On JEPA Isotropy

Thomas Y.L. Lin[1] **Han Liu**[2,3] **Jerry Yao-Chieh Hu**[2,4]

[1] Department of Physics, University of Washington, Seattle
[2] Center for Foundation Models and Generative AI &
Department of Computer Science, Northwestern University
[3] Department of Statistics and Data Science, Northwestern University
[4] Department of Computer Science, College of Connected Computing, Vanderbilt University

thermaus@uw.edu, hanliu@northwestern.edu, jhu@u.northwestern.edu

Abstract

We study the empirical risk minimization (ERM) of linear Joint-Embedding Predictive Architecture (JEPA) and establish its key conditions for optimality. By casting the embedding bottleneck as a rank constraint, we formulate JEPA as a reduced-rank regression problem and study ERM via standard oracle inequality. We show that JEPA's oracle risk decreases as target heterogeneity raises the rank until it saturates at the bottleneck. Moreover, JEPA's excess risk is governed by the spectral conditioning of the JEPA prediction. At any fixed oracle risk, it reaches optimality at a flat spectrum, termed *predictive isotropy*. We show that this condition, enforced by a regularizer, subsumes canonical JEPA isotropic embedding designs and hence justifies their empirical successes through ERM. Synthetic experiments corroborate our theory. Realistic experiments show our proposed regularizer lowers the test risk by 11% and the generalization gap by 57% from the existing paradigm.

1 Introduction

We study the empirical risk minimization (ERM) of Joint-Embedding Predictive Architecture (JEPA) [1, 2] and establish its conditions for optimality. JEPA jointly trains a context encoder and a predictor to predict the target embedding. It has become a prominent self-supervised learning paradigm [2–5]. A key recipe in JEPA is to enforce isotropy on the encoder's marginal embedding, either through architectures [2–4] or through distributional regularizations [6, 5]. Prior work justifies this recipe by appealing to typical self-supervised methods [7–9], by showing it prevents representational [10, 11] and early-training [12] collapse, or by scoping to specific objectives [5, 13]. More fundamentally, these recipes regularize only the encoder and ignore the predictor, the defining component of JEPA beyond plain Joint-Embedding Architectures (JEAs).

As a result, it remains unclear what JEPA loss is optimizing. This work targets these gaps through a finite-sample ERM analysis. Specifically, we answer:

(Q1) How does learning more features lower JEPA's risk under a fixed embedding dimension?

(Q2) What condition allows *joint* learning of JEPA's encoder and predictor to reach optimality?

(Q3) Why do encoder-isotropy recipes work in practice, and what unifies them?

Casting the embedding bottleneck as a rank constraint, we formulate JEPA as reduced-rank regression [14, 15] and study its population risk via standard oracle inequality (Theorem 3.15). When JEPA predicts K targets from one context [2, 3] through a d-dimensional embedding, its ERM $\widehat{B}_d$ satisfies,

$$\mathcal{R}_K(\widehat{B}_d) \leq \underbrace{\mathcal{R}^\star_K(d)}_{\text{oracle risk}} + \underbrace{\mathcal{E}_K(d)}_{\text{excess risk}},$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

where $\mathcal{E}_K(d)$ shrinks with the sample size and depends on the JEPA prediction through the spectrum of its covariance. We answer (Q1) by showing that $\mathcal{R}^\star_K(d)$ decreases as JEPA learns more features until it saturates at the embedding bottleneck (Propositions 3.4 and 3.9). Even after saturation, heterogeneous targets keep shrinking $\mathcal{E}_K(d)$ (Remarks 3.12 and 3.13). We answer (Q2) by showing that $\mathcal{E}_K(d)$ reaches optimality when the JEPA prediction is isotropic, a condition termed *predictive isotropy* (Lemma 4.1 and Theorem 4.2). This condition improves $\mathcal{E}_K(d)$ without raising $\mathcal{R}^\star_K(d)$, so it complements the gains from learning more features.

We answer (Q3) by showing that encoder isotropy [16, 6, 13, 5] is a special case of predictive isotropy. We introduce a unified view of JEPA isotropy as a gauge equivalence class. Specifically, JEPA's risk depends on the encoder f and predictor g only through their composition $g \circ f$. Observe that, for any invertible transformation G on the encoder f, the predictor g can absorb it by

$$(g \circ G^{-1}) \circ (G \circ f) = g \circ f.$$

We call such a G a *gauge*, and show every gauge preserves predictive isotropy and its optimal excess risk. Whitening, VICReg, Gaussian-embedding, and LeJEPA [16, 6, 13, 5] each select a gauge toward a specific marginal embedding (Theorems 4.5 and 4.6), so predictive isotropy unifies them.

Building on this unification, we derive a maximum-entropy regularizer for predictive isotropy that acts on JEPA predictions rather than the embedding. Empirically, it improves both test risk and train-test gap over encoder isotropy on MNIST and CIFAR images. Under weight decay, predictive isotropy makes the embedding isotropic (Proposition 4.8), so the regularizer also drives it closer to an isotropic Gaussian. Synthetic experiments further validate the intermediate claims of our analysis.

Contribution. Our contributions include:

- **ERM Theory of JEPA.** We recast JEPA as reduced-rank regression and derive a finite-sample guarantee on its excess risk (Theorem 3.15). We show that heterogeneous targets first lower the oracle risk by adding features, and then lower the sample complexity once the features saturate the embedding (Proposition 3.4 and Remark 3.13).
- **Identifying the Optimality Condition.** We show that, at fixed oracle risk, the excess risk reaches optimality when the JEPA predictions are isotropic over their range. This condition, predictive isotropy, constrains the encoder and predictor jointly rather than the encoder alone (Theorem 4.2).
- **Justification and Unification of Encoder Isotropy.** We give a unified risk-based explanation for empirically successful encoder-isotropy recipes, including Whitening, VICReg, Gaussian-embedding, and LeJEPA. Under predictive isotropy, these methods arise as gauge choices in the encoder-predictor factorization (Theorems 4.5 and 4.6 and Proposition 4.8).
- **Regularization Toward the Strongest Guarantee.** We translate predictive isotropy into a practical regularizer (4.3). Empirically, it reduces test risk and train-test gap, while inducing an encoder embedding closer to an isotropic Gaussian as a natural byproduct rather than a constraint.

Roadmap. Section 2 introduces preliminaries. Section 3 develops the reduced-rank regression formulation and the oracle inequality. Section 4 establishes that predictive isotropy optimizes the excess risk and shows it subsumes existing encoder-isotropy designs. Section 5 reports numerical validations. Section 6 concludes. Sections B to D contain extended discussion and related work, full proofs, and experimental details, respectively.

2 Preliminaries

In this section, we introduce the notation and background needed for our analysis.

Let $z \in \mathbb{R}^D$ denote a data sample drawn from a distribution $\mathcal{P}$ on $\mathbb{R}^D$. A *mask* is an index set $m \subseteq [D]$ selecting a subset of coordinates. A *mask distribution* $\mathcal{D}$ is a probability distribution over tuples $(m_c, m_{t,1}, \ldots, m_{t,K})$ specifying one *context block* and $K \geq 1$ *target blocks*.[1] For a sampled tuple, write $z_{\mathrm{ctx}} := z_{m_c} \in \mathbb{R}^{D_x}$ and $z^{(k)}_{\mathrm{tgt}} := z_{m_{t,k}} \in \mathbb{R}^{D_y^{(k)}}$ for the corresponding context and target subvectors. JEPA jointly trains an encoder and a predictor to predict the embedding of each target block from the embedding of the context block.

Architecture. JEPA consists of three components:

(i) A *context encoder* $f_\theta : \mathbb{R}^{D_x} \to \mathbb{R}^d$ mapping the context block to an embedding $x = f_\theta(z_{\mathrm{ctx}})$;

(ii) A *target encoder* (*teacher*) $f_{\bar{\theta}} : \mathbb{R}^{D_y^{(k)}} \to \mathbb{R}^d$ mapping each target to an embedding $y^{(k)} = f_{\bar{\theta}}(z^{(k)}_{\mathrm{tgt}})$;

[1] We call a mask a *block mask* when it selects a contiguous index set. The target blocks need not be disjoint.

(iii) A *predictor* $g_W : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ that predicts the k-th target embedding $y^{(k)}$ from the context embedding x and target-side metadata $q_k \in \mathcal{Z}$, i.e., $\widehat{y}^{(k)} = g_W(x, q_k)$.

Throughout this work, we assume a frozen teacher and a linear end-to-end map.

Assumption 2.1 (Frozen Teacher)**.** The target encoder $f_{\bar{\theta}}$ is fixed during training of the context encoder f_θ and predictor g_W.

Assumption 2.2 (End-to-end linearity)**.** The end-to-end map $g_W \circ f_\theta : \mathbb{R}^{D_x} \to \mathbb{R}^{Kd}$ is linear.

For explicit matrix-level analysis (Sections 3.1 to 3.5), we consider a specialized linearity assumption.

Assumption 2.3 (Linear factorization)**.** The context encoder is linear, and for each target block $k \in [K]$ with fixed metadata q_k, the predictor is linear in the context embedding:

$$f_\theta(z_{\mathrm{ctx}}) = A z_{\mathrm{ctx}}, \quad g_W(x, q_k) = W_k x, \qquad A \in \mathbb{R}^{d \times D_x}, \quad W_k \in \mathbb{R}^{d \times d}.$$

Note Assumption 2.3 implies Assumption 2.2 and serves as the working factorization. Section 4 relaxes it to nonlinear factorizations within the same end-to-end linear map (Lemma 4.4).

We freeze the teacher to turn JEPA into regression onto fixed targets. Li et al. [17] find that frozen-teacher training matches or beats EMA-based JEPA at equal compute. Linearity then gives this regression a closed-form solution, as in standard analyses of self-supervised learning [11, 10].

Under Assumption 2.3, stacking the target embeddings,

$$Y := \begin{bmatrix} (y^{(1)})^\top & \cdots & (y^{(K)})^\top \end{bmatrix}^\top \in \mathbb{R}^{Kd}, \tag{2.1}$$

$$\widehat{Y} = \begin{bmatrix} \widehat{y}^{(1)} \\ \vdots \\ \widehat{y}^{(K)} \end{bmatrix} = Wx, \qquad W := \begin{bmatrix} W_1 \\ \vdots \\ W_K \end{bmatrix} \in \mathbb{R}^{Kd \times d}. \tag{2.2}$$

Let $Z := z_{\mathrm{ctx}}$ and regard the end-to-end map from the context block to the stacked target embedding as $B := WA \in \mathbb{R}^{Kd \times D_x}$. Since B factors through the d-dimensional embedding, $\mathrm{rank}(B) \le d$. Conversely, every rank-r matrix B with $r \le d$ admits a factorization $B = WA$ with $A \in \mathbb{R}^{d \times D_x}$ and $W \in \mathbb{R}^{Kd \times d}$. Thus, for the linear frozen-teacher analysis, optimizing over encoder-predictor pairs (W, A) is equivalent to optimizing over rank-constrained maps B.

Unless otherwise stated, all expectations are taken over $z \sim \mathcal{P}$ and $(m_c, m_{t,1}, \ldots, m_{t,K}) \sim \mathcal{D}$.

Definition 2.4 (JEPA's population risk)**.** Write $Z := z_{\mathrm{ctx}}$, stack the target embeddings $y^{(k)} := f_{\bar{\theta}}(z_{\mathrm{tgt}}^{(k)})$ into Y as in (2.1), and the predictor blocks W_k into W as in (2.2). JEPA's population risk is

$$\mathcal{R}_K(B) := \frac{1}{K} \mathbb{E}_{Z,Y}\left[\|Y - BZ\|^2\right], \quad B \in \mathbb{R}^{Kd \times D_x}, \tag{2.3}$$

and we write $\mathcal{R}_K(W, A) = \mathcal{R}_K(WA)$ for the risk of an encoder-predictor pair.

The composition WA acts on the context Z only through its row space, a low-dimensional subspace of $\mathbb{R}^{D_x}$ that captures the directions JEPA effectively operates on. We name it next.

Definition 2.5 (Predictive subspace)**.** $\mathcal{V}$ is a *predictive subspace* if $\mathcal{V} = \mathrm{row}(WA)$ for some WA.

Denote the second moments $\Sigma_{ZZ} := \mathbb{E}[ZZ^\top] \in \mathbb{R}^{D_x \times D_x}$, $\Sigma_{Y^{(k)}Z} := \mathbb{E}[y^{(k)} Z^\top] \in \mathbb{R}^{d \times D_x}$ for $k \in [K]$, and their stack $\Sigma_{YZ} := \mathbb{E}[YZ^\top] = [\Sigma_{Y^{(1)}Z}^\top, \ldots, \Sigma_{Y^{(K)}Z}^\top]^\top \in \mathbb{R}^{Kd \times D_x}$. Assume $\Sigma_{ZZ} \succ 0$, and for n i.i.d. samples (Z_i, Y_i) write $\widehat{\Sigma}_{ZZ} := \frac{1}{n}\sum_i Z_i Z_i^\top$ and $\widehat{\Sigma}_{YZ} := \frac{1}{n}\sum_i Y_i Z_i^\top$.

3 Main Theory

Sections 3.1 to 3.3 show how target structure and the embedding bottleneck saturate the oracle risk; Section 3.4 shows how spectral conditioning governs the excess risk beyond saturation; Section 3.5 combines both into a finite-sample guarantee via standard oracle inequality (Theorem 3.15).

3.1 JEPA as Reduced-Rank Regression

We show that, under frozen teacher and linearity (Assumptions 2.1 and 2.3), JEPA reduces to a Reduced-Rank Regression (RRR) problem. For $r \le d$, consider the rank-constrained JEPA. Its ERM objective minimizes $\widehat{\mathcal{R}}_K(B) := (nK)^{-1} \sum_{i=1}^n \|Y_i - BZ_i\|^2$, with minimizers

$$B_r^\star \in \operatorname*{argmin}_{\mathrm{rank}(B) \le r} \mathcal{R}_K(B), \quad \widehat{B}_r \in \operatorname*{argmin}_{\mathrm{rank}(B) \le r} \widehat{\mathcal{R}}_K(B), \tag{3.1}$$

where $\mathcal{R}_K(B)$ is the JEPA risk in (2.3). Recall Σ_{ZZ}, Σ_{YZ} from Section 2 and define

$$\underbrace{B_{\mathrm{OLS}}^{\mathrm{raw}} := \Sigma_{YZ}\Sigma_{ZZ}^{-1}}_{\text{unconstrained OLS map}}, \qquad \underbrace{T_K := B_{\mathrm{OLS}}^{\mathrm{raw}}\Sigma_{ZZ}^{1/2} = \Sigma_{YZ}\Sigma_{ZZ}^{-1/2}}_{\text{whitened OLS solution}}. \tag{3.2}$$

Proposition 3.1 (JEPA as reduced-rank regression)**.** Under $\Sigma_{ZZ}, \widehat{\Sigma}_{ZZ} \succ 0$, Assumptions 2.1 and 2.3, the optimization (3.1) is reduced-rank regression (Definition A.1) with $p = D_x$, $q = Kd$, covariate $Z \in \mathbb{R}^{D_x}$, and response $Y \in \mathbb{R}^{Kd}$. Consequently, Lemma A.2 gives the closed forms

$$B_r^\star = [T_K]_r \, \Sigma_{ZZ}^{-1/2}, \qquad \widehat{B}_r = [\widehat{\Sigma}_{YZ}\widehat{\Sigma}_{ZZ}^{-1/2}]_r \, \widehat{\Sigma}_{ZZ}^{-1/2}, \tag{3.3}$$

where $[\cdot]_r$ truncates the SVD of a matrix to its r largest singular values.

The singular values of T_K rank whitened context directions by how much of the targets they predict, and the rank-r solution uses only the r most predictive ones. Below we study when increasing the number of target blocks can enlarge the predictive structure accessible from the context.

3.2 Multi-Target Masking and Rank Growth

Here we characterize JEPA's rank growth. The predictor in (3.1) is subject to two rank bottlenecks: the explicit constraint $r \le d$ and an intrinsic rank fixed by data. Lemma 3.2 makes the latter precise.

Lemma 3.2 (Rank of the JEPA solution)**.** Let $B_r^\star$ be the solution to Proposition 3.1. Then

$$\operatorname{rank}(B_r^\star) = \min\{r, \operatorname{rank}(T_K)\}. \tag{3.4}$$

Proof. See Section C.1 for a detailed proof. □

By Lemma 3.2, the targets affect the rank of the JEPA solution through $\operatorname{rank}(T_K)$. By (3.2), each target block $k \in [K]$ has its own whitened OLS solution $M_k := \Sigma_{Y^{(k)}Z}\Sigma_{ZZ}^{-1/2}$, and T_K stacks them. Adding a block adds rows to T_K, so $\operatorname{rank}(T_K)$ can only grow.

We define a predictive rank (Definition 3.3, Proposition 3.4, and Corollary 3.5), which measures *how large* the predictable subspace is, and a heterogeneity level (Definition 3.6 and Proposition 3.8), which measures *how well* it can be recovered from samples.

Definition 3.3 (Predictive rank)**.** For a given K, the *predictive rank* is $r_{\mathrm{pop}}(K) := \operatorname{rank}(T_K)$.

As $T_K \in \mathbb{R}^{Kd \times D_x}$, the predictive rank obeys $r_{\mathrm{pop}}(K) \le \min\{Kd, D_x\}$. Intuitively, $r_{\mathrm{pop}}(K)$ is the predictive subspace dimension before the architectural truncation. We study how it depends on K.

Proposition 3.4 (Rank growth)**.** For every $K \ge 1$,

1. **Monotonicity.** $r_{\mathrm{pop}}(K+1) \ge r_{\mathrm{pop}}(K)$.
2. **Strict Growth under Complementarity.** $r_{\mathrm{pop}}(K+1) > r_{\mathrm{pop}}(K)$ iff $\ker(T_K) \not\subseteq \ker(M_{K+1})$.

Proof. See Section C.2 for a detailed proof. □

Corollary 3.5 (Saturation)**.** There exists $K^\star$ with $r_{\mathrm{pop}}(K) = r_{\mathrm{pop}}(K^\star) \le D_x$ for all $K \ge K^\star$.

Besides the predictive rank, it is useful to track how far T_K is from losing rank. To this end, define the predictive Gram matrix

$$H_K := T_K^\top T_K = \sum_{k=1}^{K} M_k^\top M_k \in \mathbb{R}^{D_x \times D_x}. \tag{3.5}$$

Throughout, $\lambda_i(\cdot)$, $\sigma_i(\cdot)$ and $\|\cdot\|_{\mathrm{op}}$ denote the i-th largest eigenvalue, the i-th largest singular value and the operator norm.

Definition 3.6 (Heterogeneity level)**.** The heterogeneity level λ_{het} is the smallest nonzero eigenvalue of H_K, i.e., $\lambda_{\mathrm{het}} := \lambda_{r_{\mathrm{pop}}(K)}(H_K)$.

Since $H_K = T_K^\top T_K$ shares the rank as T_K, $\sqrt{\lambda_{\mathrm{het}}}$ is the smallest nonzero singular value of T_K:

$$\sigma_{r_{\mathrm{pop}}(K)}(T_K) = \sqrt{\lambda_{\mathrm{het}}}. \tag{3.6}$$

Remark 3.7 (Heterogeneity as complementarity)**.** The rank strictly increases at step $K+1$ if and only if the new block's whitened OLS solution M_{K+1} acts nontrivially on $\ker(T_K)$, i.e., contributes a predictive direction outside the span accumulated from the first K blocks. Heterogeneity is therefore complementarity among the blockwise operators rather than their number.

The next proposition shows that λ_{het} improves how well the predictive subspace of T_K is identified.

Proposition 3.8 (Heterogeneity controls predictive-subspace identifiability)**.** Assume the family $\{M_k\}_{k=1}^K$ has heterogeneity level $\lambda_{\text{het}} > 0$ (Definition 3.6). Let $\widehat{T}_K$ be any estimator of T_K and let $\mathcal{V}_{K,r_{\text{pop}}}, \widehat{\mathcal{V}}_{K,r_{\text{pop}}}$ denote the top-$r_{\text{pop}}(K)$ right singular subspaces of T_K and $\widehat{T}_K$, respectively. Suppose the event $E_T^{\text{abs}} := \{\|\widehat{T}_K - T_K\|_{\text{op}} \le \eta_n^T(\delta)\}$ satisfies $\mathbb{P}(E_T^{\text{abs}}) \ge 1-\delta$ for some $\eta_n^T(\delta) < \sqrt{\lambda_{\text{het}}}$. Then, on E_T^{abs},

$$\|\sin\Theta(\mathcal{V}_{K,r_{\text{pop}}}, \widehat{\mathcal{V}}_{K,r_{\text{pop}}})\|_{\text{op}} \le \frac{\eta_n^T(\delta)}{\sqrt{\lambda_{\text{het}}} - \eta_n^T(\delta)}.$$

Proof. See Section C.3 for a detailed proof. □

Under the half-gap condition $\eta_n^T(\delta) \le \sqrt{\lambda_{\text{het}}}/2$, the bound is at most $2\eta_n^T(\delta)/\sqrt{\lambda_{\text{het}}}$. Thus stronger heterogeneity tightens the $\sin\Theta$ bound at a fixed whitened-operator perturbation level. The envelope $\eta_n^T(\delta)$ is realized by standard sub-Gaussian concentration (Corollary A.5).

3.3 Capacity Limits of the Embedding Bottleneck

Here we study the architectural bottleneck. Since JEPA predicts through a d-dimensional embedding $x = Az_{\text{ctx}}$, its end-to-end map has rank at most d. This limits approximation even when the context contains a richer predictive structure. The next result quantifies the irreducible approximation error.

Proposition 3.9 (Capacity decomposition from the embedding bottleneck)**.** For any $r \le d$,

$$\inf_{\text{rank}(B)\le r} \mathbb{E}\left[\|Y - BZ\|^2\right] = \mathbb{E}\left[\|Y - B_{\text{OLS}}^{\text{raw}} Z\|^2\right] + \sum_{j>r} \sigma_j(T_K)^2. \tag{3.7}$$

Proof. See Section C.4 for a detailed proof. □

Here r is the reduced-rank constraint (3.1), which the architecture fixes at the embedding dimension d. The tail $\sum_{j>d}\sigma_j(T_K)^2$ is then irreducible, since no choice of (W, A) can recover predictive content beyond the top-d singular directions of T_K. This sharpens the saturation of Section 3.2, since the embedding dimension d is a ceiling that extra target blocks cannot bypass.

3.4 Post-Saturation Gains via Eigenvalue Balancing

We now show that in the saturated regime $r_{\text{pop}}(K) = d$, extra targets no longer enlarge the predictive subspace but can still lower the excess risk by making it more robust to noise. From n samples and on the whitened context $\Sigma_{ZZ}^{-1/2} Z$, the ERM's estimate of T_K is $\widehat{T}_K\widehat{F}$ with $\widehat{T}_K := \widehat{\Sigma}_{YZ}\Sigma_{ZZ}^{-1/2}$ and $\widehat{F} := \Sigma_{ZZ}^{1/2}\widehat{\Sigma}_{ZZ}^{-1}\Sigma_{ZZ}^{1/2}$. We bound the noise in both factors by the events

$$E_T(K;\rho_n) := \left\{\|\widehat{T}_K - T_K\|_{\text{op}} \le \rho_n(\delta)\|T_K\|_{\text{op}}\right\}, \qquad E_\Sigma(\varepsilon_n) := \left\{\|\widehat{F} - I\|_{\text{op}} \le \varepsilon_n(\delta)\right\}, \tag{3.8}$$

where $\rho_n(\delta)$ bounds the relative estimation error and $\varepsilon_n(\delta)$ bounds the whitening error of $\widehat{\Sigma}_{ZZ}$ (Corollary A.6). Suppose both hold with probability at least $1-\delta$, and call $\bar{\rho}_n(\delta) := \rho_n(\delta) + 4\varepsilon_n(\delta)$ the effective noise level. For any $r \le d$, let $\mathcal{V}_{K,r}$ be the top-r right singular subspace of T_K and $\widehat{\mathcal{V}}_{K,r}$ the ERM's estimate of it (Lemma A.7). Write $\mathcal{R}_K^\star(r) := \mathcal{R}_K(B_r^\star)$ for the rank-r oracle risk.

Lemma 3.10 (Excess risk of the learned predictor)**.** For a fixed target count K and rank $r \le d$, let $L := r\,\sigma_1(T_K)^2/K$. Then the rank-$r$ ERM $\widehat{B}_r$ satisfies, if $\widehat{\rho} := \|\widehat{T}_K - T_K\|_{\text{op}}/\|T_K\|_{\text{op}} \le 1$ and $\widehat{\varepsilon} := \|\widehat{F} - I\|_{\text{op}} \le \frac{1}{2}$,

$$\mathcal{R}_K(\widehat{B}_r) - \mathcal{R}_K^\star(r) \le L\Big(\underbrace{(\widehat{\rho} + 4\widehat{\varepsilon})^2}_{\text{coefficient-estimation error}} + \underbrace{(\|\sin\Theta(\mathcal{V}_{K,r}, \widehat{\mathcal{V}}_{K,r})\|_{\text{op}} + \widehat{\varepsilon})^2}_{\text{subspace-selection error}}\Big).$$

Proof. See Section C.5 for a detailed proof. □

The subspace-selection error grows as the weakest signal $\sigma_d(T_K)$ shrinks toward the noise level $\bar{\rho}_n(\delta)\|T_K\|_{\text{op}} = \bar{\rho}_n(\delta)\sigma_1(T_K)$. This error is therefore controlled by the spectral conditioning of T_K:

$$\kappa_K^2 := \frac{\lambda_d(H_K)}{\lambda_1(H_K)} = \left(\frac{\sigma_d(T_K)}{\sigma_1(T_K)}\right)^2 \in (0,1], \qquad \text{with } \kappa_K = 1 \text{ for a flat spectrum.} \tag{3.9}$$

Proposition 3.11 shows that after saturation, extra targets lower the excess risk when they raise κ_K.

Proposition 3.11 (Post-saturation risk decomposition)**.** Fix K with $r_{\text{pop}}(K) = d$ and let $L = d\,\sigma_1(T_K)^2/K$. If the effective noise level satisfies $\bar{\rho}_n(\delta) \le \kappa_K/2$ (half-gap), then on the event

$E_T(K;\rho_n) \cap E_\Sigma(\varepsilon_n)$, assumed to hold with probability at least $1-\delta$,

$$\mathcal{R}_K(\widehat{B}_d) \le \underbrace{\mathcal{R}^\star_K(d)}_{\text{oracle risk}} + \underbrace{\frac{7L\,\bar{\rho}_n(\delta)^2}{\kappa_K^2}}_{\text{excess risk}}. \tag{3.10}$$

Proof. See Section C.6 for a detailed proof. □

The half-gap condition only excludes vacuous bounds, because without it the excess-risk term exceeds $7L/4$, which is larger than the excess risk $\frac{1}{K}\|T_K\|_F^2 \le L$ of predicting zero. The excess risk in (3.10) is governed by two finite-sample effects: shrinkage in n at fixed K (Remark 3.12) and post-saturation conditioning gains in K at fixed n (Remark 3.13).

Remark 3.12 (Sample complexity)**.** The excess risk is controlled by the effective noise level $\bar{\rho}_n(\delta)$. By Corollaries A.5 and A.6, $\bar{\rho}_n(\delta) = \widetilde{O}(\sqrt{r_{\mathrm{rel}}/n})$ for an effective dimension r_{rel}, so $\mathcal{R}_K(\widehat{B}_d) - \mathcal{R}^\star_K(d) \le \varepsilon$ is obtained at $n = \widetilde{O}(L\,r_{\mathrm{rel}}/(\kappa_K^2\varepsilon))$.

Remark 3.13 (Second saturation)**.** Beyond the population saturation of $r_{\mathrm{pop}}(K)$ in Corollary 3.5, Proposition 3.11 reveals a second, finite-sample saturation once $r_{\mathrm{pop}}(K) = d$. Blocks added along weak directions keep raising κ_K, and hence lowering the sample complexity, until $\kappa_K = 1$.

3.5 Risk Reduction in Multi-Target JEPA

We now bound the risk of the learned predictor $\widehat{B}_r$ at every rank $r \le d$, covering JEPA both while its rank grows and after it saturates. Noise as large as the gap $\sigma_r(T_K) - \sigma_{r+1}(T_K)$ can make $\widehat{\mathcal{V}}_{K,r}$ contain the $(r+1)$-th right singular vector of T_K instead of the r-th. Therefore, we require the effective noise level $\bar{\rho}_n(\delta)$ to stay below the relative gap $\overline{\Delta}_{K,r} := \big(\sigma_r(T_K) - \sigma_{r+1}(T_K)\big)/\|T_K\|_{\mathrm{op}}$.

Assumption 3.14 (Gap-compatible relative perturbation)**.** For the fixed target count K and rank $r \le d$, the events $E_T(K;\rho_n)$ in (3.8) and $E_\Sigma(\varepsilon_n)$ satisfy $\mathbb{P}(E_T(K;\rho_n) \cap E_\Sigma(\varepsilon_n)) \ge 1-\delta$ for some $\rho_n(\delta), \varepsilon_n(\delta)$ with $\bar{\rho}_n(\delta) < \overline{\Delta}_{K,r}$.

Theorem 3.15 (Oracle inequality)**.** Under Assumption 3.14, on $E_T(K;\rho_n) \cap E_\Sigma(\varepsilon_n)$, with $L = r\,\sigma_1(T_K)^2/K$,

$$\mathcal{R}_K(\widehat{B}_r) \le \underbrace{\frac{1}{K}\left(\mathbb{E}\left[\|Y - B^{\mathrm{raw}}_{\mathrm{OLS}} Z\|^2\right] + \sum_{j>r}\sigma_j(T_K)^2\right)}_{\mathcal{R}^\star_K(r),\ \text{rank-}r\ \text{oracle risk}} + \underbrace{\frac{3L\bar{\rho}_n(\delta)^2}{(\overline{\Delta}_{K,r} - \bar{\rho}_n(\delta))^2}}_{\text{excess risk}}. \tag{3.11}$$

Proof. See Section C.7 for a detailed proof. □

Theorem 3.15 is a high-probability oracle inequality, the standard form of ERM guarantees [18]. As Lemma 3.10 is quadratic, a relative perturbation of order $n^{-1/2}$ gives an excess risk of order n^{-1}. Once the predictive rank saturates at d, the relative gap $\overline{\Delta}_{K,d}$ equals the spectral conditioning κ_K, and Theorem 3.15 under the half-gap condition recovers Proposition 3.11 up to a constant factor.

Remark 3.16 (Risk-reduction levers)**.** Beyond the irreducible floor $\frac{1}{K}\mathbb{E}\,\|Y - B^{\mathrm{raw}}_{\mathrm{OLS}} Z\|^2$, the bound (3.11) has two levers. *Capacity.* Complementary targets raise $r_{\mathrm{pop}}(K)$ (Proposition 3.4), but the architecture keeps at most d directions, leaving the tail $\frac{1}{K}\sum_{j>d}\sigma_j(T_K)^2$ when $r_{\mathrm{pop}}(K) > d$. *Conditioning.* The excess risk depends on the relative gap $\overline{\Delta}_{K,r}$.

4 Predictive Isotropy

Theorem 3.15 decomposes JEPA's population risk into an oracle risk and an excess risk governed by the conditioning of T_K. We show that predictive isotropy, a flat spectrum of T_K, optimizes it (Section 4.1) and constrains only the end-to-end map, so whitening, VICReg, Gaussian-embedding, and LeJEPA [16, 6, 13, 5] arise as its gauge choices that make the embedding isotropic (Section 4.2).

4.1 Optimality Condition

JEPA's risk depends on the embedding only through the end-to-end map WA. To impose a flat spectrum of T_K through distributional regularization, we therefore state it as a condition on the JEPA prediction $P_K := T_K\widetilde{Z}$ with whitened context $\widetilde{Z} := \Sigma_{ZZ}^{-1/2} Z$.

Lemma 4.1 (Predictive isotropy)**.** Assuming $\Sigma_{ZZ} \succ 0$, $\mathbb{E}[\widetilde{Z}] = 0$, and $\mathrm{rank}(T_K) = d$, the JEPA prediction $P_K = T_K \widetilde{Z}$ has isotropic covariance on its rank-d predictive range iff

$$\lambda_1(H_K) = \lambda_2(H_K) = \cdots = \lambda_d(H_K), \qquad \text{where } H_K := T_K^\top T_K. \tag{4.1}$$

Proof. See Section C.8 for a detailed proof. □

At $r = d$, the excess risk in (3.11) depends on T_K only through $L = d\,\lambda_1(H_K)/K$ and $\overline{\Delta}_{K,d} = \kappa_K$, and at a fixed trace of H_K, a flat spectrum minimizes the first and maximizes the second.

Theorem 4.2 (Optimality of predictive isotropy)**.** Fix $\tau > 0$ and $\bar{\rho}_n(\delta) > 0$, and consider operators T_K satisfying $\mathrm{rank}(T_K) = d$, $\mathrm{tr}(H_K) = \tau$ and $\bar{\rho}_n(\delta) < \kappa_K$. Among all such operators, the isotropic spectrum $\lambda_1(H_K) = \cdots = \lambda_d(H_K) = \tau/d$ uniquely minimizes the excess risk of (3.11) at $r = d$.

Proof. See Section C.9 for a detailed proof. □

At fixed trace τ and target second moment $\mathbb{E}\|Y\|^2$, every operator in Theorem 4.2 has the same oracle risk $\mathcal{R}_K^\star(d) = \frac{1}{K}(\mathbb{E}\|Y\|^2 - \tau)$, so predictive isotropy strengthens the guarantee without raising the oracle risk. This optimum is an instance of a broader majorization principle (Section A.3). Beyond the guarantee, isotropy also minimizes the expected excess risk itself in a Gaussian model.

Proposition 4.3 (Predictive isotropy minimizes the expected risk)**.** Let $d, K \geq 2$, $D_x > d$, $\sigma > 0$, $Z \sim N(0, I_{D_x})$ and $Y = T_K Z + \epsilon$ with independent $\epsilon \sim N(0, \sigma^2 I_{Kd})$, where $\mathrm{rank}(T_K) = d$ and $\mathrm{tr}(H_K) = \tau$. Write $\widehat{B}_d$ for the ERM (3.3) on n i.i.d. draws $\mathcal{S}$ of (Z, Y). For all $n \geq N(d, K, D_x, \tau, \sigma)$ and all such T_K, the expected excess risk $\mathbb{E}_{\mathcal{S}}[\mathcal{R}_K(\widehat{B}_d)] - \mathcal{R}_K^\star(d)$ is minimal exactly when $\lambda_1(H_K) = \cdots = \lambda_d(H_K)$. For each such T_K, with $\widehat{B}_d^{\mathrm{iso}}$ the ERM at isotropy,

$$\mathbb{E}_{\mathcal{S}}[\mathcal{R}_K(\widehat{B}_d)] - \mathbb{E}_{\mathcal{S}}[\mathcal{R}_K(\widehat{B}_d^{\mathrm{iso}})] = \frac{3d(K-1)(D_x-d)\sigma^4}{n^2 K}\Big(\sum_{i=1}^{d} \frac{1}{\lambda_i(H_K)} - \frac{d^2}{\tau}\Big) + o(n^{-2}). \tag{4.2}$$

Unlike the guarantee, which sees the spectrum only through $\lambda_1(H_K)$ and κ_K, the gap (4.2) weighs every direction by $1/\lambda_i(H_K)$, so the weakest directions dominate it.

4.2 Gauge Choices and Regularization

Predictive isotropy constrains only the end-to-end map and leaves its split into encoder and predictor free. We now use this freedom to explain existing encoder-isotropy designs as particular splits and to derive a regularizer that acts on the map directly.

Within Lemma 4.1's condition, the encoder-predictor factorization remains a structural choice. We specify one that bridges the predictive isotropy with the embedding isotropy of canonical JEPA paradigms [5]. Write the rank-d SVD of the whitened OLS solution as $T_K = U_K S_K V_K^\top$, and let $V_K^\top \widetilde{Z} \in \mathbb{R}^d$ be the *SVD coordinate* of the whitened context.

Lemma 4.4 (Gauge invariance of the rank-d optimum)**.** Let a gauge $G : \mathbb{R}^d \to \mathbb{R}^d$ be any measurable map with an a.e. inverse G^{-1} on the support. Define an encoder-predictor factorization

$$\Lambda_G := G \circ V_K^\top \Sigma_{ZZ}^{-1/2}, \qquad F_G := U_K S_K \circ G^{-1}.$$

Then $F_G \circ \Lambda_G$ realizes the population rank-d prediction P_K for every choice of G.

By Lemma 4.4, predictive isotropy (Lemma 4.1) is gauge-invariant, so it constrains T_K without committing to any embedding distribution. We can thus choose a gauge G for downstream purposes. In particular, it can realize either of the two encoder regularization paradigms in JEPA: covariance isotropy (whitening [16] or VICReg-style [6]) or Gaussian isotropy (LeJEPA [5], SIGReg-style).

Theorem 4.5 (Subsumption of covariance isotropy)**.** Assume $\mathbb{E}[Z] = 0$. For any orthogonal gauge $G : \mathbb{R}^d \to \mathbb{R}^d$, the encoder-predictor factorization (Λ_G, F_G) in Lemma 4.4 induces an isotropic covariance embedding $\mathrm{Cov}(\Lambda_G(Z)) = I_d$.

Proof. See Section C.11 for a detailed proof. □

Theorem 4.5 realizes covariance isotropy with a linear gauge. We next show that a nonlinear gauge can realize the strictly stronger property of isotropic-Gaussian embedding under a mild assumption.

Theorem 4.6 (Subsumption of Gaussian isotropy)**.** Assume $V_K^\top \widetilde{Z}$ has an absolutely continuous distribution on $\mathbb{R}^d$. Then there exists a gauge $G^\star : \mathbb{R}^d \to \mathbb{R}^d$ such that the encoder-predictor factorization $(\Lambda_{G^\star}, F_{G^\star})$ in Lemma 4.4 induces an isotropic Gaussian embedding $\Lambda_{G^\star}(Z) \sim N(0, I_d)$.

Proof Sketch. Since the distribution of $V_K^\top \widetilde{Z}$ has a density by assumption, Brenier's theorem [19] gives a map $G^\star$ pushing it to $N(0, I_d)$, and since $N(0, I_d)$ also has a density, $G^\star$ is invertible almost everywhere. Using an a.e. inverse on the support, the encoder-predictor pair $(\Lambda_{G^\star}, F_{G^\star})$ then satisfies the conclusion by construction (Lemma 4.4). See Section C.12 for a detailed proof. □

When the context is Gaussian, the gauge in Theorem 4.6 collapses to the identity.

Corollary 4.7 (Linear-Gaussian special case)**.** If the context follows a Gaussian distribution $Z \sim N(0, \Sigma_{ZZ})$, the encoder-predictor factorization reduces to the linear pair

$$\Lambda_I = V_K^\top \Sigma_{ZZ}^{-1/2}, \qquad F_I = U_K S_K.$$

Proof. See Section C.13 for a detailed proof. □

Subsumption in Theorems 4.5 and 4.6 is one-way. Encoder isotropy alone does not imply predictive isotropy, since it leaves the predictor free to give the end-to-end map any conditioning. Moreover, by Lemma 4.4, every choice of G realizes the same prediction $P_K = T_K \widetilde{Z}$ with the same optimality condition in Theorem 4.2, so selecting a gauge in Theorems 4.5 and 4.6 costs no excess risk.

Below we turn this principle into a regularizer. Theorem 4.2 and Proposition 4.3 show that the excess risk and its guarantee are smallest under predictive isotropy, so our regularizer pushes the end-to-end map toward it. Theorems 4.5 and 4.6 and Proposition 4.8 further show that, under such regularization, encoder isotropy need not be imposed separately, nor must a gauge in Lemma 4.4 be chosen manually.

Maximum-entropy regularization induces isotropic-Gaussian embedding. The excess risk prescribes only the spectral condition (4.1) on H_K, equivalently covariance isotropy of P_K on its rank-d predictive range. This places candidate regularizers on a two-axis frontier: *fidelity*, whether $\mathcal{R} = 0$ implies (4.1), and *commitment*, how much distributional structure $\mathcal{R} = 0$ enforces beyond it. The least-committing faithful choice is the maximum-entropy satisfying predictive isotropy. At trace τ, Lemma A.10 shows that this is uniquely $N(0, \frac{\tau}{d} I_d)$ on the intrinsic predictive range.

Therefore, we propose a maximum-entropy regularizer for the intrinsic predicted output $\widehat{P}_K^{\text{int}} := \widehat{U}_K^\top \widehat{W}\widehat{A} Z \in \mathbb{R}^d$, where $\widehat{U}_K \in \mathbb{R}^{Kd \times d}$ spans the learned rank-d predictive range. By the Cramér–Wold theorem [20], in its hyperspherical form [5, Lemma 3], the condition $\widehat{P}_K^{\text{int}} \sim N(0, I_d)$ (trace $\tau = d$) is equivalent to $u^\top \widehat{P}_K^{\text{int}} \sim N(0, 1)$ for every $u \in \mathbb{S}^{d-1}$. We approximate this all-direction condition by sampling $|\mathcal{A}|$ random unit vectors $\mathcal{A} \subset \mathbb{S}^{d-1}$, projecting the predictor output onto each direction, and averaging the Epps–Pulley standard-normal statistic EP [21]:

$$\mathcal{R}_{\text{PI}}(\widehat{P}_K^{\text{int}}; \mathcal{A}) := \frac{1}{|\mathcal{A}|} \sum_{u \in \mathcal{A}} \text{EP}\left(\{u^\top \widehat{p}_i\}_{i=1}^N\right), \tag{4.3}$$

where $\widehat{p}_i := \widehat{U}_K^\top \widehat{W}\widehat{A} Z_i$ is the intrinsic predicted output on the i-th of N samples in a minibatch.

As in Lemma 4.1, the covariance part of $N(0, I_d)$ enforces predictive isotropy on the learned end-to-end map $\widehat{W}\widehat{A}\Sigma_{ZZ}^{1/2}$. The Gaussian part is its maximum-entropy completion. (4.3) is therefore the predictive analogue of LeJEPA's SIGReg [5], constraining $\mathcal{L}(\widehat{P}_K^{\text{int}})$ rather than $\mathcal{L}(AZ)$.

Weight decay selects the gauge. The loss and any regularizer of the end-to-end map are invariant to the gauge, so they do not determine which embedding training reaches. JEPA training typically uses weight decay [22, 2, 5] that penalizes the encoder and predictor separately and thereby selects the gauge up to rotation. In this gauge, an isotropic embedding is equivalent to predictive isotropy. Write the encoder in whitened coordinates as $\widetilde{A} := A\Sigma_{ZZ}^{1/2}$, so that $AZ = \widetilde{A}\widetilde{Z}$. Given samples $\{(Z_i, Y_i)\}_{i=1}^n$ with $\widetilde{Z}_i := \Sigma_{ZZ}^{-1/2} Z_i$, training minimizes the JEPA loss with a regularizer $\mathcal{R}$ of the end-to-end map $W\widetilde{A}$ and ℓ_2 weight decay $\beta > 0$,

$$\min_{W, \widetilde{A}} \frac{1}{nK} \sum_{i=1}^n \|Y_i - W\widetilde{A}\widetilde{Z}_i\|^2 + \mathcal{R}(W\widetilde{A}) + \frac{\beta}{2}\left(\|W\|_F^2 + \|\widetilde{A}\|_F^2\right). \tag{4.4}$$

Proposition 4.8 (Isotropy under weight decay)**.** Assume $\mathbb{E}[Z] = 0$ and that $\mathcal{R}$ is differentiable, and let $s > 0$. At any minimizer of (4.4), $\text{Cov}(AZ) = sI_d$ if and only if $\sigma_1(W\widetilde{A}) = \cdots = \sigma_d(W\widetilde{A}) = s$. If moreover $U^\top W A Z \sim N(0, I_d)$ for an orthonormal basis U of $\text{col}(W)$, then $AZ \sim N(0, I_d)$.

Proof. See Section C.14 for a detailed proof. □

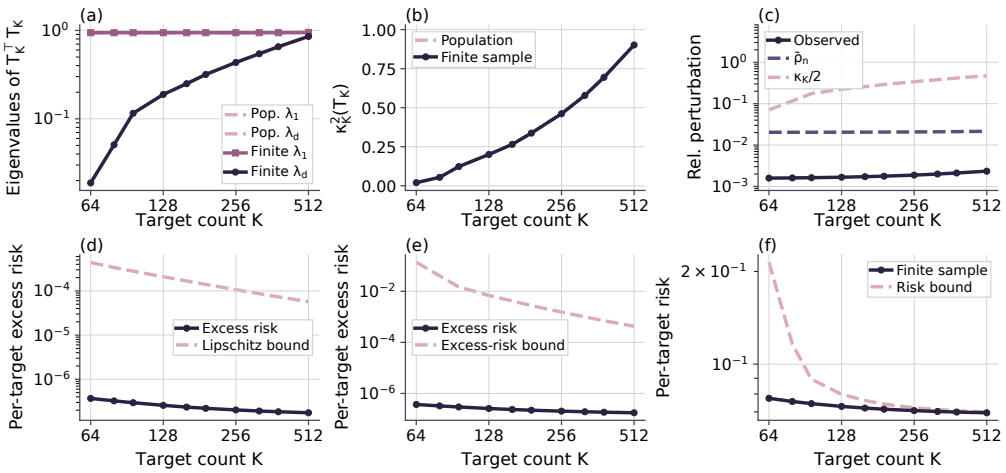

Figure 1: **Post-saturation conditioning** ($r_{\mathrm{pop}}(K) = d$)**.** Adding target blocks along weak directions raises κ_K^2 and lowers the excess and population risks (Proposition 3.11). Against K: (a) extreme eigenvalues of H_K, (b) κ_K^2, (c) relative perturbation, $\bar{\rho}_n$, and $\kappa_K/2$, (d) excess risk and its Lipschitz bound (Lemma 3.10), (e) excess risk and its bound (Proposition 3.11), and (f) population risk and its bound.

At convergence under weight decay, the embedding is isotropic exactly when the end-to-end map has a flat spectrum. An isotropic embedding therefore implies predictive isotropy, which optimizes the excess risk at fixed oracle risk (Theorem 4.2).

By Theorems 4.5 and 4.6 and Proposition 4.8, the predictive isotropy that (4.3) enforces justifies and unifies the two encoder regularization paradigms in JEPA [16, 6, 5] as factorized instances.

5 Experimental Studies

5.1 Synthetic Experiment

Setup. We first run frozen-teacher linear experiments to isolate the mechanisms in the theory from nonlinear optimization effects. The data follow a linear-Gaussian model, so the population operators are known. Full setup details are in Section D.1.

Models and baselines. We fit rank-d RRR on finite samples, where d is the embedding dimension. We compare it with its population counterpart and with unconstrained OLS.

Metrics. We measure the predictive rank, the heterogeneity level, and the conditioning κ_K^2.

Results. In Figure 1, once the rank saturates at d, adding target blocks along weak directions raises κ_K^2 and lowers both the excess and population risks, as Proposition 3.11 predicts. The appendix figures confirm the reduction to RRR, rank growth, identifiability, the bottleneck floor, the oracle inequality, predictive isotropy, and the gauge factorization (Figures 3 to 8 and 10).

5.2 Realistic Experiment

Setup. We test the practical regularizer (4.3) in nonlinear JEPA tasks on images. The context is the left half of an image and the target is the right half. The predictor targets a frozen 16-dimensional embedding of the right half, its top 16 canonical variates with the left half, each divided by its canonical correlation, so that the best linear prediction has identity covariance. A pixel head reads the detached predictions and renders completions without affecting the encoder or predictor (Figure 2).

Data. The small-data diagnostic uses the `sklearn` 8×8 digit dataset. The scaled diagnostic uses standard 28×28 MNIST. CIFAR-10 and CIFAR-100 provide 32×32 RGB natural images of different diversity, and labels are never used. All pixels are normalized to $[0, 1]$.

Models. The 8×8 digits task uses an MLP encoder, predictor, and pixel head. MNIST and CIFAR-10/100 use a convolutional encoder with a 16-dimensional embedding, a linear predictor, and a convolutional pixel head. Architectures and optimizer settings are in Table 3 and Section D.2.

Baselines. We compare four objectives: no regularization, encoder-side regularization, prediction-side regularization, and both combined. Encoder-side regularization applies the Gaussianity penalty to the encoder embedding, as SIGReg does in LeJEPA [5]. Prediction-side regularization applies

Table 1: **Regularizer comparison on real images.** Mean ± standard error over ten seeds, with each regularizer weight chosen by latent risk. Pixel and foreground (FG) MSE are in units of 10^{-2}, and FG MSE excludes background pixels. Time is wall-clock training time, with the ratio to no regularization in parentheses.

Task	Objective	Latent Risk	Test–Train Gap	Pixel MSE	FG MSE	Time
8×8 digits	No reg.	5.656±0.142	5.649±0.143	10.23±0.19	-	2.16±0.01 s (1.00×)
8×8 digits	Enc. reg.	4.303±0.099	3.291±0.080	10.42±0.19	-	2.38±0.01 s (1.10×)
8×8 digits	Pred. reg.	3.575±0.105	**1.506±0.080**	**7.69±0.09**	-	2.39±0.01 s (1.11×)
8×8 digits	Pred.+Enc.	**3.571±0.104**	1.530±0.078	7.88±0.14	-	2.93±0.02 s (1.36×)
MNIST CNN	No reg.	0.357±0.001	0.250±0.003	3.571±0.008	12.01±0.02	12.84±0.03 s (1.00×)
MNIST CNN	Enc. reg.	0.365±0.002	0.190±0.002	3.554±0.009	11.97±0.03	13.85±0.18 s (1.08×)
MNIST CNN	Pred. reg.	**0.334±0.001**	**0.075±0.001**	**3.531±0.009**	**11.91±0.03**	14.39±0.18 s (1.12×)
MNIST CNN	Pred.+Enc.	0.360±0.002	0.138±0.002	3.541±0.009	11.94±0.03	15.47±0.06 s (1.20×)
CIFAR-10 CNN	No reg.	0.538±0.004	0.448±0.006	3.395±0.003	-	14.60±0.02 s (1.00×)
CIFAR-10 CNN	Enc. reg.	0.525±0.003	0.340±0.006	3.378±0.003	-	14.97±0.04 s (1.02×)
CIFAR-10 CNN	Pred. reg.	**0.477±0.002**	**0.143±0.005**	3.377±0.002	-	16.00±0.13 s (1.10×)
CIFAR-10 CNN	Pred.+Enc.	0.514±0.003	0.264±0.007	**3.370±0.003**	-	17.20±0.21 s (1.18×)
CIFAR-100 CNN	No reg.	0.541±0.003	0.455±0.005	3.673±0.003	-	14.49±0.03 s (1.00×)
CIFAR-100 CNN	Enc. reg.	0.526±0.004	0.337±0.005	3.649±0.002	-	14.93±0.04 s (1.03×)
CIFAR-100 CNN	Pred. reg.	**0.480±0.003**	**0.150±0.003**	3.641±0.003	-	16.30±0.12 s (1.13×)
CIFAR-100 CNN	Pred.+Enc.	0.515±0.003	0.263±0.004	**3.640±0.003**	-	17.00±0.24 s (1.17×)

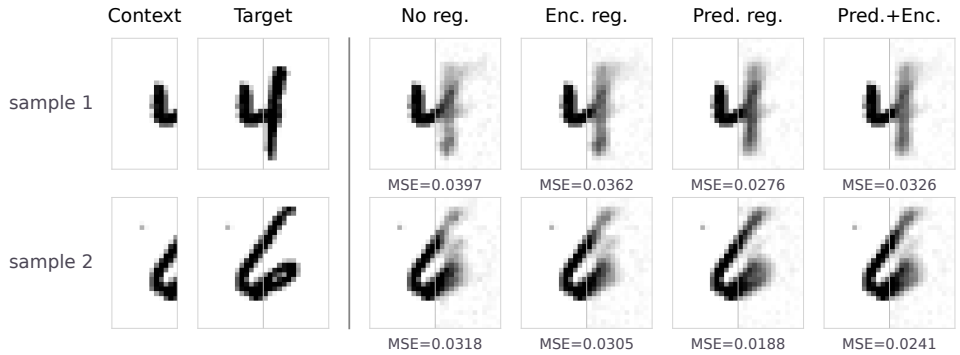

Figure 2: **Half-completions on MNIST.** Cell labels give the target-half pixel MSE of each held-out sample, which is lowest in every row with the predictive regularizer.

our regularizer (4.3) to the predictions. The combined objective is an ablation testing whether encoder-side regularization adds anything beyond prediction-side regularization.

Metrics. We report the latent test risk, the MSE between the predictions and the teacher embedding, and its train–test gap (test minus train). We also report the pixel MSE of the completions and, for MNIST, the foreground MSE, since most target pixels are background.

Results. On all tasks, prediction-side regularization lowers the latent test risk below encoder-side regularization on every seed (Table 1). Relative to encoder-side regularization, it lowers the latent test risk by 17%, 9%, 9%, and 9% and the train–test gap by 54%, 60%, 58%, and 55% on 8×8 digits, MNIST, and CIFAR-10/100. On the small-data 8×8 digits, it also lowers the pixel MSE of the completions by 26%. Encoder-side regularization makes the embedding more isotropic than prediction-side regularization, yet leaves its predictions less isotropic and less Gaussian. This matches Theorems 4.5 and 4.6, where predictive isotropy subsumes encoder isotropy as a gauge choice. Prediction-side regularization alone lowers the Gaussianity distance of the whitened encoder embedding by 35% and 37% on CIFAR-10/100, while encoder-side regularization does not (Figures 16 and 17), showing that the embedding moves closer to an isotropic Gaussian as a byproduct.

6 Concluding Remarks

We give a finite-sample ERM analysis of JEPA and identify an optimality condition. Casting JEPA as reduced-rank regression (Proposition 3.1), we derive a guarantee on its excess risk (Theorem 3.15). The oracle risk decreases with the rank, which grows with target heterogeneity and saturates at the embedding bottleneck (Propositions 3.4 and 3.9). At fixed oracle risk, the excess risk reaches optimality under a flat-spectrum condition termed *predictive isotropy* (Theorem 4.2). Because the risk is gauge-invariant, predictive isotropy constrains the spectrum of the end-to-end map but leaves its encoder–predictor factorization free. Existing encoder-isotropy recipes each fix a gauge realizing it (Lemma 4.4 and Theorems 4.5 and 4.6). We derive a predictive-isotropy regularizer (4.3) that subsumes isotropic-Gaussian embedding without fixing a gauge. Empirically, it lowers test risk over encoder-side regularization on all image tasks. We defer the discussion of related work to Section B.

Acknowledgments and Disclosure of Funding

JH would like to thank Dino Feng and Andrew Chen for valuable conversations. The authors would like to thank the anonymous reviewers and program chairs for constructive comments.

JH is partially supported by Northwestern University's Walter P. Murphy Fellowship and Terminal Year Fellowship (Paul K. Richter Memorial Award). Han Liu is partially supported by NIH R01LM1372201, NSF AST-2421845, Simons Foundation MPS-AI-00010513, AbbVie, Dolby and Chan Zuckerberg Biohub Chicago Spoke Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. All authors declare no competing interests.

Reproducibility

All reported experiments use ten deterministic seeds. Code and configurations for every figure and table are available at https://github.com/thermau5/on-jepa-isotropy.

References

[1] Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

[3] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

[4] Adrien Bardes, Jean Ponce, and Yann LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023.

[5] Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

[6] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[7] Jean-Bastien Grill, Florian Strub, Florian Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[10] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[12] Sofiane Ennadir, Levente Zólyomi, and Oleg Smirnov. Understanding early collapse in predictive world-model pretraining. In *ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling*, 2026.

[13] Randall Balestriero, Nicolas Ballas, Mike Rabbat, and Yann LeCun. Gaussian embeddings: How JEPAs secretly learn your data density. *arXiv preprint arXiv:2510.05949*, 2025.

[14] Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351, 1951.

[15] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[16] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, 2021.

[17] Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Josh Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video SSL with frozen teachers. *arXiv preprint arXiv:2509.24317*, 2025.

[18] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, 2011.

[19] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[20] Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, 1936.

[21] T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983.

[22] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, 1991.

[23] Gregory C Reinsel and Raja P Velu. *Multivariate reduced-rank regression*. Springer, 1998.

[24] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

[25] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[26] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[27] A Wayne Roberts and Dale E Varberg. *Convex Functions*, volume 57 of *Pure and Applied Mathematics*. Academic Press, 1974.

[28] Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

[29] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pages 5628–5637. PMLR, 2019.

[30] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *Journal of Machine Learning Research*, 22(281):1–31, 2021.

[31] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

[32] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

[33] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

[34] Bobak T Kiani, Randall Balestriero, Yubei Chen, Seth Lloyd, and Yann LeCun. Joint embedding self-supervised learning in the kernel regime. *arXiv preprint arXiv:2209.14884*, 2022.

[35] Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International conference on machine learning*, pages 3252–3298. PMLR, 2023.

[36] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems*, 37:91300–91336, 2024.

[37] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

Appendix

A Supplementary Theoretical Backgrounds **15**
A.1 Reduced-Rank Regression 15
A.2 Useful Lemmas 15
A.3 Maximum-Entropy Distributions 17

B Related Works and Discussion **18**

C Proofs of Main Text **19**
C.1 Proof of Rank of the JEPA Solution 19
C.2 Proof of Rank Growth and Saturation 19
C.3 Proof of Heterogeneity and Identifiability 20
C.4 Proof of Capacity Decomposition 20
C.5 Proof of Excess Risk of the Learned Predictor 20
C.6 Proof of Post-Saturation Risk Decomposition 21
C.7 Proof of the Oracle Inequality 21
C.8 Proof of Predictive Isotropy 22
C.9 Proof of the Optimality of Predictive Isotropy 22
C.10 Proof of Gauge Invariance 22
C.11 Proof of Covariance Isotropy Subsumption 22
C.12 Proof of Gaussian Isotropy Subsumption 22
C.13 Proof of Linear Gaussian Gauge 23
C.14 Proof of Isotropy under Weight Decay 23

D Additional Experimental Details **23**
D.1 Synthetic Experiment Details 23
D.2 Real-Image Regularizer Details 28

Impact Statement

By the theoretical nature of this work, we do not anticipate any negative social impact.

LLM Usage Disclosure

We used large language models (LLMs) to aid and polish writing, such as improving clarity, grammar, and conciseness. We also used LLMs for retrieval and discovery, for example, literature-search assistance to identify potential missing related work. In addition, LLM-based coding assistants were used as basic code assistance to help implement the numerical validation experiments (e.g., plotting and data pipelines). All technical content, proofs, experiments, and results are original contributions by the authors.

A Supplementary Theoretical Backgrounds

A.1 Reduced-Rank Regression

Given paired random vectors $x \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, the population squared-loss regression problem

$$\min_{B \in \mathbb{R}^{q \times p}} \mathbb{E}\big[\|y - Bx\|^2\big] \tag{A.1}$$

admits the closed-form ordinary least squares (OLS) minimizer $B_{\mathrm{OLS}} := \Sigma_{yx}\Sigma_{xx}^{-1}$ when $\Sigma_{xx} \succ 0$.

Reduced-rank regression (RRR) [15, 23] restricts the minimization in (A.1) to $\mathrm{rank}(B) \le r$.

Definition A.1 (Reduced-rank regression problem)**.** Given paired $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$ as predictor and response random vectors, for a prescribed rank $r \le \min(p, q)$, the reduced-rank regression is

$$B_{\mathrm{RRR}} \in \arg\min_{B \in \mathbb{R}^{q \times p}} \mathbb{E}\big[\|y - Bx\|^2\big] \quad \text{subject to} \quad \mathrm{rank}(B) \le r. \tag{A.2}$$

The constraint $\mathrm{rank}(B) \le r$ means the linear predictor factors through an r-dimensional representation. The following lemma gives a closed-form solution.

Lemma A.2 (RRR solution, Theorem 2.2 of Reinsel and Velu [23])**.** Assume $\Sigma_{xx} \succ 0$. Let $V_r \in \mathbb{R}^{p \times r}$ collect eigenvectors corresponding to the top r eigenvalues of $\Sigma_{xx}^{1/2} B_{\mathrm{OLS}}^\top B_{\mathrm{OLS}} \Sigma_{xx}^{1/2}$. A solution to (A.2) is

$$B_{\mathrm{RRR}} = B_{\mathrm{OLS}} \Sigma_{xx}^{1/2} V_r V_r^\top \Sigma_{xx}^{-1/2}. \tag{A.3}$$

Equivalently, the RRR solution is the best rank-r approximation of $B_{\mathrm{OLS}}\Sigma_{xx}^{1/2}$, unwhitened by $\Sigma_{xx}^{-1/2}$. RRR thus retains the r predictor-space directions most useful for linearly predicting y.

Remark A.3 (Connection to CCA and PCA)**.** Reduced-rank regression is closely related to canonical correlation analysis (CCA). With the response also whitened, the rank-r solution spans the leading r canonical directions linking the predictor and response spaces. As a special case, when $y = x$, the corresponding reduced-rank self-prediction problem recovers the principal subspace of Σ_{xx}, linking RRR to PCA.

A.2 Useful Lemmas

Lemma A.4 (Joint-covariance control of cross-covariance noise [24])**.** Let $(Z_i, Y_i)_{i=1}^n$ be i.i.d. samples with $Z \in \mathbb{R}^{D_x}$ and $Y \in \mathbb{R}^{Kd}$, and assume $U := (Y^\top, Z^\top)^\top$ is mean-zero and sub-Gaussian with second moment $M := \mathbb{E}[UU^\top]$. Let $\widehat{M} := n^{-1}\sum_{i=1}^n U_i U_i^\top$ and $\widehat{\Sigma}_{YZ} := n^{-1}\sum_{i=1}^n Y_i Z_i^\top$. Then $\widehat{\Sigma}_{YZ} - \Sigma_{YZ}$ is an off-diagonal block of $\widehat{M} - M$. Consequently, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{\mathrm{op}} \le C\|M\|_{\mathrm{op}} \left(\sqrt{\frac{r(M) + \log(1/\delta)}{n}} + \frac{r(M) + \log(1/\delta)}{n} \right),$$

where $C > 0$ depends only on the sub-Gaussian constant of U and $r(M) := \mathrm{tr}(M)/\|M\|_{\mathrm{op}}$ is the effective rank of the joint second-moment matrix.

Proof. Since $U = (Y^\top, Z^\top)^\top$,

$$M = \begin{pmatrix} \mathbb{E}[YY^\top] & \mathbb{E}[YZ^\top] \\ \mathbb{E}[ZY^\top] & \mathbb{E}[ZZ^\top] \end{pmatrix} = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix},$$

and hence

$$\widehat{M} - M = \begin{pmatrix} \widehat{\Sigma}_{YY} - \Sigma_{YY} & \widehat{\Sigma}_{YZ} - \Sigma_{YZ} \\ \widehat{\Sigma}_{ZY} - \Sigma_{ZY} & \widehat{\Sigma}_{ZZ} - \Sigma_{ZZ} \end{pmatrix}.$$

Let $E \coloneqq \widehat{\Sigma}_{YZ} - \Sigma_{YZ}$. For any unit vectors a, b, define $u = (a^\top, 0^\top)^\top$ and $v = (0^\top, b^\top)^\top$. Then $a^\top E b = u^\top(\widehat{M} - M)v$, so

$$\|E\|_{\mathrm{op}} \le \|\widehat{M} - M\|_{\mathrm{op}}.$$

Applying the effective-rank sample-covariance concentration bound for sub-Gaussian vectors [24] to $\widehat{M} - M$ gives

$$\|\widehat{M} - M\|_{\mathrm{op}} \le C\|M\|_{\mathrm{op}} \left(\sqrt{\frac{r(M) + \log(1/\delta)}{n}} + \frac{r(M) + \log(1/\delta)}{n} \right),$$

which proves the claim. □

Corollary A.5 (Whitened-operator envelope)**.** Let $\Sigma_{ZZ} \succ 0$ and $\widehat{T}_K = \widehat{\Sigma}_{YZ}\Sigma_{ZZ}^{-1/2}$. With probability at least $1 - \delta$,

$$\|\widehat{T}_K - T_K\|_{\mathrm{op}} \le \eta_n^T(\delta) \coloneqq \frac{\eta_n^{YZ}(\delta)}{\sqrt{\lambda_{\min}(\Sigma_{ZZ})}},$$

where $\eta_n^{YZ}(\delta)$ denotes the right-hand side of Lemma A.4.

Proof. Since $T_K = \Sigma_{YZ}\Sigma_{ZZ}^{-1/2}$, we have $\widehat{T}_K - T_K = (\widehat{\Sigma}_{YZ} - \Sigma_{YZ})\Sigma_{ZZ}^{-1/2}$, so submultiplicativity and $\|\Sigma_{ZZ}^{-1/2}\|_{\mathrm{op}} = \lambda_{\min}(\Sigma_{ZZ})^{-1/2}$ give

$$\|\widehat{T}_K - T_K\|_{\mathrm{op}} \le \|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{\mathrm{op}} \, \|\Sigma_{ZZ}^{-1/2}\|_{\mathrm{op}} = \frac{\|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(\Sigma_{ZZ})}}.$$

Lemma A.4 bounds the numerator by $\eta_n^{YZ}(\delta)$ with probability at least $1 - \delta$. □

Corollary A.6 (Whitening-error envelope)**.** Assume $\widetilde{Z} \coloneqq \Sigma_{ZZ}^{-1/2} Z$ is mean-zero and sub-Gaussian, and for $\delta \in (0, 1)$ let

$$\eta_n^Z(\delta) \coloneqq C\Big(\sqrt{\frac{D_x + \log(1/\delta)}{n}} + \frac{D_x + \log(1/\delta)}{n} \Big),$$

where $C > 0$ depends only on the sub-Gaussian constant of $\widetilde{Z}$. If $\eta_n^Z(\delta) \le \frac{1}{2}$, then with probability at least $1 - \delta$,

$$\|\widehat{F} - I\|_{\mathrm{op}} \le \varepsilon_n(\delta) \coloneqq 2\eta_n^Z(\delta).$$

Proof. Let $\widehat{W} \coloneqq \Sigma_{ZZ}^{-1/2}\widehat{\Sigma}_{ZZ}\Sigma_{ZZ}^{-1/2}$ be the sample second moment of $\widetilde{Z}$. Its population counterpart is I, with $\|I\|_{\mathrm{op}} = 1$ and effective rank $r(I) = D_x$. With $M = I$, the concentration inequality of Koltchinskii and Lounici [24] behind Lemma A.4 reads $\|\widehat{W} - I\|_{\mathrm{op}} \le \eta_n^Z(\delta)$ with probability at least $1 - \delta$, and $\eta_n^Z(\delta) \le \frac{1}{2}$ by assumption.

As $\widehat{W}$ is symmetric, each eigenvalue w satisfies $|w-1| \le \|\widehat{W} - I\|_{\mathrm{op}} \le \frac{1}{2}$, so $w \ge \frac{1}{2}$, $\widehat{W}$ is invertible, and $\widehat{\Sigma}_{ZZ} = \Sigma_{ZZ}^{1/2}\widehat{W}\Sigma_{ZZ}^{1/2} \succ 0$. $\widehat{F} = \widehat{W}^{-1}$ has eigenvalues $1/w$, and $w \ge \frac{1}{2}$ gives $|1/w - 1| = |w - 1|/w \le 2|w - 1|$. Hence, maximizing over w, $\|\widehat{F} - I\|_{\mathrm{op}} \le 2\|\widehat{W} - I\|_{\mathrm{op}} \le 2\eta_n^Z(\delta)$. □

Lemma A.7 (ERM under sample whitening)**.** Let $\widehat{\Sigma}_{ZZ} \succ 0$. With $\widehat{T}_K$ and $\widehat{F}$ from Section 3.4, let $\widehat{S}_K \coloneqq \widehat{T}_K \widehat{F}^{1/2}$, whose top-$r$ right singular subspace is $\widehat{\mathcal{V}}_{K,r}$.

(i) The ERM in (3.1) can also be written as

$$\widehat{B}_r = [\widehat{S}_K]_r \, \widehat{F}^{1/2}\Sigma_{ZZ}^{-1/2}.$$

(ii) If $\|\widehat{T}_K - T_K\|_{\mathrm{op}} \le \|T_K\|_{\mathrm{op}}$, then

$$\|\widehat{S}_K - T_K\|_{\mathrm{op}} \le \|\widehat{T}_K - T_K\|_{\mathrm{op}} + 2\|T_K\|_{\mathrm{op}}\|\widehat{F} - I\|_{\mathrm{op}}.$$

Proof. Part (i) reduces ERM to a truncated SVD in a suitable sample whitening, and part (ii) is a perturbation bound.

Part (i). Let $\widehat{B}_{\mathrm{OLS}} := \widehat{\Sigma}_{YZ}\widehat{\Sigma}_{ZZ}^{-1}$ be the sample OLS map. Expanding $\|Y_i - BZ_i\|^2$ around $\widehat{B}_{\mathrm{OLS}}Z_i$, the cross term vanishes because $\widehat{\Sigma}_{YZ} - \widehat{B}_{\mathrm{OLS}}\widehat{\Sigma}_{ZZ} = 0$, so for every R with $RR^\top = \widehat{\Sigma}_{ZZ}$,

$$\frac{1}{nK}\sum_{i=1}^{n}\|Y_i - BZ_i\|^2 = \frac{1}{nK}\sum_{i=1}^{n}\|Y_i - \widehat{B}_{\mathrm{OLS}}Z_i\|^2 + \frac{1}{K}\|(B - \widehat{B}_{\mathrm{OLS}})R\|_F^2.$$

Only the last term depends on B. We choose $R := \Sigma_{ZZ}^{1/2}\widehat{F}^{-1/2}$, which satisfies $RR^\top = \widehat{\Sigma}_{ZZ}$ because $\widehat{F}^{-1} = \Sigma_{ZZ}^{-1/2}\widehat{\Sigma}_{ZZ}\Sigma_{ZZ}^{-1/2}$. It turns the OLS map into $\widehat{S}_K$,

$$\widehat{B}_{\mathrm{OLS}}R = \widehat{\Sigma}_{YZ}\big(\Sigma_{ZZ}^{-1/2}\widehat{F}\,\Sigma_{ZZ}^{-1/2}\big)\Sigma_{ZZ}^{1/2}\widehat{F}^{-1/2} = \widehat{S}_K.$$

Because R is invertible, $\mathrm{rank}(BR) = \mathrm{rank}(B)$, so the Eckart–Young theorem [25] gives $\widehat{B}_r R = [\widehat{S}_K]_r$. Multiplying by $R^{-1} = \widehat{F}^{1/2}\Sigma_{ZZ}^{-1/2}$ gives (i). The choice $R = \widehat{\Sigma}_{ZZ}^{1/2}$ instead gives the form in Proposition 3.1.

Part (ii). Write $\widehat{S}_K - T_K = (\widehat{T}_K - T_K) + \widehat{T}_K(\widehat{F}^{1/2} - I)$. Each eigenvalue $f > 0$ of $\widehat{F}$ satisfies $|\sqrt{f} - 1| = |f - 1|/(\sqrt{f} + 1) \le |f - 1|$, so $\|\widehat{F}^{1/2} - I\|_{\mathrm{op}} \le \|\widehat{F} - I\|_{\mathrm{op}}$. The assumption gives $\|\widehat{T}_K\|_{\mathrm{op}} \le 2\|T_K\|_{\mathrm{op}}$, and the triangle inequality gives (ii). □

Lemma A.8 (Wedin-type subspace perturbation [26])**.** Let $M \in \mathbb{R}^{m\times p}$, let $r \le \min(m, p)$, and let $\widehat{M}$ be a perturbation with $\|\widehat{M} - M\|_{\mathrm{op}} \le \epsilon$. Denote by $\mathcal{V}_r, \widehat{\mathcal{V}}_r$ the top-r right singular subspaces of M and $\widehat{M}$, and define the *Wedin gap at level* r as $g_r(M) := \sigma_r(M) - \sigma_{r+1}(M)$, with the convention $\sigma_{r+1}(M) := 0$ when $\mathrm{rank}(M) \le r$. Provided $\epsilon < g_r(M)$,

$$\|\sin\Theta(\mathcal{V}_r, \widehat{\mathcal{V}}_r)\|_{\mathrm{op}} \;\le\; \frac{\epsilon}{g_r(M) - \epsilon}.$$

Proof. By Weyl's inequality, $\sigma_r(\widehat{M}) - \sigma_{r+1}(M) \ge g_r(M) - \epsilon > 0$. Wedin's $\sin\Theta$ theorem [26] bounds $\|\sin\Theta(\mathcal{V}_r, \widehat{\mathcal{V}}_r)\|_{\mathrm{op}}$ by $\|\widehat{M} - M\|_{\mathrm{op}} \le \epsilon$ divided by this gap, which gives the claim. □

A.3 Maximum-Entropy Distributions

We recall two standard entropy facts used in the maximum-entropy regularization argument. Let X be an absolutely continuous random vector in $\mathbb{R}^d$ with density p_X. Its differential entropy is

$$h(X) := -\int_{\mathbb{R}^d} p_X(x)\log p_X(x)\,dx,$$

whenever the integral is well-defined.

Lemma A.9 (Gaussian maximum entropy under fixed covariance)**.** Let X be an absolutely continuous random vector in $\mathbb{R}^d$ with mean μ and covariance $\Sigma \succ 0$. Then

$$h(X) \le \frac{1}{2}\log\big((2\pi e)^d \det\Sigma\big),$$

with equality if and only if $X \sim N(\mu, \Sigma)$.

Proof. Let $G \sim N(\mu, \Sigma)$. Since $D_{\mathrm{KL}}(\mathcal{L}(X)\,\|\,\mathcal{L}(G)) \ge 0$,

$$-h(X) - \mathbb{E}\log p_G(X) \ge 0.$$

Because X and G have the same mean and covariance,

$$-\mathbb{E}\log p_G(X) = \frac{1}{2}\log\big((2\pi)^d \det\Sigma\big) + \frac{1}{2}\mathbb{E}[(X-\mu)^\top\Sigma^{-1}(X-\mu)] = \frac{1}{2}\log\big((2\pi e)^d \det\Sigma\big).$$

Hence the claimed bound follows. Equality holds iff the KL divergence is zero, i.e. iff X has the Gaussian distribution. □

Lemma A.10 (Trace-constrained maximum entropy)**.** Let X be an absolutely continuous random vector in $\mathbb{R}^d$ with mean zero, covariance $C \succ 0$, and $\mathrm{tr}(C) = \tau$. Let $h(X)$ denote its differential entropy. Then

$$h(X) \le \frac{1}{2}\log\big((2\pi e)^d \det C\big) \le \frac{d}{2}\log\left(2\pi e\,\frac{\tau}{d}\right).$$

Equality holds if and only if $X \sim N(0, \frac{\tau}{d} I_d)$.

Proof. By Lemma A.9, applied with $\Sigma = C$ and $\mu = 0$,

$$h(X) \le \frac{1}{2}\log\big((2\pi e)^d \det C\big),$$

with equality iff $X \sim N(0, C)$. For the second inequality, arithmetic-geometric mean (AM-GM) inequality gives

$$\det C = \prod_{i=1}^{d}\lambda_i(C) \le \left(\frac{1}{d}\sum_{i=1}^{d}\lambda_i(C)\right)^d = \left(\frac{\tau}{d}\right)^d,$$

with equality iff $\lambda_1(C) = \cdots = \lambda_d(C) = \tau/d$, i.e., $C = \frac{\tau}{d} I_d$. Combining the equality conditions gives $X \sim N(0, \frac{\tau}{d} I_d)$. □

Connection to predictive isotropy. Theorem 4.2 is an instance of the same fixed-trace principle. At fixed trace $\mathrm{tr}(H_K) = \tau$, the flat spectrum $(\tau/d, \ldots, \tau/d)$ is the unique majorization-minimum among positive rank-d spectra with total mass τ [27, 28]. In particular, by AM-GM, it maximizes the log-determinant term $\frac{1}{2}\sum_{i=1}^{d}\log\lambda_i(H_K)$ among fixed-trace spectra. Thus $\kappa_K^2 = 1$ means the d nonzero eigenvalues of H_K are equal.

B Related Works and Discussion

Finite-sample theory of joint-embedding learning. Theory for self-supervised representation learning has largely focused on downstream transfer, contrastive objectives, or kernel-regime joint-embedding models. Contrastive analyses give downstream guarantees under latent-class or multi-view assumptions, including labeled-sample-complexity improvements in latent-class settings [29, 30]. Spectral analyses further show how augmentation structure can induce useful representation subspaces with linear-probe guarantees [31]. Related pretext-task and non-contrastive analyses explain how paired views, conditional independence, predictors, and training dynamics support downstream representations and collapse avoidance [32, 33]. Kernel-regime analyses characterize how augmentations, inductive bias, and regularization shape SSL generalization and optimal joint-embedding representations [34, 35]. These results establish finite-sample theory for broad SSL settings, but they do not analyze the JEPA predictive objective as an empirical-risk problem with an explicit low-rank predictor bottleneck. Closer to JEPA, recent deep-linear work explains why latent prediction can avoid noisy reconstruction targets by studying training-dynamics implicit bias [36].

In contrast, our analysis targets the ERM of JEPA itself. Under linearization, the predictive objective becomes reduced-rank regression (Proposition 3.1), yielding a finite-sample guarantee whose excess risk is governed by spectral conditioning (Theorem 3.15).

Encoder isotropy and anti-collapse. A central design principle in non-contrastive SSL is to prevent collapse by controlling the encoder embedding distribution. Whitening-based SSL scatters latent features through batch whitening, Barlow Twins drives the cross-correlation matrix toward identity, and VICReg combines invariance with variance and covariance regularization [16, 37, 6]. Recent

JEPA-specific work strengthens this encoder-side perspective by promoting isotropic-Gaussian embeddings as the optimal representation distribution [5, 13].

Our result changes the primitive object. For JEPA, the excess-risk bound is uniquely controlled by the spectrum of the whitened end-to-end map (Lemma 4.1 and Theorem 4.2). Encoder isotropy is therefore a gauge choice in the encoder-predictor factorization satisfying predictive isotropy (Lemma 4.4 and Theorems 4.5 and 4.6).

Entropy and Gaussianity. Our maximum-entropy argument separates the risk condition from its distributional completion. The finite-sample bound requires spectral balancing of the predictive covariance (Theorem 4.2), while Gaussianity enters as the least-committing distribution satisfying this covariance-level condition. Thus isotropic-Gaussian embeddings arise as the entropy-optimal completion of predictive isotropy, realizable through a suitable encoder-predictor gauge (Lemma 4.4 and Theorem 4.6). This reframes Gaussian embedding regularization as an encoder-side instance of an ERM condition, specifically via our regularizer (4.3). Stronger distributional constraints should be justified by the loss and task rather than by collapse prevention alone.

Limitations. Our analysis is deliberately scoped to a linearized JEPA model with a frozen target encoder. The linear assumption makes the predictor bottleneck explicit and allows the JEPA objective to reduce to reduced-rank regression, but it does not capture representation-dependent feature learning in deep encoders. Similarly, the frozen-teacher assumption isolates the ERM geometry of the context encoder and predictor, but omits the coupled training dynamics induced by EMA or jointly updated target networks. Extending predictive isotropy to adaptive teachers and nonlinear feature learning remains an important direction for future work.

C Proofs of Main Text

C.1 Proof of Rank of the JEPA Solution

Proof of Lemma 3.2. By Proposition 3.1, $B_r^\star = [T_K]_r \Sigma_{ZZ}^{-1/2}$, and $\Sigma_{ZZ}^{-1/2}$ is invertible because $\Sigma_{ZZ} \succ 0$, so $\operatorname{rank}(B_r^\star) = \operatorname{rank}([T_K]_r)$. The truncation $[T_K]_r$ keeps the $\min\{r, \operatorname{rank}(T_K)\}$ nonzero singular values of T_K, which gives (3.4). □

C.2 Proof of Rank Growth and Saturation

Proof of Proposition 3.4. We compare the kernels of T_K and T_{K+1}, and then turn the comparison into a statement about ranks.

By construction, T_{K+1} stacks the new block M_{K+1} below T_K. Hence, for every $v \in \mathbb{R}^{D_x}$,

$$T_{K+1} v = \begin{pmatrix} T_K v \\ M_{K+1} v \end{pmatrix},$$

which vanishes if and only if both $T_K v$ and $M_{K+1} v$ vanish. Therefore

$$\ker(T_{K+1}) = \ker(T_K) \cap \ker(M_{K+1}) \subseteq \ker(T_K). \tag{C.1}$$

Every T_K has D_x columns, so rank-nullity gives

$$r_{\text{pop}}(K) = \operatorname{rank}(T_K) = D_x - \dim \ker(T_K).$$

The inclusion in (C.1) gives $\dim \ker(T_{K+1}) \le \dim \ker(T_K)$, and hence $r_{\text{pop}}(K+1) \ge r_{\text{pop}}(K)$. This proves (1).

Given this inclusion, rank-nullity also shows that the rank grows strictly if and only if $\ker(T_{K+1}) \neq \ker(T_K)$. By (C.1), this holds if and only if some $v \in \ker(T_K)$ satisfies $M_{K+1} v \neq 0$, which is exactly the condition $\ker(T_K) \not\subseteq \ker(M_{K+1})$. This proves (2). □

Proof of Corollary 3.5. By (1), $r_{\text{pop}}(K)$ is a nondecreasing integer sequence bounded by D_x, so it is eventually constant. □

C.3 Proof of Heterogeneity and Identifiability

Proof of Proposition 3.8. On the assumed perturbation event, $\|\widehat{T}_K - T_K\|_{\mathrm{op}} \le \eta_n^T(\delta)$. Since $\mathrm{rank}(T_K) = r_{\mathrm{pop}}(K)$, applying Lemma A.8 with $r = r_{\mathrm{pop}}(K)$, $\epsilon = \eta_n^T(\delta)$, and Wedin gap $g_{r_{\mathrm{pop}}}(T_K) = \sqrt{\lambda_{\mathrm{het}}}$ (3.6) yields the claim. □

C.4 Proof of Capacity Decomposition

Proof of Proposition 3.9. By population least squares, $B_{\mathrm{OLS}}^{\mathrm{raw}} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}$ is the unconstrained minimizer of $\mathbb{E}\|Y - BZ\|_2^2$. Therefore, for every $B \in \mathbb{R}^{Kd \times D_x}$,

$$\mathbb{E}\left[\|Y - BZ\|_2^2\right] = \mathbb{E}\left[\|Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z\|_2^2\right] + \left\|(B - B_{\mathrm{OLS}}^{\mathrm{raw}})\Sigma_{ZZ}^{1/2}\right\|_F^2.$$

Thus, for any r, since $B_{\mathrm{OLS}}^{\mathrm{raw}}\Sigma_{ZZ}^{1/2} = T_K$,

$$\inf_{\mathrm{rank}(B)\le r} \mathbb{E}\left[\|Y - BZ\|_2^2\right] = \mathbb{E}\left[\|Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z\|_2^2\right] + \inf_{\mathrm{rank}(B)\le r} \left\|B\Sigma_{ZZ}^{1/2} - T_K\right\|_F^2.$$

Since $\Sigma_{ZZ}^{1/2}$ is invertible, the rank constraint remains $\mathrm{rank}(B\Sigma_{ZZ}^{1/2}) \le r$. By the Eckart-Young theorem [25] for Frobenius-norm low-rank approximation,

$$\inf_{\mathrm{rank}(B)\le r} \left\|B\Sigma_{ZZ}^{1/2} - T_K\right\|_F^2 = \sum_{j>r} \sigma_j(T_K)^2. \tag{C.2}$$

This proves (3.7). □

C.5 Proof of Excess Risk of the Learned Predictor

Proof of Lemma 3.10. We write both risks in terms of T_K, then compare the learned predictor with the oracle. Recall the unconstrained OLS map $B_{\mathrm{OLS}}^{\mathrm{raw}} = \Sigma_{YZ}\Sigma_{ZZ}^{-1}$ and $T_K = B_{\mathrm{OLS}}^{\mathrm{raw}}\Sigma_{ZZ}^{1/2}$ from (3.2). For an end-to-end map B, expanding the risk (2.3) around $B_{\mathrm{OLS}}^{\mathrm{raw}}$ gives

$$\begin{aligned} K\,\mathcal{R}_K(B) &= \mathbb{E}\big[\|Y - BZ\|^2\big] \\ &= \mathbb{E}\big[\|Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z\|^2\big] + \mathbb{E}\big[\|(B_{\mathrm{OLS}}^{\mathrm{raw}} - B)Z\|^2\big] + 2\,\mathbb{E}\big[\langle Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z,\, (B_{\mathrm{OLS}}^{\mathrm{raw}} - B)Z\rangle\big] \\ &= \mathbb{E}\big[\|Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z\|^2\big] + \|B\Sigma_{ZZ}^{1/2} - T_K\|_F^2, \end{aligned} \tag{C.3}$$

where the cross term vanishes because $\mathbb{E}[(Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z)Z^\top] = \Sigma_{YZ} - B_{\mathrm{OLS}}^{\mathrm{raw}}\Sigma_{ZZ} = 0$, and the middle term uses $\Sigma_{ZZ} = \mathbb{E}[ZZ^\top]$. We now apply (C.3) to the learned predictor. By Lemma A.7, $\widehat{C} := \widehat{B}_r\Sigma_{ZZ}^{1/2} = \widehat{S}_K\widehat{Q}\widehat{F}^{1/2}$, where $\widehat{Q}$ is the projection onto $\widehat{\mathcal{V}}_{K,r}$. Because $\widehat{F}$ is symmetric, the rows of $\widehat{C}$ lie in $\widehat{F}^{1/2}\widehat{\mathcal{V}}_{K,r}$, and we write $\widehat{P}$ for the projection onto this subspace. Hence $\widehat{C} = \widehat{C}\widehat{P}$, so

$$\widehat{C} - T_K = (\widehat{C} - T_K)\widehat{P} - T_K(I - \widehat{P}).$$

These terms are orthogonal because $(I - \widehat{P})\widehat{P} = 0$, so their squared Frobenius norms add. Taking $B = \widehat{B}_r$ in (C.3) with this split, and using Proposition 3.9 for the oracle,

$$\begin{aligned} K\,\mathcal{R}_K(\widehat{B}_r) &= \mathbb{E}\big[\|Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z\|^2\big] + \|(\widehat{C} - T_K)\widehat{P}\|_F^2 + \|T_K(I - \widehat{P})\|_F^2, \\ K\,\mathcal{R}_K^\star(r) &= \mathbb{E}\big[\|Y - B_{\mathrm{OLS}}^{\mathrm{raw}} Z\|^2\big] + \sum_{j>r}\sigma_j(T_K)^2. \end{aligned}$$

Subtracting gives

$$K\big(\mathcal{R}_K(\widehat{B}_r) - \mathcal{R}_K^\star(r)\big) = \|(\widehat{C} - T_K)\widehat{P}\|_F^2 + \Big(\|T_K(I - \widehat{P})\|_F^2 - \sum_{j>r}\sigma_j(T_K)^2\Big). \tag{C.4}$$

The first term is the coefficient-estimation error, and the second is the subspace-selection error.

Because $\widehat{P}$ has rank r, the first term of (C.4) is at most $r\|(\widehat{C}-T_K)\widehat{P}\|_{\mathrm{op}}^2$. To bound this norm, take a unit vector $u=\widehat{F}^{1/2}w$ in the range of $\widehat{P}$, with $w\in\widehat{\mathcal{V}}_{K,r}$. Because $\widehat{Q}w=w$ and $\widehat{S}_K=\widehat{T}_K\widehat{F}^{1/2}$,

$$(\widehat{C}-T_K)u=\widehat{S}_K\widehat{Q}\widehat{F}w-T_Ku=(\widehat{T}_K-T_K)u+\widehat{S}_K\widehat{Q}(\widehat{F}-I)w.$$

Here $\|w\|\le(1-\widehat{\varepsilon})^{-1/2}$ and $\|\widehat{S}_K\|_{\mathrm{op}}\le(1+\widehat{\rho})(1+\widehat{\varepsilon})^{1/2}\|T_K\|_{\mathrm{op}}$, so $\widehat{\rho}\le 1$ and $\widehat{\varepsilon}\le\frac{1}{2}$ give $\|(\widehat{C}-T_K)\widehat{P}\|_{\mathrm{op}}\le(\widehat{\rho}+4\widehat{\varepsilon})\|T_K\|_{\mathrm{op}}$. For the second term, let $v_1,v_2,\dots$ be the right singular vectors of T_K, so that $V_{K,r}=[v_1,\dots,v_r]$. Because $T_K^\top T_K=\sum_j\sigma_j(T_K)^2v_jv_j^\top$, we have $\|T_K(I-\widehat{P})\|_F^2=\mathrm{tr}\big((I-\widehat{P})T_K^\top T_K(I-\widehat{P})\big)=\sum_j\sigma_j(T_K)^2\|(I-\widehat{P})v_j\|^2$, so

$$\begin{aligned}\|T_K(I-\widehat{P})\|_F^2-\sum_{j>r}\sigma_j(T_K)^2&=\sum_{j\le r}\sigma_j(T_K)^2\|(I-\widehat{P})v_j\|^2-\sum_{j>r}\sigma_j(T_K)^2\|\widehat{P}v_j\|^2\\&\le\sigma_1(T_K)^2\,\|(I-\widehat{P})V_{K,r}\|_F^2.\end{aligned}$$

Because $V_{K,r}$ has orthonormal columns, the singular values of $(I-\widehat{P})V_{K,r}$ are the sines $\sin\theta_1,\dots,\sin\theta_r$ of the principal angles between $\mathcal{V}_{K,r}$ and $\widehat{F}^{1/2}\widehat{\mathcal{V}}_{K,r}$, so $\|(I-\widehat{P})V_{K,r}\|_F^2=\sum_{i=1}^r\sin^2\theta_i\le r\big(\|\sin\Theta(\mathcal{V}_{K,r},\widehat{\mathcal{V}}_{K,r})\|_{\mathrm{op}}+\widehat{\varepsilon}\big)^2$, where the last step uses the triangle inequality for $\|\sin\Theta\|_{\mathrm{op}}$ and $\|(I-\widehat{P})\widehat{Q}\|_{\mathrm{op}}=\|(I-\widehat{P})(I-\widehat{F}^{1/2})\widehat{Q}\|_{\mathrm{op}}\le\|\widehat{F}^{1/2}-I\|_{\mathrm{op}}\le\widehat{\varepsilon}$. Inserting these bounds into (C.4) and using $\sigma_1(T_K)=\|T_K\|_{\mathrm{op}}$,

$$K\big(\mathcal{R}_K(\widehat{B}_r)-\mathcal{R}_K^\star(r)\big)\le r\,\sigma_1(T_K)^2\Big((\widehat{\rho}+4\widehat{\varepsilon})^2+\big(\|\sin\Theta(\mathcal{V}_{K,r},\widehat{\mathcal{V}}_{K,r})\|_{\mathrm{op}}+\widehat{\varepsilon}\big)^2\Big).$$

Dividing by K gives the claim with $L=r\,\sigma_1(T_K)^2/K$. □

C.6 Proof of Post-Saturation Risk Decomposition

Proof of Proposition 3.11. On $E_T(K;\rho_n)\cap E_\Sigma(\varepsilon_n)$, the half-gap condition gives $\widehat{\rho}+4\widehat{\varepsilon}\le\bar{\rho}_n(\delta)\le\frac{1}{2}$, so Lemma 3.10 applies, and Lemmas A.7 and A.8 with $\sigma_{d+1}(T_K)=0$ and $\sigma_d(T_K)=\kappa_K\|T_K\|_{\mathrm{op}}$ bound the subspace error by $\bar{\rho}_n(\delta)/(\kappa_K-\bar{\rho}_n(\delta))\le 2\bar{\rho}_n(\delta)/\kappa_K$. Substituting both into Lemma 3.10 yields

$$\begin{aligned}\mathcal{R}_K(\widehat{B}_d)-\mathcal{R}_K^\star(d)&\le L\Big((\widehat{\rho}+4\widehat{\varepsilon})^2+\big(\|\sin\Theta(\mathcal{V}_{K,d},\widehat{\mathcal{V}}_{K,d})\|_{\mathrm{op}}+\widehat{\varepsilon}\big)^2\Big)\\&\le L\bar{\rho}_n(\delta)^2\Big(1+\frac{81/16}{\kappa_K^2}\Big)\\&\le\frac{7L\bar{\rho}_n(\delta)^2}{\kappa_K^2},\end{aligned}\tag{C.5}$$

where the last two inequalities use $\widehat{\varepsilon}\le\bar{\rho}_n(\delta)/4$ and $\kappa_K\le 1$. Adding $\mathcal{R}_K^\star(d)=\frac{1}{K}\big(\mathbb{E}[\|Y-B_{\mathrm{OLS}}^{\mathrm{raw}}Z\|^2]+\sum_{j>d}\sigma_j(T_K)^2\big)$ from Proposition 3.9 gives (3.10). □

C.7 Proof of the Oracle Inequality

Proof of Theorem 3.15. On $E_T(K;\rho_n)\cap E_\Sigma(\varepsilon_n)$, Assumption 3.14 gives $\widehat{\rho}+4\widehat{\varepsilon}\le\bar{\rho}_n(\delta)<\overline{\Delta}_{K,r}\le 1$, so Lemma 3.10 applies and Lemma A.7 gives $\|\widehat{S}_K-T_K\|_{\mathrm{op}}\le\bar{\rho}_n(\delta)\|T_K\|_{\mathrm{op}}$. Hence Lemma A.8 bounds the subspace error by

$$\|\sin\Theta(\mathcal{V}_{K,r},\widehat{\mathcal{V}}_{K,r})\|_{\mathrm{op}}\le\frac{\bar{\rho}_n(\delta)\|T_K\|_{\mathrm{op}}}{\sigma_r(T_K)-\sigma_{r+1}(T_K)-\bar{\rho}_n(\delta)\|T_K\|_{\mathrm{op}}}=\frac{\bar{\rho}_n(\delta)}{\overline{\Delta}_{K,r}-\bar{\rho}_n(\delta)}.$$

Substituting this and $\widehat{\varepsilon}\le\bar{\rho}_n(\delta)/4$ into Lemma 3.10 yields

$$\mathcal{R}_K(\widehat{B}_r)-\mathcal{R}_K^\star(r)\le L\bar{\rho}_n(\delta)^2\Big(1+\Big(\frac{1}{\overline{\Delta}_{K,r}-\bar{\rho}_n(\delta)}+\frac{1}{4}\Big)^2\Big)\le\frac{3L\bar{\rho}_n(\delta)^2}{(\overline{\Delta}_{K,r}-\bar{\rho}_n(\delta))^2},$$

where the last inequality uses $0 < \overline{\Delta}_{K,r} - \bar{\rho}_n(\delta) \le 1$. Adding $\mathcal{R}^\star_K(r) = \frac{1}{K}\big(\mathbb{E}[\|Y - B^{\mathrm{raw}}_{\mathrm{OLS}} Z\|^2] + \sum_{j>r} \sigma_j(T_K)^2\big)$ from Proposition 3.9 gives (3.11). □

C.8 Proof of Predictive Isotropy

Proof of Lemma 4.1. Let $T_K = U_K S_K V_K^\top$ be a compact singular value decomposition, where $U_K \in \mathbb{R}^{Kd \times d}$ and $V_K \in \mathbb{R}^{D_x \times d}$ have orthonormal columns, and $S_K = \mathrm{diag}(\sigma_1(T_K), \dots, \sigma_d(T_K))$. Since $\mathrm{rank}(T_K) = d$, $P_K = T_K \widetilde{Z}$ lies in $\mathrm{col}(T_K) = \mathrm{col}(U_K)$. Thus $U_K^\top P_K$ is the intrinsic coordinate vector of P_K on its rank-d predictive range. Covariance isotropy of P_K on this range is therefore equivalent to covariance isotropy of $U_K^\top P_K$ on $\mathbb{R}^d$. Using $P_K = T_K \widetilde{Z}$ and $\mathrm{Cov}(\widetilde{Z}) = I_{D_x}$ (as $\mathbb{E}[\widetilde{Z}] = 0$ and $\Sigma_{ZZ} = \mathbb{E}[ZZ^\top]$),

$$\mathrm{Cov}(U_K^\top P_K) = \mathrm{Cov}(U_K^\top T_K \widetilde{Z}) = \mathrm{Cov}(S_K V_K^\top \widetilde{Z}) = S_K^2.$$

Thus P_K has isotropic covariance on its rank-d range if and only if $S_K^2 = c I_d$ for some $c > 0$. The nonzero eigenvalues of $H_K = T_K^\top T_K$ are the diagonal entries of S_K^2, so this is exactly (4.1). □

C.9 Proof of the Optimality of Predictive Isotropy

Proof of Theorem 4.2. Let $\lambda_1 \ge \cdots \ge \lambda_d > 0$ denote the nonzero eigenvalues of H_K, with $\sum_{i=1}^d \lambda_i = \tau$. At fixed $\mathrm{tr}(H_K) = \tau$, $\lambda_d \le \tau/d \le \lambda_1$, with both equalities if and only if the spectrum is flat; hence the isotropic spectrum simultaneously minimizes λ_1 and maximizes $\lambda_d = \sigma_d(T_K)^2$, which uniquely maximizes $\kappa_K^2 = \lambda_d/\lambda_1 \le 1$. Because $\mathrm{rank}(T_K) = d$, we have $\sigma_{d+1}(T_K) = 0$, so the relative gap is $\overline{\Delta}_{K,d} = \sigma_d(T_K)/\|T_K\|_{\mathrm{op}} = \kappa_K$. For a fixed effective noise level $\bar{\rho}_n(\delta) > 0$, the excess-risk factor $\bar{\rho}_n(\delta)^2/(\kappa_K - \bar{\rho}_n(\delta))^2$ is strictly decreasing in κ_K whenever $\bar{\rho}_n(\delta) < \kappa_K$, so it is at least $\bar{\rho}_n(\delta)^2/(1 - \bar{\rho}_n(\delta))^2$, with equality if and only if $\kappa_K = 1$. Also, $L = d\lambda_1/K \ge \tau/K$, with equality if and only if $\lambda_1 = \tau/d$. Both factors are positive, so the excess-risk term $3L\bar{\rho}_n(\delta)^2/(\kappa_K - \bar{\rho}_n(\delta))^2$ is at least $3\tau\bar{\rho}_n(\delta)^2/\big(K(1-\bar{\rho}_n(\delta))^2\big)$, with equality if and only if $\kappa_K = 1$ and $\lambda_1 = \tau/d$, that is, if and only if the spectrum is flat. □

C.10 Proof of Gauge Invariance

Proof of Lemma 4.4. Since G^{-1} inverts G almost everywhere on the support of $V_K^\top \widetilde{Z}$, almost surely

$$F_G(\Lambda_G(Z)) = U_K S_K\, G^{-1}\big(G(V_K^\top \widetilde{Z})\big) = U_K S_K V_K^\top \widetilde{Z} = T_K \widetilde{Z} = P_K,$$

which does not depend on G. □

C.11 Proof of Covariance Isotropy Subsumption

Proof of Theorem 4.5. By Lemma 4.4, $\Lambda_G(Z) = G V_K^\top \Sigma_{ZZ}^{-1/2} Z$, and $\mathrm{Cov}(Z) = \Sigma_{ZZ}$ since $\mathbb{E}[Z] = 0$, so

$$\mathrm{Cov}(\Lambda_G(Z)) = G V_K^\top \Sigma_{ZZ}^{-1/2}\, \Sigma_{ZZ}\, \Sigma_{ZZ}^{-1/2} V_K G^\top = G V_K^\top V_K G^\top = G G^\top = I_d,$$

using $V_K^\top V_K = I_d$ from the SVD and $GG^\top = I_d$ since G is orthogonal. □

C.12 Proof of Gaussian Isotropy Subsumption

Proof of Theorem 4.6. We take $G^\star$ to be the optimal transport map from the distribution of $V_K^\top \widetilde{Z}$ to $N(0, I_d)$ and check that it is a gauge in the sense of Lemma 4.4.

Let $\mu := \mathcal{L}(V_K^\top \widetilde{Z})$ and $\gamma := N(0, I_d)$. By assumption, μ is absolutely continuous. It also has finite second moments, since

$$\mathbb{E}\big[V_K^\top \widetilde{Z}\widetilde{Z}^\top V_K\big] = V_K^\top \Sigma_{ZZ}^{-1/2}\, \mathbb{E}[ZZ^\top]\, \Sigma_{ZZ}^{-1/2} V_K = V_K^\top V_K = I_d.$$

Brenier's theorem [19] then gives a convex function φ whose gradient, defined μ-almost everywhere, pushes μ forward to γ. Let $G^\star := \nabla\varphi$ where φ is differentiable and $G^\star := 0$ elsewhere. This map is Borel measurable.

Since γ is also absolutely continuous, the gradient of the convex conjugate φ^* pushes γ back to μ and inverts $G^\star$,

$$\nabla\varphi^*\big(G^\star(x)\big) = x \qquad \text{for } \mu\text{-almost every } x.$$

Hence Lemma 4.4 applies, so $F_{G^\star} \circ \Lambda_{G^\star}$ realizes P_K. Moreover,

$$\Lambda_{G^\star}(Z) = G^\star\big(V_K^\top \widetilde{Z}\big) \sim G^\star_{\#}\mu = \gamma = N(0, I_d).$$

□

C.13 Proof of Linear Gaussian Gauge

Proof of Corollary 4.7. The whitened context $\widetilde{Z} = \Sigma_{ZZ}^{-1/2} Z \sim N(0, I_{D_x})$ under Gaussian context, and the rows of $V_K^\top$ are orthonormal, so $\Lambda_I Z = V_K^\top \widetilde{Z} \sim N(0, I_d)$. Hence $G^\star = I$ realizes Theorem 4.6, and the encoder-predictor factorization specializes to (Λ_I, F_I). □

C.14 Proof of Isotropy under Weight Decay

Proof of Proposition 4.8. Since a symmetric matrix equals sI_d exactly when all its eigenvalues equal s, it suffices to show that the eigenvalues of $\mathrm{Cov}(AZ)$ are $\sigma_1(W\widetilde{A}), \ldots, \sigma_d(W\widetilde{A})$.

Write $M := W\widetilde{A}$. The loss and $\mathcal{R}$ in (4.4) depend on $(W, \widetilde{A})$ only through M, so let G be the gradient of their sum in M. At a stationary point of (4.4), the gradients in W and $\widetilde{A}$ vanish,

$$G\widetilde{A}^\top + \beta W = 0, \qquad W^\top G + \beta \widetilde{A} = 0.$$

Multiplying the first identity by $W^\top$ on the left and the second by $\widetilde{A}^\top$ on the right, and subtracting, gives $\beta(W^\top W - \widetilde{A}\widetilde{A}^\top) = 0$. Since $\beta > 0$ and $\mathbb{E}[Z] = 0$,

$$W^\top W = \widetilde{A}\widetilde{A}^\top = A\Sigma_{ZZ}A^\top = \mathrm{Cov}(AZ).$$

Hence $M^\top M = \widetilde{A}^\top W^\top W \widetilde{A} = (\widetilde{A}^\top \widetilde{A})^2$, so the singular values of M are the eigenvalues of $\widetilde{A}^\top \widetilde{A}$. Because $\widetilde{A}^\top \widetilde{A}$ and $\widetilde{A}\widetilde{A}^\top \in \mathbb{R}^{d\times d}$ share their nonzero eigenvalues, the d largest of these are the eigenvalues of $\widetilde{A}\widetilde{A}^\top = \mathrm{Cov}(AZ)$, which proves the equivalence.

For the Gaussian claim, we show that $Q := U^\top W$ is orthogonal, so that AZ is an orthogonal image of $U^\top W A Z \sim N(0, I_d)$. Since U has $\mathrm{rank}(W)$ columns and $U^\top W A Z$ is d-dimensional, $\mathrm{rank}(W) = d$, so $Q \in \mathbb{R}^{d\times d}$ is invertible. Write $C := \mathrm{Cov}(AZ)$. Because $UU^\top$ projects onto $\mathrm{col}(W)$, the identity $W^\top W = C$ above gives the first relation below, and $U^\top W A Z \sim N(0, I_d)$ gives the second,

$$Q^\top Q = C, \qquad QCQ^\top = I_d.$$

Inverting the second relation and using the first gives $C = (Q^\top Q)^{-1} = C^{-1}$, so $C^2 = I_d$, and since $C \succeq 0$, $C = I_d$. Hence $Q^\top Q = I_d$, so Q is orthogonal and $AZ = Q^\top(U^\top W A Z) \sim N(0, I_d)$. □

D Additional Experimental Details

Visualization. Finite-sample estimates are plotted as solid curves with standard-error bands. Population predictions are plotted as dashed curves with standard-error bands when the population quantity varies across deterministic seeds. Bar charts show standard-error caps and seed-level points. In Figure 6(c), values below 10^{-4} are floored for display.

D.1 Synthetic Experiment Details

Synthetic generator. We draw data from a linear-Gaussian model that instantiates the frozen-teacher linear setting of Section 2. With a signal $s \sim N(0, I_{r_\star})$, the context and the kth target are

$$z = Cs + \sigma_z \epsilon_z, \qquad y_k = a_k^\top s + \sigma_y \epsilon_k,$$

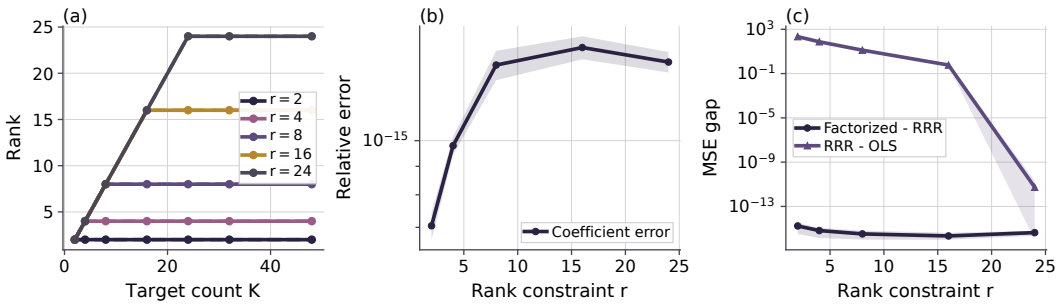

Figure 3: **Reduction to RRR (Proposition 3.1 and Lemma 3.2).** Panel (a) shows that the rank of the closed-form RRR coefficient equals $\min\{r, \operatorname{rank}(T_K)\}$ (Lemma 3.2) at every plotted r and K. Panels (b,c) show that an explicit WA factorization reproduces the RRR coefficient and loss up to roundoff.

Table 2: Synthetic experiment configurations.

Experiment	n_{train}	n_{test}	D_x	$r_\star$	K	d	$\sigma_z = \sigma_y$
K saturation	1024	8192	160	64	sweep	64	0.12
Heterogeneity	32768	8192	64	24	48	24	0.10
Bottleneck	128	8192	64	24	48	sweep	0.16
Oracle inequality	sweep	–	80	32	64	16	0.25
Post-saturation	2^{25}	–	160	64	sweep	64	0.25
Predictive isotropy	1024	8192	64	24	24	24	0.25
Gauge factorization	8192	8192	64	24	24	24	0.25

where $C \in \mathbb{R}^{D_x \times r_\star}$ is random with orthonormal columns, $a_k \in \mathbb{R}^{r_\star}$ is the kth target direction, and the noise terms are independent standard Gaussians. Table 2 lists the sample sizes, dimensions, and noise level of each experiment. For $Z = z$ and the stacked targets $Y = A_K s + \sigma_y \epsilon \in \mathbb{R}^K$, where $A_K = [a_1, \ldots, a_K]^\top$, the whitened map in (3.2) is

$$T_K = (1 + \sigma_z^2)^{-1/2} A_K C^\top.$$

Hence $r_{\text{pop}}(K) = \operatorname{rank}(A_K)$, which saturates at $r_\star$ once the a_k span $\mathbb{R}^{r_\star}$ (Corollary 3.5), and the nonzero spectrum of $H_K = T_K^\top T_K$ is that of $A_K^\top A_K/(1 + \sigma_z^2)$. The target stacks below therefore set rank, heterogeneity, and conditioning through A_K alone.

Target stacks. In the saturation and bottleneck experiments (Figures 4 and 6), the target stack is complementary (Proposition 3.4) and progressively covers the $r_\star$ signal directions, so $r_{\text{pop}}(K)$ grows with K until it saturates at $r_\star$ (Corollary 3.5). In the heterogeneity experiment (Figure 5) that tests Proposition 3.8, each target direction mixes a shared direction with its own basis direction, and a parameter $\alpha \in (0, 1]$ controls the weight on the latter, moving the stack from nearly redundant ($\alpha \to 0$) to complementary ($\alpha = 1$). We choose the values of α so that λ_{het} (Definition 3.6) is spread roughly evenly, and plot against its measured value.

Models and baselines. By Proposition 3.1, the population JEPA predictor is the closed form (3.3) with $r = d$, and the learned predictor is the same formula with the sample moments $\widehat{\Sigma}_{ZZ}$ and $\widehat{\Sigma}_{YZ}$. Unrestricted OLS, $B_{\text{OLS}}^{\text{raw}}$ in (3.2), removes the bottleneck and is the baseline (Proposition 3.9). Figure 3 checks Proposition 3.1 and Lemma 3.2 by comparing the rank of the closed-form RRR coefficient with $\min\{r, \operatorname{rank}(T_K)\}$ and by factorizing it as WA.

Metrics. Test MSE is averaged over targets, matching $\mathcal{R}_K$ in (2.3). Recovered rank is the number of singular values of the learned predictor above 0.05 times the largest, an estimate of $\operatorname{rank}(B_d^\star) = \min\{d, r_{\text{pop}}(K)\}$ from Lemma 3.2. The heterogeneity level λ_{het} is the smallest nonzero eigenvalue of H_K (Definition 3.6), which here equals $\lambda_{r_{\text{pop}}(K)}(A_K^\top A_K)/(1 + \sigma_z^2)$. Subspace error is $\|\sin\Theta(\mathcal{V}_{K,d}, \widehat{\mathcal{V}}_{K,d})\|_{\text{op}}$, the sine of the largest principal angle between the top-d right singular subspaces of T_K and of its sample version $\widehat{T}_K$ (Proposition 3.8). Reference-stack MSE is the test MSE of a linear probe from the learned embedding to a fixed target stack shared by all conditions, so unlike $\mathcal{R}_K$ it is comparable across target stacks.

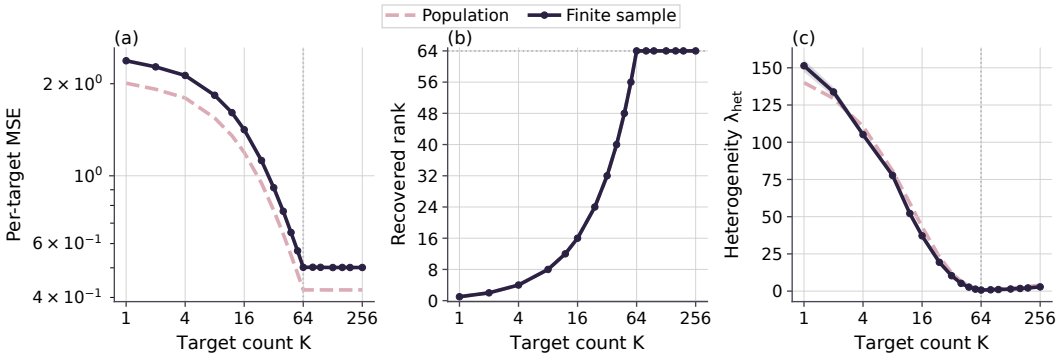

Figure 4: **Rank saturation (Proposition 3.4 and Corollary 3.5).** Complementary targets reduce per-target MSE and increase recovered rank until predictive rank saturates.

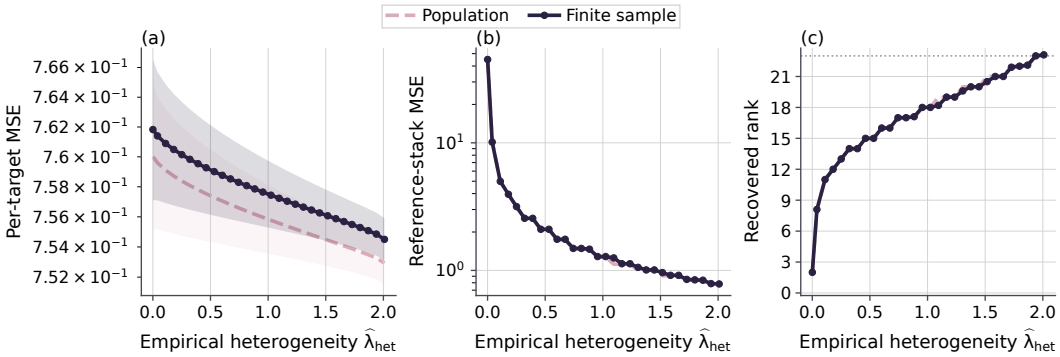

Figure 5: **Target heterogeneity (Proposition 3.8).** As $\widehat{\lambda}_{\mathrm{het}}$ grows, the model recovers more signal directions, and a linear probe on a fixed target stack improves sharply. Per-target MSE on the training targets barely changes, since they depend mainly on the recovered directions.

Heterogeneity. In Figure 5, per-target $\mathcal{R}_K$ is computed on each condition's own training stack, and the reference-stack MSE on a fixed stack shared by all conditions. The bound of Proposition 3.8 is in Figure 8 (c), where $\sigma_d(T_K)^2 = \lambda_{\mathrm{het}}$, and its concentration input on this sweep in Figure 9 (a).

Embedding bottleneck. We sweep d with the target stack fixed. By Proposition 3.9, the rank-d oracle risk exceeds the OLS risk by the spectral tail $\sum_{j>d} \sigma_j(T_K)^2$. Panel (c) of Figure 6 plots this tail and its plug-in version from $\widehat{T}_K$, together with the rank-d estimator's excess test MSE over OLS, all divided by $\sum_j \sigma_j(T_K)^2$. Panel (b) compares the recovered rank with $\min\{d, r_\star\}$ (Lemma 3.2).

Oracle inequality. Figure 7 tests Theorem 3.15 with an active spectral tail (zero in Figure 1). The target stack has $K = 64$ targets and signal dimension $r_\star = 32$, and we fix $r = d = 16$ with a controlled relative gap $\overline{\Delta}_{K,r} = 0.408$. The rank-$r$ oracle risk is then $\mathcal{R}^\star_K(r) \approx 0.365$, the OLS floor (0.121) plus the per-target spectral tail (0.244). We sweep $n_{\mathrm{train}} \in \{2^{19}, 2^{20}, \ldots, 2^{25}\}$ over ten seeds. We draw the sample second moments of (Z, Y) exactly from their Wishart distribution and compute every risk in closed form. At each n we set ρ_n and ε_n to the largest values of $\|\widehat{T}_K - T_K\|_{\mathrm{op}}/\|T_K\|_{\mathrm{op}}$ and $\|\widehat{F} - I\|_{\mathrm{op}}$ over seeds, so both events in (3.8) hold on every run. Over the sweep, $\bar{\rho}_n$ falls from 0.122 at $n = 2^{19}$ to 0.015 at $n = 2^{25}$ and stays below $\overline{\Delta}_{K,r}$, as Assumption 3.14 requires. We use $L = r\,\sigma_1(T_K)^2/K = 0.697$ from Theorem 3.15. The population risk stays below the bound (3.11) on every run, and its excess over $\mathcal{R}^\star_K(r)$ falls as $1/n$, from 2.4×10^{-5} to 3.7×10^{-7}.

Post-saturation conditioning. Figure 1 tests the relative-conditioning result of Proposition 3.11. The target stack is full rank from $K^\star = d = 64$ onward, so additional targets do not change $r_{\mathrm{pop}}(K)$. We start from an imbalanced stack, $A_{K^\star}^\top A_{K^\star} = Q\,\mathrm{diag}(c_1, \ldots, c_{64})Q^\top$ with $c_i = 0.02^{(i-1)/63}$ and a fixed random rotation Q, so $\kappa^2_{K^\star} \approx 0.02$ in (3.9). To raise κ^2_K, each target beyond $K^\star$ adds 0.1 to H_K along its current weakest eigendirection. We fix $n = 2^{25}$ and draw the moments as above, with every K using the leading targets of the same draw, so added targets keep the samples. At each K

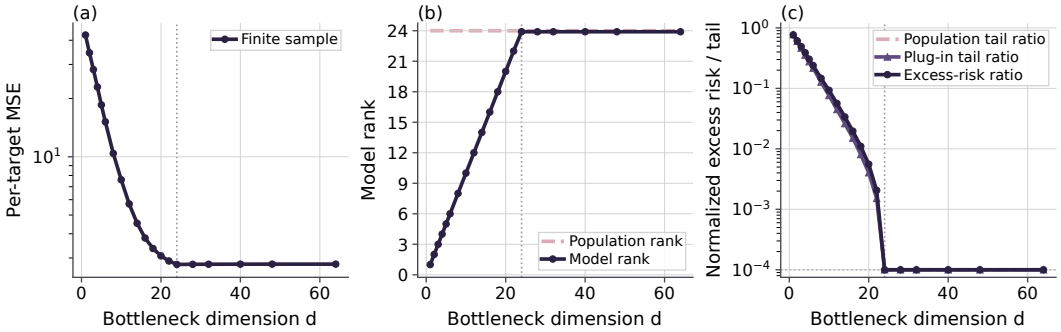

Figure 6: **Embedding bottleneck (Lemma 3.2 and Proposition 3.9).** The recovered rank follows $\min\{d, r_\star\}$ and the risk above OLS matches the spectral tail, so the per-target MSE stops improving once the tail vanishes at $d = r_\star$ (dotted).

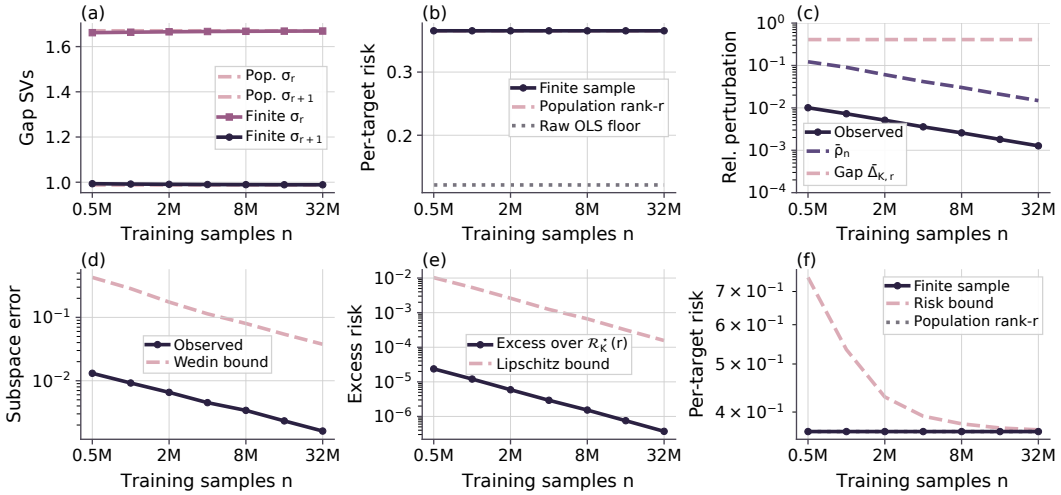

Figure 7: **Oracle inequality with an active spectral tail ($r = 16 < r_\star = 32$).** The population risk stays below the bound of Theorem 3.15 on every run and approaches the rank-r oracle risk $\mathcal{R}^\star_K(r)$. Against the training size n: (a) population and finite-sample singular values at ranks r and $r+1$, (b) population risk with $\mathcal{R}^\star_K(r)$ and the OLS floor, (c) relative perturbation with the effective noise level $\bar{\rho}_n$ and the gap $\overline{\Delta}_{K,r}$, (d) subspace error and the Wedin bound, (e) excess risk over $\mathcal{R}^\star_K(r)$ and its Lipschitz bound (Lemma 3.10), and (f) population risk between $\mathcal{R}^\star_K(r)$ and the bound.

we set ρ_n and ε_n to the largest values over seeds. We use $L = d\,\sigma_1(T_K)^2/K$ from Proposition 3.11. Panel (e) plots the excess-risk term $7L\bar{\rho}_n^2/\kappa_K^2$ of (3.10), and panel (f) adds it to $\mathcal{R}^\star_K(d)$, whose spectral tail is zero since $r_{\mathrm{pop}}(K) = d$. Panel (c) shows $\bar{\rho}_n \le \kappa_K/2$ at every K, so the half-gap condition of Proposition 3.11 holds across the window. As κ_K^2 rises from 0.02 to 0.90, the excess-risk term falls from 0.14 to 4.3×10^{-4}. Both Figures 1 and 7 use the sample-whitened RRR fit and evaluate every bound with the theory's constants and no calibration.

Predictive isotropy. Figure 8 tests the second-moment prediction of Lemma 4.1 and the bound minimization of Theorem 4.2, with rank and trace held fixed. For a decay parameter $\rho \in (0, 1]$, we set $A_K^\top A_K = Q\,\mathrm{diag}(c_1, \ldots, c_d)Q^\top$ with $c_i = \tau\rho^{i-1}/\sum_{j=1}^d \rho^{j-1}$, a random rotation Q, and τ fixed so that $\mathrm{tr}(H_K) = 22.59$. This fixes $\mathrm{rank}(A_K) = d$ and the trace while κ_K^2 in (3.9) varies. We use 20 values of ρ in $[0.78, 1]$, which span $\kappa_K^2 \in [3.3 \times 10^{-3}, 1]$ and are denser where $\sigma_d(T_K)$ approaches $\|\widehat{T}_K - T_K\|_{\mathrm{op}}$. Panel (a) plots the eigenvalues of H_K, which are the nonzero covariance eigenvalues of $P_K = T_K\widetilde{Z}$ in Lemma 4.1. Panel (b) plots $\sigma_d(T_K)$ and $\|\widehat{T}_K - T_K\|_{\mathrm{op}}$, and panel (c) compares the subspace error with the bound of Proposition 3.8 built from them, $\|\widehat{T}_K - T_K\|_{\mathrm{op}}/(\sigma_d(T_K) - \|\widehat{T}_K - T_K\|_{\mathrm{op}})$, clipped at one when the denominator is nonpositive. Panel (d) reports the reference-stack excess MSE on a balanced target stack, which weights all signal

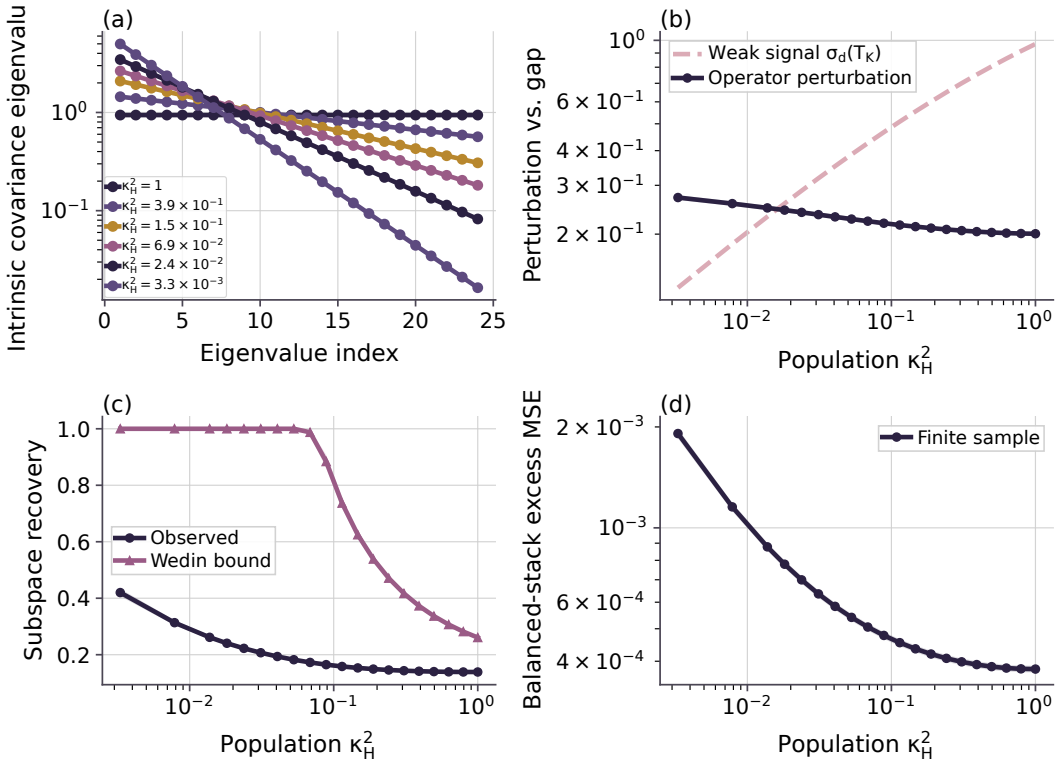

Figure 8: **Predictive isotropy (Lemma 4.1 and Theorem 4.2).** At fixed rank and trace, a flatter spectrum of H_K raises the weakest singular value $\sigma_d(T_K)$ relative to the estimation noise, which improves subspace recovery and lowers the balanced-stack excess MSE. (a) Eigenvalues of H_K for selected κ_K^2. Against κ_K^2: (b) $\sigma_d(T_K)$ and $\|\widehat{T}_K - T_K\|_{\mathrm{op}}$, (c) subspace error and the bound (Proposition 3.8), (d) balanced-stack excess MSE.

directions equally. For 16 of the 20 settings the denominator is positive for all ten seeds, and 10 of the 20 satisfy the half-gap condition $\|\widehat{T}_K - T_K\|_{\mathrm{op}} \le \sigma_d(T_K)/2$ for all ten seeds.

Concentration and perturbation diagnostics. Figure 9 plots each perturbation against its concentration bound. Panel (a) compares $\|\widehat{\Sigma}_{YZ} - \Sigma_{YZ}\|_{\mathrm{op}}$ with the bound of Lemma A.4 on the heterogeneity sweep. Panel (b) compares $\|\widehat{T}_K - T_K\|_{\mathrm{op}}$ with its analogue

$$\widehat{C}_{0.9}\|T_K\|_{\mathrm{op}}\left(\sqrt{\frac{r(T_K)+\log(1/\delta)}{n}}+\frac{r(T_K)+\log(1/\delta)}{n}\right)$$

on the post-saturation window. Panel (c) repeats (b) on the real 8×8 digit halves, with the operator estimated from the full dataset in place of T_K. We use $\delta = 0.1$ and set $\widehat{C}_{0.9}$ to the smallest constant for which each bound holds on 90% of runs, so the figure tests the predicted scaling of each bound.

Gauge factorization. Figure 10 tests Lemma 4.4 and Theorems 4.5 and 4.6. We construct the population whitened map $T_K = U_K S_K V_K^\top$ and draw the SVD coordinate $V_K^\top \widetilde{Z}$ of the whitened context from a standard Gaussian distribution (Corollary 4.7) or a standardized Laplace distribution. We compare the pairs (Λ_G, F_G) of Lemma 4.4 for four gauges G, namely the identity, a random orthogonal matrix, an anisotropic diagonal matrix as a non-orthogonal control, and a nonlinear map sending each coordinate to a standard-normal marginal. Panel (a) checks Lemma 4.4 with the relative error of $F_G \circ \Lambda_G$ against $P_K = T_K\widetilde{Z}$. Panel (b) checks Theorem 4.5 with the anisotropy of $\Lambda_G(Z)$, the relative Frobenius distance of its sample covariance from the nearest multiple of I_d. Panel (c) checks Theorem 4.6 and Corollary 4.7 with the projected Gaussianity statistic (4.3) of $\Lambda_G(Z)$, which averages the Epps–Pulley standard-normal statistic over the coordinate axes and random directions.

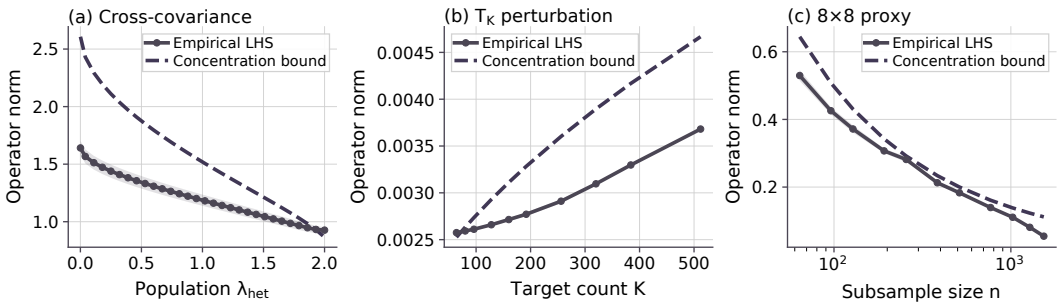

Figure 9: **Concentration bounds (Lemma A.4).** With a calibrated constant, each bound follows the trend of its perturbation on (a) the heterogeneity sweep, (b) the post-saturation window, and (c) 8×8 digit halves.

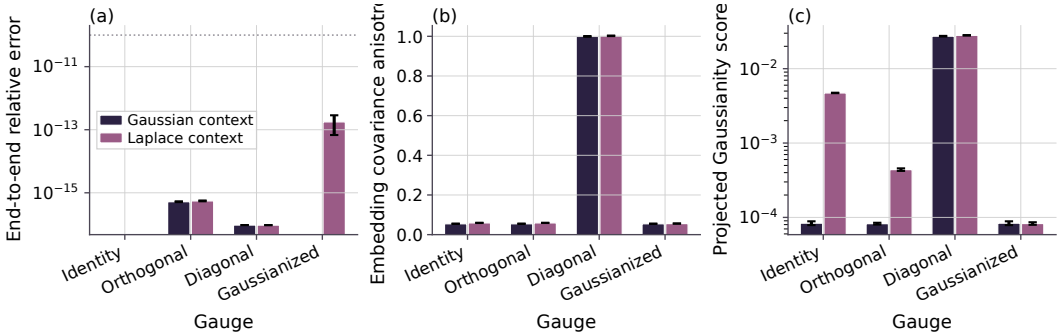

Figure 10: **Gauge factorization (Lemma 4.4 and Theorems 4.5 and 4.6).** Every gauge gives the same prediction up to roundoff (a), and only the diagonal gauge breaks covariance isotropy beyond finite-sample error (b). The identity gauge already gives a Gaussian embedding under Gaussian context (Corollary 4.7), while under Laplace context only the Gaussianizing gauge does (c).

D.2 Real-Image Regularizer Details

Figures 11 to 13 show completions for 8×8 digits, CIFAR-10 and CIFAR-100.

Objectives. The four objectives in Table 1 are special cases of

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{emb}} \mathcal{L}_{\text{emb-iso}} + \lambda_{\text{pred}} \mathcal{R}_{\text{PI}},$$

where $\mathcal{L}_{\text{pred}}$ is the latent MSE between the predictions and the teacher embedding, $\mathcal{R}_{\text{PI}}$ is the regularizer (4.3) on the predictions, and $\mathcal{L}_{\text{emb-iso}}$ is the same penalty on the encoder embedding. Both penalties also pull the sample covariance of their representation toward I. Encoder-side regularization sets $\lambda_{\text{pred}} = 0$, prediction-side regularization sets $\lambda_{\text{emb}} = 0$, and the combined objective uses both. The pixel head trains on the detached predictions with its own pixel MSE, so its loss does not reach the encoder or predictor.

Architectures and training. The 8×8 task trains on a stratified 8% split, chooses regularizer weights on 500 held-out images, and tests on the rest. Its encoder, predictor, and pixel head are MLPs with hidden width 512, and the embedding and prediction are 16-dimensional. It trains with AdamW for 400 epochs with batch size 32, learning rate 10^{-3}, no weight decay, covariance weight 10, and 32 random projection directions plus the coordinate axes, with regularizer weights from eleven values between 0.05 and 5000.

The MNIST task uses a convolutional encoder with channels $(64, 128, 256)$ and a 16-dimensional embedding, a linear $16 \to 16$ predictor, and a convolutional pixel head. It trains on 10000 images with AdamW for 50 epochs with batch size 256, learning rate 3×10^{-4} with cosine decay, weight decay 2 on the encoder and predictor and 10^{-4} on the pixel head, covariance weight 1.0, and 128 random projection directions plus the coordinate axes, and it is tested on the standard split. Regularizer weights are chosen from $\{0.08, 0.24, 0.8, 2.4, 8, 24\}$ by validation latent risk on 5000 images drawn from the training images outside the training split.

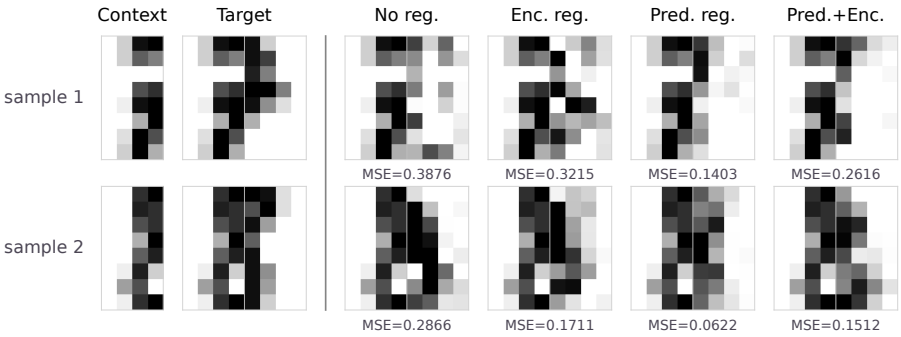

Figure 11: 8×8 digit completion examples. The left half is observed, and the pixel head renders each objective's prediction of the right half.

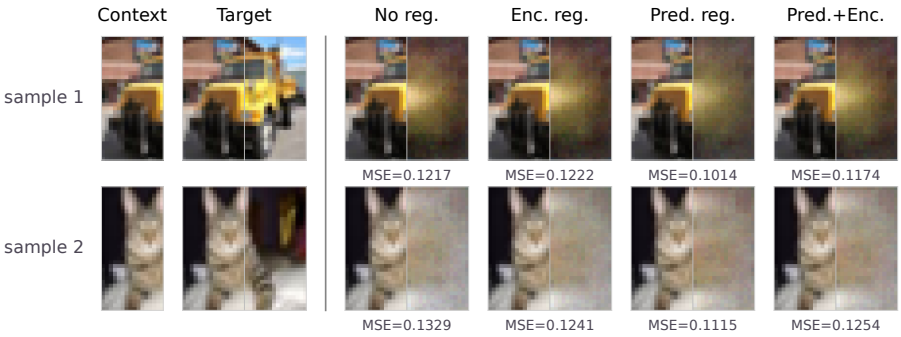

Figure 12: CIFAR-10 completion examples. The left half is observed, and the pixel head renders each objective's prediction of the right half.

The CIFAR-10 and CIFAR-100 tasks use the MNIST architecture and training settings, adapted to 32×32 color images. The first encoder layer reads three color channels instead of one, and the output layer produces three. They train on 10000 stratified images from the 50000-image training split and are tested on the standard 10000-image split. Table 3 specifies all three architectures layer by layer.

Metrics. Foreground MSE is the MSE over target pixels with intensity above 0.05. The Gaussianity distance is $\mathcal{R}_{\mathrm{PI}}$ evaluated on the test-set predictions, and the predictive effective rank is $r(\widehat{C})$ (Lemma A.4) for their sample covariance $\widehat{C}$. We also evaluate the effective rank of the encoder embedding, and its Gaussianity distance after whitening it. Whitening is a linear gauge (Lemma 4.4) that folds into the last encoder layer and the first predictor layer, so it leaves every prediction unchanged and makes the embedding isotropic.

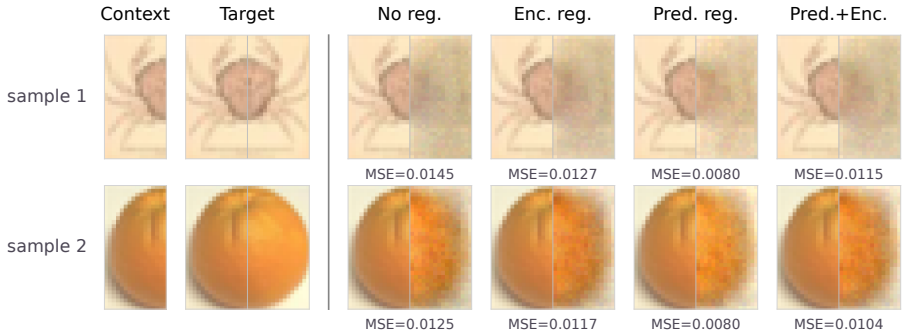

Figure 13: CIFAR-100 completion examples. The left half is observed, and the pixel head renders each objective's prediction of the right half.

Table 3: **Encoder, predictor, and pixel-head architectures.** Convolutions are listed as kernel/stride and input→output channels, and ConvT is a transposed convolution.

Component	8×8 digits	MNIST CNN	CIFAR-10/100 CNN
Context / target	32 / 32	1×28×14	3×32×16
Encoder input layer	Linear 32→512	Conv 3×3, 1→64	Conv 3×3, 3→64
Encoder stage 1	–	Conv 3×3/2, 64→128	Conv 3×3/2, 64→128
Encoder stage 2	–	Conv 3×3/2, 128→256	Conv 3×3/2, 128→256
Spatial path	–	28×14→14×7→7×4	32×16→16×8→8×4
Encoder head	Linear 512→16	Linear 7168→512→16	Linear 8192→512→16
Embedding dim	16	16	16
Predictor	Linear 16→512→16	Linear 16→16	Linear 16→16
Predictor dim	16	16	16
Pixel head projection	–	Linear 16→64·7·4	Linear 16→64·8·4
Pixel head stage 1	–	ConvT 4×3/2, 64→128	ConvT 4×4/2, 64→128
Pixel head stage 2	–	ConvT 4×4/2, 128→256	ConvT 4×4/2, 128→256
Readout	Linear 16→512→32	Conv 1×1, 256→1	Conv 1×1, 256→3
Norm / activation	GELU	GroupNorm(8) + GELU	GroupNorm(8) + GELU
Parameters	6.7×10^4	4.7×10^6	5.3×10^6

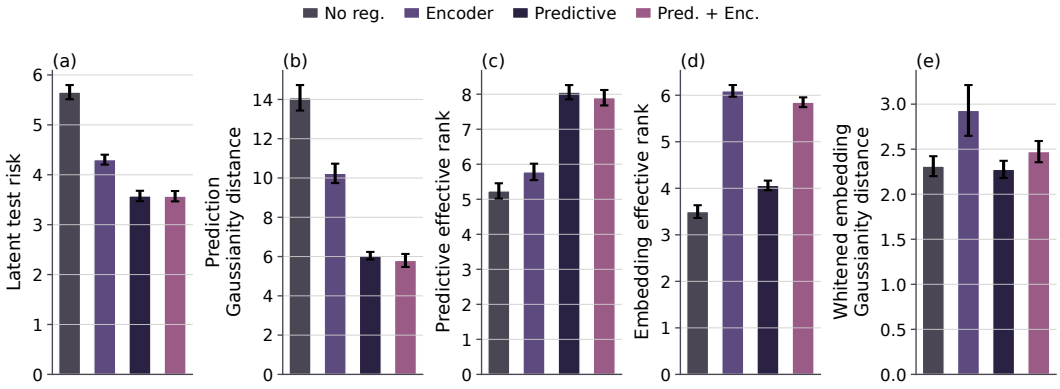

Figure 14: **Regularizer** (4.3) **on** 8×8 **digits.** Gaussianity distances are in units of 10^{-3}. Relative to no regularization, prediction-side regularization lowers the Gaussianity distance from 14.1 to 6.04 and raises the predictive effective rank from 5.24 to 8.06, beyond encoder-side regularization (10.2 and 5.78). Encoder-side regularization instead raises the embedding effective rank from 3.50 to 6.09.

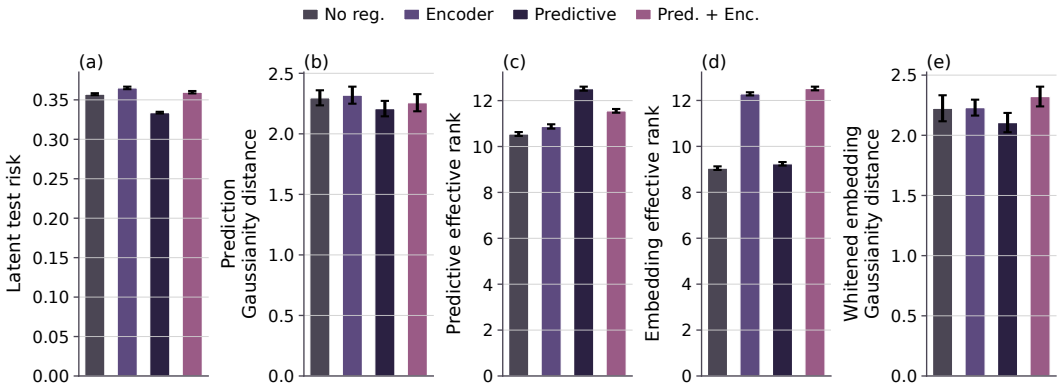

Figure 15: **Regularizer** (4.3) **on MNIST.** Gaussianity distances are in units of 10^{-3}. Relative to no regularization, prediction-side regularization lowers the Gaussianity distance from 2.30 to 2.21, which encoder-side regularization barely changes, and raises the predictive effective rank from 10.55 to 12.52. Encoder-side regularization instead raises the embedding effective rank from 9.05 to 12.29 but the predictive effective rank only to 10.87.

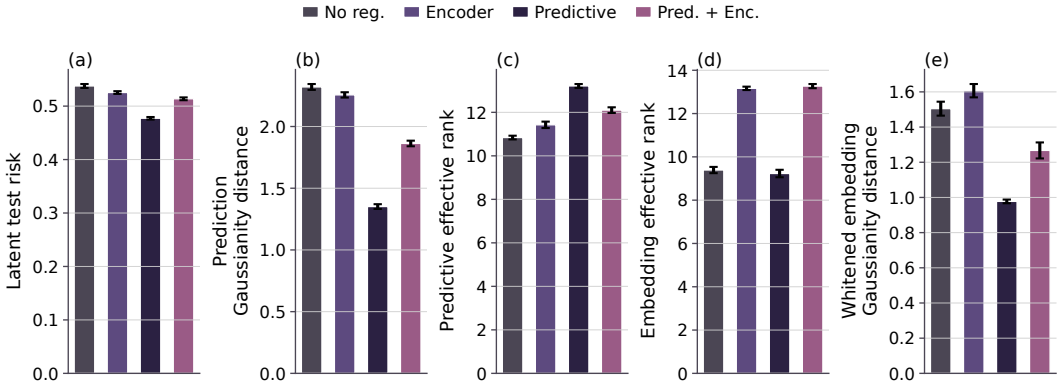

Figure 16: **Regularizer** (4.3) **on CIFAR-10.** Gaussianity distances are in units of 10^{-3}. Relative to no regularization, prediction-side regularization lowers the Gaussianity distance from 2.32 to 1.35, which encoder-side regularization barely changes, and raises the predictive effective rank from 10.84 to 13.21. Encoder-side regularization instead raises the embedding effective rank from 9.39 to 13.16 but the predictive effective rank only to 11.43. After whitening, prediction-side regularization lowers the encoder embedding's Gaussianity distance from 1.50 to 0.98, which encoder-side regularization raises to 1.61.

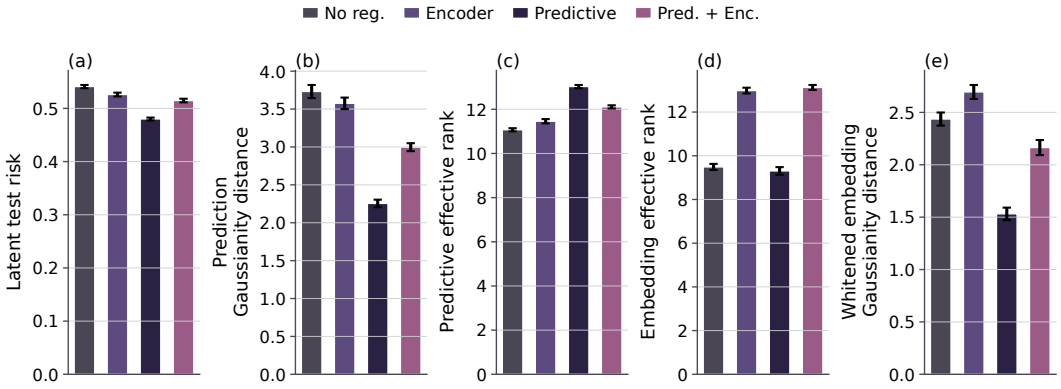

Figure 17: **Regularizer** (4.3) **on CIFAR-100.** Gaussianity distances are in units of 10^{-3}. Relative to no regularization, prediction-side regularization lowers the Gaussianity distance from 3.73 to 2.25, which encoder-side regularization barely changes, and raises the predictive effective rank from 11.07 to 13.03. Encoder-side regularization instead raises the embedding effective rank from 9.49 to 12.98 but the predictive effective rank only to 11.46. After whitening, prediction-side regularization lowers the encoder embedding's Gaussianity distance from 2.44 to 1.53, which encoder-side regularization raises to 2.70.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are scoped to the linearized/frozen-teacher setting and matched to the stated theorems.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see the Limitations paragraph at the end of Appendix B (Related Works and Discussion).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please see our main text and appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see our experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: We release a public code repository (https://github.com/thermau5/on-jepa-isotropy) containing experiment scripts, configurations, plotting scripts, tests, and reproduction instructions. The experiments use public datasets (`sklearn` digits, MNIST, CIFAR-10, and CIFAR-100) and synthetic generators included in the repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see our experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see our experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see our experimental section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: All authors read through carefully.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see our impact statement.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no pretrained models, datasets, or other assets that could pose misuse or dual-use risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see our experimental section.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code repository is documented with a README covering dependencies, reproduction commands, generated outputs, and dataset usage. The paper does not introduce a new dataset or pretrained model.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects, so no IRB approval was required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core contributions of this paper are theoretical (a reduced-rank regression formulation of JEPA, the predictive isotropy condition, and accompanying proofs). LLM-based coding assistants were used only as basic code assistance for implementing the numerical validation experiments (e.g., plotting and data pipelines), and did not impact the core methodology, scientific rigor, or originality of the research. Per NeurIPS policy, declaration is therefore not required.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.