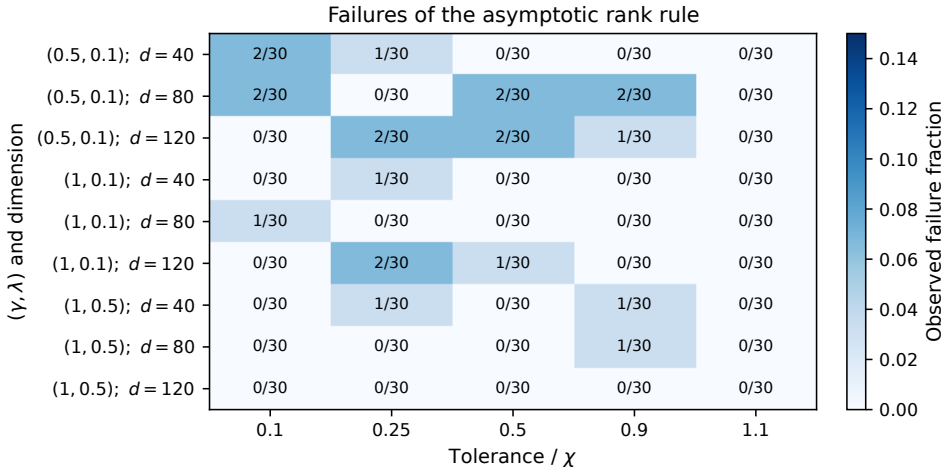


Figure 9: Failure counts of the asymptotic rank rule for every tested setting. Each cell has 30 independent seeds. Cells share samples across tolerances and some parameter settings.

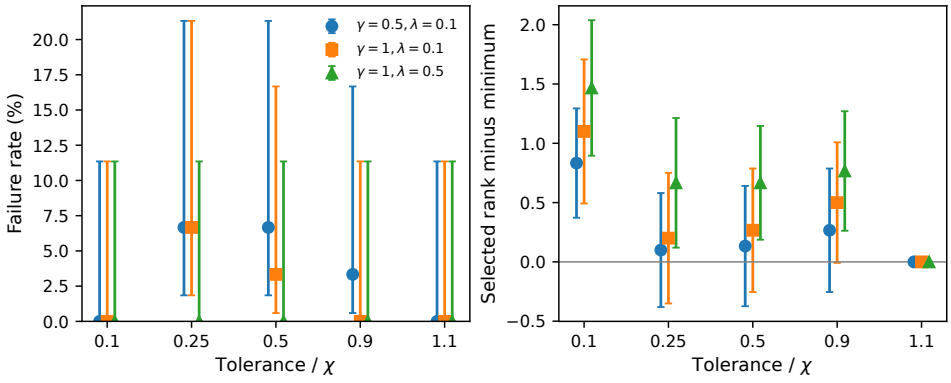


Figure 10: Rank decisions at $d = 120$. Left: failure frequencies with Wilson 95% intervals over 30 seeds per parameter pair. Right: selected rank minus the exact minimum, with mean $\pm$ one sample standard deviation.

### K.5 Curvature and the gradient direction

To vary curvature while holding the rank-one query and sample fixed, write $S = n^{-1} \sum_{i \in R} \operatorname{vec}(X_i) \operatorname{vec}(X_i)^\top$ and consider

$$H_t = \lambda I + (1 - t) \operatorname{diag}(S) + tS, \qquad t \in \{0, 0.25, 0.5, 1\}. \tag{K.2}$$

Here $\operatorname{diag}(S)$ denotes the diagonal matrix with the same diagonal as $S$. We use $d \in \{40, 80\}$, $\gamma = 1$, $\lambda = 0.5$, and seeds 2701 through 2720. The $t = 1$ results reuse the rank-decision grid, so this adds 120 solves. At $d = 80$, mean rank-one errors are respectively $0.000260$, $0.0206$, $0.0683$, and $0.2357$. The diagonal operator is not asserted to preserve rank exactly. Condition numbers are not held fixed here, unlike in the construction with identical eigenvalues.

We also compare the rank-one SVD oracle with three approximations to the exact $M$: its best scalar multiple of the raw query $uv^\top$, the diagonal-Hessian solution, and the solution of $\lambda A + \widehat{\Sigma}_x A \widehat{\Sigma}_z =$

$uv^\top$, where each empirical covariance uses denominator $N$. The last two candidates can have rank larger than one. They assess approximate curvature, not additional rank-one algorithms. None is an implementation of LoKU or Obliviate.

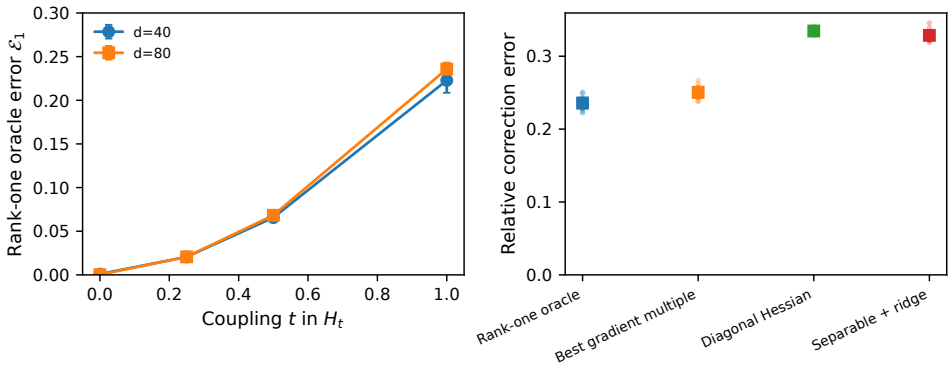


Figure 11: Curvature comparisons using 20 paired seeds. Left: adding off-diagonal entries to the Hessian changes the rank-one approximation error. Right: individual errors and mean ± sample standard deviation for correction approximations at $d = 80$.

### K.6 Other feature distributions and the prediction metric

We add 240 solves with standardized Rademacher, uniform on $[-\sqrt{3}, \sqrt{3}]$, Student with five degrees of freedom scaled by $\sqrt{3/5}$, and correlated Gaussian features. The latter have covariance $\Sigma_{ij} = 0.7^{|i-j|}$. In all cases $x$ and $z$ remain independent. We use $d \in \{40, 80, 120\}$, $\gamma = 1$, $\lambda = 0.5$, and seeds 2701 through 2720. Gaussian controls reuse the corresponding 20 seeds from the rank-decision grid.

For non-isotropic features, the prediction metric is not the unweighted Frobenius norm. If $\Sigma = LL^\top$, the exact best relative rank-$r$ prediction error is obtained from the singular values of $L^\top ML$, because multiplication by invertible $L$ preserves rank. The correlated curves in Figure 12 use this metric. The Gaussian formula is shown only for comparison with the other feature distributions, since our theorem does not cover them.

At $d = 120$, mean rank-one prediction errors are $0.2391$ for Gaussian features, $0.2411$ for Rademacher, $0.2397$ for uniform, $0.2343$ for Student, and $0.1994$ for correlated Gaussian features. Positive errors persist in all tested settings. These finite grids neither establish universality nor quantify performance in neural networks.

### K.7 Larger dimensions and absolute correction size

Twenty additional solves use $d \in \{192, 256\}$, $\gamma = 1$, $\lambda \in \{0.1, 0.5\}$, and seeds $4701$ through $4705$. At $d = 256$, $\lambda = 0.5$, the ratio $\mathcal{E}_1/\chi$ has mean $0.9829$ and standard deviation $0.00665$. For each draw, we also take the largest absolute difference between the oracle error curve and $\chi\tau(r/d)$. Across draws, this maximum has mean $0.00172$ and standard deviation $0.000763$. Figure 13 reports both regularizations, including all draws at larger dimensions.

On the same draws, we set the deleted residual to one and compute $\|\Delta\|_{\mathrm{F}}^2 = \|x_{i_0}\|^2 \|z_{i_0}\|^2 \|M\|_{\mathrm{F}}^2 / n^2$. Figure 14 illustrates its decreasing absolute size and the normalization by $\gamma m_2/n$. This does not simulate the distribution of fitted residuals. It separates a positive relative rank cost from the absolute scale of one deletion.

### K.8 Numerical verification of the algebra

Nine dense instances use block dimensions $k \in \{8, 12, 20\}$ and pairs $(\gamma, \lambda) \in \{(0.5, 0.1), (1, 0.5), (2, 0.1)\}$. The seeds are $6101$ through $6109$. The numerical comparisons concern the Schur system, reconstruction of the remaining block, conditional covariance, both partial

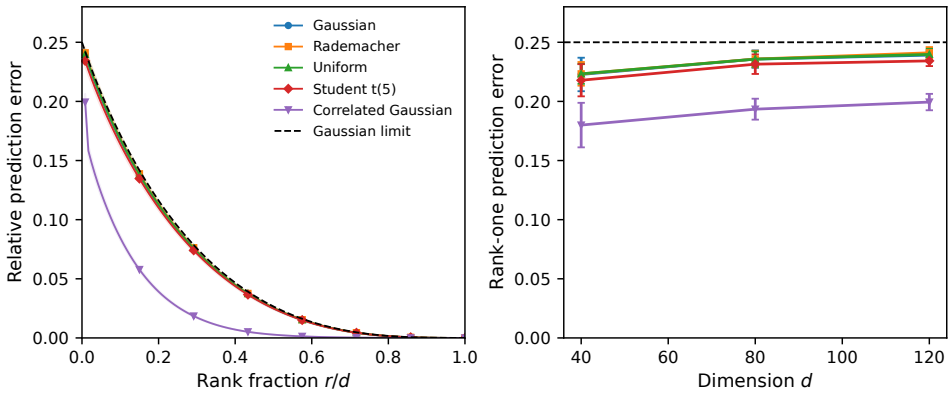


Figure 12: Prediction errors for five feature distributions. Left: full rank curves at $d = 120$. Right: rank-one errors versus dimension. Curves and bands or bars are means $\pm$ sample standard deviations over 20 seeds. The correlated case uses its exact covariance-weighted metric.

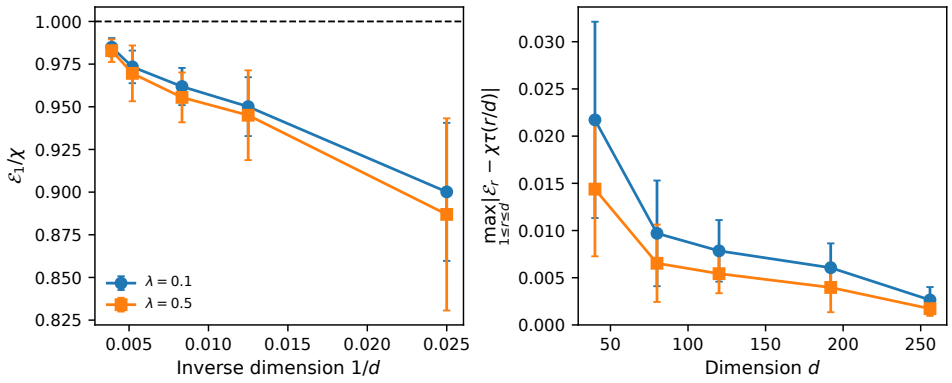


Figure 13: Finite-dimensional approach to the Gaussian law at $\gamma = 1$. Error bars are one sample standard deviation. There are 30 seeds per setting up to $d = 120$ and five seeds at each larger dimension.

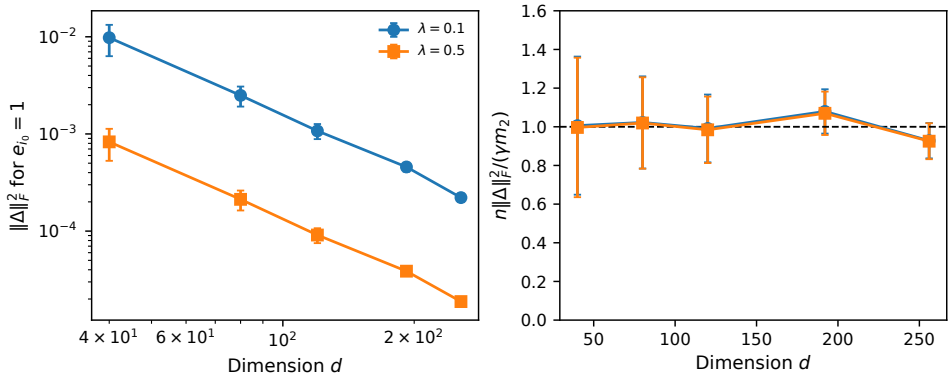


Figure 14: Absolute correction size with deleted residual fixed to one, and the same quantity divided by its predicted asymptotic scale. Sample sizes and uncertainty conventions match Figure 13.

traces, and the derivative in Lemma S5. The largest relative algebraic discrepancy is $6.47 \times 10^{-14}$. A central directional finite difference with step $10^{-5}$ differs from the analytic derivative by at most $3.95 \times 10^{-8}$. These computations verify algebra and indexing on the sampled instances, not asymptotic statements. The independent identity test also compares the Hessian with a dense Kronecker

representation and verifies spectral-tail integration, the construction with identical eigenvalues, and the covariance-weighted prediction metric.

The additional validation in Sections L–M uses 540 new Gaussian correction solves, independently of the 30 calibration draws in each setting. Its protocol was fixed before those new simulations. Thus the two studies together contain 1284 correction solves, alongside the 21 direct retraining comparisons and nine dense algebra verifications. The largest relative residual among the new solves is $1.00 \times 10^{-11}$ at the displayed precision.

# L Absolute Error and Interpretation

The rank law describes the fraction of the correction missed by a low-rank update. Its practical meaning also depends on the absolute size of the correction. In the Gaussian setting with fixed $\lambda > 0$, the exact deletion identity and the product-Gaussian norm law give

$$\Delta = \frac{e_{i_0} \|x_{i_0}\| \|z_{i_0}\|}{n} M, \qquad \frac{\|x_{i_0}\|^2 \|z_{i_0}\|^2}{d^2} \xrightarrow{\mathbb{P}} 1. \tag{L.1}$$

Therefore, wherever $e_{i_0} \neq 0$,

$$\frac{n \|\Delta\|_{\mathrm{F}}^2}{e_{i_0}^2} = \frac{d^2}{n} \frac{\|x_{i_0}\|^2 \|z_{i_0}\|^2}{d^2} \|M\|_{\mathrm{F}}^2 \xrightarrow{\mathbb{P}} \gamma m_2. \tag{L.2}$$

At $e_{i_0} = 0$, define the quotient by the cancelled expression $\|x_{i_0}\|^2 \|z_{i_0}\|^2 \|M\|_{\mathrm{F}}^2 / n$. The exact correction itself is then zero. The ratio does not require independence of $e_{i_0}$ and $M$, because the residual cancels exactly. If $e_{i_0} = O_{\mathbb{P}}(1)$, the squared effect is $O_{\mathbb{P}}(n^{-1})$. A positive relative lower bound on the approximation error can thus coexist with vanishing absolute prediction mismatch.

The target is obtained by retraining over all $d \times d$ matrices. If $W_D - W_0$ and $W_R - W_0$ are each constrained to rank $r_0$ on the same frozen base, their difference has rank at most $2r_0$. The results do not contradict that elementary bound. They concern low-rank edits to a model whose retrained optimum is not itself restricted to that adapter class.

The rank requirement applies when one seeks to reproduce a sufficiently accurate fraction of the correction produced by this retraining. It is not a universal lower bound for every definition of forgetting. A distributional comparison must specify what is observed. Guarantees for released weights, membership information, or downstream generative capabilities require their own mechanisms and metrics (Guo et al., 2020; Sekhari et al., 2021). The example below makes such a choice explicitly for noisy predictions.

The interpretation uses the theorem for Gaussian feature pairs in Section G, with its assumptions of fixed positive ridge, Gaussian features, and a deletion selected independently of the sample. The separate invariant-curvature model in Section J is not needed here.

## L.1 Choosing a tolerance from an observable prediction requirement

An application may specify an absolute root mean squared prediction tolerance $a > 0$. Under the isotropic test-feature model, this requires $\|B - \Delta\|_{\mathrm{F}} \leqslant a$. For $\Delta \neq 0$, the relevant relative squared tolerance is

$$\varepsilon := \frac{a^2}{\|\Delta\|_{\mathrm{F}}^2}. \tag{L.3}$$

If $\|\Delta\|_{\mathrm{F}} \leqslant a$, the zero update already suffices. Thus a fixed relative tolerance is justified only when the required accuracy resolves the deletion effect. Its value should come from the application rather than from the asymptotic curve.

A requirement can also compare the laws of several observable responses. Fix the training sample and consider $q$ independent test pairs $(x_j, z_j)$, with independent centered factors of covariance $I_d$. Here $q$ denotes the number of queries. A response is $x_j^\top W z_j + \xi_j$, where the noises $\xi_j \sim \mathcal{N}(0, \sigma^2)$ are mutually independent and independent of the features. Let $P_B^{(q)}$ and $P_R^{(q)}$ denote the joint distributions of all observed feature pairs and responses, using $W_D + B$ and $W_R$, respectively. Both mechanisms use the same noise variance and query distribution.

Table 1: Noisy-response example on 60 new draws per dimension, $\gamma = 1$, $\lambda = 0.5$, deleted residual one, $q = n$, and $\sigma = 1$. Entries are mean $\pm$ sample standard deviation. The target divergence is $0.05$. The last column is a sufficient rank chosen with the calibrated rule of Section M and the gradient norm bound, without using the new correction to choose the rank.

| $d$ | No-update divergence | Rank-one divergence | Minimum rank | Selected rank |
|---|---|---|---|---|
| 40 | $0.700 \pm 0.238$ | $0.153 \pm 0.052$ | $11.5 \pm 3.0$ | $33.8 \pm 6.9$ |
| 80 | $0.651 \pm 0.157$ | $0.153 \pm 0.037$ | $22.8 \pm 4.4$ | $48.2 \pm 5.0$ |
| 120 | $0.689 \pm 0.127$ | $0.165 \pm 0.030$ | $36.1 \pm 4.8$ | $65.3 \pm 3.9$ |

**Proposition S12** (Discrepancy of noisy prediction laws)**.** *For $\sigma > 0$ and fixed matrices $W_D, W_R, B$,*

$$\mathrm{KL}\Big(P_B^{(q)} \,\|\, P_R^{(q)}\Big) = \frac{q}{2\sigma^2}\,\|B - \Delta\|_{\mathrm{F}}^2\,. \tag{L.4}$$

*Consequently, for $\Delta \neq 0$, the minimum divergence over corrections of rank at most $r$ is $q\,\|\Delta\|_{\mathrm{F}}^2\,\mathcal{E}_r/(2\sigma^2)$. A required divergence at most $\zeta > 0$ corresponds to the relative squared tolerance*

$$\varepsilon := \frac{2\sigma^2\zeta}{q\,\|\Delta\|_{\mathrm{F}}^2}. \tag{L.5}$$

*Proof.* Condition on all query pairs. The response laws are Gaussian with common covariance $\sigma^2 I_q$, so their divergence is

$$\frac{1}{2\sigma^2}\sum_{j=1}^{q}\big(x_j^\top (B - \Delta) z_j\big)^2.$$

The two mechanisms have identical feature marginals. The chain rule for divergence therefore averages this conditional expression over the features. Independence and identity covariance give $\mathbb{E}[(x^\top A z)^2] = \|A\|_{\mathrm{F}}^2$. Minimizing over rank uses the singular-value truncation identity. □

For a fixed number of responses and bounded deleted residual, even the zero update has divergence tending to zero. If instead $q = n$, $e_{i_0} = 1$, and $\sigma = 1$, its divergence converges to $\gamma m_2/2$. At $(\gamma, \lambda) = (1, 0.5)$, this limit is $2/3$, and the best rank-one divergence tends to $1/6$. For a fixed requirement $\zeta = 0.05$, the minimum sufficient rank fraction therefore converges to

$$\tau^{-1}\left(\frac{2\zeta}{\gamma m_2 \chi}\right) = \tau^{-1}(0.3) \simeq 0.302.$$

The single-response discrepancy vanishes while the law of the complete response sequence remains sensitive to the approximation. This example specifies an observable, a noise level, a number of responses, and a fixed tolerance. It does not assert that these choices describe any particular deployed system. The query pairs here are independent draws, rather than adaptively chosen requests, and the comparison concerns response laws rather than released parameters.

Table 1 uses the independent test draws specified in Section M. For one response, all no-update divergences are below $0.00077$, already below the chosen $0.05$ tolerance. For $n$ responses, every best rank-one correction misses that tolerance. The last column is conservative. Its oracle corrections meet the requirement in all 180 draws, with maximum observed divergence $0.0141$. Fixing $e_{i_0} = 1$ normalizes the deletion size and does not estimate the distribution of fitted residuals.

## M Finite-Dimensional Rank Selection by Independent Calibration

Theorem S3 gives a limit but no finite-dimensional probability bound on its error. We obtain such a bound by measuring the approximation error on independent complete datasets at the same dimension. The construction uses the standard order-statistic argument underlying conformal calibration (Vovk, 2012). We apply that established principle to an entire rank-approximation curve so that one calibrated bound can support every tolerance choice.

### M.1 A SIMULTANEOUS UPPER BOUND FOR ALL RANKS

Fix $d \geqslant 2$, $n$, $\lambda > 0$, and a deletion rule that does not depend on the observed sample. Let $b$ calibration datasets and one new dataset be independent with the same feature distribution. Each dataset gives a nonzero correction direction and errors $\mathcal{E}_r^{(j)}$, for $j \in [\![1, b+1]\!]$. In the paired model these errors depend only on the features, so calibration does not require simulated labels. To approximate the error at each rank, define the fixed curve

$$g_r := \chi(d^2/n, \lambda)\tau(r/d), \qquad 1 \leqslant r < d.$$

The argument applies to any fixed curve and requires no known finite-dimensional accuracy for $g_r$. To control every rank simultaneously, assign dataset $j$ a score equal to its largest positive discrepancy above this curve. For the $b$ calibration datasets, compute upper bounds $U_{j,r} \geqslant \mathcal{E}_r^{(j)}$ and use their largest discrepancy as the adjustment. Thus define, with $S_j$ also defined for the new dataset $j = b+1$,

$$S_j := \max_{1 \leqslant r < d} (\mathcal{E}_r^{(j)} - g_r)_+, \qquad c_b := \max_{j \leqslant b} \max_{1 \leqslant r < d} (U_{j,r} - g_r)_+. \tag{M.1}$$

Exact SVD errors may be used for $U_{j,r}$. The residual-based upper bound from Section C is also valid and accounts for an approximate correction solve. For a tolerance $\varepsilon \geqslant 0$, select the first rank whose adjusted curve value meets the tolerance. If no rank below $d$ does, use full rank. Formally, define

$$\widehat{r}(\varepsilon) := \min\big(\{r \in [\![1, d-1]\!] : g_r + c_b \leqslant \varepsilon\} \cup \{d\}\big). \tag{M.2}$$

The fallback $d$ uses the exact identity $\mathcal{E}_d = 0$ and is available even when $\varepsilon < c_b$.

**Proposition S13** (Calibrated rank probability)**.** *Under the preceding assumptions, with probability over the calibration datasets and the independent new dataset,*

$$\mathbb{P}\Big(\exists \varepsilon \geqslant 0 : \mathcal{E}_{\widehat{r}(\varepsilon)}^{(b+1)} > \varepsilon\Big) \leqslant \frac{1}{b+1}. \tag{M.3}$$

*Moreover, define $p(c) := \mathbb{P}(S_{b+1} > c)$, with probability taken over a new dataset at a fixed value of $c$. For every $\alpha \in ]0, 1[$,*

$$\mathbb{P}_{\mathrm{cal}}\big(p(c_b) > \alpha\big) \leqslant (1-\alpha)^b. \tag{M.4}$$

*The guarantees hold with ties and without continuity assumptions.*

*Proof.* On $S_{b+1} \leqslant c_b$, the upper bound $\mathcal{E}_r^{(b+1)} \leqslant g_r + c_b$ holds simultaneously for every rank below $d$. Equation (M.2) then meets every tolerance, including one chosen after inspecting the new dataset. The rank-$d$ case follows from its zero error. Since $c_b \geqslant \max_{j \leqslant b} S_j$, failure requires the new score to exceed every calibration score. Among $b+1$ exchangeable scores, the probability that a specified one is the strict largest is at most $1/(b+1)$. Ties can only reduce this probability.

For the second claim, let $F$ be the common cumulative distribution function of the exact scores. The random variable $F(S_j)$ satisfies $\mathbb{P}(F(S_j) < t) \leqslant t$ for all $t \in [0, 1]$, including distributions with atoms. If $p(c_b) = 1 - F(c_b) > \alpha$, then all $S_j \leqslant c_b$ obey $F(S_j) < 1 - \alpha$. Independence bounds this event's probability by $(1-\alpha)^b$. □

With $b = 30$, the failure probability, averaged over both the calibration datasets and the independent new dataset, is at most $1/31$, approximately $3.23\%$. A different statement holds with probability at least $95\%$ over the calibration datasets: after these datasets are fixed, the probability over an independent new dataset that its error curve exceeds the calibrated upper curve at any rank below $d$ is at most $1 - 0.05^{1/30} \simeq 9.50\%$. Both statements concern a fixed parameter setting. Neither conditions on a particular retained dataset or justifies pooling dependent deletions from the same trained model.

### M.2 ABSOLUTE REQUIREMENTS WITHOUT THE NEW CORRECTION NORM

To choose a rank before solving the new correction, its norm may be bounded by

$$\|\Delta\|_{\mathrm{F}} \leqslant A := \|G\|_{\mathrm{F}} / \lambda.$$

Table 2: Independent validation of the calibrated rule. Failures refer to 300 tolerance decisions per row. Exceedances count the 60 draws whose exact curve exceeds the calibrated upper curve at any interior rank. Extra rank is the mean increase relative to the unadjusted rule.

| $d$ | $\gamma$ | $\lambda$ | $c_b$ | Raw failures | Adjusted failures | Exceedances | Extra rank |
|---|---|---|---|---|---|---|---|
| 40 | 0.5 | 0.1 | 0.02139 | 1 | 0 | 3 | 1.6 |
| 40 | 1 | 0.1 | 0.04291 | 1 | 0 | 0 | 2.6 |
| 40 | 1 | 0.5 | 0.02577 | 1 | 0 | 0 | 5.2 |
| 80 | 0.5 | 0.1 | 0.02331 | 4 | 0 | 0 | 4.2 |
| 80 | 1 | 0.1 | 0.01749 | 3 | 0 | 0 | 1.8 |
| 80 | 1 | 0.5 | 0.01262 | 3 | 0 | 1 | 2.8 |
| 120 | 0.5 | 0.1 | 0.00822 | 11 | 0 | 3 | 2.0 |
| 120 | 1 | 0.1 | 0.00676 | 6 | 0 | 3 | 1.0 |
| 120 | 1 | 0.5 | 0.00423 | 3 | 0 | 1 | 1.2 |

The quantity $A$ uses only the deleted-example gradient and the regularization parameter. If $A = 0$, the correction is zero. More generally, if $qA^2/(2\sigma^2) \leqslant \zeta$, the zero update already meets the response-law requirement of Section L. Otherwise $A > 0$, and a sufficient relative tolerance is

$$\varepsilon_A := \frac{2\sigma^2\zeta}{qA^2}. \tag{M.5}$$

On the simultaneous event in the proof, the best rank-$\widehat{r}(\varepsilon_A)$ correction has divergence at most $q\,\|\Delta\|_{\mathrm{F}}^2\,\varepsilon_A/(2\sigma^2) \leqslant \zeta$. The simultaneous event covers this choice even though $\varepsilon_A$ depends on the new data through $A$. Rank selection therefore avoids solving the new correction. Producing the corresponding best update remains a separate optimization problem.

### M.3 Independent numerical validation

For each of the nine original $(d, \gamma, \lambda)$ settings, the 30 draws with seeds 2701–2730 provide calibration. We then generate 60 new draws with seeds 6701–6760 per setting, for 540 new correction solves. The curve, adjustment based on the maximum score, five tolerances $\chi \times \{0.1, 0.25, 0.5, 0.9, 1.1\}$, and test seeds were fixed before these new runs. Each row of Table 2 uses 60 independent datasets and 300 dependent rank decisions. Shared seeds across settings also prevent treating the complete table as one independent binomial sample.

The original rule misses 33 of 2700 tested tolerances, while the adjusted rule misses none. However, 11 of the 540 complete curves exceed their calibrated upper curve somewhere. These observations concern different events: the five tested tolerances do not exhaust all possible rank decisions. The probability bound allows failures, so the observed zero count at those tolerances does not establish success for every tolerance. The adjustment increases the selected rank, especially when the desired tolerance is below $c_b$. At $(d, \gamma, \lambda) = (40, 1, 0.5)$, the smallest tolerance invokes full rank on all 60 test draws. No such fallback occurs at the tested tolerances in the other settings.

Calibration requires full correction curves for independent datasets from the same distribution. This cost can be shared across repeated uses of a fixed setting. For a single deletion, calibration adds these full correction solves. The guarantee concerns the oracle error at the selected rank, not the error of an arbitrary factor optimization algorithm. The recorded bounds use floating-point residuals without outward interval rounding. An explicit analytic rate in $(d, n, \lambda)$ for the original Gaussian approximation, without a calibration sample, remains open.

## N Experiments on Frozen Representations

We first study bilinear heads on pretrained text representations, then a softmax head on image representations in Section N.9. Feature extractors remain fixed. The text experiments measure score differences and binary decisions after retraining. The score comparisons cover individual and grouped deletions and define the finite metric used in the lower bound below. Binary decisions use separate single deletions and evaluation pairs.

### N.1 MODELS, DATA AND PREPROCESSING

We use the public, unquantized ONNX versions of `all-MiniLM-L6-v2`, `bge-small-en-v1.5`, and `e5-small-v2`, with CPU inference and 384-dimensional output representations (Sentence Transformers, n.d.a; BAAI, n.d.; Intfloat, n.d.). MiniLM and E5 use attention-masked mean pooling, while BGE uses the CLS token. All embeddings are normalized to unit Euclidean norm. E5 uses the documented `query:` prefix on both texts. The token limits are 256 for MiniLM and 512 for BGE and E5. Immutable model and dataset revisions, download instructions and inference settings are recorded in the code archive.

For STSB, we use the 5749 training, 1500 validation and 1379 test pairs in the public release (Sentence Transformers, n.d.b). Scores are restored to the original $[0, 5]$ scale. The primary test evaluation excludes seven ordered pairs also present in training, leaving 1372 pairs. We also report results on the full test set to assess the effect of this filtering. This initial filter does not guarantee removal of every reversed duplicate.

The SICK release (Marelli et al., 2014) contains 4439 training, 495 validation and 4906 test pairs, with scores in $[1, 5]$. The supplied split labels are preserved. Removing 39 test pairs that occur in training or validation, in either order, leaves 4867 primary test pairs. Individual sentences remain shared across splits. In particular, 4831 of the 4906 original test pairs contain at least one sentence found in training or validation.

For the QQP score comparisons (Wang et al., 2018), we select 8000 training pairs and 1000 separate preprocessing pairs from the official training split, and 2000 evaluation pairs from the official validation split. The public test labels are unavailable, so these are not test-leaderboard results. For PAWS-Wiki (Zhang et al., 2019), the same sample sizes use the official training, validation and test roles. Sampling uses fixed permutations without looking at predictions. Exact or reversed duplicate pairs are removed within source splits before sampling. Conflicting labels occur for 31 canonical training pairs and seven test pairs in the PAWS source data. The first occurrence is retained under the fixed rule. The selected evaluation pairs have no exact or reversed overlap with the training or preprocessing source partitions. The selected QQP and PAWS pairs do not overlap each other or STSB and SICK by this criterion. On QQP, 200 of 2000 evaluation pairs share an individual sentence with the selected training or preprocessing data. This count is zero on PAWS. These comparisons do not establish absence from encoder pretraining.

For each corpus, the preprocessing pairs fit a single PCA basis and whitening transform using the two sentence positions together. The same pairs fit an affine function of the original embedding cosine. Writing this fixed baseline as $b_i$, the head prediction is $b_i + x_i^\top W z_i$. The matrix $W$ minimizes the quadratic objective in Section 3, with response $y_i - b_i$. Thus the exact deletion identity still applies. QQP and PAWS have binary targets. Their scores are neither clipped to $[0, 1]$ nor interpreted as calibrated probabilities. Accuracy uses a fixed threshold of $0.5$.

### N.2 HEADS, DELETION REQUESTS AND CANDIDATE UPDATES

Each corpus uses dimensions $d \in \{16, 32\}$, sample size $n = 2d^2$, regularization $\lambda \in \{0.1, 1\}$ and three training subsamples, for 12 heads. The training and deletion seeds are fixed in the archive. For each head we select 20 single deletions. Ten of these indices anchor groups of size 4 and 16. Groups can overlap. The score comparison has 480 requests per corpus, hence 1920 over the four encoder–corpus settings. The experiments change encoder and corpus together, so differences between settings cannot be attributed to the encoder alone.

The denominator $n$, regularization, encoder, preprocessing and affine baseline remain fixed after deletion. We compare no update, exact head retraining, truncated SVD of its correction, and two factorized candidates at ranks 1, 2, 4 and 8. One candidate is fitted by numerically minimizing the retained quadratic objective over its factors. The other matches the exact retrained predictions on 2048 input pairs outside the head-fitting subset. These pairs are drawn from the training pool and exclude pair duplicates of fitted observations according to the dataset's pairing convention. Matching requires no labels, but does require oracle access to the retrained head. It is unrelated to the calibration across independent datasets in Section M.

Both factorized objectives use three initializations: balanced factors from truncated SVD, a 10% Gaussian perturbation of those factors, and random factors. L-BFGS-B uses at most 500 iterations,

Table 3: Single deletions at $d = 32$, $\lambda = 1$. Relative squared prediction discrepancies (%), averaged within each head and then across three heads, with 60 requests per dataset. All selected retained-objective candidates shown have successful optimizer status. Complete configurations and optimization diagnostics are given below.

| Update | MiniLM/STSB | BGE/SICK | E5/QQP |
|---|---|---|---|
| SVD, rank 1 | 36.0 | 26.9 | 27.4 |
| Retained objective, rank 1 | 20.3 | 16.2 | 18.0 |
| Retained objective, rank 2 | 11.7 | 9.4 | 11.0 |
| Retained objective, rank 4 | 6.1 | 4.8 | 5.7 |
| Retained objective, rank 8 | 2.6 | 2.0 | 2.4 |
| Prediction matching, rank 1 | 20.8 | 16.1 | 18.0 |

40 line-search steps, gradient tolerance $10^{-8}$, function tolerance $10^{-12}$ and memory size 10. The best candidate is selected by its optimization objective, including the SVD initialization as a fallback. The test set never selects a candidate. For score comparisons, each corpus therefore has 11520 optimizer starts and 3840 selected optimized candidates. An unsuccessful convergence status remains recorded even when a finite candidate can be evaluated.

### N.3 Metrics, aggregation and complete results

Equation (21) measures relative squared prediction discrepancy on evaluation pairs. The corresponding absolute measure is the root mean square of $x_i^\top (B - \Delta) z_i$ over the same pairs. Frobenius error, retained-objective excess and prediction-matching error are stored separately. Truncated SVD is an oracle for the Frobenius metric, not for a general empirical prediction metric. Factorized numerical candidates do not determine the minimum achievable rank.

Relative errors are averaged over requests within a head and then over three heads. Ranges between head means describe these three subsamples, not confidence intervals. Absolute RMS summaries below are the median of the three within-head medians. Counts report the actual requests, without treating overlapping groups or shared corpora as independent observations. The raw RMS values use score ranges of five for STSB, four for SICK and one for QQP and PAWS. A threshold of $0.01$ therefore has different normalized meanings across corpora.

At $d = 32$ and $\lambda = 1$, the numbers of single-deletion requests meeting relative tolerance $0.1$ with retained-objective factors of rank one are $2/60$, $5/60$, $1/60$ and $1/60$ on STSB, SICK, QQP and PAWS. At rank four, the counts are $60/60$, $60/60$, $60/60$ and $51/60$. The absolute no-update RMS summaries are $0.002058$, $0.001280$, $0.001743$ and $0.002088$. Already $55/60$, $60/60$, $60/60$ and $60/60$ requests meet an absolute tolerance of $0.01$ without an update.

For groups of 16 in the same setting, no-update RMS summaries are $0.01705$, $0.01280$, $0.009750$ and $0.009710$. Rank-eight retained-objective updates reduce them to $0.00478$, $0.00348$, $0.002958$ and $0.002818$. Their relative squared errors average $8.5\%$, $7.1\%$, $9.1\%$ and $9.1\%$. The corresponding counts below relative tolerance $0.1$ are $21/30$, $30/30$, $23/30$ and $21/30$. Group gradients have numerical rank 15 or 16 in these size-16 examples. They do not isolate the individual-deletion mechanism of the Gaussian theorem.

### N.4 Task performance across datasets and regularization levels

We report task performance alongside agreement with retraining. On STSB at $d = 32$, $\lambda = 1$, mean head RMSE is approximately $0.858$, versus $0.860$ for the affine cosine baseline. On SICK it is $0.57025$, versus $0.56888$. On QQP, mean head accuracy, AUROC and RMSE are $77.88\%$, $0.8549$ and $0.4016$, versus $77.60\%$, $0.8603$ and $0.4049$ for the baseline. Majority accuracy is $62.75\%$. The QQP head has higher accuracy and lower AUROC than the affine cosine baseline. These are descriptive differences, without a claim of statistical significance.

On PAWS, head accuracy is $52.90\%$, AUROC $0.5229$ and RMSE $0.5068$. Majority accuracy is $55.00\%$, and the affine cosine baseline reaches $55.15\%$ accuracy and AUROC $0.5858$. Prediction

agreement with retraining improves with update rank here as well. The task metrics and approximation errors quantify different properties of the fitted head.

We also evaluate $\lambda = 0.1$. At $d = 32$, STSB head RMSE is approximately $0.981$ and SICK RMSE is $0.63619$. QQP accuracy is $72.58\%$ with RMSE $0.4614$. PAWS accuracy is $51.60\%$ and AUROC $0.5035$. These settings assess how regularization affects both task performance and agreement with retraining. Complete tables include every dimension, regularization and group size.

### N.5 Numerical accuracy and optimizer convergence

Each corpus has 72 direct retained-data refits, covering both dimensions, both regularization values, all three heads, and all deletion sizes. Maximum relative errors against the exact correction identity are $1.83 \times 10^{-12}$, $4.17 \times 10^{-12}$, $1.41 \times 10^{-12}$ and $4.98 \times 10^{-13}$ for STSB, SICK, QQP and PAWS. These 288 refits assess numerical agreement using ordinary floating-point arithmetic. They do not provide rigorous interval bounds. Analytic factor gradients are compared with directional finite differences. Truncated SVD has the smallest Frobenius error among the tested candidates, and each selected optimized candidate has an objective no larger than its SVD initialization.

Of the 3840 selected candidates per corpus in the score comparison, 20 on STSB, nine on SICK, three on QQP and 268 on PAWS lack successful optimizer status. The STSB and SICK flags all concern prediction matching. On QQP, two concern prediction matching and one the retained objective, all at rank eight. On PAWS, 265 concern prediction matching at ranks four or eight, and three concern the retained objective at rank eight. No selected rank-one candidate is flagged. All selected retained-objective candidates at $d = 32$, $\lambda = 1$ used in the central comparison have successful status. Local convergence does not establish global optimality.

The same factorized objectives, initializations and candidate selection rule are used across the four encoder–corpus settings. Retraining changes the head while the encoder and preprocessing remain fixed. These empirical findings concern the studied heads and corpora. The exact Gaussian rank law applies under Assumption 1.

### N.6 Minimum Frobenius errors on the observed features

To relate the Gaussian rank law directly to the text representations, we compute the minimum error in the same Frobenius metric as the theorem. We retain the four encoder–corpus pairs, preprocessing, training subsets, $d \in \{16, 32\}$, $\lambda \in \{0.1, 1\}$ and three seeds used above. Every fitting observation is deleted in turn, giving 61440 single deletions across 48 heads. The original sample size remains the denominator after deletion.

Section B shows that a nonzero single-deletion correction is a scalar multiple of $\mathrm{mat}(H_D^{-1}\phi_i)$. This scalar cancels from the relative squared error. We therefore factor $H_D$ once per head, solve for each feature direction, and compute all rank minima from the singular values using (5). All computed residuals for the real-data fits are nonzero. The relative linear-system residuals are below $6.34 \times 10^{-15}$. The resulting error bound, divided by the computed direction's Frobenius norm, is below $4.0 \times 10^{-14}$. These numerical values use ordinary floating-point arithmetic.

We compare each feature sample with a random permutation of its second feature vectors and with independent standard Gaussian vectors of the same dimensions and sample size. The permutation preserves both empirical marginal distributions and changes their pairing. The permutation and Gaussian draws remain fixed across the two regularization values. These comparisons use the same direction calculation and give 184320 directions in total. We average errors within each sample and then across its three seeds.

At the setting of Figure 15, the mean rank-one errors are $12.35\%$, $11.45\%$, $10.67\%$ and $8.60\%$ for MiniLM/STSB, BGE/SICK, E5/QQP and E5/PAWS, respectively. Rank eight reduces them to $2.49\%$, $2.49\%$, $2.55\%$ and $1.93\%$. The independent Gaussian comparisons give rank-one means between $8.46\%$ and $8.51\%$, close to the theoretical value $8.53\%$. Shuffling the real pairs reduces the rank-one means to $6.55\%$, $5.98\%$, $6.22\%$ and $3.38\%$. This reduction occurs in all 16 combinations of corpus, dimension and regularization, showing that the observed pairing contributes to the rank-one approximation error in this experiment.

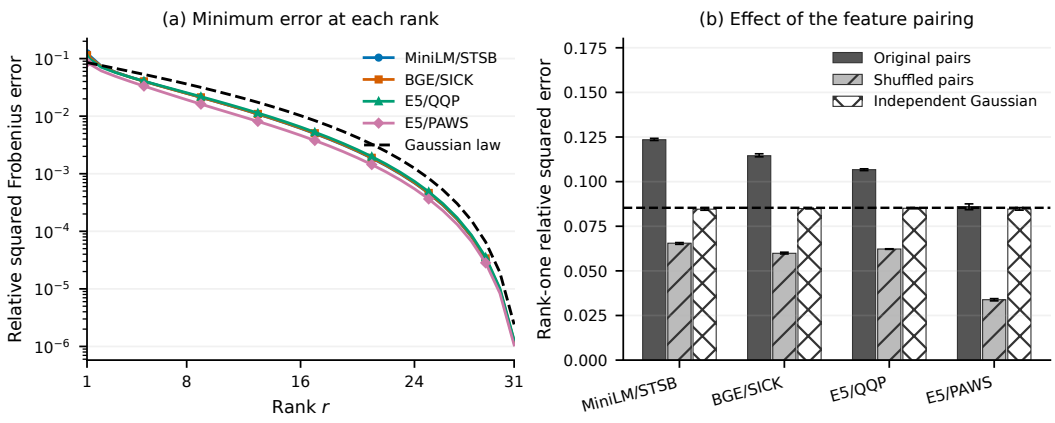


Figure 15: Minimum relative squared Frobenius errors at $d = 32$, $n = 2048$ and $\lambda = 1$. Left: means over all single deletions and three training subsets, compared with $\chi(0.5, 1)\tau(r/32)$ for $1 \leqslant r < 32$. At rank 32 the error is zero. Right: rank-one errors for the original pairs, shuffled pairs and independent Gaussian features. The dashed line is the Gaussian law at rank one. Shading and vertical intervals show the range of the three sample means, not confidence intervals.

These curves quantify the minimum error over all matrices satisfying the rank constraint, without fitting low-rank factors. The Gaussian law provides a comparison whose differences from the real-feature curves vary with the rank and setting. The metric here is Frobenius error. Prediction discrepancies on real evaluation pairs are reported separately above. The complete grid, per-deletion curves, numerical diagnostics and reproduction commands are in `code_lora.zip` and its `real_rank/README.md`.

### N.7 Absolute prediction errors and binary decisions

A relative discrepancy does not indicate how much predictions change in absolute terms. We therefore measure both the RMS prediction change caused by retraining, $a_T$, and the RMS discrepancy left by a correction $B$, $e_T(B)$. Define

$$a_T := \left( \frac{1}{|T|} \sum_{i \in T} (x_i^\top \Delta z_i)^2 \right)^{1/2}, \qquad e_T(B) := \left( \frac{1}{|T|} \sum_{i \in T} (x_i^\top (B - \Delta) z_i)^2 \right)^{1/2}. \tag{N.1}$$

Thus $e_T(0) = a_T$ and $\widehat{\mathcal{E}}_T(B) = e_T(B)^2 / a_T^2$ when $a_T > 0$. The absolute metric uses score units. To compare corpora, let $s$ be the target-score range: five on STSB, four on SICK, and one on QQP and PAWS. We require $e_T(B) \leqslant 10^{-3} s$, which sets a numerical tolerance of one thousandth of the target range for the root mean squared score discrepancy from retraining. It does not guarantee identical rounded scores or decisions. The scores for the binary tasks can fall outside $[0, 1]$.

This numerical tolerance provides a common comparison in score units. An application can specify a different requirement. Figure 16 gives the complete tolerance curves, and the raw-score threshold $0.01$ provides a looser comparison. No update is needed whenever $a_T$ already meets the chosen tolerance.

At $d = 32$, $\lambda = 1$, the tolerance requires an update for 11, 12, 43 and 58 of the 60 single deletions on STSB, SICK, QQP and PAWS. Among these requests, the retained-objective rank-one candidate meets it for 9, 11, 20 and 22 requests, respectively. Rank eight meets it for all requests requiring an update. Thus rank one fails and rank eight passes for 2, 1, 23 and 36 of all 60 requests. These are candidate comparisons. They do not rule out a better rank-one correction.

For groups of four, the corresponding counts where rank one fails and rank eight passes are 15, 8, 21 and 13 out of 30. For groups of 16 they are 17, 24, 0 and 0 out of 30. In particular, a rank-eight candidate need not meet the absolute tolerance even when it meets a relative tolerance. All 1920 requests in the score comparison and all its settings are displayed below and included as machine-

readable values in the archive. The plots sort by $a_T$ within each panel, using the same order for every rank. They do not select the largest effects.

**Changes in binary decisions.** QQP and PAWS assign class one at score at least $0.5$. Each corpus uses three E5 heads with $d = 32$ and $n = 2048$. For each head, we select regularization by classification accuracy on 2000 validation pairs over the grid $\{0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$. Ties are resolved by choosing the largest value. The selected values are 3 for all QQP heads and $3, 30, 30$ for the PAWS heads with training seeds $1701, 1702, 1703$, respectively.

Validation pairs come from the official QQP training and PAWS validation splits. They exclude every unordered pair in the training pool, preprocessing sample and score- and decision-evaluation samples. Individual sentences can remain shared. The encoder, preprocessing and affine baseline are fixed across regularization values. Selection uses validation accuracy alone, before evaluating deletion outcomes.

For each head, we sample 20 distinct training indices uniformly, excluding its 20 single-deletion indices from the score comparisons. We evaluate decisions on 5000 pairs from the official QQP validation and PAWS test splits. Each sample contains no duplicate unordered pairs and excludes the corpus's 8000 training, 1000 preprocessing and 2000 score-evaluation pairs. Individual sentences remain shared with those sets in 525 QQP and 3073 PAWS pairs. The fixed-regularization and validation-selected comparisons use the same evaluation pairs. The selected regularization, original denominator $n$, encoder, preprocessing and decision threshold remain fixed after deletion.

Retraining changes 268 QQP and 340 PAWS decisions over 300000 comparisons of requests and evaluation pairs per corpus. A recovered change matches retraining on a pair whose original decision changes. A missed change remains different from retraining. An introduced disagreement occurs where retraining leaves the original decision unchanged. Table 4 includes every request and pair. Recovery fractions use the 268 or 340 changed decisions as denominators. Total disagreements are the sum of missed changes and introduced disagreements.

Table 4: Retained-objective updates, with 60 single deletions and 5000 evaluation pairs per corpus. Recovered plus missed equals the number of decisions changed by retraining. Total disagreements equal missed plus introduced.

| Corpus | Rank | Recovered | Missed | Introduced | Total disagreements |
|---|---|---|---|---|---|
| QQP | 1 | 199 | 69 | 66 | 135 |
| QQP | 2 | 212 | 56 | 52 | 108 |
| QQP | 4 | 234 | 34 | 34 | 68 |
| QQP | 8 | 259 | 9 | 21 | 30 |
| PAWS | 1 | 249 | 91 | 121 | 212 |
| PAWS | 2 | 268 | 72 | 98 | 170 |
| PAWS | 4 | 284 | 56 | 66 | 122 |
| PAWS | 8 | 308 | 32 | 32 | 64 |

Rank-eight retained-objective updates give fewer disagreements than rank-one updates for 38 QQP and 40 PAWS requests, and the same number for the remaining 22 and 20 requests. At ranks 1, 2, 4 and 8, SVD gives totals 162, 127, 93 and 50 on QQP, and 239, 211, 161 and 101 on PAWS.

Mean initial head accuracy is $79.51\%$ on QQP and $56.39\%$ on PAWS. The affine baselines reach $76.52\%$ and $55.76\%$, and majority accuracy is $62.98\%$ and $55.62\%$, respectively. With fixed $\lambda = 1$, mean initial accuracy on these same pairs is $78.35\%$ and $54.67\%$. These descriptive comparisons reuse training observations and evaluation pairs, so the comparisons are dependent.

All 480 selected retained-objective candidates have successful optimizer status with the same settings as the score comparisons. Each Cholesky refit uses one refinement step. Strong convexity bounds Frobenius fit error by gradient norm divided by $\lambda$. The sum of the two fit-error estimates, divided by the computed correction norm, is below $8 \times 10^{-12}$. For each fixed computed update, multiplying the original-fit estimate by $\|xz^\top\|_{\mathrm{F}}$ gives its score-error estimate. All original, retrained and candidate margins exceed their corresponding estimates. These floating-point residual estimates exclude

roundoff in features and scores and are not interval bounds. Optimizer convergence does not establish global optimality.

### N.8 A LOWER BOUND FOR EVERY RANK-ONE CORRECTION ON REAL PAIRS

A failed candidate does not establish that rank is the obstacle. For this purpose we need a lower bound for every matrix of that rank in the actual empirical prediction metric. Let $v := \mathrm{vec}(B - \Delta)$, with the column-stacking convention of Section 3, and define

$$Q_T := \frac{1}{|T|} \sum_{i \in T} \mathrm{vec}(x_i z_i^\top)\, \mathrm{vec}(x_i z_i^\top)^\top. \tag{N.2}$$

Then $e_T(B)^2 = v^\top Q_T v$. Paired text representations need not be independent or isotropic, so replacing $Q_T$ by a product of marginal covariances is not an identity.

To obtain a lower bound that can be minimized by singular-value truncation, we compare the empirical prediction error with a squared Frobenius norm after invertible left and right transformations. Choose invertible matrices $L_x, L_z \in \mathbb{R}^{d \times d}$ and a scalar $c > 0$ satisfying

$$Q_T \succeq c\big[(L_z L_z^\top) \otimes (L_x L_x^\top)\big]. \tag{N.3}$$

This matrix inequality is a sufficient condition that can be verified numerically. It implies the following deterministic bound, which uses the classical Eckart–Young theorem rather than an assumption about the feature distribution.

**Proposition S14** (An empirical prediction lower bound)**.** *Under* (N.3)*, for every* $r \in \{0, \dots, d\}$*, every matrix* $B$ *with* $\mathrm{rg}(B) \leqslant r$ *satisfies*

$$e_T(B)^2 \geqslant c \sum_{j > r} s_j(L_x^\top \Delta L_z)^2. \tag{N.4}$$

*Proof.* The matrix inequality gives $e_T(B)^2 \geqslant c\|L_x^\top (B - \Delta) L_z\|_{\mathrm{F}}^2$. Invertibility makes $B \mapsto L_x^\top B L_z$ a bijection preserving rank. Minimizing the right-hand side over matrices of rank at most $r$ therefore gives the stated singular-value tail by Eckart and Young (1936). □

The bound remains valid for dependent, anisotropic feature pairs. Its usefulness depends on the minorant. It need not equal the minimum empirical error, and a weak bound does not show that a low-rank correction succeeds.

**Feasibility survey and selection.** We first evaluated isotropic and marginal-covariance minorants for all 960 requests with $d = 16$. For the 60 single deletions at $\lambda = 1$ in each corpus, these floating-point bounds exceeded the explicit rank-eight error for 39 STSB, 57 SICK, 58 QQP and zero PAWS requests. None crossed the numerical tolerance. At $\lambda = 0.1$, they crossed it while rank eight passed for one STSB, three SICK, two QQP and zero PAWS requests. The task-performance subsection reports the corresponding metrics at this regularization.

We then refined the minorant on all 60 QQP single deletions with $d = 16$, $\lambda = 1$. The factors $L_x, L_z$ are positive diagonal matrices with entries $\exp(\theta_i)$, with each $\theta_i \in [-2, 2]$. For each request, L-BFGS-B starts at zero and runs for at most 150 iterations to increase the bound in (N.4), taking $c$ from the generalized minimum eigenvalue. All runs reach this iteration cap. This search does not establish optimality of the minorant and need not do so: any verified feasible minorant gives a valid bound. It uses the evaluation metric and the correction to establish a lower bound, and does not refit or select the rank-eight candidate. All 60 numerical lower bounds exceed that candidate's error, but only one crosses $0.001$. These survey counts are floating-point feasibility results. The following selected instance receives a separate interval verification.

**Verified instance.** The instance is E5/QQP, $d = 16$, $n = 512$, $\lambda = 1$, training seed 1701 and the ninth stored single-deletion request, which removes local training row 23 with zero-based indexing. Its original head accuracy is $75.85\%$, versus $62.75\%$ for the majority class. The instance is selected from the survey, not an independent replication. On its 2000 evaluation pairs, the verified bounds are

$$\inf_{\mathrm{rg}(B) \leqslant 1} e_T(B) > 0.001088 \quad > \quad 0.001 \quad > \quad e_T(B_8), \qquad e_T(B_8) < 0.000760. \tag{N.5}$$

Here $B_8 = AC^\top$ is the stored rank-eight retained-objective candidate. Without updating, $a_T \simeq 0.0103202$. Thus a correction is necessary under the stated tolerance, every rank-one correction fails, and an explicit correction of rank at most eight succeeds. The argument does not identify the minimum sufficient rank between two and eight. The verified inequalities concern this finite empirical metric. They do not bound future-pair error or establish encoder unlearning or privacy. One of the 60 surveyed requests satisfies these verified inequalities at the fixed threshold.

**Interval verification and the quantities it covers.** The archive supplies the projected training and evaluation feature vectors, residual training responses, deleted index, factors $A, C$ and the diagonal minorant. Each stored binary floating-point number is treated as an exact rational input defining the finite instance. At 128-bit precision, Arb ball arithmetic (Johansson, 2017) encloses every subsequent feature product, Gram matrix, original fit and retained-data fit. Thus $\Delta$ is computed by verified linear solves, rather than taken as an exact floating-point approximation. Its Frobenius distance from the stored correction is less than $1.30 \times 10^{-17}$.

For the supplied diagonal matrices, the coefficient $c = 0.10931760$ is rounded down before verification. An approximate inverse Cholesky factor $T_0$ is used only as a change of coordinates. Ball arithmetic verifies that $T_0(Q_T - cK)T_0^\top$ is strictly diagonally dominant with positive diagonal, where $K$ is the Kronecker matrix in (N.3). This symmetric matrix is positive definite, which also implies that $T_0$ is invertible and proves (N.3). The proof does not rely on unverified numerical eigenvalues.

To bound the rank-one tail, set $D := L_x^\top \Delta L_z$. We use $\sum_{j>1} s_j(D)^2 = \|D\|_{\mathrm{F}}^2 - \lambda_{\max}(D^\top D)$. An approximate eigenvector matrix, inverted with ball arithmetic, gives a similar matrix whose Gershgorin bounds enclose the largest eigenvalue. For an eigenvector, the coordinate of largest absolute value bounds its eigenvalue by the corresponding diagonal entry plus the off-diagonal absolute row sum. Orthogonality of the numerical eigenvectors is not assumed. Direct ball evaluation of $AC^\top$ and its prediction errors gives the upper bound, preserving the factorization's exact rank constraint. The verifier evaluates the inequalities in (N.5) with outward rounding and reports failure if any inequality cannot be established. The code uses row-major vectorization, which permutes the Kronecker factors consistently and leaves the metric unchanged. This verifies the stored finite instance, not the encoder or preprocessing computations that originally produced its inputs.

### N.9 A SOFTMAX HEAD ON FIXED IMAGE FEATURES

We next examine whether a rank-one update can approximate retraining accurately when the loss is not quadratic. We train a CNN on the first 30000 Fashion-MNIST training images and freeze its 64-dimensional features. The other 30000 images supply the head-fitting pool. Feature vectors are rescaled to have mean squared norm 64 on this pool, then augmented with a constant coordinate, giving $h_i \in \mathbb{R}^{65}$. The head matrix is $W \in \mathbb{R}^{10\times 65}$. We minimize

$$L_T(W) := \frac{1}{n}\sum_{i\in T} -\log\big(\mathrm{softmax}(Wh_i)_{y_i}\big) + \frac{\lambda}{2}\|W\|_{\mathrm{F}}^2 .$$

The original denominator $n$ and $\lambda > 0$ stay fixed after deletion. Regularization makes the objective strongly convex. An individual cross-entropy term has the rank-one gradient $(p_i - e_{y_i})h_i^\top$, where $p_i := \mathrm{softmax}(W_D h_i)$ and $e_{y_i}$ is the coordinate vector for its class. We use Newton–CG with exact Hessian–vector products to compute the full retained fit. For its nonzero correction, (5) applies with ten singular values. The Frobenius error measures parameter agreement here, rather than error in class probabilities.

We use $n \in \{650, 1300, 3250, 6500, 13000, 30000\}$, $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, and subsampling seeds 0 and 1. Each fit contributes 12 randomly selected single deletions and the four observations of largest training loss, with possible overlap. At $n = 30000$ both seeds use the same fitting pool. Table 5 summarizes all 18 settings, with 32 requests per setting. The largest recorded retained-gradient norm is $7.30 \times 10^{-10}$. These are numerical comparisons, not interval bounds or an asymptotic result. The code, complete stored summaries and reproduction instructions are in `code_lora.zip`, under `softmax/`.

**Relative numerical accuracy.** A small stationarity residual must be compared with the deletion correction itself. Let $\widehat{W}_D, \widehat{W}_R$ denote the computed original and retained fits, and set $\widehat{\Delta} := \widehat{W}_R - \widehat{W}_D$.

Table 5: Softmax head. Ranges across the six sample sizes of mean minimum relative squared Frobenius error (%). All random and largest-loss requests are included.

| $\lambda$ | Rank 1 | Rank 4 |
|---|---|---|
| $10^{-3}$ | 5.77–7.05 | 0.206–0.410 |
| $10^{-2}$ | 4.04–4.47 | 0.113–0.148 |
| $10^{-1}$ | 2.39–2.89 | 0.046–0.060 |

Both objectives are $\lambda$-strongly convex. Their gradient norms therefore bound the distance from each computed fit to its exact minimizer. The triangle inequality gives

$$\left\|\widehat{\Delta} - \Delta\right\|_{\mathrm{F}} \leqslant \eta, \qquad \eta := \frac{\left\|\nabla L_D(\widehat{W}_D)\right\|_{\mathrm{F}} + \left\|\nabla L_R(\widehat{W}_R)\right\|_{\mathrm{F}}}{\lambda}. \tag{N.6}$$

To compare this uncertainty with the rank-approximation error, define $a := \left\|\widehat{\Delta}\right\|_{\mathrm{F}}$ and $t_r^2 := \sum_{j>r} s_j(\widehat{\Delta})^2$. If $a > \eta$, the Lipschitz bound (C.3) and the reverse triangle inequality yield

$$\left(\frac{(t_r - \eta)_+}{a + \eta}\right)^2 \leqslant \mathcal{E}_r(\Delta) \leqslant \min\left\{1, \left(\frac{t_r + \eta}{a - \eta}\right)^2\right\}. \tag{N.7}$$

For the 16 requests used to assess numerical accuracy, both gradients are evaluated at the returned matrices, including when the solver reaches its iteration limit.

Numerical accuracy is assessed at $n \in \{650, 30000\}$, $\lambda \in \{10^{-3}, 10^{-1}\}$ and subsampling seed zero. The feature extractor for these measurements uses the four-epoch training protocol and seed zero. Each of the four settings includes the first two of the twelve random deletions and the first two largest-loss deletions, giving 16 requests selected before evaluating the bounds. Across these requests, $\eta/a \leqslant 4.83 \times 10^{-8}$. The maximum widths of the bounds in (N.7) are $8.12 \times 10^{-8}$ at rank one and $2.23 \times 10^{-8}$ at rank four. These estimates use float64 gradients and singular values, without outward interval rounding. They concern these 16 requests. Table 5 reports approximation errors across the full 18 settings.

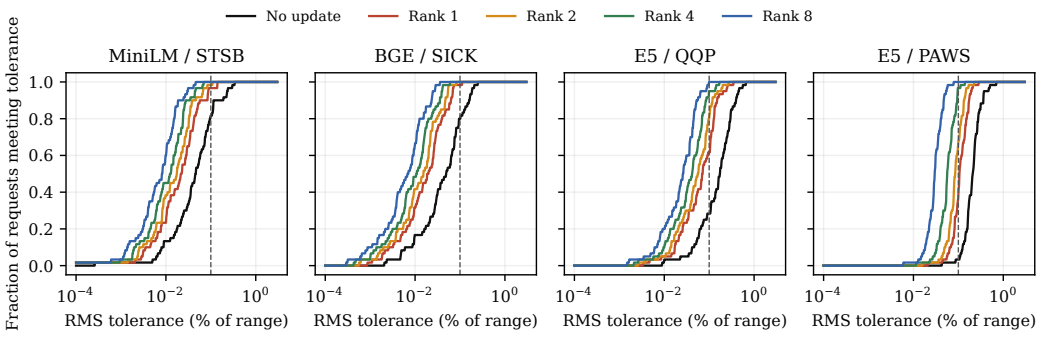


Figure 16: Complete tolerance curves for single deletions at $d = 32$, $\lambda = 1$. Each curve gives the fraction of all 60 requests meeting an absolute RMS requirement. The vertical line is the stated $0.1\%$ tolerance. It is not a threshold fitted to these curves.

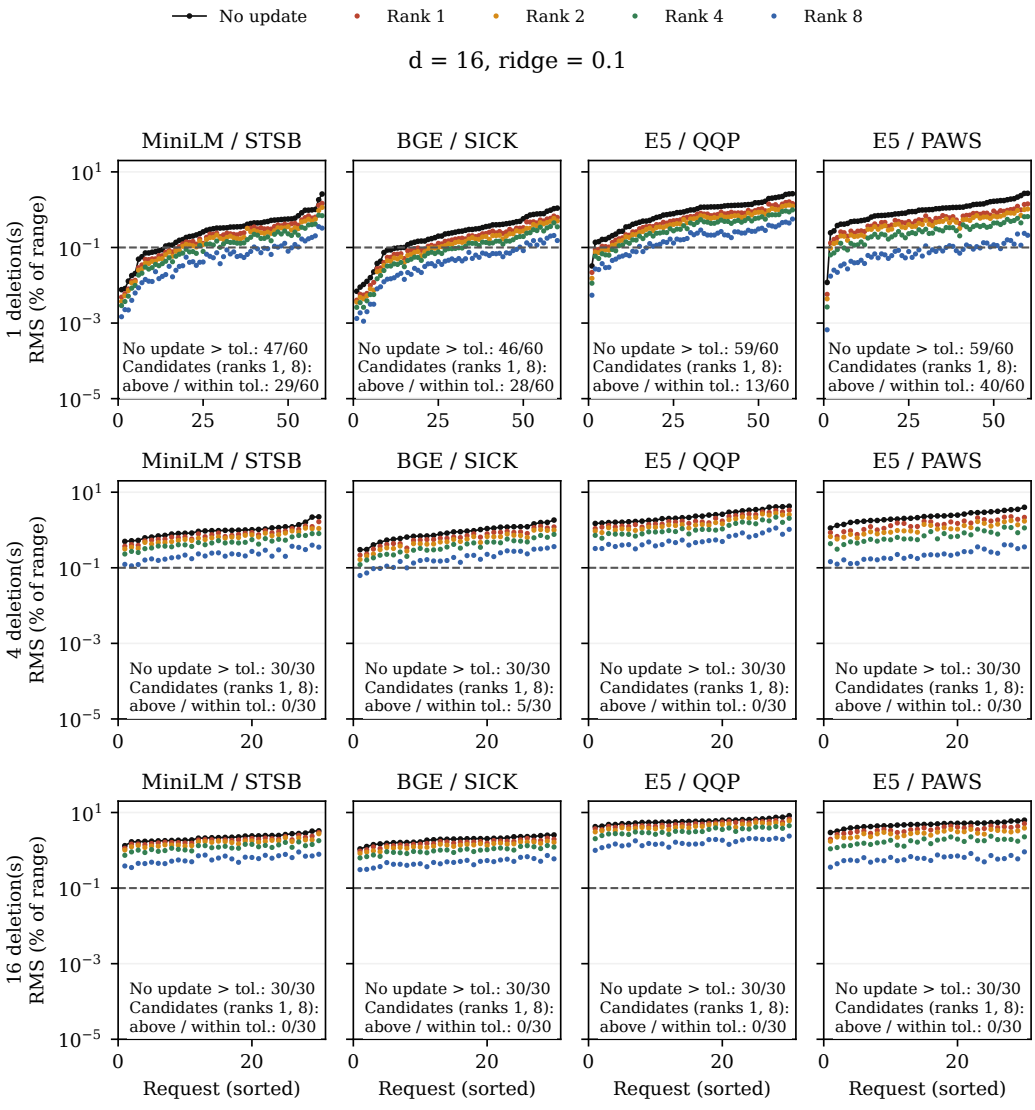


Figure 17: Every stored request at $d = 16$, $\lambda = 0.1$. Rows give deletion sizes 1, 4 and 16. Each panel uses its own no-update ordering. The dashed horizontal line is the fixed numerical tolerance. All four encoder–dataset settings are shown.

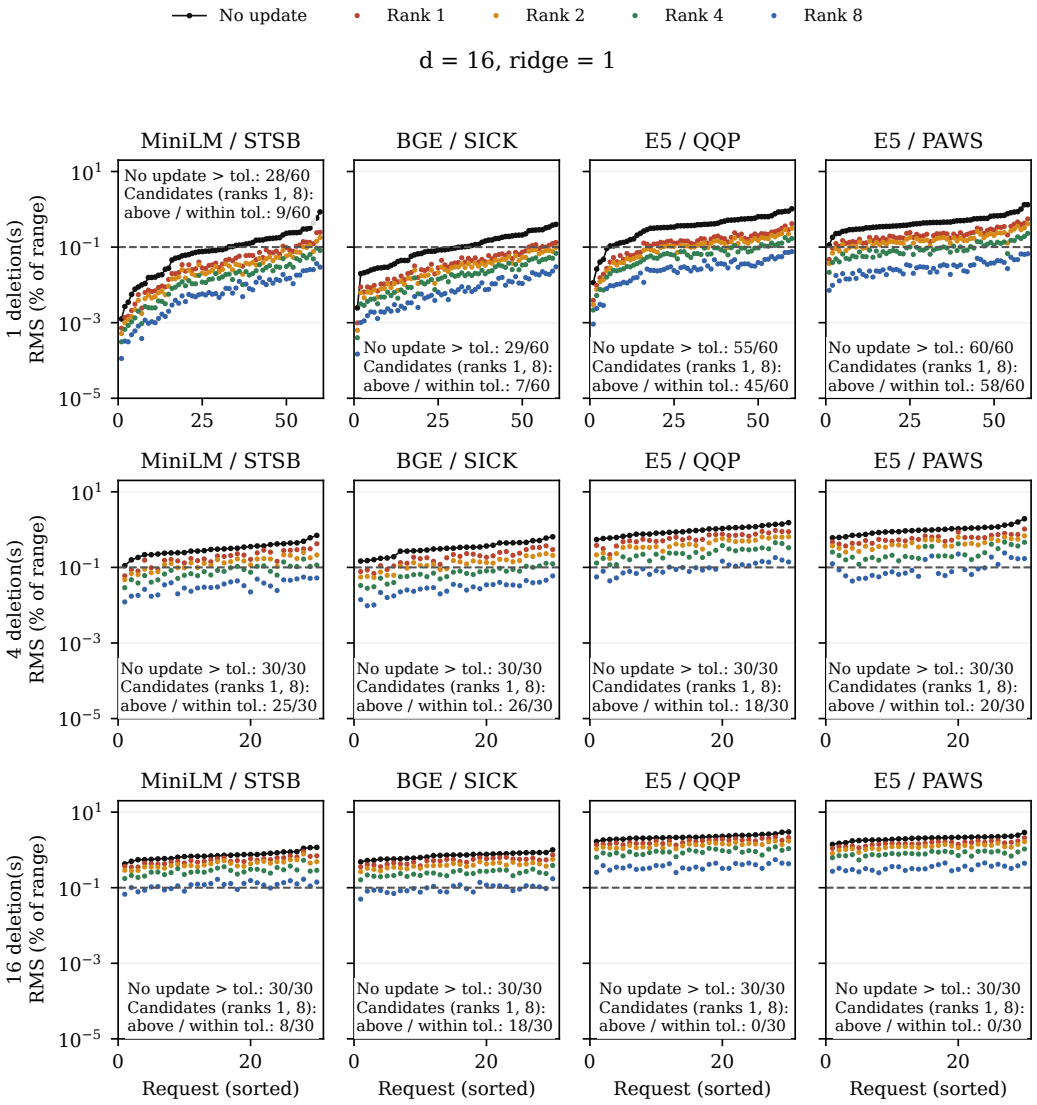


Figure 18: Every stored request at $d = 16$, $\lambda = 1$, with the same conventions. The verified QQP lower bound concerns a single request in the third panel of the first row.

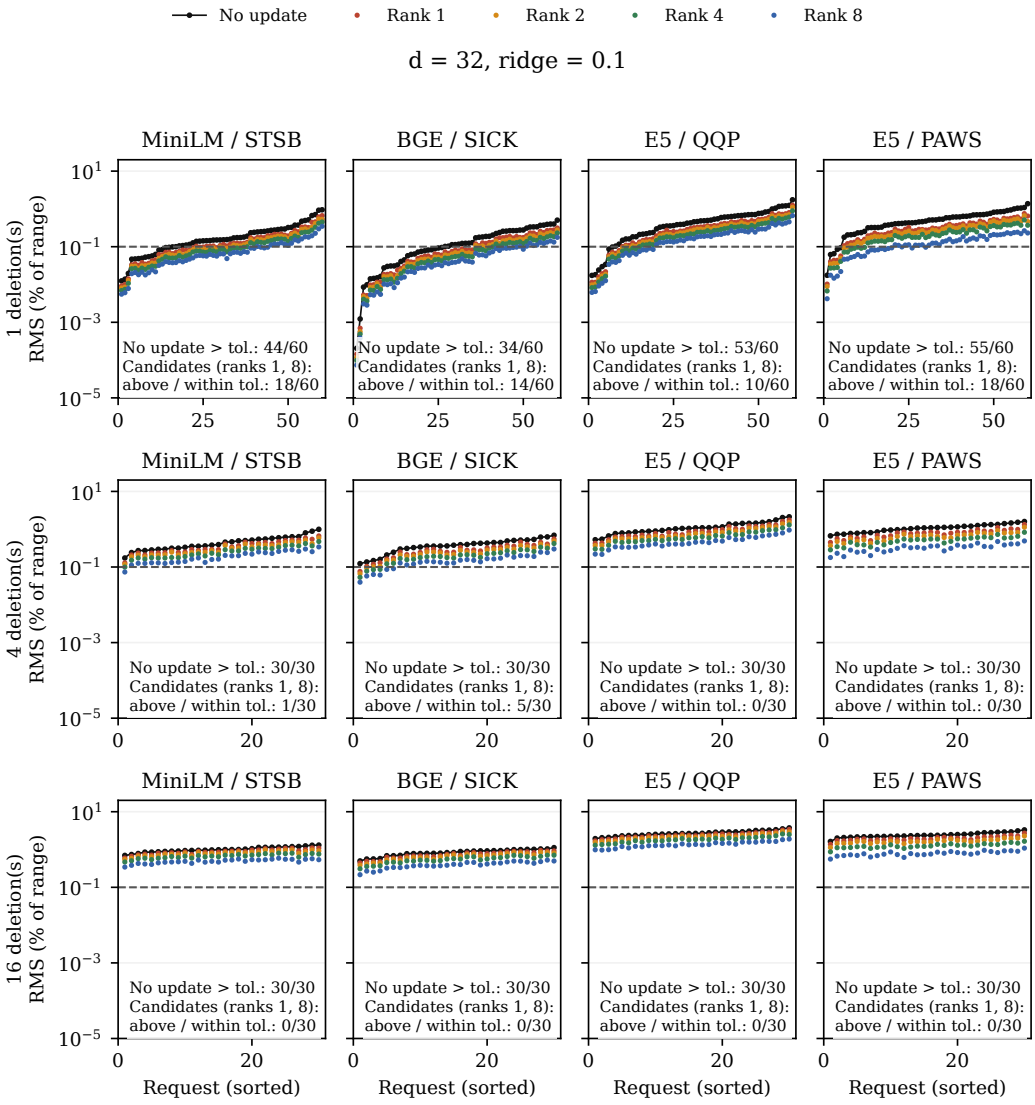


Figure 19: Every stored request at $d = 32$, $\lambda = 0.1$, for all deletion sizes and datasets.

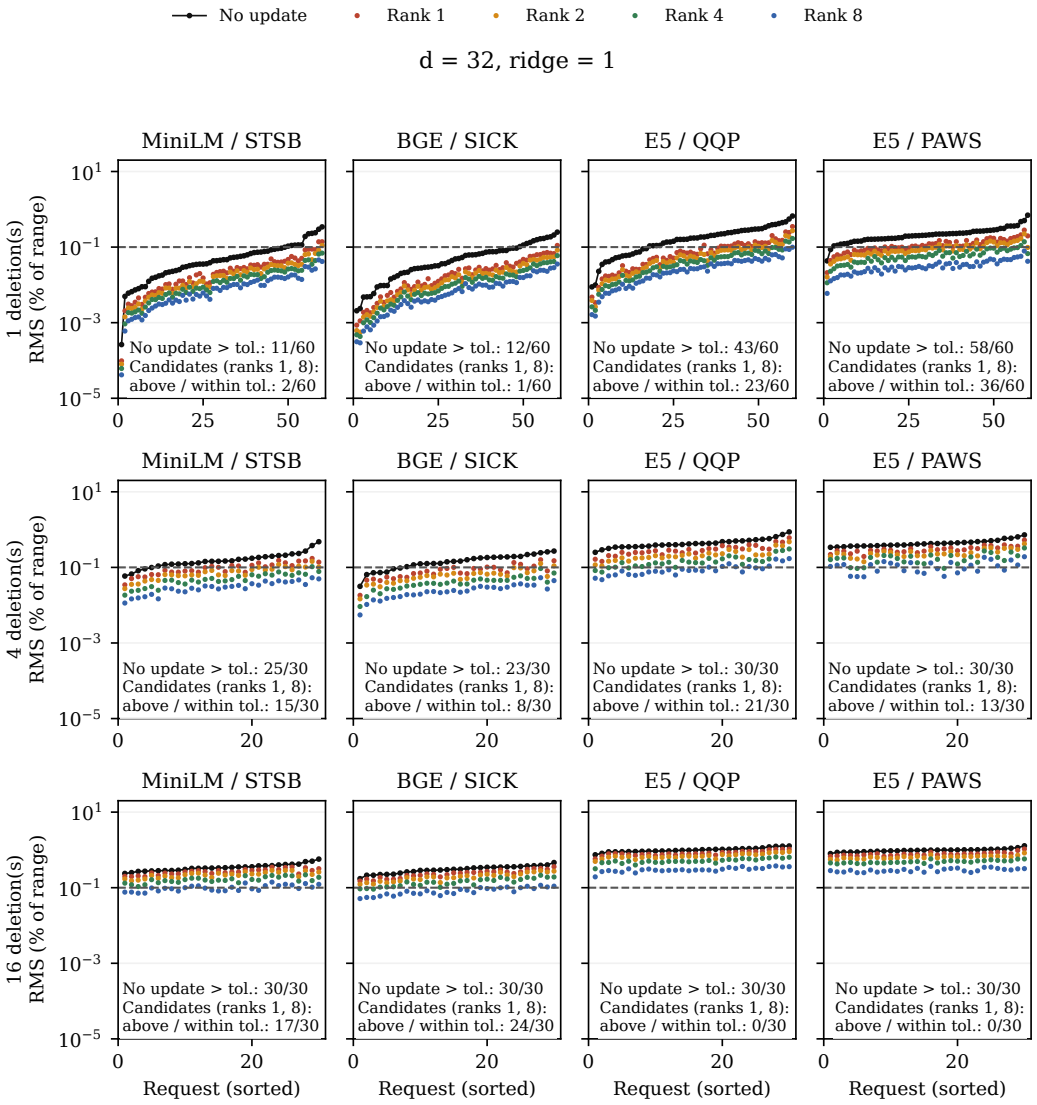


Figure 20: Every stored request at $d = 32$, $\lambda = 1$, including grouped deletions and PAWS. Rank eight does not meet the fixed absolute tolerance for any size-16 QQP or PAWS request.

Table 6: MiniLM/STSB. Mean relative squared prediction discrepancy (%) for all configurations. Each entry aggregates 60 single deletions or 30 groups. $|F|$ is the deletion size. RL denotes retained-objective fitting and PM denotes prediction matching.

| | | | $d = 16$ | | | | | | | $d = 32$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ | $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ |
| 0.1 | 1 | SVD | 68.8 | 32.7 | 16.7 | 4.5 | 0.1 | 1 | SVD | 67.4 | 39.1 | 26.7 | 15.8 |
| 0.1 | 1 | RL | 46.0 | 31.6 | 16.4 | 3.9 | 0.1 | 1 | RL | 49.9 | 37.6 | 27.1 | 15.4 |
| 0.1 | 1 | PM | 39.7 | 24.6 | 12.2 | 2.8 | 0.1 | 1 | PM | 46.7 | 34.8 | 26.1 | 16.9 |
| 0.1 | 4 | SVD | 72.6 | 50.2 | 25.5 | 6.4 | 0.1 | 4 | SVD | 75.0 | 57.0 | 36.6 | 19.1 |
| 0.1 | 4 | RL | 61.1 | 43.4 | 23.3 | 5.5 | 0.1 | 4 | RL | 62.0 | 48.0 | 32.9 | 18.2 |
| 0.1 | 4 | PM | 54.1 | 35.2 | 17.3 | 3.9 | 0.1 | 4 | PM | 60.2 | 45.9 | 32.3 | 19.7 |
| 0.1 | 16 | SVD | 76.7 | 56.1 | 32.5 | 9.2 | 0.1 | 16 | SVD | 82.8 | 69.5 | 49.5 | 26.6 |
| 0.1 | 16 | RL | 71.8 | 54.5 | 30.6 | 7.6 | 0.1 | 16 | RL | 76.8 | 63.0 | 44.4 | 24.2 |
| 0.1 | 16 | PM | 62.6 | 45.0 | 23.2 | 5.8 | 0.1 | 16 | PM | 77.5 | 63.2 | 46.2 | 27.9 |
| 1 | 1 | SVD | 37.6 | 14.5 | 4.6 | 0.8 | 1 | 1 | SVD | 36.0 | 17.0 | 8.5 | 3.4 |
| 1 | 1 | RL | 19.1 | 9.5 | 3.3 | 0.6 | 1 | 1 | RL | 20.3 | 11.7 | 6.1 | 2.6 |
| 1 | 1 | PM | 17.6 | 8.6 | 3.0 | 0.5 | 1 | 1 | PM | 20.8 | 12.4 | 6.8 | 3.3 |
| 1 | 4 | SVD | 48.7 | 29.2 | 11.7 | 1.8 | 1 | 4 | SVD | 51.4 | 34.9 | 16.9 | 5.7 |
| 1 | 4 | RL | 35.6 | 18.0 | 6.6 | 1.1 | 1 | 4 | RL | 39.0 | 22.4 | 10.0 | 3.8 |
| 1 | 4 | PM | 33.8 | 16.8 | 5.9 | 1.0 | 1 | 4 | PM | 39.2 | 23.0 | 11.0 | 4.7 |
| 1 | 16 | SVD | 64.9 | 44.8 | 21.3 | 4.3 | 1 | 16 | SVD | 67.4 | 52.7 | 31.7 | 13.7 |
| 1 | 16 | RL | 54.6 | 35.1 | 15.4 | 2.7 | 1 | 16 | RL | 60.8 | 42.1 | 23.2 | 8.5 |
| 1 | 16 | PM | 51.0 | 31.0 | 13.2 | 2.3 | 1 | 16 | PM | 61.1 | 43.8 | 24.9 | 10.7 |

Table 7: BGE/SICK. Mean relative squared prediction discrepancy (%) for all configurations. Each entry aggregates 60 single deletions or 30 groups. $|F|$ is the deletion size. RL denotes retained-objective fitting and PM denotes prediction matching.

| | | | $d = 16$ | | | | | | | $d = 32$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ | $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ |
| 0.1 | 1 | SVD | 51.1 | 27.9 | 15.1 | 4.1 | 0.1 | 1 | SVD | 65.2 | 36.5 | 24.0 | 14.5 |
| 0.1 | 1 | RL | 40.5 | 26.3 | 13.8 | 3.4 | 0.1 | 1 | RL | 44.8 | 31.9 | 22.2 | 12.3 |
| 0.1 | 1 | PM | 35.4 | 21.6 | 10.8 | 2.6 | 0.1 | 1 | PM | 41.8 | 29.2 | 21.4 | 13.3 |
| 0.1 | 4 | SVD | 61.8 | 40.8 | 19.6 | 5.2 | 0.1 | 4 | SVD | 66.1 | 52.1 | 32.1 | 17.1 |
| 0.1 | 4 | RL | 55.0 | 36.0 | 18.2 | 4.3 | 0.1 | 4 | RL | 55.2 | 41.1 | 27.3 | 14.7 |
| 0.1 | 4 | PM | 49.1 | 30.1 | 14.4 | 3.3 | 0.1 | 4 | PM | 52.6 | 38.5 | 26.2 | 15.6 |
| 0.1 | 16 | SVD | 72.6 | 54.7 | 32.0 | 8.2 | 0.1 | 16 | SVD | 80.7 | 67.6 | 51.3 | 27.3 |
| 0.1 | 16 | RL | 68.4 | 49.2 | 27.0 | 6.5 | 0.1 | 16 | RL | 75.0 | 60.6 | 42.3 | 22.1 |
| 0.1 | 16 | PM | 62.1 | 42.9 | 21.8 | 5.1 | 0.1 | 16 | PM | 73.6 | 59.2 | 42.5 | 23.8 |
| 1 | 1 | SVD | 19.8 | 10.6 | 3.9 | 0.6 | 1 | 1 | SVD | 26.9 | 14.0 | 6.8 | 2.7 |
| 1 | 1 | RL | 13.8 | 7.4 | 2.7 | 0.5 | 1 | 1 | RL | 16.2 | 9.4 | 4.8 | 2.0 |
| 1 | 1 | PM | 12.9 | 6.8 | 2.5 | 0.4 | 1 | 1 | PM | 16.1 | 9.4 | 5.1 | 2.3 |
| 1 | 4 | SVD | 40.4 | 19.9 | 6.6 | 1.1 | 1 | 4 | SVD | 41.1 | 28.2 | 13.0 | 4.4 |
| 1 | 4 | RL | 32.5 | 14.5 | 4.5 | 0.8 | 1 | 4 | RL | 33.1 | 18.5 | 7.8 | 2.9 |
| 1 | 4 | PM | 29.7 | 13.0 | 4.1 | 0.7 | 1 | 4 | PM | 32.6 | 18.2 | 8.1 | 3.4 |
| 1 | 16 | SVD | 58.4 | 35.7 | 16.1 | 2.8 | 1 | 16 | SVD | 65.1 | 48.7 | 30.3 | 12.5 |
| 1 | 16 | RL | 53.0 | 30.5 | 12.1 | 1.9 | 1 | 16 | RL | 59.4 | 40.1 | 21.4 | 7.1 |
| 1 | 16 | PM | 49.1 | 27.3 | 10.3 | 1.7 | 1 | 16 | PM | 58.9 | 40.0 | 22.4 | 8.2 |

Table 8: E5/QQP. Mean relative squared prediction discrepancy (%) for all configurations. Each entry aggregates 60 single deletions or 30 groups. $|F|$ is the deletion size. RL denotes retained-objective fitting and PM denotes prediction matching.

| | | | $d = 16$ | | | | | | | $d = 32$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ | $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ |
| 0.1 | 1 | SVD | 47.3 | 27.6 | 15.6 | 4.2 | 0.1 | 1 | SVD | 55.5 | 35.4 | 24.4 | 14.2 |
| 0.1 | 1 | RL | 41.3 | 28.0 | 14.9 | 3.8 | 0.1 | 1 | RL | 45.9 | 34.9 | 25.0 | 14.3 |
| 0.1 | 1 | PM | 36.0 | 22.8 | 11.8 | 3.0 | 0.1 | 1 | PM | 42.5 | 31.4 | 23.4 | 14.8 |
| 0.1 | 4 | SVD | 62.1 | 43.5 | 22.7 | 5.8 | 0.1 | 4 | SVD | 68.6 | 54.3 | 35.6 | 18.3 |
| 0.1 | 4 | RL | 58.5 | 39.6 | 21.1 | 5.2 | 0.1 | 4 | RL | 63.1 | 47.4 | 32.1 | 17.4 |
| 0.1 | 4 | PM | 52.8 | 33.7 | 16.7 | 4.0 | 0.1 | 4 | PM | 60.9 | 45.1 | 31.2 | 18.6 |
| 0.1 | 16 | SVD | 73.9 | 54.9 | 31.6 | 8.3 | 0.1 | 16 | SVD | 82.9 | 69.6 | 49.8 | 27.0 |
| 0.1 | 16 | RL | 70.8 | 53.4 | 30.6 | 7.7 | 0.1 | 16 | RL | 78.6 | 64.1 | 45.8 | 25.1 |
| 0.1 | 16 | PM | 64.1 | 46.2 | 25.2 | 6.1 | 0.1 | 16 | PM | 77.5 | 62.7 | 46.7 | 27.2 |
| 1 | 1 | SVD | 19.7 | 11.3 | 4.3 | 0.8 | 1 | 1 | SVD | 27.4 | 16.3 | 7.9 | 3.0 |
| 1 | 1 | RL | 14.4 | 8.1 | 3.2 | 0.6 | 1 | 1 | RL | 18.0 | 11.0 | 5.7 | 2.4 |
| 1 | 1 | PM | 13.7 | 7.6 | 3.0 | 0.5 | 1 | 1 | PM | 18.0 | 11.3 | 6.2 | 2.9 |
| 1 | 4 | SVD | 47.7 | 26.7 | 10.4 | 1.6 | 1 | 4 | SVD | 52.2 | 33.5 | 20.0 | 6.8 |
| 1 | 4 | RL | 41.3 | 19.3 | 6.7 | 1.1 | 1 | 4 | RL | 44.3 | 23.7 | 10.8 | 4.0 |
| 1 | 4 | PM | 38.5 | 17.6 | 5.9 | 1.0 | 1 | 4 | PM | 44.1 | 24.1 | 11.4 | 4.9 |
| 1 | 16 | SVD | 64.1 | 42.3 | 19.5 | 3.6 | 1 | 16 | SVD | 72.7 | 55.5 | 32.6 | 13.2 |
| 1 | 16 | RL | 58.1 | 36.7 | 16.4 | 2.8 | 1 | 16 | RL | 66.3 | 47.5 | 26.1 | 9.1 |
| 1 | 16 | PM | 53.2 | 33.0 | 14.6 | 2.5 | 1 | 16 | PM | 65.2 | 47.2 | 27.8 | 11.0 |

Table 9: E5/PAWS. Mean relative squared prediction discrepancy (%) for all configurations. Each entry aggregates 60 single deletions or 30 groups. $|F|$ is the deletion size. RL denotes retained-objective fitting and PM denotes prediction matching.

| | | | $d = 16$ | | | | | | | $d = 32$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ | $\lambda$ | $\lvert F\rvert$ | Method | $r = 1$ | $r = 2$ | $r = 4$ | $r = 8$ |
| 0.1 | 1 | SVD | 52.1 | 32.8 | 12.4 | 2.4 | 0.1 | 1 | SVD | 73.6 | 43.8 | 23.4 | 9.5 |
| 0.1 | 1 | RL | 30.4 | 18.8 | 7.1 | 0.7 | 0.1 | 1 | RL | 37.9 | 28.1 | 16.8 | 5.9 |
| 0.1 | 1 | PM | 23.8 | 13.7 | 4.8 | 0.3 | 0.1 | 1 | PM | 34.3 | 25.1 | 15.3 | 5.5 |
| 0.1 | 4 | SVD | 63.1 | 43.0 | 21.3 | 3.7 | 0.1 | 4 | SVD | 76.2 | 65.8 | 55.8 | 22.0 |
| 0.1 | 4 | RL | 40.1 | 22.7 | 9.2 | 1.0 | 0.1 | 4 | RL | 53.3 | 37.1 | 22.6 | 9.1 |
| 0.1 | 4 | PM | 33.6 | 16.4 | 5.5 | 0.4 | 0.1 | 4 | PM | 50.2 | 34.8 | 20.5 | 7.9 |
| 0.1 | 16 | SVD | 73.8 | 55.5 | 30.8 | 7.1 | 0.1 | 16 | SVD | 80.8 | 66.7 | 46.6 | 24.8 |
| 0.1 | 16 | RL | 56.6 | 35.4 | 13.2 | 1.7 | 0.1 | 16 | RL | 65.2 | 47.3 | 27.2 | 10.6 |
| 0.1 | 16 | PM | 51.0 | 28.3 | 8.6 | 0.6 | 0.1 | 16 | PM | 63.2 | 44.7 | 24.0 | 8.5 |
| 1 | 1 | SVD | 34.4 | 20.0 | 6.9 | 0.6 | 1 | 1 | SVD | 45.2 | 28.9 | 14.6 | 3.5 |
| 1 | 1 | RL | 22.4 | 12.4 | 3.9 | 0.4 | 1 | 1 | RL | 28.4 | 17.5 | 7.8 | 2.1 |
| 1 | 1 | PM | 14.0 | 7.8 | 2.6 | 0.2 | 1 | 1 | PM | 20.2 | 13.6 | 7.0 | 2.2 |
| 1 | 4 | SVD | 55.1 | 31.7 | 18.4 | 3.3 | 1 | 4 | SVD | 65.8 | 51.0 | 57.8 | 22.1 |
| 1 | 4 | RL | 36.6 | 19.7 | 6.8 | 1.3 | 1 | 4 | RL | 48.5 | 31.1 | 16.5 | 6.6 |
| 1 | 4 | PM | 27.4 | 9.8 | 3.1 | 0.4 | 1 | 4 | PM | 42.4 | 23.9 | 12.0 | 4.8 |
| 1 | 16 | SVD | 71.0 | 49.8 | 26.6 | 5.5 | 1 | 16 | SVD | 78.2 | 59.2 | 38.1 | 18.6 |
| 1 | 16 | RL | 54.8 | 35.7 | 15.9 | 2.9 | 1 | 16 | RL | 62.9 | 43.5 | 25.0 | 9.1 |
| 1 | 16 | PM | 47.9 | 25.1 | 6.1 | 0.4 | 1 | 16 | PM | 59.5 | 36.9 | 16.2 | 4.2 |