Representation Capacity in In-Context Learning of Exponential-Family Tasks

Anonymous Author
Anonymous Institution

Abstract

We study how the number of shared query features and the task structure influence in-context learning of exponential-family classification tasks. In this setting, class scores depend on a common sufficient statistic, providing a framework for examining how a limited set of features is allocated across tasks. We develop an eigenvalue-based analysis that combines variation across inputs with relevance across tasks and investigate its connections to query-key prediction, sufficient-statistic recovery, and classification performance. We also consider prediction from finitely many examples, where uncertainty about the task may affect which features are useful. Controlled synthetic experiments explore how these considerations relate to learned representations and context-dependent prediction. This framework offers a way to examine the interplay between the available features, task inference, and optimization in simplified in-context learning settings.

1 INTRODUCTION

In-context learning (ICL) allows a model to adapt its predictions using examples while keeping its parameters fixed. Previous work offers several explanations for this ability. Bayesian approaches describe how examples help the model infer the underlying task (Xie et al., 2022). Other studies show how transformers can implement gradient descent and ridge regression in their forward pass (Akyürek et al., 2023; Von Oswald et al., 2023). We ask a complementary representation question:

Which shared query features are needed to express the prediction rules selected by different contexts?

This question builds on work that studies shared representations in ICL. Existing studies consider prediction using fixed nonlinear features, the learning of features shared across tasks, and regression over learned basis functions (**????**). Related work in multitask and meta-learning examines when shared representations help models learn new tasks and how these representations can be learned (Maurer et al., 2016; Tripuraneni et al., 2021; **?**). We focus on the limitation imposed by the number of shared features. Even when context provides the best possible weights for combining them, a small set of features may not be enough to express every task's prediction rule.

2 PROBLEM SETUP

In-context learning requires a model to adapt its predictions to a task using examples. We study the representation that supports this adaptation. The central distinction is between the features available for describing a query and the coefficients used to combine those features. Context determines the coefficients, while the number of shared query features limits the query functions those coefficients can act on.

Setup. A task $\tau \sim \rho$ specifies a classification problem with K classes. Examples from class k are drawn from an exponential family

$$p(x \mid Y = k, \tau) = h(x) \exp\left\{\eta_{\tau,k}^\top T(x) - A(\eta_{\tau,k})\right\}.$$

The sufficient statistic $T(x) \in \mathbb{R}^d$, base density h, and normalizing function A are shared while the parameters $\eta_{\tau,k}$ vary across tasks and classes.

For each task τ, we draw n independent examples from each class to form the context $C = \{(X_{k,i}, k) : k = 1, \ldots, K,\ i = 1, \ldots, n\}$.

We then generate a query (X, Y) by choosing its label $Y = k$ with probability $\pi_{\tau,k} > 0$, then drawing X from

Preliminary work. Under review by AISTATS 2027. Do not distribute.

that class's density. The learner observes C and X and predicts Y without knowing τ.

The learner maps the query to m features $\phi(x) \in \mathbb{R}^m$ that depend only on x. The weights $w_k \in \mathbb{R}^m$ and biases b_k depend on the context C and combine these features into a score for each class:

$$\widehat{z}_k(x) = w_k^\top \phi(x) + b_k, \qquad k = 1, \dots, K. \tag{1}$$

Applying softmax gives the predicted class probabilities

$$\widehat{p}_k(x) = \frac{\exp(\widehat{z}_k(x))}{\sum_{\ell=1}^K \exp(\widehat{z}_\ell(x))}.$$

Denote $\widehat{p}(x) = (\widehat{p}_1(x), \dots, \widehat{p}_K(x))^\top$ and $\widehat{z}(x) = (\widehat{z}_1(x), \dots, \widehat{z}_K(x))^\top$. Without loss of generality, we can take the scores to sum to zero by subtracting their average across classes. The scores and probabilities therefore also depend on both the query x and the context C. The functions that compute the features, weights, and biases are learned during training and remain fixed at evaluation.

The model adapts through the weights and biases while using the same query features across contexts. This restriction is very important. With arbitrary context-dependent features, a single feature could already encode the optimal probability of one of the two classes. We ask how many features are needed when they must be shared across contexts.

The Bayes-optimal probability of class k accounting for uncertainty of task τ given the query and context is

$$p_k(x) = \mathbb{P}(Y = k \mid X = x, C),$$

and the corresponding scores is

$$z_k(x) = \log p_k(x) - \frac{1}{K} \sum_{\ell=1}^K \log p_\ell(x)$$

Denote $p(x) = (p_1(x), \dots, p_K(x))^\top$ and $z(x) = (z_1(x), \dots, z_K(x))^\top$.

To compare the prediction rules on the same inputs, we evaluate their scores under a common query distribution ν, independently of the task and context. For example, ν can be the query distribution averaged over tasks. We minimize the squared score error using m shared features

$$\mathcal{R}_m = \inf_{\phi, w, b} \mathbb{E}_C \mathbb{E}_{X \sim \nu} \left\| z(X) - \widehat{z}(X) \right\|_2^2, \tag{2}$$

where the predictor $\widehat{z}(X)$ has the form in Equation (1).

3 OPTIMAL SHARED FEATURES

Each context determines a score function $z(x)$. For each context, the bias can represent the average class score

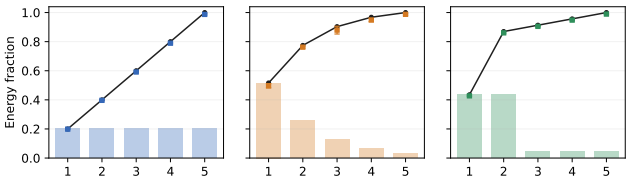


Figure 1: Predicted and held-out captured energy.

$\mathbb{E}_{X \sim \nu}[z_k(X)]$. The weighted query features therefore approximate the remaining score function

$$g_k(x) = z_k(x) - \mathbb{E}_{X \sim \nu}[z_k(X)].$$

Assume $\mathbb{E}_C \mathbb{E}_\nu \|z(X)\|_2^2 < \infty$, and define

$$(\mathcal{K}_n f)(x) = \mathbb{E}_C \sum_{k=1}^K g_k(x)\, \mathbb{E}_{X \sim \nu}[g_k(X) f(X)].$$

For a unit-norm feature f, the quantity $\mathbb{E}_\nu[f(X)(\mathcal{K}_n f)(X)]$ is the squared score contribution it captures, averaged over contexts and summed over classes.

Theorem 1. *Let $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ be the positive eigenvalues of $\mathcal{K}_n$. Then*

$$\mathcal{R}_m(n) = \sum_{j > m} \lambda_j.$$

An optimal representation consists of eigenfunctions corresponding to the largest m eigenvalues.

Each eigenvalue gives the contribution of one feature, so the best width-m representation keeps the largest contributions. In Fig. 1, the bars show the individual contributions, the black curve their predicted cumulative sum, and the squares the energy captured on held-out data. Their agreement shows that the spectrum predicts the cost of limiting width. The largest fraction captured by m shared features is

$$F_m^\star(n) = \frac{\sum_{j=1}^m \lambda_j}{\sum_{j \geq 1} \lambda_j} = 1 - \frac{\mathcal{R}_m(n)}{\mathbb{E}_C \mathbb{E}_\nu \|g(X)\|_2^2},$$

where $g = (g_1, \dots, g_K)^\top$. This fraction measures how much of the query-dependent score variation the features can represent after the biases account for the query averages.

Corollary 2. *Let $\widehat{p}_m$ be the probabilities obtained by applying softmax to an optimal score approximation using m shared features. Then*

$$\mathbb{E}_C \mathbb{E}_\nu \operatorname{KL}(p(X) \,\|\, \widehat{p}_m(X)) \leq \frac{1}{4} \sum_{j > m} \lambda_j.$$

4 FEATURES AND TASK UNCERTAINTY

The spectral characterization and the KL bound hold for any model whose Bayes scores satisfy the stated integrability assumption. We now study how the class distributions and the number of examples per class determine which shared features suffice for accurate prediction.

The exponential family has a convenient form for this problem. The query enters the log density through $\eta_{\tau,k}^\top T(x)$, while the common factor $h(x)$ cancels from the class probabilities. Consequently, when the task is known, the Bayes scores are affine functions of $T(x)$. When the task is unknown, averaging over its parameters can make these scores nonlinear in $T(x)$. We show that products of the coordinates of $T(x)$ approximate these nonlinear terms, with an error that decreases as the context grows.

Suppose $\theta = (\eta_1, \dots, \eta_K) \in \mathbb{R}^{Kd}$ has a joint density $q(\theta)$, and define

$$a_k(\theta) = \mathbb{E}[\pi_{\tau,k} \mid \theta], \qquad V_k(\theta) = -\log\big(q(\theta)a_k(\theta)\big).$$

Theorem 3. *Fix an integer $r \geq 1$. Suppose the natural parameter space is $\mathbb{R}^d$, $A \in C^{r+3}$, and*

$$\alpha I \preceq \nabla^2 A(\eta) \preceq \beta I \qquad \text{for all } \eta,$$

where $0 < \alpha \leq \beta < \infty$. Assume the derivatives of A of orders $3, \dots, r+3$ are bounded, and each $V_k \in C^{r+3}(\mathbb{R}^{Kd})$ has bounded derivatives of orders $2, \dots, r+3$. If $\mathbb{E}_\nu \|T(X)\|^{2r+2} < \infty$, and $\mathbb{E}_C \mathbb{E}_\nu \|z(X)\|_2^2 < \infty$, then the monomials in $T(x)$ of degrees one through r achieve

$$\mathcal{R}_{M_r}(n) = O(n^{-2r}), \qquad M_r = \binom{d+r}{r} - 1.$$

Here the task distribution and d, K, r, ν remain fixed as $n \to \infty$.

The additional features account for uncertainty about the class parameters. With more examples, the context identifies these parameters more accurately, so the scores approach functions that are affine in $T(x)$. Under the assumptions above, terms of higher degree become small enough that retaining degrees through r gives the stated error.

This provides an explicit choice of shared features without computing the eigenfunctions of $\mathcal{K}_n$. The coordinates of T alone give squared score error $O(n^{-2})$; adding quadratic products gives $O(n^{-4})$, and adding cubic products gives $O(n^{-6})$. The same rates bound the expected KL error under ν. Thus the result relates the number of features and the context length to approximation accuracy. These are sufficient feature counts. Fewer may suffice when terms are dependent or when they cancel across classes.

5 INFORMATION RETAINED BY THE FEATURES

The preceding theorem relies on products of the coordinates of $T(x)$ to approximate Bayes scores. Even though $T(x)$ already includes all relevant query information, these products help because the model evaluates query features through linear combinations. Define

$$S_k(C) = \sum_{i=1}^{n} T(X_{k,i}), \qquad S(C) = (n, S_1(C), \dots, S_K(C)).$$

Proposition 4. *Under the setup in Section 2, the Bayes probabilities can be computed from $T(x)$ and $S(C)$ as*

$$p_k(x \mid C) = p_k(T(x) \mid S(C))$$

The observed examples provide information about the class parameters through the sums $S_k(C)$ and the example count n. The query enters the class probabilities through $T(x)$, since the common factor $h(x)$ cancels. Thus the raw context can be replaced by Kd statistic sums and the example count without changing the Bayes prediction.

However, recovering the sufficient statistic does not guarantee that the prediction head can use it appropriately. For example, knowing $T(x)$ determines $T_1(x)^2$, but a linear head cannot generally form this square from the coordinates of $T(x)$. Adding the square as a feature lets the head use information that was already present.

This explains why a representation can perfectly recover the sufficient statistic and still give suboptimal predictions from a finite context. Uncertainty about the task can make the Bayes scores nonlinear in that statistic. Recovery measures what the representation retains, while prediction also depends on how the head combines it.

This suggests comparing prediction heads while keeping the exact query statistic and context summaries fixed. An improvement from additional products or a more flexible head then occurs without retaining any additional information from the observations.

6 ERRORS IN THE CONTEXT REPRESENTATION

The preceding results assume that the sufficient statistic is available. We now examine how errors in computing

this statistic affect the benefit of additional context. In an exponential family, the class mean of $T(X)$ equals $\nabla A(\eta)$, so averaging context features provides a direct way to estimate the class parameter. This lets us distinguish errors that average out from errors that persist as the context grows.

Consider one class with parameter η. Suppose a fixed encoder, learned independently of the evaluation context, produces $\widetilde{T}(x) \in \mathbb{R}^d$. Define the parameter estimate by

$$\nabla A(\widehat{\eta}_n) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{T}(X_i), \qquad X_i \overset{\text{iid}}{\sim} p_\eta.$$

When $\widetilde{T} = T$, this is the maximum-likelihood estimate. We isolate errors in context encoding by evaluating the resulting density $p_{\widehat{\eta}_n}$, using the exact statistic in the query likelihood.

Define the mean encoding error and the variation of the encoded features within the class:

$$b_\eta = \mathbb{E}_\eta[\widetilde{T}(X) - T(X)], \qquad V_\eta = \mathrm{Cov}_\eta(\widetilde{T}(X)).$$

Proposition 5. *Suppose the natural parameter space is $\mathbb{R}^d$, $A \in C^2$, and*

$$\alpha I \preceq \nabla^2 A(\eta) \preceq \beta I \qquad \text{for all } \eta,$$

where $0 < \alpha \le \beta < \infty$. If $\mathbb{E}_\eta\|\widetilde{T}(X)\|^2 < \infty$, then $\widehat{\eta}_n$ exists uniquely and

$$\frac{\|b_\eta\|^2 + \mathrm{tr}(V_\eta)/n}{2\beta} \le \mathbb{E}_{C|\eta}\,\mathrm{KL}(p_\eta \,\|\, p_{\widehat{\eta}_n}) \le \frac{\|b_\eta\|^2 + \mathrm{tr}(V_\eta)/n}{2\alpha}$$

The two terms have different consequences. Variation between encoded examples contributes an error that decreases as $1/n$. A nonzero mean encoding error contributes a positive lower bound that additional examples cannot remove. Thus an encoder need not compute $T(x)$ exactly at every input, but its average error within each class matters for prediction. This indicates that one should check the encoder's mean error within each class. When the error is negligible, incorporating additional examples enhances prediction. Conversely, when the error dominates, optimizing or calibrating the encoder is more useful than increasing the number of examples in the context.

7 EXPERIMENTS

Unless stated otherwise, conditions in this section use three paired seed bundles with held-out evaluation; means and sample standard deviations are descriptive because three seeds do not support narrow uncertainty estimates. Operator targets are estimated by Monte Carlo on a context split disjoint from the evaluation split. Appendix D reports protocols, the appendix diagnostics, and the complete inventory.

7.1 Context-length transfer and head adaptation

We ask whether query features learned from long contexts remain useful when fewer examples are available. We use binary classification tasks with equal class probabilities and independent parameters $\eta_1, \eta_2 \sim \mathcal{N}(0,1)$. Each class follows

$$p_\eta(x) = w_\eta\,\mathcal{N}(x;\eta{+}2,1){+}(1{-}w_\eta)\,\mathcal{N}(x;\eta{-}2,1), \qquad w_\eta = \frac{1}{1+e^{-4\eta}}.$$

This is an exponential family with

$$h(x) = \tfrac{1}{2}\mathcal{N}(x;2,1){+}\tfrac{1}{2}\mathcal{N}(x;{-}2,1), \qquad A(\eta) = \frac{\eta^2}{2}{+}\log\cosh(2\eta).$$

Its sufficient statistic is $T(x) = x$. When the class parameters are known, the Bayes scores are affine in x. With a finite context, uncertainty about these parameters can make the scores nonlinear. This lets us examine whether a representation retains the relevant statistic but fails to provide the functions needed by a linear head.

We train the same one-feature encoder separately at $n \in \{2, 4, 8, 16, 32, 64\}$. We also train a mixed-length model by sampling every integer length from 2 through 64 uniformly. Every model receives the exact statistic sums for each class and the context length, so it does not need to learn how to summarize the examples. We repeat training with five seeds.

For each evaluation length, we freeze the query encoder and train a replacement head. Figure 2 shows the full training–evaluation grid. With the original heads, performance is best near the training length and deteriorates away from it. At $n = 2$, for example, the excess loss of the model trained at $n = 64$ falls from 0.02558 to 0.00212 after head replacement. After replacement, the six encoder rows differ by at most 4.5×10^{-5} at any evaluation length. Thus the transfer gap largely reflects how the head uses the features, rather than a need to relearn them.

The mixed-length model is not included in the figure. With its original head, its mean excess loss across the six displayed evaluation lengths is 0.00748, slightly below the best fixed-length average of 0.00779. This robustness comes with a short-context compromise: its loss at $n = 2$ is 0.01471, compared with 0.00192 for training only at $n = 2$.

This distinction matters because context length changes the uncertainty about the class parameters. Features that remain useful across lengths may still require different context-dependent weights and biases. Appending $T(x)$ and $T(x)^2$ to the frozen long-trained encoder reduces its remaining gap by a further 9.4%, showing

Table 1: Width-two decomposition for Gaussian location–scale tasks (nats). The first three columns use the generative query law; the last two use the common reference measure.

n	Task unc.	Repr.	Total	M_n^{mean} tail	KL/tail
2	0.1332	0.0390	0.1722	0.2134	0.151
4	0.0832	0.0516	0.1356	0.3341	0.131
8	0.0486	0.0634	0.1117	0.4651	0.119
16	0.0268	0.0725	0.0993	0.5797	0.111
64	0.0076	0.0836	0.0913	0.7159	0.103

a smaller benefit from expanding the available query functions.

For comparison, we repeat the experiment with Gaussian classes of known variance, using identical Gaussian priors on their means. In this setting, the finite-context Bayes scores are exactly affine in x, so no additional query features are needed.

Adapting the head accounts for most of the improvement when context length changes. Additional query features help, but their effect is smaller. The results suggest checking whether the head uses an existing representation effectively before retraining the encoder.

7.2 Width–context decomposition

For every (m, n) cell we compute independently the three terms of (**??**): $\mathbb{E}\,\mathrm{KL}(p_\tau \| p_m)$, $\mathbb{E}\,\mathrm{KL}(p_\tau \| p_{C_n})$, and $\mathbb{E}\,\mathrm{KL}(p_{C_n} \| p_m)$, where p_m is the spectral projection of the predictive with the context-dependent class bias retained. The independently estimated identity closes to within 0.0039 nats over all cells and within 7×10^{-4} nats at width two. Table 1 reports width two for the Gaussian location–scale family. Task uncertainty falls from 0.133 to 0.0076 nats, approximately as $1/n$. The representation term rises from 0.039 to 0.084 nats for the reason **??** gives: the predictable tail $\sum_{j>2} \lambda_{n,j}^{\text{mean}}$ grows from 0.213 to 0.716 as posterior shrinkage relaxes. Total excess falls throughout, from 0.172 to 0.091 nats, toward the floor set by the oracle tail. Under the common reference measure of **??**, the ratio of representation KL to predictable tail lies between 0.10 and 0.15 in every cell, inside the universal ceiling 1/4 of (**??**); the lower comparison of (**??**) is vacuous here because Gaussian queries are unbounded (**??**).

7.3 A rank-constrained head reads the context

We instantiate **??** for the Gaussian family. Two hard-masked label-keyed softmax heads recover the per-class sample means and sums to floating-point precision (maximum errors 2.4×10^{-7} and 1.9×10^{-6} over three seeds). Applying the Gaussian posterior natural map and the bilinear query readout matches direct plug-in logits to a maximum centered-logit error of 3.8×10^{-6}. The plug-in–Bayes gap, measured as $\mathbb{E}\,\mathrm{KL}(p_{C_n} \| \text{plug-in})$, falls from 0.0104 nats at $n = 2$ to 2.3×10^{-5} at $n = 64$, consistent with the $O(n^{-2})$ rate. Table 2 sweeps the final query–key rank at $n = 8$. The captured context-head energy follows the predictable spectrum, and the full-rank remainder is precisely the plug-in–Bayes gap. A structured model of the same shape trained from scratch reaches a Bayes gap of 0.0141 ± 0.0040 nats after 700 steps while its first-layer states recover the per-class statistic sums with affine $R^2 > 1 - 10^{-13}$; the residual is downstream optimization, not failure to read the context.

Table 2: Query–key rank sweep of the constructed head at $n = 8$.

QK rank	Captured energy	CE gap to p_{C_n} (nats)
1	0.564	0.2154
2	0.832	0.0678
3	0.953	0.0183
4	0.991	0.0041
5	1.000	0.0012

7.4 Corrupted demonstrations

Label swaps and observation outliers are injected into the contexts at corruption fractions up to 0.375. The perturbed operator is estimated on one context split and its top-m subspace and tail are evaluated on an independent split. Predicted operator tail versus held-out corrupted residual has correlation 0.9986, mean absolute error 0.0496, and median relative error 2.7%. Across 90 width–severity–seed–mechanism conditions there are zero violations of the sandwich (5) and zero violations of the subspace bound (6). The bounds are conservative, as the theorem's worst-case geometry implies: the largest distance-to-bound ratio is 0.026 for label noise and 0.21 for outliers (Fig. 5).

7.5 Population capacity does not guarantee end-to-end recovery

A harder experiment hides five polynomial statistics of a two-dimensional signal inside a 12-dimensional invertible nonlinear mixing with ten nuisance variables. Fisher balancing reduces the task–statistic condition number from 11.84 to 1.62. With true task coefficients supplied, the encoder recovers approximately 0.91 in both designs. End-to-end recovery remains only 0.53–0.57, and balancing does not rescue it (appendix D). The result separates the population theorem from the optimization problem: the optimal width-five subspace exists, but joint learning of the encoder and context estimator does not reliably find it.

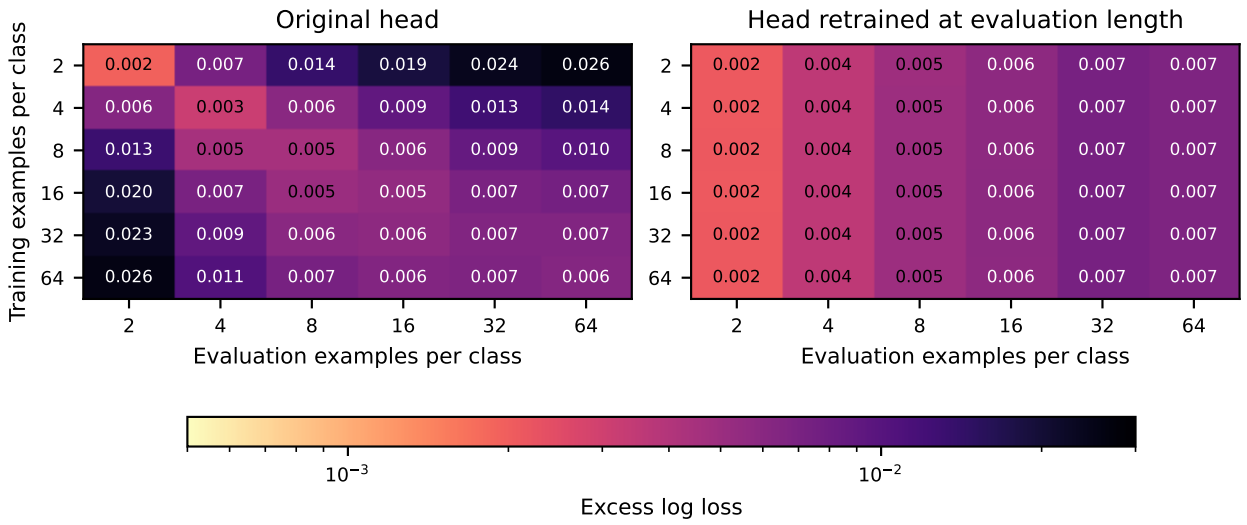


Figure 2: Context-length transfer in the bimodal-base family. The original head is sensitive to the training length, whereas replacement heads make the six encoders nearly indistinguishable. Cells show mean excess log loss over five seeds.

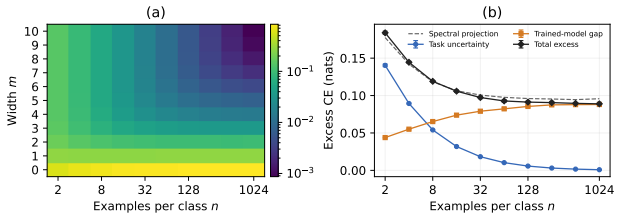


Figure 3: Width–context excess cross-entropy. Task uncertainty falls with context while the width-limited representation term rises; their sum decreases toward the oracle tail.

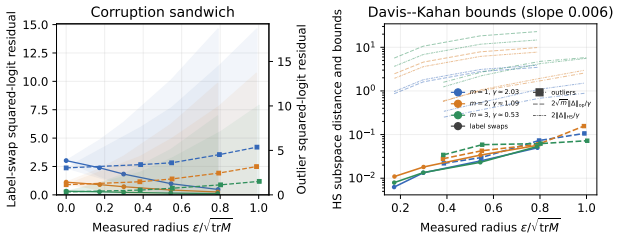


Figure 4: Corruption through the context. Left: predicted perturbed tail versus held-out residual. Right: subspace error stays below both branches of (6) across widths and eigengaps.

Code availability. Code, data, and reproduction instructions are available at https://anonymous.4open.science/r/representation-capacity-in-context-learning-34C1.

AI use statement

Generative AI tools were used to assist with manuscript formatting and editing. All AI-assisted changes are subject to author review. The authors take responsibility for the final content of this work.

References

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

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32.

Tripuraneni, N., Jin, C., and Jordan, M. (2021). Provable meta-learning of linear representations. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10434–10443. PMLR.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T.

(2022). An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Section 2 and the result statements.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable] No new algorithm is proposed.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] See `https://anonymous.4open.science/r/representation-capacity-in-context-learning-34C1`

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 2 and each result.
   (b) Complete proofs of all theoretical results. [Yes] See Appendix A and later appendices.
   (c) Clear explanations of any assumptions. [Yes] Their scope is discussed in the main text.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The linked repository regenerates all figures.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Appendix D.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Captions define metrics and seed variation.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Runs use CPU by default.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable] Only synthetic data are used.
   (b) The license information of the assets, if applicable. [Not Applicable] No third-party assets are used.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] Code and data are linked above.
   (d) Information about consent from data providers/curators. [Not Applicable] No external data are used.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable] No sensitive data are used.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable] No human subjects.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable] No human subjects.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable] No participants.

Representation Capacity in In-Context Learning of Exponential-Family Tasks: Supplementary Materials

A PROOFS

A.1 Proof of ??

The function-space notation used only in the proofs is as follows. Let $\mathcal{H} = L_2^0(\nu)$ with $\langle f, g\rangle_{\mathcal{H}} = \mathbb{E}_\nu[f(X)g(X)]$, and let

$$\mathcal{U}_\phi = \operatorname{span}\{\phi_j - \mathbb{E}_\nu \phi_j\}_{j=1}^m.$$

Define $L : \mathbb{R}^d \to \mathcal{H}$ by $(La)(x) = a^\top t(x)$, $f_{\tau,k} = L\theta_{\tau,k}$, and

$$\mathcal{C} = \mathbb{E}_\tau \sum_k f_{\tau,k} \otimes f_{\tau,k} = L\Gamma L^*, \qquad \mathcal{R}_2(\mathcal{U}) = \mathbb{E}_\tau \sum_k \|(I - P_{\mathcal{U}})f_{\tau,k}\|_{\mathcal{H}}^2.$$

Here $P_{\mathcal{U}}$ is orthogonal projection and $(f \otimes g)h = \langle g, h\rangle_{\mathcal{H}} f$. Optimizing over representations in **??** is equivalent to optimizing $\mathcal{R}_2(\mathcal{U})$ over subspaces of dimension at most m.

Proof. Let $P = P_{\mathcal{U}}$. Then

$$\begin{aligned}
\mathcal{R}_2(\mathcal{U}) &= \mathbb{E}_\tau \sum_{k=1}^K \|(I-P)L\theta_{\tau,k}\|_{\mathcal{H}}^2 \\
&= \mathbb{E}_\tau \sum_{k=1}^K \langle L\theta_{\tau,k}, (I-P)L\theta_{\tau,k}\rangle_{\mathcal{H}} \\
&= \operatorname{tr}((I-P)L\Gamma L^*) \\
&= \operatorname{tr}((I-P)\mathcal{C}) \\
&= \operatorname{tr}(\mathcal{C}) - \operatorname{tr}(P\mathcal{C}).
\end{aligned}$$

Choose an orthonormal eigenbasis (e_j) of $\mathcal{C}$ and set $a_j = \|Pe_j\|_{\mathcal{H}}^2$. Thus

$$\begin{aligned}
\operatorname{tr}(P\mathcal{C}) &= \sum_j \lambda_j a_j, \\
0 \le a_j &\le 1, \\
\sum_j a_j &= \operatorname{rank}(P) \le m, \\
\sum_j \lambda_j a_j &\le \sum_{j=1}^m \lambda_j.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathcal{R}_2(\mathcal{U}) &\ge \sum_j \lambda_j - \sum_{j=1}^m \lambda_j \\
&= \sum_{j>m} \lambda_j.
\end{aligned}$$

For $\mathcal{U} = \operatorname{span}\{e_1, \ldots, e_m\}$, equality holds. If $\lambda_m > \lambda_{m+1}$, equality requires

$$a_j = 1 \quad (j \le m), \qquad a_j = 0 \quad (j > m),$$

so the optimizer is unique.

For completeness, the finite-dimensional spectrum also admits the alternative forms

$$\operatorname{spec}_+(\mathcal{C}) = \operatorname{spec}_+(\Gamma^{1/2}\Sigma\Gamma^{1/2}) = \operatorname{spec}_+(M) = \operatorname{spec}_+(\Gamma\Sigma). \tag{3}$$

To verify these equivalences, put $A = L\Gamma^{1/2}$. Then

$$\begin{aligned} \mathcal{C} &= AA^*, \\ A^*A &= \Gamma^{1/2}L^*L\Gamma^{1/2} \\ &= \Gamma^{1/2}\Sigma\Gamma^{1/2}. \end{aligned}$$

The AA^*–A^*A identity gives the first equality of nonzero spectra in Equation (3). The AB–BA identity gives

$$\operatorname{spec}_+(\Gamma^{1/2}\Sigma\Gamma^{1/2}) = \operatorname{spec}_+(\Sigma^{1/2}\Gamma\Sigma^{1/2}).$$

Sylvester's identity gives

$$\det(tI - \Gamma\Sigma) = \det(tI - \Sigma^{1/2}\Gamma\Sigma^{1/2}),$$

including multiplicities. Therefore

$$r = \operatorname{rank}(\mathcal{C}) = \operatorname{rank}(\Sigma^{1/2}\Gamma\Sigma^{1/2}),$$

which need not equal $\operatorname{rank}(\Gamma)$ or $\operatorname{rank}(\Sigma)$. Finally,

$$\begin{aligned} \mathbb{E}_{\tau,X} \sum_k (\theta_{\tau,k}^\top t(X))^2 &= \operatorname{tr}(L\Gamma L^*) \\ &= \operatorname{tr}(\mathcal{C}) \\ &= \operatorname{tr}(M) \\ &= \sum_{j=1}^{r} \lambda_j. \end{aligned}$$

□

A.2 Proof of Proposition 6

Proposition 6. *Let $A \in \mathbb{R}^{d\times d}$ be invertible and $a \in \mathbb{R}^d$, and reparameterize*

$$T'(x) = AT(x) + a, \qquad \theta'_{\tau,k} = A^{-\top}\theta_{\tau,k}, \tag{4}$$

*which leaves every task logit unchanged. Then $\Sigma' = A\Sigma A^\top$ and $\Gamma' = A^{-\top}\Gamma A^{-1}$, and the covariance operator is exactly invariant: $\mathcal{C}' = \mathcal{C}$. Consequently the spectral tail **??**, the captured fraction **??**, the optimal subspaces in $\mathcal{H}$, and the eigenvalues of M are invariant.*

Proof. Centering Equation (4) gives

$$\begin{aligned} t' &= At, \\ L' &= LA^\top, \\ \Sigma' &= \mathbb{E}[t't'^\top] = A\Sigma A^\top, \\ \Gamma' &= A^{-\top}\Gamma A^{-1}. \end{aligned}$$

Therefore

$$\begin{aligned}\mathcal{C}' &= L'\Gamma'(L')^* \\ &= LA^\top A^{-\top}\Gamma A^{-1}AL^* \\ &= L\Gamma L^* \\ &= \mathcal{C}.\end{aligned}$$

Thus the spectral tail, its optimizing subspaces, and $F_m^\star$ are unchanged. Also

$$\begin{aligned}\Gamma'\Sigma' &= A^{-\top}\Gamma A^{-1}A\Sigma A^\top \\ &= A^{-\top}(\Gamma\Sigma)A^\top.\end{aligned}$$

Hence $\Gamma'\Sigma'$ and $\Gamma\Sigma$ are similar. By **??**, M' and M have the same eigenvalues. □

A.3 Proof of Corollary 7

Let $e_1, e_2, \ldots$ be orthonormal eigenfunctions of $\mathcal{C}$ in decreasing eigenvalue order and set $\mathcal{U}_m^\star = \operatorname{span}\{e_1, \ldots, e_m\}$.

Corollary 7. *If $\gamma_m = \lambda_m - \lambda_{m+1} > 0$, then every m-dimensional subspace $\mathcal{U}$ satisfies*

$$\mathcal{R}_2(\mathcal{U}) - \sum_{j>m}\lambda_j \geq \gamma_m \big\|(I - P_{\mathcal{U}})P_{\mathcal{U}_m^\star}\big\|_{\mathrm{HS}}^2.$$

Proof. Set $a_j = \|P_{\mathcal{U}}e_j\|^2$. Since $\dim\mathcal{U} = m$,

$$\begin{aligned}\sum_j a_j &= m, \\ \sum_{j\leq m}(1 - a_j) &= m - \sum_{j\leq m} a_j \\ &= \sum_{j>m} a_j \\ &= \big\|(I - P_{\mathcal{U}})P_{\mathcal{U}_m^\star}\big\|_{\mathrm{HS}}^2.\end{aligned}$$

Moreover,

$$\begin{aligned}\mathcal{R}_2(\mathcal{U}) - \sum_{j>m}\lambda_j &= \sum_{j\leq m}\lambda_j(1 - a_j) - \sum_{j>m}\lambda_j a_j \\ &\geq \lambda_m\sum_{j\leq m}(1 - a_j) - \lambda_{m+1}\sum_{j>m} a_j \\ &= (\lambda_m - \lambda_{m+1})\sum_{j\leq m}(1 - a_j) \\ &= \gamma_m\big\|(I - P_{\mathcal{U}})P_{\mathcal{U}_m^\star}\big\|_{\mathrm{HS}}^2.\end{aligned}$$

□

A.4 Perturbed-logit theorem and proof

Absorb each residual's ν-mean into the free bias and write $r_{\tau,k}$ for its centered part. Define the perturbed residual objective

$$\mathcal{R}_2^\varepsilon(\mathcal{U}) := \mathbb{E}_\tau\sum_k \|(I - P_{\mathcal{U}})(L\theta_{\tau,k} + r_{\tau,k})\|_{\mathcal{H}}^2.$$

Theorem 8. *Suppose the true logits satisfy $z_{\tau,k}(x) = \theta_{\tau,k}^\top t(x) + b_{\tau,k} + r_{\tau,k}(x)$ with $\mathbb{E}_{\tau,X}\|r_\tau(X)\|_2^2 \leq \varepsilon^2$. Then for every m,*

$$\left(\sqrt{\sum_{j>m}\lambda_j} - \varepsilon\right)_+^2 \leq \inf_{\dim\mathcal{U}\leq m}\mathcal{R}_2^\varepsilon(\mathcal{U}) \leq \left(\sqrt{\sum_{j>m}\lambda_j} + \varepsilon\right)^2. \tag{5}$$

Moreover the perturbed covariance operator satisfies $\mathcal{C}^\varepsilon = \mathcal{C} + \Delta$ *with* $\|\Delta\|_{\mathrm{tr}} \le 2\varepsilon\sqrt{\operatorname{tr} M} + \varepsilon^2$*; consequently* $|\lambda_j^\varepsilon - \lambda_j| \le \|\Delta\|_{\mathrm{op}}$ *for all* j*, and if* $\gamma_m > 0$ *the perturbed optimal subspace obeys*

$$\left\|(I - P_{\mathcal{U}_m^\varepsilon})P_{\mathcal{U}_m^\star}\right\|_{\mathrm{HS}} \le \frac{2\min\left(\sqrt{m}\,\|\Delta\|_{\mathrm{op}},\, \|\Delta\|_{\mathrm{HS}}\right)}{\gamma_m}. \tag{6}$$

The Hilbert–Schmidt term in Equation (6) can be bounded using the trace-norm bound above. It is the compact-self-adjoint extension of the Yu–Wang–Samworth Davis–Kahan variant (**?**), not the operator-norm $\sin\Theta$ bound. The same eigenvalue gap γ_m that gives uniqueness in Corollary 7 also controls how much the optimal feature span changes when the scores depart from the assumed form.

Proof. Put

$$\mathcal{G} = L^2(\rho \otimes \nu; \mathbb{R}^K), \qquad F(\tau, x) = (\theta_{\tau,k}^\top t(x))_{k=1}^K,$$

and absorb the ν-mean of r_τ into the free bias. Denote the remaining part by R and let $Q_{\mathcal{U}}$ apply $P_{\mathcal{U}}$ componentwise. Then

$$\begin{aligned} \|R\|_{\mathcal{G}} &\le \varepsilon, \\ \left|\|(I - Q_{\mathcal{U}})(F + R)\|_{\mathcal{G}} - \|(I - Q_{\mathcal{U}})F\|_{\mathcal{G}}\right| &\le \|(I - Q_{\mathcal{U}})R\|_{\mathcal{G}} \\ &\le \|R\|_{\mathcal{G}} \\ &\le \varepsilon. \end{aligned}$$

Taking infima over $\dim\mathcal{U} \le m$ gives

$$\left|\inf_{\dim\mathcal{U}\le m} \|(I - Q_{\mathcal{U}})(F + R)\|_{\mathcal{G}} - \inf_{\dim\mathcal{U}\le m} \|(I - Q_{\mathcal{U}})F\|_{\mathcal{G}}\right| \le \varepsilon,$$

$$\inf_{\dim\mathcal{U}\le m} \|(I - Q_{\mathcal{U}})F\|_{\mathcal{G}} = \sqrt{\sum_{j>m} \lambda_j}.$$

Therefore

$$\left(\sqrt{\sum_{j>m} \lambda_j} - \varepsilon\right)_+ \le \inf_{\dim\mathcal{U}\le m} \|(I - Q_{\mathcal{U}})(F + R)\|_{\mathcal{G}} \le \sqrt{\sum_{j>m} \lambda_j} + \varepsilon.$$

Squaring gives Equation (5).

Write

$$f_{\tau,k} = L\theta_{\tau,k}, \qquad r_{\tau,k} = R_k(\tau, \cdot).$$

Then

$$\begin{aligned} \mathcal{C}^\varepsilon &= \mathbb{E}_\tau \sum_k (f_{\tau,k} + r_{\tau,k}) \otimes (f_{\tau,k} + r_{\tau,k}) \\ &= \mathcal{C} + \Delta, \end{aligned}$$

where

$$\Delta = \mathbb{E}_\tau \sum_k \left(f_{\tau,k} \otimes r_{\tau,k} + r_{\tau,k} \otimes f_{\tau,k} + r_{\tau,k} \otimes r_{\tau,k}\right).$$

Using $\|g \otimes h\|_{\mathrm{tr}} = \|g\|_{\mathcal{H}}\|h\|_{\mathcal{H}}$,

$$\begin{aligned} \|\Delta\|_{\mathrm{tr}} &\le 2\mathbb{E}_\tau \sum_k \|f_{\tau,k}\|_{\mathcal{H}}\|r_{\tau,k}\|_{\mathcal{H}} + \mathbb{E}_\tau \sum_k \|r_{\tau,k}\|_{\mathcal{H}}^2 \\ &\le 2\left(\mathbb{E}_\tau \sum_k \|f_{\tau,k}\|_{\mathcal{H}}^2\right)^{1/2} \left(\mathbb{E}_\tau \sum_k \|r_{\tau,k}\|_{\mathcal{H}}^2\right)^{1/2} + \varepsilon^2 \\ &\le 2\varepsilon\sqrt{\operatorname{tr} M} + \varepsilon^2. \end{aligned}$$

Thus $\mathcal{C}^\varepsilon$ is positive and trace-class, hence compact and self-adjoint. Weyl's inequality gives

$$|\lambda_j^\varepsilon - \lambda_j| \le \|\Delta\|_{\mathrm{op}}.$$

For symmetric matrices, Yu–Wang–Samworth give (**?**)

$$\|\sin\Theta(\widehat{U}, U)\|_F \le \frac{2\min(\sqrt{m}\,\|E\|_{\mathrm{op}}, \|E\|_F)}{\lambda_m - \lambda_{m+1}}.$$

Let Q_n be increasing finite-rank orthogonal projections such that

$$Q_n \to I \quad \text{strongly}, \qquad \mathrm{range}(\mathcal{C}) \subseteq \mathrm{range}(Q_n).$$

Since Δ is Hilbert–Schmidt,

$$\|Q_n\Delta Q_n - \Delta\|_{\mathrm{HS}} \longrightarrow 0,$$
$$\|Q_n\Delta Q_n - \Delta\|_{\mathrm{op}} \longrightarrow 0.$$

Apply the matrix inequality to

$$Q_n\mathcal{C}Q_n, \qquad Q_n\mathcal{C}^\varepsilon Q_n.$$

The compact spectral theorem and the min–max principle give, along a convergent subsequence and with a consistent choice inside any tied perturbed eigenspace,

$$\lambda_j(Q_n\mathcal{C}^\varepsilon Q_n) \longrightarrow \lambda_j(\mathcal{C}^\varepsilon),$$
$$P_{\mathcal{U}^\varepsilon_{m,n}} \longrightarrow P_{\mathcal{U}^\varepsilon_m},$$
$$\|(I - P_{\mathcal{U}^\varepsilon_{m,n}})P_{\mathcal{U}^\star_m}\|_{\mathrm{HS}} \longrightarrow \|(I - P_{\mathcal{U}^\varepsilon_m})P_{\mathcal{U}^\star_m}\|_{\mathrm{HS}}.$$

Passing to the limit yields

$$\begin{aligned}\|(I - P_{\mathcal{U}^\varepsilon_m})P_{\mathcal{U}^\star_m}\|_{\mathrm{HS}} &= \|\sin\Theta(\mathcal{U}^\varepsilon_m, \mathcal{U}^\star_m)\|_F \\ &\le \frac{2\min(\sqrt{m}\,\|\Delta\|_{\mathrm{op}}, \|\Delta\|_{\mathrm{HS}})}{\gamma_m}.\end{aligned}$$

This is Equation (6). Finally,

$$\|\Delta\|_{\mathrm{HS}} \le \|\Delta\|_{\mathrm{tr}} \le 2\varepsilon\sqrt{\mathrm{tr}\,M} + \varepsilon^2.$$

□

A.5 When task query distributions differ slightly

Proposition 9. *Suppose each task-conditional query law satisfies $1 - \varepsilon \le dP_\tau/d\nu \le 1 + \varepsilon$ pointwise, and let $\mathcal{R}_2^{\mathrm{mix}}(\mathcal{U})$ denote the task-weighted objective in which each task chooses its own coefficients to minimize squared error in $L^2(P_\tau)$. Then for every $\mathcal{U}$,*

$$(1-\varepsilon)\,\mathcal{R}_2(\mathcal{U}) \;\le\; \mathcal{R}_2^{\mathrm{mix}}(\mathcal{U}) \;\le\; (1+\varepsilon)\,\mathcal{R}_2(\mathcal{U}),$$

and hence $(1-\varepsilon)\sum_{j>m}\lambda_j \le \inf_{\dim\mathcal{U}\le m}\mathcal{R}_2^{\mathrm{mix}}(\mathcal{U}) \le (1+\varepsilon)\sum_{j>m}\lambda_j$.

Proof. For the upper bound, evaluate each task using the coefficients that minimize its error under ν and use $dP_\tau/d\nu \le 1 + \varepsilon$. For the lower bound, apply $dP_\tau/d\nu \ge 1 - \varepsilon$ to the residual minimized under P_τ, then lower bound its $L^2(\nu)$ error by the smallest error attainable under ν. Taking the width-m infimum and applying **??** completes the claim. □

A.6 Proof of Lemma 10

Lemma 10. *For any $p \in \Delta_K$ and $u \in \mathbb{R}^K$, the quadratic form obeys*

$$u^\top H(p)u = \min_{c\in\mathbb{R}} \sum_{k=1}^{K} p_k(u_k - c)^2 = \sum_{i<j} p_i p_j (u_i - u_j)^2.$$

If p is α-interior, then

$$\alpha\|P_K u\|_2^2 \le u^\top H(p)u \le \beta_K(\alpha)\|P_K u\|_2^2$$

where $\beta_K(\alpha) = \min\{1/2,\, 1-(K-1)\alpha\}$. Universally, $H(p) \preceq \frac{1}{2}P_K$.

Proof. Let $\bar{u}_p = p^\top u$. Then

$$\begin{aligned} u^\top H(p)u &= \sum_k p_k u_k^2 - \left(\sum_k p_k u_k\right)^2 \\ &= \sum_k p_k (u_k - \bar{u}_p)^2 \\ &= \min_{c\in\mathbb{R}} \sum_k p_k (u_k - c)^2 \\ &= \frac{1}{2}\sum_{i,j} p_i p_j (u_i - u_j)^2 \\ &= \sum_{i<j} p_i p_j (u_i - u_j)^2. \end{aligned}$$

If $p_k \ge \alpha$, then

$$\begin{aligned} u^\top H(p)u &= \min_c \sum_k p_k (u_k - c)^2 \\ &\ge \alpha \min_c \sum_k (u_k - c)^2 \\ &= \alpha\|P_K u\|_2^2. \end{aligned}$$

With $\bar{u} = K^{-1}\mathbf{1}^\top u$,

$$\begin{aligned} u^\top H(p)u &\le \sum_k p_k (u_k - \bar{u})^2 \\ &\le \max_k p_k \sum_k (u_k - \bar{u})^2 \\ &\le [1-(K-1)\alpha]\|P_K u\|_2^2. \end{aligned}$$

Put $v = P_K u$. Since $v_i - v_j = u_i - u_j$,

$$\begin{aligned} u^\top H(p)u &= \sum_{i<j} p_i p_j (v_i - v_j)^2 \\ &\le 2\sum_k p_k(1-p_k)v_k^2 \\ &\le \frac{1}{2}\sum_k v_k^2 \\ &= \frac{1}{2}\|P_K u\|_2^2. \end{aligned}$$

Combining the last two upper bounds gives $\beta_K(\alpha)$. □

A.7 KL–centered-logit equivalence

Theorem 11. *For $p, q \in \Delta_K^\circ$, let $\Delta = P_K(\log q - \log p)$. Then*

$$\mathrm{KL}(p\|q) \le \frac{1}{4}\|\Delta\|_2^2.$$

If both p and q are α-interior, then

$$\frac{\alpha}{2}\|\Delta\|_2^2 \le \mathrm{KL}(p\|q) \le \frac{\beta_K(\alpha)}{2}\|\Delta\|_2^2.$$

Proof. Choose logits a, b with $p = \mathrm{softmax}(a)$ and $q = \mathrm{softmax}(b)$. Then

$$\mathrm{KL}(p\|q) = \sum_k p_k \log \frac{p_k}{q_k} = \mathrm{lse}(b) - \mathrm{lse}(a) - p^\top (b - a).$$

Common shifts cancel, so put

$$\Delta = P_K(b - a) = P_K(\log q - \log p).$$

Taylor's formula gives

$$\mathrm{KL}(p\|q) = \int_0^1 (1-t)\Delta^\top H(p_t)\Delta \, dt, \tag{7}$$

where

$$p_t = \mathrm{softmax}((1-t)a + tb),$$
$$(p_t)_k = \frac{p_k^{1-t} q_k^t}{\sum_j p_j^{1-t} q_j^t}.$$

Hölder's inequality gives

$$\sum_j p_j^{1-t} q_j^t \le \left(\sum_j p_j\right)^{1-t} \left(\sum_j q_j\right)^t = 1.$$

If $p_k, q_k \ge \alpha$, then

$$(p_t)_k \ge p_k^{1-t} q_k^t \ge \alpha.$$

Thus Lemma 10 and Equation (7) give

$$\mathrm{KL}(p\|q) \ge \int_0^1 (1-t)\alpha\|\Delta\|_2^2 \, dt = \frac{\alpha}{2}\|\Delta\|_2^2,$$
$$\mathrm{KL}(p\|q) \le \int_0^1 (1-t)\beta_K(\alpha)\|\Delta\|_2^2 \, dt = \frac{\beta_K(\alpha)}{2}\|\Delta\|_2^2.$$

Without interiority,

$$\mathrm{KL}(p\|q) \le \int_0^1 (1-t)\frac{1}{2}\|\Delta\|_2^2 \, dt$$
$$= \frac{1}{4}\|\Delta\|_2^2.$$

□

A.8 Affine identifiability

Theorem 12. *Suppose $\Sigma \succ 0$ and $\phi : \mathcal{X} \to \mathbb{R}^d$. If d linearly independent task contrasts θ_{τ_j,k_j} satisfy $w_j^\top \phi(X) + c_j = \theta_{\tau_j,k_j}^\top t(X)$ almost surely, then $\phi(X) = Bt(X) + a$ almost surely for an invertible matrix B.*

Proof. Let

$$V = [\theta_{\tau_1,k_1}, \dots, \theta_{\tau_d,k_d}], \qquad W = \begin{bmatrix} w_1^\top \\ \vdots \\ w_d^\top \end{bmatrix}, \qquad c = (c_1, \dots, c_d)^\top.$$

Stacking these equations gives

$$W\phi + c = V^\top t.$$

The matrix V is invertible because its columns are linearly independent. If W were singular, there would be a nonzero $u \in \mathbb{R}^d$ such that $u^\top W = 0$. The stacked identity would then imply $u^\top V^\top t = u^\top c$ almost surely. But $Vu \neq 0$ and

$$\mathrm{Var}(u^\top V^\top t) = u^\top V^\top \Sigma V u > 0,$$

contradicting constancy. Hence W is invertible, and

$$\phi = W^{-1}V^\top t - W^{-1}c.$$

Both W^{-1} and $V^\top$ are invertible, which proves the claim with $B = W^{-1}V^\top$ and $a = -W^{-1}c$. □

A.9 Verification of the optimal whitened features

Proof. Let $e_j(x) = u_j^\top Z(x) = u_j^\top \Sigma^{-1/2} t(x)$. For i, j,

$$\begin{aligned} \langle e_i, e_j \rangle_{\mathcal{H}} &= \mathbb{E}[(u_i^\top Z)(u_j^\top Z)] \\ &= u_i^\top \mathbb{E}[ZZ^\top] u_j \\ &= u_i^\top u_j \\ &= \delta_{ij}. \end{aligned}$$

Also,

$$\begin{aligned} L^* e_j &= \mathbb{E}[t(X) e_j(X)] \\ &= \mathbb{E}[t(X)t(X)^\top] \Sigma^{-1/2} u_j \\ &= \Sigma^{1/2} u_j. \end{aligned}$$

Hence, for every x,

$$\begin{aligned} (\mathcal{C} e_j)(x) &= (L \Gamma L^* e_j)(x) \\ &= t(x)^\top \Gamma \Sigma^{1/2} u_j \\ &= u_j^\top M \Sigma^{-1/2} t(x) \\ &= \lambda_j u_j^\top \Sigma^{-1/2} t(x) \\ &= \lambda_j e_j(x). \end{aligned}$$

Thus e_j is an eigenfunction with eigenvalue λ_j. By **??**, the optimal subspaces are spans of these functions, which verifies **??**. □

B FINITE-CONTEXT RISK DECOMPOSITION

In ICL, task parameters are not explicitly provided but inferred from a finite context C. To separate the effect of limited shared features from finite-sample task estimation, we decompose the total excess risk into three nonnegative terms.

Proposition 13. *Let τ be a latent task, C its context, and (X, Y) a query pair with $Y \perp C \mid (X, \tau)$. For a deterministic representation $\Phi = \phi(X)$ and predictor $q(Y \mid \Phi, C)$,*

$$\mathbb{E}[-\log q(Y \mid \Phi, C)] - H(Y \mid X, \tau) = \underbrace{\mathbb{E}\,\mathrm{KL}(p(\cdot \mid \Phi, C) \,\|\, q(\cdot \mid \Phi, C))}_{\textit{Decoder Gap}} + \underbrace{I(Y; X \mid \Phi, C)}_{\textit{Representation Insufficiency}} + \underbrace{I(Y; \tau \mid X, C)}_{\textit{Finite-Context Task Uncertainty}}.$$

All three terms are nonnegative.

The second term, $I(Y; X \mid \Phi, C)$, isolates information lost by the chosen representation. Width alone need not control this term without restrictions on the encoder or decoder. The third term, $I(Y; \tau \mid X, C)$, reflects task ambiguity from finite contexts and is independent of the number of shared query features.

Because estimating $H(Y \mid \Phi, C)$ by fitting a decoder on one data split and evaluating it on another can introduce estimation variance, we group the decoder gap and representation insufficiency in empirical evaluations into an exact nonnegative two-term split. This is the compact identity in **??**, with $R = \mathbb{E}[-\log q(Y \mid \Phi, C)]$, $R_C = H(Y \mid X, C)$, and $R_\tau = H(Y \mid X, \tau)$. The full three-term identity separates the grouped term conceptually, while the two-term split avoids this decoder-estimation error in the quantitative accounting below.

For these finite-context risk estimates, we use a binary latent Gaussian location–scale task with eight labeled context examples per class, independent priors $m_z \sim \mathcal{N}(0, 1.5^2)$ and $\log_2 v_z \sim \mathrm{Unif}[-1, 1]$, and a 160×120 Gauss–Hermite/Gauss–Legendre product quadrature over $(m_z, \log_2 v_z)$. Using numerical integration to evaluate the Bayes classifier given the finite context, the empirical risks are

$$\begin{aligned} R_\tau &= 0.434365, \\ R_C &= 0.466990, \\ R &= 0.470207. \end{aligned}$$

Hence the exact nonnegative decomposition **??** evaluates to

$$\begin{aligned} R - R_\tau &= 0.035842 \\ &= 0.003217 + 0.032625 \\ &= (R - R_C) \\ &\quad + (R_C - R_\tau). \end{aligned}$$

Approximately 91% of the excess risk over the optimal predictor that knows the task is irreducible uncertainty about the task after observing a finite context. This supports separating limitations of the shared features from context inference.

The context-length sweeps agree with this separation: discrete task uncertainty decays rapidly, continuous-parameter uncertainty approximately as $1/n$, and encoder recovery changes much less. The excess risk of the optimal predictor given the examples, evaluated by quadrature, falls from 0.0787 ± 0.0022 nats at context size 2 to 0.0051 ± 0.0052 at size 64. Appendix D also reports the Gamma–Poisson study, where two or three directions capture most of the predictive-score variation on the sampled domain.

B.1 Proof of Proposition 13

Proof. Let

$$R_q = \mathbb{E}[-\log q(Y \mid \Phi, C)].$$

Conditional cross-entropy gives

$$R_q = H(Y \mid \Phi, C) + \mathbb{E}\,\mathrm{KL}(p(\cdot \mid \Phi, C) \,\|\, q(\cdot \mid \Phi, C)).$$

Because Φ is a function of X,

$$\begin{aligned} I(Y; X \mid \Phi, C) &= H(Y \mid \Phi, C) - H(Y \mid X, \Phi, C) \\ &= H(Y \mid \Phi, C) - H(Y \mid X, C). \end{aligned}$$

Also,

$$\begin{aligned} I(Y;\tau \mid X,C) &= H(Y \mid X,C) - H(Y \mid X,C,\tau) \\ &= H(Y \mid X,C) - H(Y \mid X,\tau), \end{aligned}$$

where the last equality uses $Y \perp C \mid (X,\tau)$. Hence

$$\begin{aligned} R_q - H(Y \mid X,\tau) &= R_q - H(Y \mid \Phi,C) \\ &\quad + H(Y \mid \Phi,C) - H(Y \mid X,C) \\ &\quad + H(Y \mid X,C) - H(Y \mid X,\tau) \\ &= \mathbb{E}\,\mathrm{KL}(p(\cdot \mid \Phi,C) \,\|\, q(\cdot \mid \Phi,C)) \\ &\quad + I(Y;X \mid \Phi,C) + I(Y;\tau \mid X,C). \end{aligned}$$

□

B.2 Proof of ??

Proof. Conditioning on the observed context gives

$$p(Y=k, X=x \mid C) = h(x)\,\mathbb{E}\big[\pi_{\tau,k} \exp\big\{\eta_{\tau,k}^\top T(x) - A(\eta_{\tau,k})\big\} \,\big|\, C\big]\,.$$

Let the posterior expectation on the right, without $h(x)$, be $R_k(T(x))$. Normalizing over classes gives $p_k(x) = R_k(T(x))/\sum_\ell R_\ell(T(x))$. Centering the logarithms across classes cancels the common normalizing term and yields $z_k(x) = F_k(T(x))$ with F_k as in **??**. □

B.3 Proof of Theorem 1

Proof. For a closed subspace $\mathcal{U} \subseteq \mathcal{H}$, define

$$\mathcal{R}_{2,n}(\mathcal{U}) = \mathbb{E}_C \sum_k \|(I - P_{\mathcal{U}})g_k\|_{\mathcal{H}}^2.$$

For every closed subspace $\mathcal{U} \subseteq \mathcal{H}$,

$$\mathcal{R}_{2,n}(\mathcal{U}) = \mathrm{Tr}(\mathcal{K}_n) - \mathrm{Tr}(P_{\mathcal{U}}\mathcal{K}_n).$$

The integrability assumption gives $\mathrm{Tr}(\mathcal{K}_n) = \mathbb{E}_C \sum_k \|g_k\|_{\mathcal{H}}^2 < \infty$, so $\mathcal{K}_n$ is positive and trace-class. The leading-eigenspace variational principle gives the stated tail and the usual qualifications for rank and ties.

For the exponential-family consequence stated in the main text, let $\mathcal{S}_T \subseteq L^2(\nu)$ be the closed subspace of functions measurable with respect to $T(X)$. By **??**, every g_k belongs to $\mathcal{S}_T$, so the range of $\mathcal{K}_n$ is contained in $\mathcal{S}_T$. Every eigenfunction with a positive eigenvalue is therefore a function of $T(X)$. □

C WHEN THE ASSUMPTIONS DO NOT HOLD EXACTLY

Average task risk versus tasks that span every statistic direction. The identification theorem in Theorem 12 requires exact prediction for a finite set of tasks whose coefficients span every statistic direction. If training risk is averaged over a continuous task distribution, small average risk does not guarantee small error on a measure-zero basis. One remedy is a condition ensuring that the task distribution uses every statistic direction, together with least-squares recovery over task coefficients. Another is uniform risk control.

D ADDITIONAL EXPERIMENTS AND PROTOCOLS

D.1 Protocol

The experiments that vary feature width and task eigenvalues, the condition number, and the number of training tasks use ten random seeds matched across conditions. The risk decomposition integrates over the task prior and predictive distribution. Architecture comparisons report recovery using both a linear transformation plus a

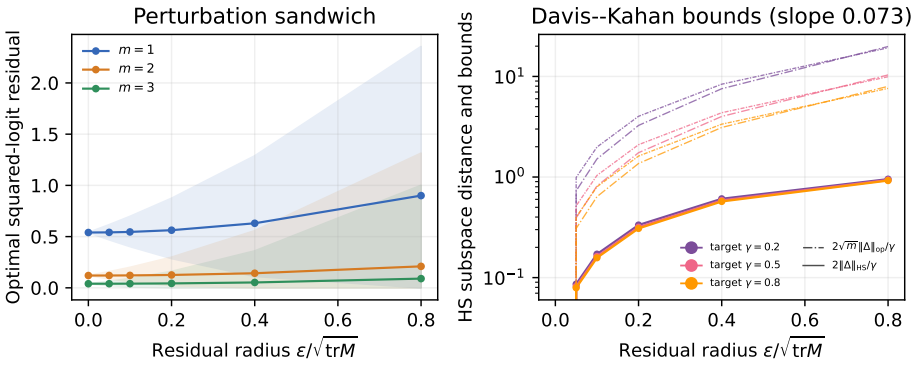


Figure 5: Numerical check under departures from the assumed score form. Left: optimal residuals remain inside the theoretical bands. Right: subspace error stays below both Davis–Kahan bounds across three eigenvalue gaps.

constant and a nonlinear decoder. We measure how much information about irrelevant variables the representation retains. Comparisons of prediction heads that account for uncertainty about the task are interpreted using the number of directions needed to capture most of the sampled predictive-score variation; the residual errors are reported below. Additional controls use the sample sizes stated in their respective subsections. Original training code for some supporting controls is not included in the experimental pipeline.

Every task is synthetic and has a known sufficient statistic and exact class score. Unless stated otherwise, representations are evaluated on held-out episodes by linear-plus-constant R^2, canonical correlations, and prediction relative to either the optimal predictor that knows the task or the optimal predictor given the examples, evaluated analytically or by accurate numerical integration. "Recovered fraction" is the mean squared canonical correlation over the target statistic dimension. Task-weighted logit energy is reported separately when varying feature width and task eigenvalues while giving the model the true task coefficients (Table 3). Raw-input and initialized-encoder decoders provide baselines for feature recovery. Nonlinear decoders are fitted on one data split and evaluated on another to assess recoverable information. Results are reported as means and sample SD across seeds. Evaluation-set bootstrap intervals quantify sampling noise but do not replace optimization-seed replication.

For the width–context study in **??**, task means have prior variances $(2, 1, 0.5, 0.2, 0.08, 0.04, 0.02, 0.01, 0.005, 0.0025)$ and the context grid is $n = 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024$ examples per class. Panel (a) evaluates the best analytic width-m score projection for $m = 0, \ldots, 10$. Its predictive coefficient is $H_n^{-1}(m_{1,n} - m_{0,n})$, and its squared-logit calculation uses the summed centered-logit normalization and independent $\nu = \mathcal{N}(0, I)$. The displayed excess cross-entropy is instead evaluated under the query distribution produced by the task model. Panel (b) trains a rank-two context model for 700 steps at every n and seed. Its evaluation makes no parameter updates and samples fresh task means. Because label-keyed averaging is fixed, we sample each Gaussian class mean directly from its exact distribution rather than materializing up to 1024 individual examples. The learned encoder, bilinear factors, and bias network are unchanged.

D.2 Checks under departures from the assumed score form

The numerical check in Figure 5 evaluates the analytic optima and bounds. It checks the implementation rather than testing learned-model behavior.

Corrupting examples does not alter the known-task Bayes logits when the latent task is unchanged. The check with corrupted examples therefore studies a different object: the score produced by a fixed finite-context estimator. Let $C = (X_{k,i})$ be a clean binary Gaussian context with n observations per class, let $\widetilde{C}$ be its label-swap or outlier-corrupted version, and let $v_1, \ldots, v_d$ be the prior variances of the class means. With classwise sample means $\bar{X}_k(C)$, the check uses the posterior-mean contrast

$$D_n = \operatorname{diag}\left(\frac{nv_\ell}{1 + nv_\ell}\right)_{\ell=1}^{d}, \qquad a(C) = D_n\big(\bar{X}_1(C) - \bar{X}_0(C)\big).$$

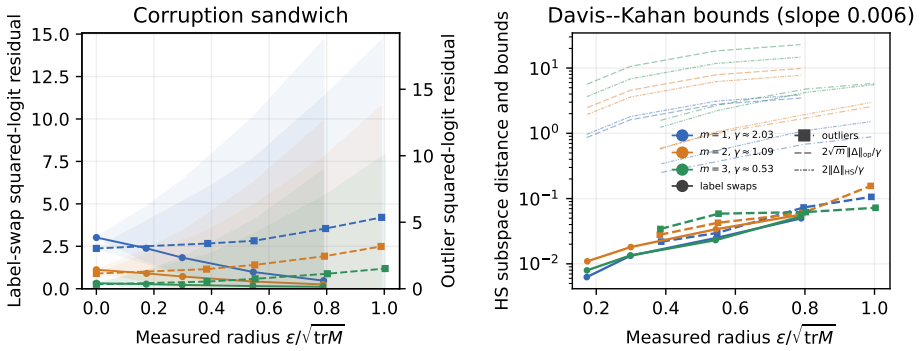


Figure 6: How corrupted examples change the scores. Left: optimal residuals for scores derived from label-swapped and outlier-corrupted contexts remain inside the perturbation sandwich. Right: subspace errors stay below both Davis–Kahan bounds for widths one through three.

Its centered context-dependent score and corrupted counterpart are

$$g_C(x) = a(C)^\top \big(x - \mathbb{E}_\nu X\big), \qquad \widetilde{g}_C(x) = a(\widetilde{C})^\top \big(x - \mathbb{E}_\nu X\big).$$

For two classes, $g \mapsto (-g/2, g/2)$ converts this scalar log-odds score into centered logits; it changes every squared energy by the same fixed factor. Thus the additive residual being measured is $r_C = \widetilde{g}_C - g_C$, with radius

$$\varepsilon^2_{\mathrm{ctx}} = \mathbb{E}_{C,\widetilde{C}} \|r_C\|^2_{L^2(\nu)} = \mathbb{E}_{C,\widetilde{C}} \left[(a(\widetilde{C}) - a(C))^\top \Sigma_\nu (a(\widetilde{C}) - a(C)) \right], \qquad \Sigma_\nu = \mathrm{Cov}_\nu(X).$$

The clean and corrupted operators are respectively $\mathbb{E}[g_C \otimes g_C]$ and $\mathbb{E}[\widetilde{g}_C \otimes \widetilde{g}_C]$; the perturbation bounds apply to these two families of scores that vary with context, not to the unchanged known-task logits. The check uses the same query law $\nu = \mathcal{N}(0, I_d)$ for every context and corruption level, independently of the context. In computation, the same centered query sample is used for every context within a condition. The objective therefore has the product form required by **??**.

D.3 Task-cardinality construction and protocol

For statistic dimension $d = 5$, let $U \in \mathbb{R}^{6\times 5}$ have orthonormal columns spanning the orthogonal complement of the all-ones vector and set $A_0 = \sqrt{6}U$. The six rows of A_0 form a centered regular simplex with empirical second moment I_5. For each seed, we independently rotate 16 copies of this simplex using orthogonal matrices obtained by QR factorization of standard Gaussian matrices, then stack the copies. The task pools are the first $N \in \{6, 12, 24, 36, 48, 60, 72, 96\}$ rows. Each nested pool is therefore a union of complete simplex blocks and has exactly zero empirical mean and identity empirical second moment.

For the representation fit, we draw 1200 training and 3000 test statistics from $\mathcal{N}(0, I_5)$. Observations contain a randomly rotated statistic with additive Gaussian noise of standard deviation 0.055, together with seven independent standard-normal nuisance coordinates. Ridge regression with penalty 10^{-5} maps observations to the task logits $T^\top a/\sqrt{5}$; the top five left singular vectors of its coefficient matrix define the encoder. Recovery is the mean of the five squared canonical correlations on the test set. For adaptation, a task vector a produces a 14-dimensional context signature $\tanh(0.72Ga) + \epsilon$, where $G_{ij} \sim \mathcal{N}(0, 1/5)$ and $\epsilon \sim \mathcal{N}(0, 0.075^2 I)$. We distribute 480 noisy signatures as evenly as possible across the training tasks and fit a ridge map with penalty 0.025 from the coordinatewise cubic features $(1, c, c^{\odot 2}, c^{\odot 3})$ to a. Cross-entropy gaps use 600 held-out task vectors drawn uniformly from the sphere of radius $\sqrt{5}$ and 800 Gaussian query statistics. All conditions share the random draws within each of ten random seeds matched across conditions.

D.4 Cross-family statistic recovery

The cross-family control repeats sufficient-statistic recovery in Gaussian, Bernoulli, and Poisson families. The model accuracies and the accuracies of the optimal predictor that knows the task are respectively 0.749/0.756, 0.636/0.647, and 0.709/0.720. This control uses one seed per family.

Table 3: Varying feature width and task eigenvalues while giving the model the true task coefficients. "Pred." is $F_m^\star$ from **??**. Learned energy is measured by fitting a linear decoder on one data split and evaluating it on another.

Eigenvalue pattern	m	Predicted fraction	Held-out fraction	CE gap
All equal	1	0.200	0.199	0.0726
All equal	2	0.400	0.396	0.0548
All equal	3	0.600	0.591	0.0377
All equal	4	0.800	0.789	0.0195
All equal	5	1.000	0.987	0.0013
Successive halves	1	0.516	0.497	0.0490
Successive halves	2	0.774	0.762	0.0242
Successive halves	3	0.903	0.879	0.0130
Successive halves	4	0.968	0.949	0.0054
Successive halves	5	1.000	0.988	0.0014
Two dominant	1	0.435	0.427	0.0535
Two dominant	2	0.870	0.860	0.0153
Two dominant	3	0.913	0.905	0.0102
Two dominant	4	0.957	0.948	0.0054
Two dominant	5	1.000	0.990	0.0016

Table 4: Full architecture comparison. Nonlinear decoders recover the quadratic statistic even when coordinates recoverable by a linear transformation plus a constant vary.

Architecture	Accuracy	Affine $R^2(s^2)$	Nonlinear $R^2(s^2)$	Quadratic-logit R^2
Single head	0.755 ± 0.007	0.881 ± 0.023	0.999	0.982
Deep Sets	0.750 ± 0.005	0.493 ± 0.374	0.999	0.971
Set Transformer	0.756 ± 0.006	0.942 ± 0.016	0.999	0.951
Transformer, no position	0.753 ± 0.006	0.949 ± 0.035	1.000	0.881
Transformer, positional	0.751 ± 0.005	0.915 ± 0.031	1.000	0.883

D.5 Observation-level, task-relevant, and context-level statistics

The observation- and context-level comparison tests task-relevant statistic recovery. Across three seeds, the full task has linear-plus-constant $R^2(s) = 0.980 \pm 0.002$ and $R^2(s^2) = 0.920 \pm 0.017$. A mean-only width-one model has $R^2(s) = 0.998$ but $R^2(s^2) = 0.0006$, showing that recovery is task relevant rather than indiscriminate inversion of the hidden flow. At the context level, the pooled representation linearly decodes $\sum_i s_i$ with $R^2 = 0.991$ and $\sum_i s_i^2$ with $R^2 = 0.958$. Fitting a conditional decoder on one split and evaluating it on another gives $\mathrm{CE}(\phi) - \mathrm{CE}(X) = -0.00346 \pm 0.00135$ nats. The near-zero value supports predictive sufficiency, but its slight negativity is decoder-estimation error and must not be interpreted as an estimate of information loss, which cannot be negative.

The observation encoder retains $T(X_i)$, whereas the context mechanism aggregates approximately $\sum_i T(X_i)$ to infer episode-specific task coefficients. This comparison does not isolate a unique aggregation mechanism.

D.6 Controlled task eigenvalues and condition numbers

The encoder in the task-eigenvalue experiment is trained without access to the eigenvectors or their optimal projection. When all eigenvalues are equal, optimal subspaces below full rank are nonunique; evaluation therefore uses captured energy rather than alignment to a selected eigenbasis.

The implementation uses the shorthand names *balanced*, *decaying*, and *gapped*. Concretely, their normalized eigenvalue profiles are $(.2, .2, .2, .2, .2)$, approximately $(.516, .258, .129, .065, .032)$, and approximately $(.435, .435, .0435, .0435, .0435)$, respectively. The descriptive names in Table 3 emphasize what these profiles mean.

D.7 Architecture comparison and initialization controls

Raw inputs and initialized encoders already decode substantial signal under the one-dimensional monotone flow: averaged across architectures, raw X linearly predicts s with $R^2 = 0.972$, while initialized encoders predict s and

Table 5: Gamma–Poisson prediction from finite examples. Bayes prediction given the examples matters when the context is small, while a few directions capture most of the sampled score variation.

Context/class	CE given context	Posterior-mean CE	MLE CE	Bilinear gap	Nonlinear gap
1	0.52218	0.54453	1.41722	0.00049	0.00006
3	0.46651	0.47115	0.71884	0.00077	0.00010
10	0.43698	0.43760	0.48648	≈ 0	≈ 0
30	0.42590	0.42594	0.43438	≈ 0	≈ 0

s^2 with $R^2 = 0.955$ and 0.754. Consequently, absolute decoding scores in these studies are not by themselves evidence of emergence. To determine what training changes, we vary which statistics the tasks require, remove selected representation components, and test whether the representation discards irrelevant input variables.

D.8 Prediction from examples and the number of important directions

Over the sampled domain, approximating the scores of the optimal predictor given the examples with two shared query functions gives a relative error of 0.0027–0.0037, and using four functions reduces it below 9×10^{-5}. Here relative error is the Frobenius norm of the residual score matrix divided by the Frobenius norm of the full score matrix. Thus, formal infinite-rank structure alone does not force a nonlinear query head in this experiment. To distinguish the models, we need tasks whose predictive scores are poorly approximated by the number of features available to the bilinear model.

D.9 Design for varying the task condition number

The latent signal $s = (s_1, s_2)$ follows the full bivariate Gaussian exponential family with sufficient statistic

$$T(s) = (s_1, s_2, s_1^2, s_1 s_2, s_2^2)^\top$$

and reference natural parameter $\eta_0 = (0, 0, -0.75, 0, -0.75)^\top$. For each binary task, set $v_\tau = Ba_\tau$ and use

$$\eta_{\tau,1} = \eta_0 - \tfrac{1}{2} v_\tau, \qquad \eta_{\tau,2} = \eta_0 + \tfrac{1}{2} v_\tau. \tag{8}$$

The observed input is $X = f(s, n)$, where $n \sim \mathcal{N}(0, I_{10})$ is nuisance variation and f is a fixed invertible nonlinear coupling flow on $\mathbb{R}^{12}$.

To vary the condition number without changing the sum of the eigenvalues, let $\Sigma = \mathrm{Cov}_{\eta_0}[T(s)]$ and use $N = 500$ task coefficients whose entries form a cosine tight frame,

$$(a_\tau)_j = \sqrt{2} \cos\left(\frac{\pi(\tau + \frac{1}{2})j}{N} \right), \quad \tau = 0, \dots, N-1, \quad j = 1, \dots, 5. \tag{9}$$

Hence $N^{-1} \sum_\tau a_\tau = 0$ and $N^{-1} \sum_\tau a_\tau a_\tau^\top = I_5$. For $\kappa \in \{1, 2, 4, 8, 16\}$, the implementation uses the scalar log-odds spectrum

$$\widetilde{\lambda}_j(\kappa) = 0.4 \, \frac{\kappa^{-(j-1)/4}}{\sum_{\ell=0}^{4} \kappa^{-\ell/4}}, \qquad j = 1, \dots, 5, \tag{10}$$

and set

$$B_\kappa = \Sigma^{-1/2} Q \, \mathrm{diag}\big(\sqrt{\widetilde{\lambda}_1(\kappa)}, \dots, \sqrt{\widetilde{\lambda}_5(\kappa)} \big), \tag{11}$$

where the same orthogonal matrix Q is used across conditions. Writing $\widetilde{\Gamma}_\kappa = N^{-1} \sum_\tau v_\tau v_\tau^\top$, the scalar log-odds operator is

$$\widetilde{M}_\kappa = \Sigma^{1/2} \widetilde{\Gamma}_\kappa \Sigma^{1/2} = Q \, \mathrm{diag}(\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_5) Q^\top.$$

For the binary parameters in Equation (8), the centered contrasts used by the theory are $\theta_{\tau,1} = -v_\tau/2$ and $\theta_{\tau,2} = v_\tau/2$. Hence

$$\Gamma_\kappa = \frac{1}{2} \widetilde{\Gamma}_\kappa, \qquad M_\kappa = \frac{1}{2} \widetilde{M}_\kappa, \qquad \lambda_j = \frac{1}{2} \widetilde{\lambda}_j.$$

Thus $\mathrm{Tr}(M_\kappa) = 0.2$ and $\mathrm{cond}(M_\kappa) = \lambda_1 / \lambda_5 = \kappa$ exactly. The factor of two changes neither the generated tasks nor any normalized spectral fraction. The statistic, nonlinear mixer architecture and mixer random seed, number of representation coordinates (five), 24 context examples per class, task count, and optimization budgets are held fixed.

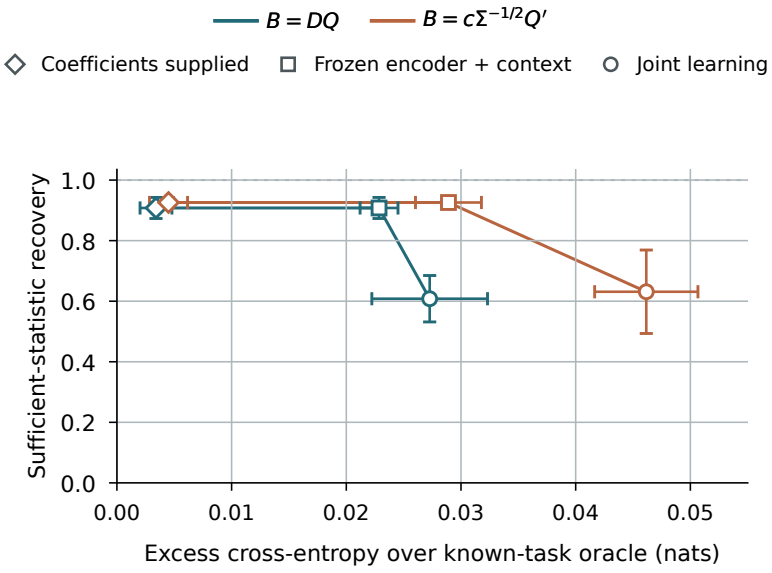


Figure 7: Staged comparison of statistic recovery and prediction for the two coefficient maps in Equation (12). Excess cross-entropy is measured against the optimal predictor that knows the task. The first segment changes the context estimator while holding the encoder fixed; the second also relearns the representation. Horizontal and vertical error bars show sample standard deviations over ten random seeds matched across conditions. Lines connect experimental conditions and are not training trajectories.

D.10 Additional results for nonlinear input mixing

An additional experiment with the nonlinear mixer finds that raw observations decode nuisance coordinates with $R^2 \simeq 0.979$, whereas a learned width-three representation has nuisance R^2 near zero. Thus training discards nuisance variation.

The separate staged comparison draws $a_\tau \sim \mathcal{N}(0, 0.6^2 I_5)$ and compares

$$B_{\mathrm{sc}} = DQ, \quad D = \mathrm{diag}(1, 1, 0.35, 0.35, 0.35), \qquad B_{\mathrm{wh}} = c\Sigma^{-1/2}Q', \tag{12}$$

where Q, Q' are orthogonal and c matches $\mathrm{Tr}(\Sigma^{1/2} B B^\top \Sigma^{1/2})$ between the two maps. This staged comparison is distinct from the exact fixed-trace sweep in Equations (9) and (10).

Using the two coefficient maps in Equation (12), the staged comparison separates learned context inference at a fixed encoder from joint representation and context learning. Across ten random seeds matched across conditions, statistic recovery when the true task coefficients are given is 0.908 ± 0.035 and 0.926 ± 0.016 for B_{sc} and B_{wh}, respectively, followed by end-to-end recovery 0.608 ± 0.077 and 0.631 ± 0.138. In Figure 7, the frozen stage holds the encoder trained with the true task coefficients fixed while learning the context estimator, which locates the predictive gap without changing representation recovery. Figure 8 compares the averages obtained with the true task coefficients against those from joint learning, separately for each canonical direction: the encoder trained with the true task coefficients recovers all five directions strongly, whereas joint learning loses most of the fourth and especially the fifth direction.

E ADDITIONAL EXPERIMENTAL TABLES

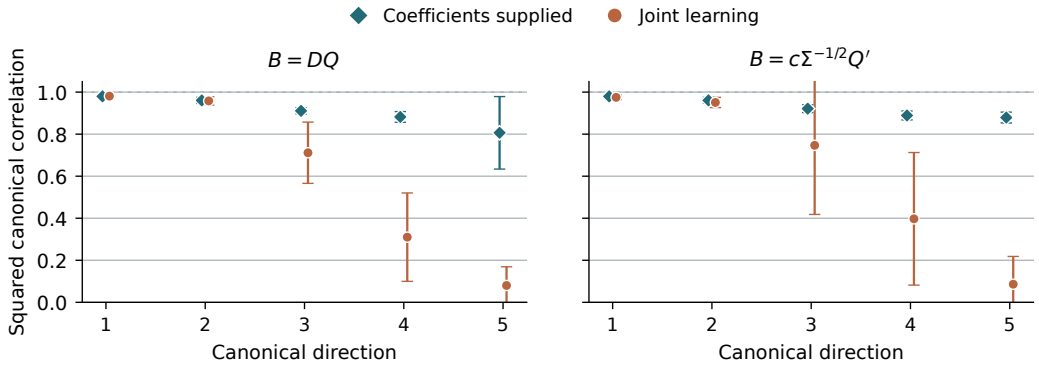


Figure 8: Direction-wise sufficient-statistic recovery for the two task maps in Equation (12). Each point is the mean squared canonical correlation ρ_j^2, and error bars are sample standard deviations over ten random seeds matched across conditions. The frozen-context model is omitted because its encoder is the encoder trained with the true task coefficients and held fixed, so its representation recovery is identical by construction.

Table 6: Condition-number sweep with a fixed eigenvalue sum. The five entries in the spectrum column are the prescribed centered-logit eigenvalues of M_κ in descending order; each underlying row sums to 0.2 before rounding. The implementation's scalar log-odds eigenvalues are twice these values. Cross-entropy and accuracy are for the optimal predictor that knows the task. All entries are means ± sample SD over ten random seeds matched across conditions.

κ	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$	Known-task CE	Known-task accuracy	Recovery with true coefficients	Joint recovery
1	$(0.0400, 0.0400, 0.0400, 0.0400, 0.0400)$	0.6448 ± 0.0023	0.6119 ± 0.0029	0.927 ± 0.016	0.456 ± 0.080
2	$(0.0549, 0.0462, 0.0388, 0.0326, 0.0275)$	0.6450 ± 0.0023	0.6110 ± 0.0033	0.926 ± 0.018	0.626 ± 0.077
4	$(0.0712, 0.0503, 0.0356, 0.0252, 0.0178)$	0.6453 ± 0.0023	0.6103 ± 0.0033	0.923 ± 0.019	0.699 ± 0.081
8	$(0.0876, 0.0521, 0.0310, 0.0184, 0.0109)$	0.6455 ± 0.0023	0.6097 ± 0.0044	0.900 ± 0.064	0.713 ± 0.081
16	$(0.1032, 0.0516, 0.0258, 0.0129, 0.0065)$	0.6456 ± 0.0023	0.6093 ± 0.0053	0.889 ± 0.072	0.708 ± 0.092

Table 7: Nonlinear input-mixing experiment with separate stages for learning the representation and the context-dependent weights. Entries are means and sample SD over ten random seeds matched across conditions. CE entries are excess cross-entropy over the optimal predictor that knows the task.

Task map	cond(M)	Recovery, coefficients given	CE, coefficients given	CE with fixed encoder	Joint recovery	Joint C
$B_{\text{sc}} = DQ$	12.37 ± 1.06	0.908 ± 0.035	0.0034 ± 0.0014	0.0229 ± 0.0017	0.608 ± 0.077	0.0273 ± 0.005
$B_{\text{wh}} = c\Sigma^{-1/2}Q'$	1.579 ± 0.076	0.926 ± 0.016	0.0045 ± 0.0017	0.0289 ± 0.0029	0.631 ± 0.138	0.0462 ± 0.004