SHARP SCALAR RECONSTRUCTION RATES AND FUSION SEPARATION UNDER LIPSCHITZ CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In multimodal prediction, observations may be compressed separately before fusion or jointly after fusion. We ask whether the Euclidean dimension of a nonlinear representation alone imposes a nontrivial information constraint. Without a fixed Lipschitz or network-complexity budget, scalar neural coding extends to arbitrary finite-dimensional observation laws and square-integrable responses. The joint risk infimum equals the full-observation Bayes risk, while separate encoding reduces the dimension budget to modality selection. Under a Lipschitz product budget, however, the approximation problem becomes quantitative. For a uniform input on $[0,1]^d$ with $d \geq 2$, the optimal mean squared reconstruction error through a scalar bottleneck is of order $\Lambda^{-2/(d+1)}$ when the product of the encoder and decoder Lipschitz constants is at most Λ. The lower bound holds for all Lipschitz maps, and finite ReLU networks achieve the same rate. For two independent uniform three-dimensional modalities and an additive response, separate encoding with total latent dimension two has risk of order $\Lambda^{-1/2}$, whereas joint encoding under the same budgets attains $O(\Lambda^{-2/3})$. Linear Gaussian frontiers, exact continuous factorization dimensions, and matching finite-bit Gaussian risk exponents provide complementary characterizations of representation cost. Seven controlled experiments examine the constructions and Gaussian benchmarks, showing a decreasing attained joint-to-separate risk ratio under matched Lipschitz budgets and finite-bit decay rates close to the rank-based predictions.

1 INTRODUCTION

Multimodal prediction combines observations from sensing systems with different measurement mechanisms and resolutions. In precision agriculture, satellite, aerial, handheld, and vehicle-mounted platforms provide heterogeneous measurements (Mulla, 2013); neural predictors are widely studied (Kamilaris & Prenafeta-Boldú, 2018) and can incorporate process-based knowledge (von Bloh et al., 2024). Multimodal learning distinguishes fusion strategies by when information is integrated (Baltrusaitis et al., 2019). We study a specific representation constraint: modalities may be compressed separately before fusion or jointly after fusion, under a common total dimension budget. Linear encoders then induce different admissible subspaces. For nonlinear neural encoders, it is less clear whether dimension alone restricts predictive information or whether additional regularity constraints are needed for fusion placement to matter.

Under joint encoding, predictive directions may combine information across modalities before compression, whereas block-separable encoding must compress each modality within its own observation space. Thus, even under the same total representation dimension, the two architectures can access different predictive subspaces and induce different risk frontiers. This distinction provides the linear benchmark for asking what survives under nonlinear encoding.

A continuous map cannot injectively compress a set with nonempty interior into fewer Euclidean coordinates. Population squared loss, however, requires small average error under a fixed distribution. Neural encoders can approximate finite partitions on high-probability sets while confining the transitions required by continuity to sets of small probability. Exact continuous factorization can therefore remain impossible even when the risk infimum equals the Bayes risk. This leaves a quantitative question: how costly is it to narrow those transition regions? A bound on the product of encoder and decoder Lipschitz constants limits this construction without depending on an arbitrary rescaling

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

of the latent coordinate. We use this constraint to derive a sharp scalar risk rate and an architecture separation that persists when both modalities receive a positive-dimensional representation.

Our contributions are fourfold. **First**, we prove matching scalar reconstruction bounds under a Lipschitz product budget, using encoder level sets for the lower bound and a finite ReLU construction for the upper bound. This yields a fusion separation under a common dimension and Lipschitz budget. **Second**, we extend scalar neural coding to arbitrary finite-dimensional probability laws and square-integrable targets, obtaining the joint and block-separable risk infima. **Third**, linear Gaussian frontiers and exact continuous factorization distinguish dimension-only approximation from subspace and rank constraints. For correlated Gaussian modalities, we prove matching finite-bit risk exponents governed by the joint regression rank and the sum of its modality-specific ranks. **Fourth**, a solvable Gaussian model, a shifted-lattice realization, and seven controlled experiments illustrate these comparisons (Section 10). The experiments examine prescribed constructions, Gaussian bounds, bounded-network training, fixed-code quantization, and attained risk decay under Lipschitz and finite-bit budgets.

2 Relation to Existing Theory

Scalar coding is a direct antecedent of our approximation result. Park et al. (2021) encode multivariate inputs into scalar codewords before vector decoding. More specifically, the ICLR 2024 construction of Kim et al. (2024, Section 4, Lemmas 5–6) covers almost all of the unit cube, in Lebesgue measure, by disjoint small-diameter sets. Successive half-space projections collapse these sets to distinct points, which are then mapped to distinct scalar codes; a ReLU decoder approximates the prescribed vector value at each code. This is already a continuous scalar encoder–decoder construction, not merely a width bound. We extend this existing coding principle to arbitrary probability laws, including singular and atomic laws, and derive the block-separable modality-selection frontier.

Stable nonlinear approximation has also been studied under Lipschitz constraints on encoding and decoding maps (Cohen et al., 2022), and Lipschitz widths quantify approximation through a Lipschitz decoder from a bounded latent set (Petrova & Wojtaszczyk, 2021). Latent rescaling makes a product constraint equivalent to a common bound on the two constants (Cohen et al., 2022, Remark 1.1). Those widths use worst-case error over a compact set. Our result concerns mean squared error under a fixed uniform law, with a scalar encoder and a scale-invariant product budget. Its lower bound combines a level-set estimate for the encoder with a volume bound for the decoder curve. The resulting rate also yields a joint–separate comparison under the same constraint.

Other relevant comparisons involve different notions of representation complexity. Reduced-rank regression and distributed linear estimation impose linear subspace restrictions (Izenman, 1975; Zhu et al., 2005; Schizas et al., 2005); topological autoencoding studies obstructions to exact reconstruction and approximation outside small exceptional sets (Batson et al., 2021; Kvalheim & Sontag, 2024). Our linear Gaussian frontiers and exact factorization ranks distinguish these constraints from dimension-only neural risk infima. Classical quantization and distributed functional quantization already establish nontrivial distortion costs for finite-rate descriptions (Gray & Neuhoff, 1998; Misra et al., 2011). We use this perspective to identify matching exponents for joint and separate Gaussian prediction under one common fixed-length bit budget, and relate them to the two factorization ranks. We do not propose a new general rate–distortion theory or infer optimal finite-codebook behavior from a particular quantized neural construction. Appendix J gives broader comparisons with multimodal learning, sufficient dimension reduction, information bottlenecks, analog compression, and sampling.

3 Prediction Frontiers and Factorization Dimensions

Write $\mathbb{N} = \{1, 2, \ldots\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and $[a] = \{1, \ldots, a\}$ for $a \in \mathbb{N}$. There are $M \in \mathbb{N}$ modalities, $M \geq 2$, with observations $X_m \in \mathbb{R}^{p_m}$ and $p_m \in \mathbb{N}$. Set $X = (X_1, \ldots, X_M) \in \mathbb{R}^p$, where $p = \sum_m p_m$. The response is $Y \in \mathbb{R}^q$, with $q \in \mathbb{N}$ and $\mathbb{E}\,\|Y\|_2^2 < \infty$. All conditional means use Borel versions (Kallenberg, 1997). For $J \subseteq [M]$, let $X_J = (X_m)_{m \in J}$ and define the subset Bayes risk

$$\mathcal{B}(J) = \mathbb{E}\,\|Y - \mathbb{E}[Y \mid X_J]\|_2^2 . \tag{1}$$

The empty set generates the trivial sigma-algebra, so $\mathcal{B}(\varnothing) = \mathbb{E}\,\|Y - \mathbb{E}Y\|_2^2$. The raw-observation Bayes risk is $\mathcal{B}([M])$. For every real matrix A, we index its singular values as $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq 0$ and extend the sequence by zeros beyond $\operatorname{rank}(A)$. The notation $\|A\|_{\mathrm{F}}$ denotes the Frobenius norm.

Let $\rho(t) = \max\{t, 0\}$ denote the ReLU activation function, applied coordinatewise to the hidden-layer preactivations. For integers $a, b \geq 0$, let $\mathrm{NN}_{a,b}$ be the class of all finite feedforward ReLU networks from $\mathbb{R}^a$ to $\mathbb{R}^b$, with arbitrary finite depth and width and a final affine layer. We identify $\mathbb{R}^0$ with a singleton. Affine maps and constant maps belong to the corresponding classes.

For $r \in \mathbb{N}_0$, a joint neural predictor factors through an encoder $\phi \in \mathrm{NN}_{p,r}$ and a decoder $\psi \in \mathrm{NN}_{r,q}$. A block-separable predictor uses local encoders $\phi_m \in \mathrm{NN}_{p_m,r_m}$ and one decoder $\psi \in \mathrm{NN}_{\sum_m r_m,q}$, with $r_m \in \mathbb{N}_0$ and $\sum_m r_m \leq r$. The decoder sees only the concatenated local representations. Define

$$\mathcal{R}^{\mathrm{nn}}_{\mathrm{joint}}(r) = \inf_{\substack{\phi \in \mathrm{NN}_{p,r} \\ \psi \in \mathrm{NN}_{r,q}}} \mathbb{E}\,\|Y - \psi(\phi(X))\|_2^2\,,$$

$$\mathcal{R}^{\mathrm{nn}}_{\mathrm{sep}}(r) = \inf_{\substack{r_m \in \mathbb{N}_0,\ \sum_m r_m \leq r \\ \phi_m \in \mathrm{NN}_{p_m,r_m},\ \psi \in \mathrm{NN}_{\sum_m r_m,q}}} \mathbb{E}\,\|Y - \psi(\phi_1(X_1), \ldots, \phi_M(X_M))\|_2^2\,.$$

These infima compare compression after fusion with local compression followed by a shared decoder.

Only bottleneck dimension is constrained; network sizes and parameters are unrestricted. Risks take values in $[0, \infty]$, but their infima are finite because constant predictors are admissible. Finite-risk predictors are square-integrable because Y is square-integrable.

For comparison, a joint linear encoder is a matrix $L \in \mathbb{R}^{r \times p}$. A separate linear encoder uses matrices $L_m \in \mathbb{R}^{r_m \times p_m}$ with $\sum_m r_m \leq r$. The downstream predictor is unrestricted, so

$$\mathcal{R}^{\mathrm{lin}}_{\mathrm{joint}}(r) = \inf_{\operatorname{rank}(L) \leq r} \mathbb{E}\,\|Y - \mathbb{E}[Y \mid LX]\|_2^2\,,$$

$$\mathcal{R}^{\mathrm{lin}}_{\mathrm{sep}}(r) = \inf_{\sum_m \operatorname{rank}(L_m) \leq r} \mathbb{E}\,\|Y - \mathbb{E}[Y \mid L_1X_1, \ldots, L_MX_M]\|_2^2\,.$$

Dependent output rows may be removed without changing the generated sigma-algebra, so matrix rank and effective output dimension are equivalent in these definitions.

Risk infima must be distinguished from exact factorization. For $r \in \mathbb{N}_0$, a joint continuous factorization of the Bayes predictor consists of continuous maps $\phi : \mathbb{R}^p \to \mathbb{R}^r$ and $\psi : \mathbb{R}^r \to \mathbb{R}^q$ satisfying $\psi(\phi(X)) = \mathbb{E}[Y \mid X]$ almost surely. A separate continuous factorization consists of local dimensions $r_m \in \mathbb{N}_0$, continuous maps $\phi_m : \mathbb{R}^{p_m} \to \mathbb{R}^{r_m}$, and a continuous decoder $\psi : \mathbb{R}^{\sum_m r_m} \to \mathbb{R}^q$ such that $\psi(\phi_1(X_1), \ldots, \phi_M(X_M)) = \mathbb{E}[Y \mid X]$ almost surely. Let $D^{\mathrm{exact}}_{\mathrm{joint}}$ be the minimum r over joint continuous factorizations, and let $D^{\mathrm{exact}}_{\mathrm{sep}}$ be the minimum $\sum_m r_m$ over separate continuous factorizations; assign the value $+\infty$ when the corresponding class is empty. These quantities may be much larger than the smallest dimensions at which a risk infimum equals $\mathcal{B}([M])$.

4 Collapse of the Dimension-Only Neural Frontier

The following approximation lemma shows that one scalar neural representation per input block suffices for arbitrarily accurate L^2 approximation when the encoder and decoder complexities are unrestricted.

Lemma 1 (Blockwise scalar neural approximation)**.** *Let $k \in \mathbb{N}$. For each $i \in [k]$, let $Z_i \in \mathbb{R}^{d_i}$ be an input block with $d_i \in \mathbb{N}$, and let $Z = (Z_1, \ldots, Z_k) \in \mathbb{R}^{\sum_{i=1}^k d_i}$ denote the stacked input. Let $f : \mathbb{R}^{\sum_{i=1}^k d_i} \to \mathbb{R}^q$ be a Borel-measurable target function satisfying $\mathbb{E}\,\|f(Z)\|_2^2 < \infty$. For every $\epsilon > 0$, there are networks $\phi_i \in \mathrm{NN}_{d_i,1}$ and $\psi \in \mathrm{NN}_{k,q}$ such that*

$$\mathbb{E}\,\|f(Z) - \psi(\phi_1(Z_1), \ldots, \phi_k(Z_k))\|_2^2 < \epsilon. \tag{2}$$

The lemma does not construct a continuous injection of the support of Z into $\mathbb{R}^k$. It first approximates f by a bounded function that is constant on a finite product partition. Each local partition is replaced

by disjoint compact cores with arbitrarily large marginal probability. A continuous scalar map assigns a distinct encoding to each core, and uniform neural approximation preserves encoding separation. A bounded neural decoder recovers the product-cell value. The approximation is accurate on the joint compact core; its complement has arbitrarily small probability. Appendix A gives the full argument.

Theorem 2 (Exact dimension-only neural risk curves)**.** *For every joint law of (X, Y) satisfying the assumptions of Section 3 and every $r \in \mathbb{N}_0$,*

$$\mathcal{R}^{\mathrm{nn}}_{\mathrm{joint}}(r) = \begin{cases} \mathcal{B}(\varnothing), & r = 0, \\ \mathcal{B}([M]), & r \geq 1, \end{cases} \tag{3}$$

and

$$\mathcal{R}^{\mathrm{nn}}_{\mathrm{sep}}(r) = \min_{\substack{J \subseteq [M] \\ |J| \leq r}} \mathcal{B}(J). \tag{4}$$

The same identities hold if, for each fixed encoded representation, the neural decoder class is enlarged to all measurable decoders; under squared loss, the optimal decoder is the corresponding conditional mean.

Proof sketch. For the joint upper bound, apply Lemma 1 with one input block $Z_1 = X$ and target $f(X) = \mathbb{E}[Y \mid X]$. Orthogonality of conditional expectation gives

$$\mathbb{E}\,\|Y - g(X)\|_2^2 = \mathcal{B}([M]) + \mathbb{E}\,\|\mathbb{E}[Y \mid X] - g(X)\|_2^2 \tag{5}$$

for every square-integrable X-measurable predictor g. A scalar bottleneck therefore approaches the Bayes risk. For the separate upper bound, fix J and apply the lemma to $f(X_J) = \mathbb{E}[Y \mid X_J]$, assigning one scalar coordinate to each modality in J. Conversely, a separate representation with total dimension r can use positive dimension in at most r modalities. Its prediction is measurable with respect to the full observations of those modalities, so its risk is at least the corresponding value of $\mathcal{B}(J)$. This proves the matching lower bound. □

Remark 3 (Infimum versus minimum). Theorem 2 describes infima over finite networks, not necessarily minima. Approaching these infima may require finer partitions, narrower transition regions, or larger weights. Bayes-risk attainment by finite ReLU networks requires an exact continuous factorization, but the converse need not hold: for independent X_1, X_2 uniform on $[0, 1]$ and $Y = X_1^2$, both continuous factorization dimensions are one, yet no finite ReLU predictor is exact, since its restriction to an input line is piecewise affine. The Gaussian dimensions in Theorem 9 are also sufficient because affine factorizations attain them.

Remark 4 (What remains of the architectural constraint). The joint curve loses every positive dimension distinction because its encoder can use all modalities before producing one encoding. The separate curve retains a discrete obstruction. A zero-dimensional local encoder communicates no information, so a budget of r can activate at most r modalities. Once a modality is active, however, additional real coordinates have no value for the unrestricted neural risk infimum.

5 Lipschitz Constraints and a Quantitative Fusion Gap

All Lipschitz constants in this section use Euclidean norms. Put $Q_d = [0, 1]^d$. For a Lipschitz product budget $\Lambda \geq 1$, define

$$\mathcal{E}^{\mathrm{Lip}}_d(\Lambda) = \inf_{\mathrm{Lip}(\phi)\,\mathrm{Lip}(\psi) \leq \Lambda} \int_{Q_d} \|x - \psi(\phi(x))\|_2^2 \, dx, \tag{6}$$

where $\phi : Q_d \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}^d$ are Lipschitz. Let $\mathcal{E}^{\mathrm{nn}}_d(\Lambda)$ denote the same infimum with ϕ restricted to Q_d from a finite ReLU network and ψ a finite ReLU network. The encoder Lipschitz constant is measured on Q_d, whereas the decoder Lipschitz constant is global. The product is unchanged by $\phi \mapsto a\phi$ and $\psi(t) \mapsto \psi(t/a)$ for $a > 0$.

Theorem 5 (Sharp scalar risk under a Lipschitz constraint)**.** *For every integer $d \geq 2$, there are constants $0 < c_d \leq C_d < \infty$ such that, for all $\Lambda \geq 1$,*

$$c_d \Lambda^{-2/(d+1)} \leq \mathcal{E}^{\mathrm{Lip}}_d(\Lambda) \leq \mathcal{E}^{\mathrm{nn}}_d(\Lambda) \leq C_d \Lambda^{-2/(d+1)}. \tag{7}$$

Finite ReLU constructions satisfy the upper bound with $\mathrm{Lip}(\phi) \leq \Lambda$ *and* $\mathrm{Lip}(\psi) \leq 1$.

Proof sketch. Write D for the risk and let $[a, b]$ be the interval between the $1/4$ and $3/4$ quantiles of the encoder's output distribution. A relative isoperimetric estimate (Ambrosio et al., 2016) and the coarea formula (Fleming & Rishel, 1960) give $\mathrm{Lip}(\phi)D \geq c_d(b-a)$. A fixed fraction of the input mass lies within distance $2\sqrt{D}$ of the curve $\psi([a, b])$. A curve-neighborhood volume bound therefore gives $\mathrm{Lip}(\psi)(b-a) \geq c_d D^{-(d-1)/2}$ when D is small. Combining the two inequalities eliminates $b-a$ and gives $D^{(d+1)/2} \geq c_d/(\mathrm{Lip}(\phi)\,\mathrm{Lip}(\psi)) \geq c_d/\Lambda$, and hence $D \geq c_d\Lambda^{-2/(d+1)}$. For the upper bound, order n^d grid cells along a path through adjacent centers and make the encoder constant on cell interiors, leaving transition bands of width w. A finite piecewise-affine extension gives risk at most $C_d(n^{-2} + nw)$ and Lipschitz product at most $C_d n^{d-2}/w$. Choosing n of order $\Lambda^{1/(d+1)}$ and w of order n^{d-2}/Λ proves the claim. Appendix C handles atoms, degenerate maps, the extension, and the constants. □

For fusion and $a \in \{\mathrm{joint}, \mathrm{sep}\}$, let $\mathcal{R}_a^{\mathrm{Lip}}(r, \Lambda)$ be the neural risk infimum of Section 3, with the additional constraint $\mathrm{Lip}(\Phi)\,\mathrm{Lip}(\psi) \leq \Lambda$. Here Φ is the joint encoder or the concatenation of the separate encoders; its constant is measured on the full input support, and ψ has a global constant. The same total dimension budget is used in both architectures.

Theorem 6 (Fusion separation under a common Lipschitz budget)**.** *Let X_1, X_2 be independent and uniform on Q_3, and let $Y = (X_1 + X_2)/\sqrt{2}$. There are constants $c, C > 0$ such that, for every $\Lambda \geq 1$,*

$$c\Lambda^{-1/2} \leq \mathcal{R}_{\mathrm{sep}}^{\mathrm{Lip}}(2, \Lambda) \leq C\Lambda^{-1/2},\ \mathcal{R}_{\mathrm{joint}}^{\mathrm{Lip}}(2, \Lambda) \leq C\Lambda^{-2/3}. \tag{8}$$

Thus $\mathcal{R}_{\mathrm{joint}}^{\mathrm{Lip}}(2, \Lambda)/\mathcal{R}_{\mathrm{sep}}^{\mathrm{Lip}}(2, \Lambda) \leq C'\Lambda^{-1/6}$ for a constant $C' > 0$. The exact continuous factorization dimensions are 3 and 6, respectively.

The two exponents arise from different effective reconstruction dimensions. In the separate architecture, total latent dimension two and positive use of both modalities force the allocation $(1, 1)$. Conditioning on either modality leaves the other as a three-dimensional uniform input that must be reconstructed through a scalar representation, so Theorem 5 with $d = 3$ yields the rate $\Lambda^{-1/2}$. In the joint architecture, the encoder can first form $S = (X_1 + X_2)/2$. One latent coordinate transmits one coordinate of S exactly, while the second scalar coordinate encodes the remaining two coordinates. Their joint density is bounded, so the $d = 2$ scalar construction gives risk of order $\Lambda^{-2/3}$. Thus fusion before compression reduces the scalar reconstruction problem from effective dimension three to effective dimension two under the same total latent dimension and Lipschitz product budget. Appendix C verifies the admissible Lipschitz constants and all separate allocations. The joint exponent $2/3$ is achievable; a matching lower bound is not needed for the separation.

6 The Nondegenerate Linear Gaussian Benchmark

Assume in this section that (X, Y) is centered and jointly Gaussian and that $\Sigma_X = \mathrm{Cov}(X) \succ 0$. Let $\Sigma_{YX} = \mathrm{Cov}(Y, X)$ and $\Sigma_{YY} = \mathrm{Cov}(Y)$. Covariance subscripts specify the corresponding blocks, and $\Sigma_X^{1/2}$ is the symmetric positive-definite square root. Define the regression matrix and the whitened predictive operator

$$K = \Sigma_{YX}\Sigma_X^{-1},\ T = \Sigma_{YX}\Sigma_X^{-1/2}. \tag{9}$$

Partition $K = [K_1\ \cdots\ K_M]$ according to the modalities. For a subspace $W \leq \mathbb{R}^p$, let P_W be its Euclidean orthogonal projection. Define the whitened separate family

$$\mathcal{W}_{\mathrm{sep}}(r) = \left\{ \Sigma_X^{1/2}\Big(\bigoplus_{m=1}^{M} V_m\Big) : V_m \leq \mathbb{R}^{p_m},\ \sum_m \dim(V_m) \leq r \right\}. \tag{10}$$

The whitening map is essential because cross-modal covariance generally maps a coordinate direct sum to a non-coordinate subspace.

Theorem 7 (Gaussian linear frontiers)**.** *For $0 \leq r \leq p$,*

$$\mathcal{R}_{\mathrm{joint}}^{\mathrm{lin}}(r) = \mathrm{tr}(\Sigma_{YY}) - \sum_{j=1}^{r} \sigma_j^2(T), \tag{11}$$

$$\mathcal{R}^{\mathrm{lin}}_{\mathrm{sep}}(r) = \mathrm{tr}(\Sigma_{YY}) - \max_{W \in \mathcal{W}_{\mathrm{sep}}(r)} \|TP_W\|^2_{\mathrm{F}}\,, \tag{12}$$

$$\mathcal{B}([M]) = \mathrm{tr}(\Sigma_{YY}) - \|T\|^2_{\mathrm{F}}\,. \tag{13}$$

Consequently, $\mathcal{R}^{\mathrm{lin}}_{\mathrm{joint}}(r) - \mathcal{B}([M]) = \sum_{j>r} \sigma_j^2(T)$.

The joint identity is the classical spectral solution. The separate identity is also exact, but it is a constrained subspace problem rather than a collection of independent spectral truncations. A useful lower bound follows from conditional innovations. Let X_{-m} contain every modality except m. Since $\Sigma_X \succ 0$, the principal submatrix $\Sigma_{-m,-m}$ and its Schur complement Ω_m defined below are positive definite. Define $\Omega_m = \Sigma_{mm} - \Sigma_{m,-m}\Sigma^{-1}_{-m,-m}\Sigma_{-m,m},\ \Delta_m = \Sigma_{Y,m} - \Sigma_{Y,-m}\Sigma^{-1}_{-m,-m}\Sigma_{-m,m},\ \textit{and}\ T^\circ_m = \Delta_m \Omega_m^{-1/2}$. Here Ω_m is the covariance of the residual of X_m after linear regression on X_{-m}, and T°_m is the corresponding whitened residual cross-covariance.

Proposition 8 (Conditional-innovation lower bound)**.** *Fix local rank limits $r_m \in \{0, \dots, p_m\}$. Every separate linear encoder with $\mathrm{rank}(L_m) \le r_m$ satisfies*

$$\mathbb{E}\,\|Y - \mathbb{E}[Y \mid L_1X_1, \dots, L_MX_M]\|_2^2 - \mathcal{B}([M]) \ge \max_{m \in [M]} \sum_{j > r_m} \sigma_j^2(T^\circ_m). \tag{14}$$

The proof supplies X_{-m} to the decoder at no cost. Even with this side information, the local rank-r_m representation must lose the indicated singular-value tail. The actual separate architecture has no more information. We next turn to exact continuous factorization. In contrast to the unrestricted neural risk infima of Theorem 2, the required representation dimensions retain a rank barrier.

Theorem 9 (Exact Gaussian factorization dimensions)**.** *Under the Gaussian assumptions above, $D^{\mathrm{exact}}_{\mathrm{joint}} = \mathrm{rank}(K),\ D^{\mathrm{exact}}_{\mathrm{sep}} = \sum_{m=1}^M \mathrm{rank}(K_m) = \sum_{m=1}^M \mathrm{rank}(T^\circ_m)$. The same dimensions are obtained if all continuous maps are required to be finite ReLU networks.*

Appendix D proves the upper bounds by row-space factorization and the lower bounds by invariance of domain. The separate frontier also has a geometric gap. Under a strict singular-value gap, equality with the joint frontier requires the leading predictive subspace to belong to the whitened block family. Appendix D.1 gives the quantitative bound and its generic strictness.

7 A Model with Four Closed-Form Frontiers

Let $N \in \mathbb{N}$, with $m \in [M]$ indexing modalities and $h \in [N]$ indexing response coordinates. Fix response coefficients $a_{mh} \in \mathbb{R}$, latent variances $v_{mh} > 0$, observation-noise variances $\omega_{mh} \ge 0$, and response-noise variances $\tau_h^2 \ge 0$. Let the latent variables U_{mh}, observation noises ε_{mh}, and response noises ζ_h be mutually independent centered Gaussians, allowing degenerate noise laws, with $U_{mh} \sim \mathcal{N}(0, v_{mh}),\ \varepsilon_{mh} \sim \mathcal{N}(0, \omega_{mh}),\ \textit{and}\ X_{mh} = U_{mh} + \varepsilon_{mh}$, and $Y_h = \sum_{m=1}^M a_{mh}U_{mh} + \zeta_h,\ \zeta_h \sim \mathcal{N}(0, \tau_h^2)$.

Set $X_m = (X_{mh})_{h=1}^N$ and $Y = (Y_h)_{h=1}^N$. Then (X, Y) is centered and jointly Gaussian, and $\mathrm{Cov}(X) \succ 0$ because $v_{mh} + \omega_{mh} > 0$ for every (m, h). Define the predictive energy of pair (m, h), the aggregated coordinate energy, and the total modality energy by $e_{mh} = \frac{a_{mh}^2 v_{mh}^2}{v_{mh} + \omega_{mh}},\ \kappa_h = \sum_{m=1}^M e_{mh},\ \textit{and}\ \eta_m = \sum_{h=1}^N e_{mh}$. Thus κ_h aggregates across modalities and η_m across response coordinates. Let $\kappa_{(1)} \ge \cdots \ge \kappa_{(N)}$, let $e_{[1]} \ge \cdots \ge e_{[MN]}$ be the ordered individual energies, and let $\eta_{(1)} \ge \cdots \ge \eta_{(M)}$. Write $V_Y = \mathrm{tr}\,\mathrm{Cov}(Y)$.

Theorem 10 (Four exact risk curves)**.** *For every $r \in \mathbb{N}_0$, $\mathcal{R}^{\mathrm{lin}}_{\mathrm{joint}}(r) = V_Y - \sum_{j=1}^{\min\{r,N\}} \kappa_{(j)}$, $\mathcal{R}^{\mathrm{lin}}_{\mathrm{sep}}(r) = V_Y - \sum_{j=1}^{\min\{r,MN\}} e_{[j]}$, $\mathcal{R}^{\mathrm{nn}}_{\mathrm{joint}}(r) = V_Y - \min\{r, 1\} \sum_{m=1}^M \sum_{h=1}^N e_{mh}$, and $\mathcal{R}^{\mathrm{nn}}_{\mathrm{sep}}(r) = V_Y - \sum_{j=1}^{\min\{r,M\}} \eta_{(j)}$. The raw-observation Bayes risk is*

$$\mathcal{B}([M]) = \sum_{h=1}^N \left[\tau_h^2 + \sum_{m=1}^M \frac{a_{mh}^2 v_{mh}\omega_{mh}}{v_{mh} + \omega_{mh}}\right] = V_Y - \sum_{m,h} e_{mh}. \tag{15}$$

The aggregation differs across architectures. Joint linear compression ranks the coordinate totals κ_h, separate linear compression ranks the individual e_{mh}, and separate neural compression ranks the modality totals η_m. Joint neural compression approaches Bayes risk for every $r \geq 1$.

If every $e_{mh} > 0$, Theorem 9 gives $D_{\text{joint}}^{\text{exact}} = N$ and $D_{\text{sep}}^{\text{exact}} = MN$, although the neural risk infima equal the Bayes value at dimensions 1 and M. Exact continuous factorization and risk approximation thus require different dimensions.

8 Shifted-Lattice Realization

Proposition 15, stated and proved in Appendix G, realizes the independent model on $M = s^d$ shifted lattices of $\mathbb{T}^d$, where $s \geq 2$ and $d \geq 1$ are integers. Each modality has $N = \prod_{i=1}^{d} n_i$ sites, with $n_i \in \mathbb{N}$, unit latent variance, observation-noise variance $\nu^2 \geq 0$, and coefficients $a_{mh} = M^{-1/2}$; each response coordinate has noise variance $\tau^2 \geq 0$.

Corollary 11 (Resolution–fusion separation)**.** *Under Proposition 15, let $V_0 = N(1+\tau^2)$. For every $r \in \mathbb{N}_0$, $\mathcal{R}_{\text{joint}}^{\text{lin}}(r) = V_0 - \frac{\min\{r,N\}}{1+\nu^2}$, $\mathcal{R}_{\text{sep}}^{\text{lin}}(r) = V_0 - \frac{\min\{r,MN\}}{M(1+\nu^2)}$, $\mathcal{R}_{\text{joint}}^{\text{nn}}(r) = V_0 - \frac{N\min\{r,1\}}{1+\nu^2}$, and $\mathcal{R}_{\text{sep}}^{\text{nn}}(r) = V_0 - \frac{N\min\{r,M\}}{M(1+\nu^2)}$. The Bayes risk is $\mathcal{B}([M]) = N\left(\tau^2 + \frac{\nu^2}{1+\nu^2}\right)$. The smallest dimensions at which the neural risk infima equal the Bayes risk are 1 and $M = s^d$. In contrast, $D_{\text{joint}}^{\text{exact}} = N$, $D_{\text{sep}}^{\text{exact}} = MN = s^d N$.*

Separate encoding requires s^d times the dimension of joint encoding to approach Bayes risk. Exact continuous factorization multiplies each requirement by N, the number of coarse spatial coordinates. Linear compression retains a graded dependence on r.

Approaching the dimension-only infimum still requires increasing resources. Below the exact Gaussian factorization ranks, a fixed ReLU architecture with bounded parameters has strictly positive excess risk (Proposition 14, Appendix E). Alongside the Lipschitz bounds of Section 5, we next quantify the number of transmitted bits.

9 Finite Codebooks and the Cost of Fusion Placement

We now replace unlimited real precision by a fixed-length bit budget. For $B \in \mathbb{N}_0$, a joint encoder is any Borel map $e : \mathbb{R}^p \to [2^B]$. A separate encoder uses Borel maps $e_m : \mathbb{R}^{p_m} \to [2^{b_m}]$, where $b_m \in \mathbb{N}_0$ and $\sum_m b_m \leq B$. A decoder maps the received message or message tuple to $\mathbb{R}^q$. Let $\mathcal{R}_a^{\text{fin}}(B)$ be the infimum squared-loss risk in architecture $a \in \{\text{joint}, \text{sep}\}$ and set $\mathcal{E}_a(B) = \mathcal{R}_a^{\text{fin}}(B) - \mathcal{B}([M])$. These are deterministic, noninteractive codes for one observation, not codes across repeated samples; the decoder receives no uncharged side information. Only transmitted bits are constrained.

Theorem 12 (Matching finite-bit risk exponents)**.** *Assume the Gaussian model of Section 6. If $K \neq 0$, put*

$$d_{\text{joint}} = \operatorname{rank}(K),\ d_m = \operatorname{rank}(K_m),\ d_{\text{sep}} = \sum_{m=1}^{M} d_m.$$

For each $a \in \{\text{joint}, \text{sep}\}$ there are constants $0 < c_a \leq C_a < \infty$, depending on the fixed Gaussian law but not on B, such that for every $B \in \mathbb{N}_0$,

$$c_a\, 2^{-2B/d_a} \ \leq\ \mathcal{E}_a(B) \ \leq\ C_a\, 2^{-2B/d_a}. \tag{16}$$

Explicit finite product quantizers attain the upper bounds. The same risk infima hold when each nonconstant message is produced by a scalar finite ReLU encoder followed by a fixed finite-level quantizer, and the decoder is a finite ReLU network. Network sizes and parameters remain unrestricted. If $K = 0$, both excess risks are zero for every B.

Proof sketch. A rank-d Gaussian vector with covariance C requires mean squared error at least $d \det_+(C)^{1/d} 2^{-2b/d}$ when described by b bits, where $\det_+$ multiplies the positive eigenvalues. This gives the joint lower bound for KX. For a separate encoder, supplying X_{-m} to the decoder

leaves the independent innovation $K_m(X_m - \mathbb{E}[X_m \mid X_{-m}])$ with covariance $K_m\Omega_m K_m^\top$. Its error is at least $a_m 2^{-2b_m/d_m}$ for some $a_m > 0$ whenever $d_m > 0$. Taking the geometric mean of these bounds with weights d_m/d_{sep} gives the separate exponent. For the upper bounds, quantize the Gaussian coordinates of KX, or separately those of each $K_m X_m$, using $2^{\lfloor B/d_a \rfloor}$ levels per coordinate. Appendix H gives the quantizers, all constants, and complete proofs, including the quantized ReLU realization. □

For $K \neq 0$, writing $B_a(\epsilon) = \min\{B \in \mathbb{N}_0 : \mathcal{E}_a(B) \leq \epsilon\}$, equation (16) implies

$$B_a(\epsilon) = \frac{d_a}{2}\log_2(1/\epsilon) + O(1),\ \epsilon \downarrow 0. \tag{17}$$

Thus the ranks governing exact continuous factorization also govern the finite-bit approximation cost. In the shifted-lattice model, $d_{\text{joint}} = N$ and $d_{\text{sep}} = MN$. Separate encoding therefore requires asymptotically $M = s^d$ times as many bits to reach the same small excess risk. These are matching exponents, not sharp finite-budget constants. The constants need not remain uniform as Σ_X approaches singularity, and the result imposes no bound on network size or Lipschitz constants. The finite-level quantizer is an explicit, generally discontinuous component; it is not itself a ReLU network.

10 Experiments

All seven experiments were run on a workstation with an Intel Core i9-10900X CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU. All computations used FP64 and seeds 1000–1009; uncertainties below are sample standard deviations across seeds. Appendix I reports the protocols and complete results. The Lipschitz and finite-bit experiments evaluate prescribed constructions rather than trained networks.

Scalar approximation. In two-dimensional Gaussian regression with Bayes risk 0.2, refining the explicit scalar ReLU construction from 4×4 to 64×64 cells reduced excess risk from 0.3361 to 0.002840. The largest absolute parameter increased from 64.5 to $507{,}904.5$, illustrating approximation with growing network resources rather than exact scalar recovery.

Fusion and allocation. As Table 1 shows, in the independent model, joint and separate neural constructions approached Bayes risk at budgets $r = 1$ and $r = 3$, below the respective exact factorization dimensions 2 and 6, which also mark where the linear curves reach the Bayes value (Table 1). At total budget two, allocations $(2, 0, 0)$ and $(1, 1, 0)$ gave risks 3.4887 and 1.8629, showing the benefit of retaining another modality. In the shifted-lattice model with $M = 4$ and $N = 2$, the 64×64 constructions gave joint risk 0.773450 at $r = 1$ and separate risk 0.773462 at $r = 4$, close to Bayes risk 0.771429; the theoretical exact factorization dimensions remain 2 and 8, respectively.

Table 1: Mean evaluated risks over ten seeds in the independent model ($M = 3$, $N = 2$; Bayes risk 1.4724). Neural predictors use 64×64 scalar codes where needed and exact affine maps when dimensions permit. These are achieved risks, not neural risk infima; full results and uncertainties appear in Appendix I.2.

r	Joint linear	Separate linear	Joint neural	Separate neural
1	3.4683	3.9914	1.4791	3.4927
2	1.4724	2.7042	1.4724	1.8629
3	1.4724	2.2002	1.4724	1.4791
6	1.4724	1.4724	1.4724	1.4724

Correlated Gaussian bounds. At $r = 1$, the geometric gap bound held in all 60 instances. Across correlation levels, the mean joint–separate risk gap ranged from 4.0293 to 4.6470, and the mean geometric lower bound was approximately 76–78% of the corresponding mean gap. All $15{,}000$ feasible-encoder checks satisfied the conditional-innovation bound within 10^{-8}. Ignoring whitening produced a mean absolute separate-risk error of 1.4765 at correlation 0.98.

Bounded networks and quantized representations. With width 256, four hidden layers per network and parameter magnitudes bounded by 2, the 20,000-step scalar-bottleneck model attained excess risk 0.1892 ± 0.0303. A separately trained two-dimensional control with width 256, two hidden layers per network and parameter bound 8 attained $(2.93 \pm 0.17) \times 10^{-6}$ after 2,500 steps. This comparison does not exclude optimization difficulties specific to scalar networks. For the fixed 64×64 scalar construction without retraining, excess risk was 0.002839 without quantization and 0.325812, 0.023006, and 0.002848 at 8, 10, and 12 bits, respectively. These results concern attained risk and a particular code, not certified global minima or optimal quantized encoders.

Lipschitz-constrained scalar reconstruction. We evaluated the explicit construction of Theorem 5 on uniform cubes in dimensions two and three, using $w = 1/(8n^3)$ and the analytic Lipschitz product bound $\Lambda = 16dn^{d+1}$. Increasing n from 4 to 32 in dimension two reduced mean reconstruction risk from 0.020073 to 0.0002503; increasing n from 4 to 16 in dimension three reduced it from 0.041112 to 0.002205. Rescaling risk by $\Lambda^{2/(d+1)}$ gave values within $[2.584, 3.237]$ and $[3.910, 4.605]$, respectively. Log–log fits over the four largest budgets yielded slopes -0.6885 ± 0.0013 and -0.5424 ± 0.0015. The normalized risks varied moderately over the tested ranges, although the finite-range slopes did not exactly match the asymptotic references $-2/3$ and $-1/2$.

Fusion under a common Lipschitz budget. For independent uniform inputs on Q_3 and response $Y = (X_1 + X_2)/\sqrt{2}$, both architectures used total latent dimension two and a Lipschitz product bound at most Λ. As Λ increased from 2^{12} to 2^{22}, separate construction risk decreased from 0.278433 to 0.006331, while joint construction risk decreased from 0.019596 to 0.0001669. The mean joint-to-separate risk ratio decreased from 0.07038 to 0.02636. Over the four largest budgets, separate and joint log–log slopes were -0.5295 ± 0.0005 and -0.6713 ± 0.0009; the joint risk multiplied by $\Lambda^{2/3}$ remained between 4.335 and 4.419. These measurements illustrate the advantage of the joint construction under matched budgets. They do not estimate the ratio of architecture-level risk infima or establish optimality of the joint exponent.

Finite-bit risk decay. We evaluated the product quantizers of Appendix H on a fixed correlated Gaussian model with $\mathrm{rank}(K) = 2$ and $\mathrm{rank}(K_1) = \mathrm{rank}(K_2) = 2$. The codebooks were reconstructed at each budget, rather than obtained by quantizing a fixed neural code. For $B \in \{16, 20, 24, 28, 32\}$, slopes of log-base-two excess risk against total bits were -0.99995 ± 0.00032 for joint encoding and -0.50475 ± 0.00033 for separate encoding, close to the references -1 and $-1/2$. At $B = 32$, the respective excess risks were 9.576×10^{-9} and 3.859×10^{-4}. Low-bit nonmonotonicity of the prescribed separate quantizer is reported and discussed in Appendix I.7. These experiments concern attained construction risks; the matching lower bounds are established analytically.

11 Discussion and Conclusion

The Lipschitz product budget yields a sharp scalar reconstruction rate and a fusion separation at the same total representation dimension (Theorems 5–6). Without a fixed regularity or complexity budget, scalar joint coding approaches Bayes risk, while separate coding reduces to modality selection. Linear Gaussian encoding instead retains subspace restrictions, while exact continuous factorization obeys rank requirements. The joint regression rank and the sum of its modality-specific ranks also govern the finite-bit risk exponents (Theorem 12). These results show that representation dimension, stability, and transmitted bits impose distinct constraints.

The results concern population risk for finite-dimensional observations, not finite-sample estimation, optimization guarantees, distribution shift, or sharp network-size and parameter bounds. The matching Lipschitz rate concerns scalar reconstruction on a uniform cube; higher-dimensional bottlenecks and the optimal joint fusion rate are outside the present results. The finite-bit exponents hold for a fixed nonsingular Gaussian law, without sharp constants or uniformity over covariance families. The shifted-lattice construction is a specific multiresolution example, not a universal spatial sensing model.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, to support the retrieval and discovery of related work, and to assist in checking and revising portions of the mathematical arguments and proofs. The authors are responsible for independently verifying the mathematical arguments and references before submission. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs of the theoretical results are provided in Appendices A–H. Appendix I reports the experimental protocols, implementation details, hardware and software configurations, random seeds, evaluation procedures, and detailed numerical results. The supplementary material contains the source code used to reproduce the experiments and numerical evaluations reported in the paper.

REFERENCES

Inaki Estella Aguerri and Abdellatif Zaidi. Distributed variational representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):120–138, 2019.

Luigi Ambrosio, Jean Bourgain, Haïm Brezis, and Alessio Figalli. Bmo-type norms related to the perimeter of sets. *Communications on Pure and Applied Mathematics*, 69(6):1062–1086, 2016.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2019. doi: 10.1109/TPAMI.2018.2798607. URL https://ieeexplore.ieee.org/document/8269806/.

Joshua Batson, {C. Grace} Haaf, Yonatan Kahn, and {Daniel A.} Roberts. Topological obstructions to autoencoding. *Journal of High Energy Physics*, 2021(4), April 2021. ISSN 1126-6708. doi: 10.1007/JHEP04(2021)280. Publisher Copyright: © 2021, The Author(s).

Toby Berger, Zhen Zhang, and Harish Viswanathan. The ceo problem [multiterminal source coding]. *IEEE Transactions on Information Theory*, 42(3):887–902, 1996.

L.E.J. Brouwer. Beweis der invarianz des n-dimensionalen gebiets. *Mathematische Annalen*, 71: 305–313, 1912. URL http://eudml.org/doc/158534.

Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal stable nonlinear approximation. *Foundations of Computational Mathematics*, 22(3):607–648, June 2022. ISSN 1615-3375. doi: 10.1007/s10208-021-09494-z. URL https://doi.org/10.1007/s10208-021-09494-z.

R Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954. URL https://doi.org/10.1137/S0895479895290954.

Ky Fan. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences*, 37(11):760–766, 1951.

Jun Fang and Hongbin Li. Joint dimension assignment and compression for distributed multisensor estimation. *IEEE Signal Processing Letters*, 15:174–177, 2008.

Jun Fang and Hongbin Li. Optimal/near-optimal dimensionality reduction for distributed estimation in homogeneous and certain inhomogeneous scenarios. *Trans. Sig. Proc.*, 58(8):4339–4353, August 2010. ISSN 1053-587X. doi: 10.1109/TSP.2010.2048213. URL https://doi.org/10.1109/TSP.2010.2048213.

Wendell H Fleming and Raymond Rishel. An integral formula for total gradient variation. *Archiv der Mathematik*, 11(1):218–222, 1960.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1907330.

Michael Gastpar, Pier Luigi Dragotti, and Martin Vetterli. The distributed karhunen–loève transform. *IEEE Transactions on Information Theory*, 52(12):5177–5196, 2006. doi: 10.1109/TIT.2006.885449.

Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning, 2020. URL https://hdl.handle.net/1721.1/143836.

R.M. Gray and D.L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998. doi: 10.1109/18.720541.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(91)90009-T. URL https://www.sciencedirect.com/science/article/pii/089360809190009T.

S.H. Izen. Generalized sampling expansion on lattices. *IEEE Transactions on Signal Processing*, 53(6):1949–1963, 2005. doi: 10.1109/TSP.2005.847841.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(75)90042-1. URL https://www.sciencedirect.com/science/article/pii/0047259X75900421.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2018.02.016. URL https://www.sciencedirect.com/science/article/pii/S0168169917308803.

Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using reLU networks on compact domain. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dpDw5U04SU.

Casimir Kuratowski. Sur une généralisation de la notion d'homéomorphie. *Fundamenta Mathematicae*, 22:206–220, 1934.

Matthew Kvalheim and Eduardo Sontag. Why should autoencoders work? *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=uGVFtjvI3v.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Original contribution: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. 6(6):861–867, June 1993. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80131-5. URL https://doi.org/10.1016/S0893-6080(05)80131-5.

Yi Li, Fei Song, Changwen Zheng, and Jiangmeng Li. Supporting multimodal intermediate fusion with informatic constraint and distribution coherence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5bxmmuRhO6.

Vinith Misra, Vivek K Goyal, and Lav R. Varshney. Distributed scalar quantization for computing: High-resolution analysis and extensions. *IEEE Transactions on Information Theory*, 57(8):5298–5325, 2011. doi: 10.1109/TIT.2011.2158882.

David J Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems engineering*, 114(4):358–371, 2013.

A. Papoulis. Generalized sampling expansion. *IEEE Transactions on Circuits and Systems*, 24(11):652–654, 1977. doi: 10.1109/TCS.1977.1084284.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O-XJwyoIF-k.

Vladimir Pestov. Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437, 2013.

Guergana Petrova and Przemyslaw Wojtaszczyk. Lipschitz widths. *Constructive Approximation*, 57: 759–805, 2021. URL https://api.semanticscholar.org/CorpusID:240420180.

I.D. Schizas, G.B. Giannakis, and Zhi-Quan Luo. Distributed estimation using reduced dimensionality sensor observations. In *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, pp. 1029–1033, 2005. doi: 10.1109/ACSSC.2005.1599914.

Jiawei Shao, Yuyi Mao, and Jun Zhang. Learning task-oriented communication for edge inference: An information bottleneck approach. *IEEE Journal on Selected Areas in Communications*, 40(1): 197–211, 2022. doi: 10.1109/JSAC.2021.3126087.

Heinrich Tietze. Über funktionen, die auf einer abgeschlossenen menge stetig sind. 1915.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *ArXiv*, physics/0004057, 2000. URL https://api.semanticscholar.org/CorpusID:8936496.

Matías Vera, Leonardo Rey Vega, and Pablo Piantanida. Collaborative information bottleneck. *IEEE Transactions on Information Theory*, 65(2):787–815, 2019. doi: 10.1109/TIT.2018.2883295.

Malte von Bloh, David Lobell, and Senthold Asseng. Knowledge informed hybrid machine learning in agricultural yield prediction. *Computers and Electronics in Agriculture*, 227:109606, 2024. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2024.109606. URL https://www.sciencedirect.com/science/article/pii/S0168169924009979.

Yihong Wu and Sergio Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Transactions on Information Theory*, 56:3721–3748, 2010. URL https://api.semanticscholar.org/CorpusID:206737933.

Lin-Xia Zhang, Dun-Biao Niu, Enbin Song, Jie Zhou, Qingjiang Shi, and Yunmin Zhu. Joint optimization of dimension assignment and compression in distributed estimation fusion. *IEEE Transactions on Signal Processing*, 67:2453–2468, 2019. URL https://api.semanticscholar.org/CorpusID:96432126.

Yunmin Zhu, Enbin Song, Jie Zhou, and Zhisheng You. Optimal dimensionality reduction of sensor data in multisensor estimation fusion. *Trans. Sig. Proc.*, 53(5):1631–1639, May 2005. ISSN 1053-587X. doi: 10.1109/TSP.2005.845429. URL https://doi.org/10.1109/TSP.2005.845429.

A Proof of the Product-Coding Lemma

We prove Lemma 1. Let μ_Z denote the law of $Z = (Z_1, \ldots, Z_k)$. Since measurable rectangles generate the product Borel sigma-algebra, finite linear combinations of their indicator functions are dense in $L^2(\mu_Z)$. Hence, for every $\epsilon > 0$, there exists a bounded vector-valued simple function f_0, measurable with respect to a finite product partition, such that

$$\mathbb{E}\,\|f(Z) - f_0(Z)\|_2^2 < \epsilon/8. \tag{18}$$

More explicitly, for each $i \in [k]$, choose $\ell_i \in \mathbb{N}$ and let

$$\mathbb{R}^{d_i} = A_{i1} \,\dot\cup \cdots \dot\cup\, A_{i\ell_i} \tag{19}$$

be a finite Borel partition. There are vectors $b_{j_1 \ldots j_k} \in \mathbb{R}^q$, indexed by $j_i \in [\ell_i]$ for every $i \in [k]$, such that

$$f_0(z_1, \ldots, z_k) = b_{j_1 \ldots j_k} \text{ on } A_{1j_1} \times \cdots \times A_{kj_k}. \tag{20}$$

Let $B > 1 + \max_{j_1,\dots,j_k} \|b_{j_1\dots j_k}\|_\infty$. Fix $\delta > 0$, to be specified below. By inner regularity of the marginal law of Z_i, for each $i \in [k]$ and $j \in [\ell_i]$ there exists a compact set $C_{ij} \subseteq A_{ij}$ such that

$$\Pr\{Z_i \in A_{ij} \setminus C_{ij}\} < \frac{\delta}{k\ell_i}. \tag{21}$$

Set $C_i = \bigcup_{j=1}^{\ell_i} C_{ij}$. Since the C_{ij} are pairwise disjoint compact sets, the function $g_i : C_i \to \mathbb{R}$ defined by $g_i(z) = 3j$ for $z \in C_{ij}$ is continuous. By the Tietze extension theorem (Tietze, 1915), g_i admits a continuous extension $h_i : \mathbb{R}^{d_i} \to \mathbb{R}$. Universal approximation by ReLU networks gives $\phi_i \in \mathrm{NN}_{d_i,1}$ satisfying

$$\sup_{z_i \in C_i} |\phi_i(z_i) - h_i(z_i)| < 1/2. \tag{22}$$

Therefore $\phi_i(C_{ij}) \subset \mathcal{I}_{ij}$, where $\mathcal{I}_{ij} = [3j - 1/2, 3j + 1/2]$. These intervals are pairwise disjoint.

For each multi-index $\boldsymbol{j} = (j_1, \dots, j_k)$, let $\mathcal{Q}_{\boldsymbol{j}} = \mathcal{I}_{1j_1} \times \cdots \times \mathcal{I}_{kj_k}$. The boxes $\mathcal{Q}_{\boldsymbol{j}}$ are pairwise disjoint compact sets. Let $\mathcal{Q} = \bigcup_{\boldsymbol{j}} \mathcal{Q}_{\boldsymbol{j}}$ and define $g : \mathcal{Q} \to \mathbb{R}^q$ by

$$g(x) = b_{\boldsymbol{j}},\ x \in \mathcal{Q}_{\boldsymbol{j}}. \tag{23}$$

Since the boxes $\mathcal{Q}_{\boldsymbol{j}}$ are pairwise disjoint and compact, g is continuous on $\mathcal{Q}$ and satisfies $\|g\|_\infty < B$. Applying the Tietze extension theorem coordinatewise yields a continuous map $\chi : \mathbb{R}^k \to [-B, B]^q$ satisfying $\chi = g$ on $\mathcal{Q}$. Choose a compact box containing their union. Universal approximation yields a ReLU network $\widetilde{\psi}$ whose error from χ on that box is below a number $\gamma > 0$.

The scalar clipping function

$$c_B(t) = -B + \rho(t + B) - \rho(t - B) \tag{24}$$

is represented exactly by a ReLU network and maps $\mathbb{R}$ into $[-B, B]$. Apply it coordinatewise to $\widetilde{\psi}$ and call the resulting bounded network ψ. For sufficiently small γ,

$$\sup_{(z_1,\dots,z_k) \in C_1 \times \cdots \times C_k} \|f_0(z_1, \dots, z_k) - \psi(\phi_1(z_1), \dots, \phi_k(z_k))\|_2^2 < \epsilon/8. \tag{25}$$

Let $\mathcal{E} = \bigcap_{i=1}^k \{Z_i \in C_i\}$. By the union bound and equation (21), $\Pr(\mathcal{E}^c) < \delta$. Both f_0 and ψ are bounded, so there is a constant C_B depending only on B and q such that

$$\mathbb{E}\,\|f_0(Z) - \psi(\phi_1(Z_1), \dots, \phi_k(Z_k))\|_2^2 < \epsilon/8 + C_B\delta. \tag{26}$$

Choose δ so that $C_B\delta < \epsilon/8$. Combining equations (18) and (26) with

$$\|u + v\|_2^2 \le 2\,\|u\|_2^2 + 2\,\|v\|_2^2 \tag{27}$$

gives equation (2). This completes the proof.

B Proof of the Neural Risk Curves

For any square-integrable predictor $u(X)$, conditional expectation is the orthogonal projection of Y onto the closed subspace of X-measurable square-integrable functions. This gives equation (5). Every joint neural predictor is X-measurable, so its risk is at least $\mathcal{B}([M])$. When $r \ge 1$, apply Lemma 1 with $k = 1$, $Z_1 = X$, and $f(X) = \mathbb{E}[Y \mid X]$. Append zero output coordinates when $r > 1$. By equation (5), the approximation error in Lemma 1 is exactly the excess risk above $\mathcal{B}([M])$. Hence the resulting risks can be made arbitrarily close to $\mathcal{B}([M])$. When $r = 0$, every predictor is constant and the optimal constant is $\mathbb{E}Y$, giving $\mathcal{B}(\varnothing)$.

For the separate architecture, fix local dimensions $(r_m)_m$ and let $J = \{m : r_m > 0\}$. Then $|J| \le \sum_m r_m \le r$, and every corresponding predictor is measurable with respect to X_J. Its risk is therefore at least $\mathcal{B}(J)$. Since $J \subseteq [M]$ and $|J| \le r$, we have

$$\mathcal{B}(J) \ge \min_{\substack{J' \subseteq [M] \\ |J'| \le r}} \mathcal{B}(J'). \tag{28}$$

Thus every feasible separate neural predictor has risk at least the right-hand side of equation (4).

For the matching upper bound, fix $J \subseteq [M]$ with $|J| \le r$. If J is nonempty, apply Lemma 1 to the blocks $(X_m)_{m\in J}$ and the target $\mathbb{E}[Y \mid X_J]$. Assign one scalar coordinate to each modality in J and zero coordinates to every modality outside J. The orthogonal decomposition conditional on X_J, together with Lemma 1, shows that the resulting risks can be made arbitrarily close to $\mathcal{B}(J)$. The case $J = \varnothing$ uses the constant predictor $\mathbb{E}Y$. Minimizing over J proves equation (4).

Finally, enlarge the decoder class to all measurable functions. For any fixed encoded representation H, the conditional mean $\mathbb{E}[Y \mid H]$ minimizes squared loss among all H-measurable predictors. Enlarging the decoder class can therefore only decrease the corresponding risk infimum. The lower bounds above depend only on the information available through the representation and remain valid, while the neural constructions already produce risk sequences converging to the same bounds. Hence both identities are unchanged.

C Proofs for Lipschitz-Constrained Encoding

We prove Theorems 5 and 6. Throughout this appendix, Lipschitz constants use Euclidean norms unless another norm is specified. We write $\mathring{Q}_d = (0,1)^d$, so $Q_d = \overline{\mathring{Q}_d}$, and use $|A|$ for the Lebesgue measure of a measurable subset of $\mathbb{R}^d$. The boundary of Q_d has measure zero. Constants with a subscript d depend only on d.

C.1 A perimeter estimate and a curve-neighborhood estimate

For a measurable set $E \subset \mathring{Q}_d$ of finite perimeter, let $P(E;U) = |D\mathbf{1}_E|(U)$ denote its perimeter relative to an open set $U \subset \mathring{Q}_d$, and let $\partial^* E$ denote its reduced boundary. Thus

$$P(E;U) = \mathcal{H}^{d-1}(\partial^* E \cap U).$$

We use the following relative isoperimetric inequality on the unit cube (Ambrosio et al., 2016, Appendix):

$$P(E;\mathring{Q}_d) \ge 4|E|(1-|E|). \tag{29}$$

Lemma 13 (A weighted perimeter bound)**.** *Let $d \ge 2$. There is $\beta_d > 0$ such that, for every set $E \subset \mathring{Q}_d$ of finite perimeter with $1/4 \le |E| \le 3/4$ and every $y \in \mathbb{R}^d$,*

$$\int_{\partial^* E \cap \mathring{Q}_d} |x-y|_2^2 \, d\mathcal{H}^{d-1}(x) \ge \beta_d. \tag{30}$$

Proof. Let ω_d be the volume of the unit Euclidean ball and put $\kappa = 7/16$. Choose $\rho_d > 0$ so small that

$$\omega_d(2\rho_d)^d \le \frac{1}{8}, \; d\omega_d(2\rho_d)^{d-1} \le \frac{\kappa}{2}.$$

Such a choice is possible because $d \ge 2$. For almost every $s \in (\rho_d, 2\rho_d)$, the set $F = E \setminus \overline{B}(y,s)$ has finite perimeter, the perimeter measure of E gives no mass to $\partial B(y,s)$, and the finite-perimeter product rule gives

$$P(F;\mathring{Q}_d) \le P(E;\mathring{Q}_d \setminus \overline{B}(y,s)) + \mathcal{H}^{d-1}(\partial B(y,s) \cap \mathring{Q}_d). \tag{31}$$

Fix one such s. Since $|F| \ge 1/4 - \omega_d(2\rho_d)^d \ge 1/8$ and $|F| \le 3/4$, equation (29) implies $P(F;\mathring{Q}_d) \ge \kappa$. The spherical term in equation (31) is at most $d\omega_d s^{d-1} \le \kappa/2$. Consequently,

$$P(E;\mathring{Q}_d \setminus \overline{B}(y,s)) \ge \frac{\kappa}{2}.$$

On this part of the reduced boundary, $|x-y|_2 \ge s \ge \rho_d$. Equation (30) follows with $\beta_d = \rho_d^2\kappa/2$. The argument also applies when the ball meets the boundary of the cube, since every perimeter in equation (31) is relative to $\mathring{Q}_d$. □

We will also use the following elementary estimate. If $a \le b$, $\psi : \mathbb{R} \to \mathbb{R}^d$ is L_ψ-Lipschitz, and $\Gamma = \psi([a,b])$, then for every $\epsilon > 0$,

$$\big|\{x : \operatorname{dist}_2(x,\Gamma) \le \epsilon\}\big| \le A_d\big(L_\psi(b-a)\epsilon^{d-1} + \epsilon^d\big), \tag{32}$$

where one may take $A_d = 2^{d+1}\omega_d$. Indeed, subdividing $[a, b]$ into intervals whose images have diameter at most ϵ covers Γ by at most $L_\psi(b-a)/\epsilon + 2$ balls of radius ϵ. Balls with the same centers and radius 2ϵ cover its ϵ-neighborhood. This proves equation (32), including the case of a constant curve. In particular, the image of a bounded interval under a Lipschitz map into $\mathbb{R}^d$ has zero d-dimensional measure when $d \geq 2$.

C.2 The scalar lower bound

Fix an admissible Lipschitz pair $\phi : Q_d \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}^d$, and set

$$L_\phi = \mathrm{Lip}(\phi),\ L_\psi = \mathrm{Lip}(\psi),\ D = \int_{Q_d} |x - \psi(\phi(x))|_2^2 \, dx.$$

The integral is finite because both the input and the image of ϕ are compact. We assume $\Lambda \geq 1$ and $L_\phi L_\psi \leq \Lambda$. If either Lipschitz constant is zero, the predictor is constant and $D \geq d/12$. We may therefore suppose that both constants are positive. Moreover, $D = 0$ is impossible: it would place almost every point of Q_d in $\psi([\min_{Q_d} \phi, \max_{Q_d} \phi])$, whereas this curve has zero d-dimensional measure by equation (32).

Let $F_\phi(t) = |\{x \in Q_d : \phi(x) \leq t\}|$ and define its two quantiles

$$a = \inf\{t : F_\phi(t) \geq 1/4\},\ b = \inf\{t : F_\phi(t) \geq 3/4\}.$$

They are finite, $a \leq b$, and right continuity of F_ϕ gives

$$|\{a \leq \phi \leq b\}| \geq \frac{1}{2}. \tag{33}$$

If $a = b$, this set is an atom of mass at least $1/2$. Choose $r_d > 0$ with $\omega_d r_d^d = 1/4$. At least $1/4$ of that atom lies outside $B(\psi(a), r_d)$, and hence $D \geq r_d^2/4$.

Choose $\delta_d > 0$ smaller than both $d/12$ and $r_d^2/4$, and sufficiently small that

$$A_d(2\sqrt{\delta_d})^d \leq \frac{1}{8}. \tag{34}$$

If $D \geq \delta_d$, the desired lower bound follows after decreasing its constant, since $\Lambda \geq 1$. We therefore consider $0 < D < \delta_d$, for which $a < b$.

For every $t \in (a, b)$, the sublevel set $E_t = \{x \in \mathring{Q}_d : \phi(x) < t\}$ has measure between $1/4$ and $3/4$. For almost every such t, it has finite perimeter, and continuity of ϕ places its reduced boundary in the level set $\phi^{-1}(t)$, up to an $\mathcal{H}^{d-1}$-null set. Applying Lemma 13 with $y = \psi(t)$ and then using the weighted coarea formula for a Lipschitz real-valued function (Fleming & Rishel, 1960), we obtain

$$\begin{aligned} L_\phi D &\geq \int_{\mathring{Q}_d} |x - \psi(\phi(x))|_2^2 |\nabla\phi(x)|_2 \, dx \\ &= \int_{\mathbb{R}} \int_{\phi^{-1}(t) \cap \mathring{Q}_d} |x - \psi(t)|_2^2 \, d\mathcal{H}^{d-1}(x) \, dt \\ &\geq \beta_d(b - a). \end{aligned} \tag{35}$$

Here $|\nabla\phi|_2 \leq L_\phi$ almost everywhere. All integrands are nonnegative, so restricting the last integral to (a, b) is legitimate.

Set $\epsilon = 2\sqrt{D}$. Markov's inequality and equation (33) show that a set of measure at least $1/4$ satisfies both $a \leq \phi(x) \leq b$ and $|x - \psi(\phi(x))|_2 \leq \epsilon$. This set lies in the ϵ-neighborhood of $\psi([a, b])$. Equations (32) and (34) therefore give

$$\frac{1}{4} \leq A_d\big(L_\psi(b-a)\epsilon^{d-1} + \epsilon^d\big) \leq A_d L_\psi(b-a)\epsilon^{d-1} + \frac{1}{8}.$$

Thus

$$b - a \geq \frac{1}{2^{d+2} A_d L_\psi D^{(d-1)/2}}. \tag{36}$$

Combining equations (35) and (36) yields

$$D^{(d+1)/2} \geq \frac{\beta_d}{2^{d+2} A_d L_\phi L_\psi} \geq \frac{\beta_d}{2^{d+2} A_d \Lambda}.$$

Including the preceding constant-predictor and large-error cases, we conclude that

$$D \geq c_d \Lambda^{-2/(d+1)}, \; c_d = \min\left\{\delta_d, \left(\frac{\beta_d}{2^{d+2} A_d}\right)^{2/(d+1)}\right\} > 0. \tag{37}$$

Taking the infimum proves the lower bound for $\mathcal{E}_d^{\mathrm{Lip}}(\Lambda)$. It also applies to the more restricted finite-ReLU class.

C.3 A FINITE RELU CONSTRUCTION FOR THE SCALAR UPPER BOUND

We construct networks with the stronger individual bounds $\mathrm{Lip}(\phi) \leq \Lambda$ and $\mathrm{Lip}(\psi) \leq 1$. Fix an integer $n \geq 2$ and let $\mathcal{I}_{d,n} = \{0, \ldots, n-1\}^d$. There is an ordering

$$j(0), j(1), \ldots, j(n^d - 1)$$

of this grid in which consecutive indices differ by one in exactly one coordinate. We use the reflected ordering: in dimension one, list the indices in increasing order; in each subsequent dimension, list the previous ordering in the first slice, its reversal in the next slice, and continue alternating. If $k(j)$ denotes the position of j in this ordering, then adjacent grid indices satisfy

$$|k(i) - k(j)| \leq 2n^{d-1}. \tag{38}$$

To verify this bound inductively, indices in the same slice inherit the lower-dimensional bound. Across successive slices, their positions differ by at most twice the slice size, which is n^{d-1}. The endpoints of successive slices have identical first $d-1$ coordinates, so the resulting ordering also has the stated adjacency property. Connecting arbitrary grid indices by ℓ_1-shortest grid paths gives

$$\left|\frac{k(i)}{n} - \frac{k(j)}{n}\right| \leq 2n^{d-2}|i-j|_1. \tag{39}$$

For $j \in \mathcal{I}_{d,n}$, let $v_j = (j + (1/2)\mathbf{1})/n$ be its cell center and let $\tau_j = k(j)/n$. Put $N_{d,n} = n^d$ and $\ell = (N_{d,n} - 1)/n$. The decoder ψ_n follows the centers in the reflected ordering, with $\psi_n(\tau_j) = v_j$, and is affine between consecutive parameter values. Extend it constantly on $(-\infty, 0]$ and $[\ell, \infty)$. Consecutive centers are at distance $1/n$, equal to their parameter spacing. Hence ψ_n is globally 1-Lipschitz and has image in Q_d.

This decoder is a finite ReLU network. Explicitly, write $v_k = v_{j(k)}$ and $s_k = n(v_{k+1} - v_k)$ for $0 \leq k \leq N_{d,n} - 2$. Its hinge representation is

$$\psi_n(t) = v_0 + s_0\rho(t) + \sum_{k=1}^{N_{d,n}-2} (s_k - s_{k-1})\rho(t - k/n) - s_{N_{d,n}-2}\rho(t - \ell). \tag{40}$$

The vectors s_k have Euclidean norm one, and the final term makes the slope zero after ℓ.

Choose $0 < w \leq 1/(4n)$ and define the compact cores

$$C_j = \prod_{h=1}^{d} [j_h/n + w, (j_h + 1)/n - w].$$

Assign the constant scalar value τ_j to C_j. We check the Lipschitz condition for every pair of core points. If $x \in C_i$, $y \in C_j$, and $|i_h - j_h| = m \geq 1$, then

$$|x_h - y_h| \geq \frac{m-1}{n} + 2w \geq 2wm,$$

where the last inequality uses $2w \leq 1/n$. Consequently, for $i \neq j$,

$$|i-j|_1 \leq \frac{|x-y|_1}{2w} \leq \frac{d}{2w}|x-y|_\infty.$$

Together with equation (39), this proves that the prescribed core values are $L_{n,w}$-Lipschitz in the ℓ_∞ norm for

$$L_{n,w} = \frac{2dn^{d-2}}{w}. \tag{41}$$

Pairs in the same core have equal prescribed values.

An explicit extension to all of $\mathbb{R}^d$ is

$$F_{n,w}(x) = \min_{j \in \mathcal{I}_{d,n}} \{\tau_j + L_{n,w} \operatorname{dist}_\infty(x, C_j)\},\ \phi_{n,w}(x) = \min\{\ell, \max\{0, F_{n,w}(x)\}\}. \tag{42}$$

Every term in the minimum is $L_{n,w}$-Lipschitz in ℓ_∞, as is their minimum. Clipping preserves this bound. Since $|x-y|_\infty \le |x-y|_2$, the encoder is also globally $L_{n,w}$-Lipschitz in the Euclidean norm. For $x \in C_i$, the term indexed by i equals τ_i, and the verified core Lipschitz condition gives $\tau_j + L_{n,w} \operatorname{dist}_\infty(x, C_j) \ge \tau_i$ for every j. Thus $\phi_{n,w} = \tau_i$ throughout C_i.

The encoder in equation (42) is itself a finite ReLU network, not merely a Lipschitz function to be approximated by one. For a rectangle $C_j = \prod_h [a_{jh}, b_{jh}]$,

$$\operatorname{dist}_\infty(x, C_j) = \max\{0, a_{j1} - x_1, x_1 - b_{j1}, \ldots, a_{jd} - x_d, x_d - b_{jd}\}.$$

Finite maxima and minima of scalar network outputs are exactly implementable using

$$\max\{u, v\} = v + \rho(u - v),\ \min\{u, v\} = u - \rho(u - v).$$

These formulas, affine maps, and composition implement every operation in equation (42). The Lipschitz bound is therefore preserved exactly by the neural realization.

On a core, the reconstructed point is its cell center, so the squared error is at most $d/(4n^2)$. The union of the cores has measure $(1-2nw)^d$, and its complement has measure at most $2dnw$. Both x and $\psi_n(\phi_{n,w}(x))$ lie in Q_d, so the squared error on the complement is at most d. It follows that

$$\int_{Q_d} |x - \psi_n(\phi_{n,w}(x))|_2^2 \, dx \le \frac{d}{4n^2} + 2d^2 nw. \tag{43}$$

For the choice of integer parameters, let

$$\Lambda_d = \max\{2^{d+1}, (8d)^{(d+1)/2}\}.$$

For $\Lambda \ge \Lambda_d$, take

$$n = \lfloor \Lambda^{1/(d+1)} \rfloor,\ w = \frac{2dn^{d-2}}{\Lambda}.$$

Then $n \ge 2$, $n \ge \Lambda^{1/(d+1)}/2$, and $8dn^{d-1} \le \Lambda$, which is precisely the required inequality $w \le 1/(4n)$. Equation (41) gives $L_{n,w} = \Lambda$, while equation (43) yields

$$\int_{Q_d} |x - \psi_n(\phi_{n,w}(x))|_2^2 \, dx \le (d + 4d^3)\Lambda^{-2/(d+1)}. \tag{44}$$

For $1 \le \Lambda < \Lambda_d$, the constant reconstruction $(1/2)\mathbf{1}$ has risk $d/12$ and admits constant encoder and decoder networks. Increasing the upper constant to

$$C_d = \max\{d + 4d^3, (d/12)\Lambda_d^{2/(d+1)}\}$$

covers this remaining interval. The construction always has a globally Λ-Lipschitz finite ReLU encoder and a globally 1-Lipschitz finite ReLU decoder. In conjunction with equation (37) and the inclusion of the neural class in the Lipschitz class, it proves

$$c_d \Lambda^{-2/(d+1)} \le \mathcal{E}_d^{\mathrm{Lip}}(\Lambda) \le \mathcal{E}_d^{\mathrm{nn}}(\Lambda) \le C_d \Lambda^{-2/(d+1)}.$$

This completes the proof of Theorem 5.

C.4 Fusion under a common dimension and Lipschitz budget

We now prove Theorem 6. Let X_1 and X_2 be independent and uniform on Q_3, and let $Y = (X_1 + X_2)/\sqrt{2}$. The full-observation Bayes risk is zero. Both architectures have total latent dimension at most two and satisfy $\mathrm{Lip}(\Phi)\,\mathrm{Lip}(\Psi) \le \Lambda$, where the encoder Lipschitz constant is measured on $Q_3 \times Q_3$ and the decoder is globally Lipschitz.

For a separate predictor, if either modality receives zero latent coordinates, its independence from the other modality gives risk at least

$$\frac{1}{2}\,\mathrm{tr}(\mathrm{Cov}(X_m)) = \frac{1}{8}.$$

The same bound applies if both modalities are omitted. Otherwise the only admissible allocation is $(1,1)$, with scalar encoders ϕ_1, ϕ_2 and decoder $\Psi : \mathbb{R}^2 \to \mathbb{R}^3$. Fix $x_2 \in Q_3$ and define

$$\psi_{x_2}(t) = \sqrt{2}\Psi(t, \phi_2(x_2)) - x_2.$$

Restriction to a slice gives $\mathrm{Lip}(\phi_1|_{Q_3}) \le \mathrm{Lip}(\Phi)$, and $\mathrm{Lip}(\psi_{x_2}) \le \sqrt{2}\,\mathrm{Lip}(\Psi)$. Hence their Lipschitz product is at most $\sqrt{2}\Lambda$. The scalar lower bound in dimension three yields, for every x_2,

$$\begin{aligned}\int_{Q_3} \left| \frac{x_1 + x_2}{\sqrt{2}} - \Psi(\phi_1(x_1), \phi_2(x_2)) \right|_2^2 dx_1 &= \frac{1}{2}\int_{Q_3} |x_1 - \psi_{x_2}(\phi_1(x_1))|_2^2\, dx_1 \\ &\ge \frac{c_3}{2}(\sqrt{2}\Lambda)^{-1/2}.\end{aligned}$$

Integration over x_2 proves

$$\mathcal{R}^{\mathrm{Lip}}_{\mathrm{sep}}(2,\Lambda) \ge \min\{1/8, 2^{-5/4}c_3\}\Lambda^{-1/2}. \tag{45}$$

Since the scalar theorem also allows constant maps, no additional assumption on the local encoders is needed.

For the separate upper bound, use the dimension-three construction above independently on each modality. Denote its encoder and decoder by ϕ and ψ, with $\mathrm{Lip}(\phi) \le \Lambda$, $\mathrm{Lip}(\psi) \le 1$, and uniform reconstruction risk at most $C_3\Lambda^{-1/2}$. Define

$$\Phi(x_1, x_2) = (\phi(x_1), \phi(x_2)),\ \Psi(h_1, h_2) = \frac{\psi(h_1) + \psi(h_2)}{\sqrt{2}}.$$

The product Euclidean norm gives $\mathrm{Lip}(\Phi) \le \Lambda$. Cauchy–Schwarz gives $\mathrm{Lip}(\Psi) \le 1$, since

$$\frac{|\psi(h_1) - \psi(k_1) + \psi(h_2) - \psi(k_2)|_2}{\sqrt{2}} \le \sqrt{|h_1 - k_1|^2 + |h_2 - k_2|^2}.$$

Writing $e_m = X_m - \psi(\phi(X_m))$ and using $|e_1 + e_2|_2^2 \le 2|e_1|_2^2 + 2|e_2|_2^2$ gives

$$\mathcal{R}^{\mathrm{Lip}}_{\mathrm{sep}}(2,\Lambda) \le \frac{1}{2}\mathbb{E}|e_1 + e_2|_2^2 \le 2C_3\Lambda^{-1/2}. \tag{46}$$

All maps in this construction are finite ReLU networks.

For joint encoding, put $S = (X_1 + X_2)/2 \in Q_3$, so $Y = \sqrt{2}S$. Take the dimension-two scalar construction with encoder $\phi^{(2)} : \mathbb{R}^2 \to \mathbb{R}$ and decoder $\psi^{(2)} : \mathbb{R} \to \mathbb{R}^2$, satisfying $\mathrm{Lip}(\phi^{(2)}) \le \Lambda$, $\mathrm{Lip}(\psi^{(2)}) \le 1$, and uniform risk at most $C_2\Lambda^{-2/3}$. Define

$$\begin{aligned}\Phi_{\mathrm{joint}}(x_1, x_2) &= \big(s_1, \phi^{(2)}(s_2, s_3)\big),\ s = (x_1 + x_2)/2, \\ \Psi_{\mathrm{joint}}(h_1, h_2) &= \sqrt{2}\big(h_1, \psi^{(2)}(h_2)\big).\end{aligned}$$

The linear averaging map has Euclidean operator norm $1/\sqrt{2}$. The map $s \mapsto (s_1, \phi^{(2)}(s_2, s_3))$ is at most $\max\{1, \Lambda\} = \Lambda$-Lipschitz, and Ψ_{joint} is at most $\sqrt{2}$-Lipschitz. Their product is therefore at most Λ, with exactly two latent coordinates.

Each coordinate of S has density $4t$ on $[0, 1/2]$ and $4(1-t)$ on $[1/2, 1]$. The coordinates are independent, so the density of (S_2, S_3) is bounded by four on Q_2. The first response coordinate is recovered exactly. Consequently,

$$\begin{aligned}\mathcal{R}^{\mathrm{Lip}}_{\mathrm{joint}}(2,\Lambda) &\le 2\mathbb{E}\left|(S_2, S_3) - \psi^{(2)}(\phi^{(2)}(S_2, S_3))\right|_2^2 \\ &\le 8C_2\Lambda^{-2/3}.\end{aligned} \tag{47}$$

Affine maps, parallel evaluation, and composition preserve the finite ReLU class, so this joint construction is admissible in the stated neural architecture. Equations (45)–(47) establish the risk bounds in Theorem 6. The upper bound for joint encoding is used only to establish the separation; no matching joint lower bound is asserted.

We finally verify the exact continuous factorization dimensions stated in the theorem. If continuous encoder and decoder maps reproduce Y almost surely, their composition equals $(x_1 + x_2)/\sqrt{2}$ at every point of $Q_3 \times Q_3$. Indeed, any failure of equality, including at a boundary point, would by continuity give a relatively open set of failure with positive product Lebesgue measure. For joint encoding, fix $x_2 \in \mathring{Q}_3$. The map $x_1 \mapsto \phi(x_1, x_2)$ must be injective on $\mathring{Q}_3$: two equal codes would give equal decoded values and hence equal x_1. A continuous injection from an open subset of $\mathbb{R}^3$ into $\mathbb{R}^r$ requires $r \geq 3$. Otherwise, composing it with the coordinate inclusion $\mathbb{R}^r \hookrightarrow \mathbb{R}^3$ would contradict invariance of domain, since its image would have to be open in $\mathbb{R}^3$ while lying in a proper linear subspace (Brouwer, 1912). Thus the joint exact factorization dimension is at least three.

For separate encoding, fixing the other modality in the same way shows that each local encoder must be injective on $\mathring{Q}_3$. Therefore each local dimension is at least three, and their sum is at least six. The affine joint encoder $(x_1, x_2) \mapsto (x_1 + x_2)/\sqrt{2}$ with the identity decoder attains dimension three. Two local identity encoders followed by the affine averaging decoder attain total dimension six. These maps are finite ReLU networks, proving

$$D_{\text{joint}}^{\text{exact}} = 3, \; D_{\text{sep}}^{\text{exact}} = 6 \tag{48}$$

for both continuous and finite-ReLU exact factorizations. This completes the proof of Theorem 6.

D Proofs for the Gaussian Benchmark

Let $\widetilde{X} = \Sigma_X^{-1/2} X$, so $\widetilde{X} \sim \mathcal{N}(0, I_p)$ and $\mathbb{E}[Y \mid X] = T\widetilde{X}$. For a joint linear encoder L, define

$$W_L = \operatorname{Ran}(\Sigma_X^{1/2} L^\top). \tag{49}$$

Thus W_L is the row space of the encoder expressed in the whitened input coordinates. Gaussian conditioning gives

$$\mathbb{E}[\widetilde{X} \mid L\Sigma_X^{1/2}\widetilde{X}] = P_{W_L}\widetilde{X}. \tag{50}$$

Therefore

$$\mathbb{E}\,\|Y - \mathbb{E}[Y \mid LX]\|_2^2 = \operatorname{tr}(\Sigma_{YY}) - \mathbb{E}\left\|TP_{W_L}\widetilde{X}\right\|_2^2 \tag{51}$$

$$= \operatorname{tr}(\Sigma_{YY}) - \|TP_{W_L}\|_{\mathrm{F}}^2 . \tag{52}$$

Every subspace of dimension at most r is realized by a joint encoder. The Ky Fan maximum principle (Fan, 1951) gives

$$\max_{\dim(W) \leq r} \|TP_W\|_{\mathrm{F}}^2 = \sum_{j=1}^{r} \sigma_j^2(T), \tag{53}$$

which proves equation (11). For a block-diagonal encoder, $\operatorname{Ran}(L^\top) = \oplus_m \operatorname{Ran}(L_m^\top)$, so W_L ranges exactly over $\mathcal{W}_{\text{sep}}(r)$. Compactness of the finite union of fixed-allocation Grassmannian images (Edelman et al., 1998) ensures that the maximum is attained. This proves equation (12). Taking $L = I_p$ proves equation (13).

For Proposition 8, fix m and define the Gaussian residuals

$$X_m^\circ = X_m - \Sigma_{m,-m}\Sigma_{-m,-m}^{-1} X_{-m}, \tag{54}$$

$$Y_m^\circ = Y - \Sigma_{Y,-m}\Sigma_{-m,-m}^{-1} X_{-m}. \tag{55}$$

Then X_m° is independent of X_{-m}, $\operatorname{Cov}(X_m^\circ) = \Omega_m$, and $\operatorname{Cov}(Y_m^\circ, X_m^\circ) = \Delta_m$. Once X_{-m} is known, $L_m X_m$ and $L_m X_m^\circ$ generate the same additional information. Applying the joint spectral formula to the residual pair gives

$$\mathbb{E}\,\|Y - \mathbb{E}[Y \mid X_{-m}, L_m X_m]\|_2^2 - \mathcal{B}([M]) \geq \sum_{j > r_m} \sigma_j^2(T_m^\circ). \tag{56}$$

The actual separate representation is measurable with respect to $(X_{-m}, L_m X_m)$, so it cannot have smaller risk. Taking the maximum over m proves equation (14).

We next prove Theorem 9. Since X has full support and both sides of an exact continuous factorization are continuous, almost-sure equality implies equality on all of $\mathbb{R}^p$. Let a linear map $H : \mathbb{R}^u \to \mathbb{R}^v$ have rank ℓ and suppose $H = \psi \circ \phi$ for continuous maps through $\mathbb{R}^r$. Choose an ℓ-dimensional subspace S_H on which H is injective and identify S_H linearly with $\mathbb{R}^\ell$. Then $\phi|_{S_H}$ is injective. If $r < \ell$, compose $\phi|_{S_H}$ with the standard linear embedding of $\mathbb{R}^r$ into $\mathbb{R}^\ell$. This gives a continuous injection from $\mathbb{R}^\ell$ into $\mathbb{R}^\ell$ whose image lies in a proper linear subspace. Invariance of domain makes the image open, which is impossible. Hence $r \geq \ell$.

Apply this argument to $H = K$ for the joint architecture. For the separate architecture, set every input block except X_m to zero. The map $x_m \mapsto K_m x_m$ factors through the local encoder, so $r_m \geq \operatorname{rank}(K_m)$. Row-space encoders attain all bounds with affine decoders. The Frisch–Waugh–Lovell identity (Frisch & Waugh, 1933) gives $\Delta_m = K_m \Omega_m$, and therefore $T_m^\circ = K_m \Omega_m^{1/2}$. Since $\Omega_m \succ 0$, the two matrices have the same rank.

D.1 Geometric gap and generic strictness

The objective $W \mapsto \|T P_W\|_{\mathrm{F}}^2$ is nondecreasing under subspace inclusion. Since $r \leq p$, every feasible collection $(V_m)_m$ with $\sum_m \dim(V_m) < r$ can be enlarged within the modality blocks to a feasible collection $(V'_m)_m$ with $V_m \subseteq V'_m$ and $\sum_m \dim(V'_m) = r$. Hence the maximum in equation (12) is unchanged if $\sum_m \dim(V_m) \leq r$ is replaced by $\sum_m \dim(V_m) = r$.

The linear architecture gap also has a geometric form. Fix an integer r with $1 \leq r < p$. Let E_r be the leading r-dimensional right singular subspace of T, assume $\sigma_r(T) > \sigma_{r+1}(T)$, and let $\mathcal{W}_{\mathrm{sep}}^{=r}$ denote equation (10) with total dimension exactly r. Set $\delta_{\mathrm{sep}}^2(r) = \inf_{W \in \mathcal{W}_{\mathrm{sep}}^{=r}} \left\| P_{E_r^\perp} P_W \right\|_{\mathrm{F}}^2$. Then $\mathcal{R}_{\mathrm{sep}}^{\mathrm{lin}}(r) - \mathcal{R}_{\mathrm{joint}}^{\mathrm{lin}}(r) \geq \left[\sigma_r^2(T) - \sigma_{r+1}^2(T)\right] \delta_{\mathrm{sep}}^2(r)$. For a fixed allocation $(r_m)_m$, the feasible family is an embedded copy of $\prod_m \mathrm{Gr}(r_m, p_m)$ with dimension $\sum_m r_m(p_m - r_m)$, strictly below $r(p-r) = \dim \mathrm{Gr}(r,p)$. Since there are finitely many allocations, $\mathcal{W}_{\mathrm{sep}}^{=r}$ has zero Riemannian volume and empty interior in $\mathrm{Gr}(r,p)$. Under the singular-value gap, equality of the two linear frontiers can occur only if $E_r \in \mathcal{W}_{\mathrm{sep}}^{=r}$.

Finally, let $\lambda_j = \sigma_j^2(T)$ and assume $\lambda_r > \lambda_{r+1}$. For every rank-$r$ subspace W, expansion in an eigenbasis of $T^\top T$ gives

$$\sum_{j=1}^{r} \lambda_j - \operatorname{tr}(P_W T^\top T) \geq (\lambda_r - \lambda_{r+1}) \left\| P_{E_r^\perp} P_W \right\|_{\mathrm{F}}^2 . \tag{57}$$

Taking the infimum over $\mathcal{W}_{\mathrm{sep}}^{=r}$ proves equation $\mathcal{R}_{\mathrm{sep}}^{\mathrm{lin}}(r) - \mathcal{R}_{\mathrm{joint}}^{\mathrm{lin}}(r) \geq \left[\sigma_r^2(T) - \sigma_{r+1}^2(T)\right] \delta_{\mathrm{sep}}^2(r)$. For a fixed allocation, the map

$$(V_m)_{m=1}^M \longmapsto \Sigma_X^{1/2}\Big(\bigoplus_{m=1}^{M} V_m\Big) \tag{58}$$

is a smooth embedding of $\prod_m \mathrm{Gr}(r_m, p_m)$ into $\mathrm{Gr}(r,p)$. Its dimension is $\sum_m r_m(p_m - r_m)$. The deficit from $r(p-r)$ equals

$$r(p-r) - \sum_m r_m(p_m - r_m) = \sum_{m \neq n} r_m(p_n - r_n) > 0 \tag{59}$$

for $1 \leq r < p$. Thus each fixed-allocation image, and hence their finite union, has zero Riemannian volume and empty interior.

E Nonattainment Under Fixed Architecture and Bounded Parameters

The gap between a risk infimum and exact factorization has a concrete consequence for neural-network complexity. A fixed ReLU architecture has finitely many parameters. Restricting those

parameters to a closed bounded box produces a compact model class, and compactness prevents an approximating sequence from developing arbitrarily sharp finite-partition encodings within the fixed bounded parameter class.

Proposition 14 (No bounded fixed architecture attains the collapsed frontier)**.** *Assume the Gaussian model of Section 6. Fix a joint encoder architecture with bottleneck dimension r, a decoder architecture, and a bound $0 \le C < \infty$ on the absolute value of every parameter. If $r < \operatorname{rank}(K)$, then the minimum risk over this bounded parameter set is strictly larger than $\mathcal{B}([M])$.*

For the separate architecture, fix local bottleneck dimensions $(r_m)_m$, fixed local encoder architectures, a fixed decoder architecture, and the same type of parameter bound. If $r_m < \operatorname{rank}(K_m)$ for at least one modality, then the minimum risk is again strictly larger than $\mathcal{B}([M])$.

The proof below establishes attainment by compactness and excludes Bayes-risk equality using Theorem 9. Whenever the neural risk infimum equals the Bayes risk below the exact dimension, any sequence approaching that infimum must leave every fixed architecture with uniformly bounded parameters. In the shifted-lattice model with $N > 1$, this applies at dimensions 1 and M, where the joint and separate neural risk infima equal the Bayes value: risk sequences can converge to that value, but every exact representation still requires N joint coordinates or MN local coordinates.

Consider a fixed joint encoder–decoder architecture and collect all of its parameters in a vector $\vartheta \in \mathbb{R}^P$. The box $[-C, C]^P$ is compact. For every fixed input x, the network output $\widehat{y}_\vartheta(x)$ is continuous in ϑ. Because the architecture is fixed and every affine parameter is bounded by C, induction over the layers gives a constant $A_C < \infty$ such that

$$\|\widehat{y}_\vartheta(x)\|_2 \le A_C(1 + \|x\|_2) \tag{60}$$

for all x and all $\vartheta \in [-C, C]^P$. Since X and Y are square-integrable, equation (60) provides an integrable dominating function for the squared loss. Dominated convergence shows that

$$\vartheta \longmapsto \mathbb{E}\,\|Y - \widehat{y}_\vartheta(X)\|_2^2 \tag{61}$$

is continuous. It therefore attains its minimum on the parameter box.

If this minimum equaled $\mathcal{B}([M])$, equation (5) would imply $\widehat{y}_\vartheta(X) = KX$ almost surely for a minimizing parameter vector. The encoder and decoder of that network would constitute an exact ReLU factorization through $\mathbb{R}^r$, contradicting $r < \operatorname{rank}(K)$ and Theorem 9. The separate case is identical after collecting all local and decoder parameters into one finite vector. Equality with KX would give an exact separate ReLU factorization, which is impossible when some $r_m < \operatorname{rank}(K_m)$.

F Proof of the Four Closed-Form Frontiers

For each (m, h),

$$\operatorname{Var}(X_{mh}) = v_{mh} + \omega_{mh},\ \operatorname{Cov}(Y_h, X_{mh}) = a_{mh} v_{mh}. \tag{62}$$

After whitening, row h of the predictive operator has entry $a_{mh} v_{mh} / \sqrt{v_{mh} + \omega_{mh}}$ in coordinate (m, h) and zero elsewhere. Distinct rows have disjoint supports. Hence

$$TT^\top = \operatorname{diag}(\kappa_1, \ldots, \kappa_N). \tag{63}$$

The joint linear formula follows from Theorem 7. For modality m, its predictive block is diagonal across h with squared singular values $e_{m1}, \ldots, e_{mN}$. The observation covariance is diagonal, so whitening preserves the modality blocks. Optimizing all local allocations is equivalent to retaining the r largest values from the full collection (e_{mh}), proving equation $\mathcal{R}^{\mathrm{lin}}_{\mathrm{sep}}(r) = V_Y - \sum_{j=1}^{\min\{r, MN\}} e_{[j]}$.

Gaussian conditioning gives

$$\mathbb{E}[U_{mh} \mid X_{mh}] = \frac{v_{mh}}{v_{mh} + \omega_{mh}} X_{mh} \tag{64}$$

and

$$\operatorname{Var}(a_{mh} U_{mh} \mid X_{mh}) = \frac{a_{mh}^2 v_{mh} \omega_{mh}}{v_{mh} + \omega_{mh}}. \tag{65}$$

Summing over (m, h) and adding the response-noise variances proves equation (15).

For $J \subseteq [M]$, conditioning on the full observations of the modalities in J retains exactly the energies e_{mh} with $m \in J$. Therefore

$$\mathcal{B}(J) = V_Y - \sum_{m \in J} \eta_m. \tag{66}$$

The neural formulas now follow from Theorem 2. The joint risk infimum equals the Bayes risk for $r \geq 1$, while the separate risk infimum is determined by the $\min\{r, M\}$ largest modality totals.

G Construction of the Shifted-Lattice Field

We now give a spatial realization of the independent Gaussian model in Section 7 on the d-dimensional torus $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. Let $s \in \mathbb{N}$ with $s \geq 2$, let $d \in \mathbb{N}$, and set $M = s^d$. Let $n = (n_1, \ldots, n_d) \in \mathbb{N}^d$ and $N = \prod_i n_i$. Define the coarse index set $\mathcal{J}_n = \prod_i \{0, \ldots, n_i - 1\}$ and the offset set $\mathcal{J}_s = \{0, \ldots, s-1\}^d$. Here $c \in \mathcal{J}_s$ indexes the lattice offset, and hence the modality, while $u \in \mathcal{J}_n$ indexes the spatial location within each coarse lattice. For $c \in \mathcal{J}_s$, modality c samples the periodic field at

$$x_{u,c} = \left(\frac{su_1 + c_1}{sn_1}, \ldots, \frac{su_d + c_d}{sn_d} \right), \ u \in \mathcal{J}_n. \tag{67}$$

The M shifted coarse lattices partition the fine grid with side lengths $(sn_1, \ldots, sn_d)$.

Proposition 15 (Finite periodic field realization)**.** *There is an MN-dimensional real trigonometric space on $\mathbb{T}^d$ and a centered Gaussian field F in that space such that the fine-grid values*

$$U_{c,u} = F(x_{u,c}), \ c \in \mathcal{J}_s, \ u \in \mathcal{J}_n, \tag{68}$$

are independent $\mathcal{N}(0,1)$ variables. If the observations and response are

$$X_{c,u} = U_{c,u} + \varepsilon_{c,u}, \ Y_u = M^{-1/2} \sum_{c \in \mathcal{J}_s} U_{c,u} + \zeta_u, \tag{69}$$

where $\nu^2 \geq 0$ is the observation-noise variance and $\tau^2 \geq 0$ is the response-noise variance, and the family $\{\varepsilon_{c,u} : c \in \mathcal{J}_s, \, u \in \mathcal{J}_n\} \cup \{\zeta_u : u \in \mathcal{J}_n\}$ is mutually independent and independent of F, with $\varepsilon_{c,u} \sim \mathcal{N}(0, \nu^2)$ and $\zeta_u \sim \mathcal{N}(0, \tau^2)$, then the model is the specialization of Section 7 with $v_{mh} = 1$, $\omega_{mh} = \nu^2$, $a_{mh} = M^{-1/2}$, and $\tau_h^2 = \tau^2$ after enumerating c by m and u by h.

A real Fourier interpolation basis on the fine periodic grid has a scaled orthogonal evaluation matrix. Assigning isotropic Gaussian coefficients with variance $(MN)^{-1}$ therefore makes the full fine-grid value vector standard Gaussian. The shifted lattices merely partition that vector. The following construction proves the proposition.

Let

$$\mathcal{G}_{sn} = \prod_{i=1}^{d} \{0, \ldots, sn_i - 1\} \tag{70}$$

be the finite periodic grid, whose cardinality is MN. The normalized complex characters of the finite abelian group $\prod_i \mathbb{Z}/(sn_i\mathbb{Z})$ form a unitary Fourier matrix. For each pair of nonreal conjugate grid characters, choose an integer frequency representative $\xi \in \mathbb{Z}^d$ and use $\sqrt{2}\cos(2\pi\xi^\top x)$ and $\sqrt{2}\sin(2\pi\xi^\top x)$. Extend each self-conjugate grid character by the corresponding cosine, using the constant function 1 for the trivial character. This gives real trigonometric functions $\varphi_1, \ldots, \varphi_{MN}$ whose fine-grid evaluation matrix Φ_{sn} satisfies

$$\Phi_{sn}[j,t] = \varphi_t \left(\frac{j_1}{sn_1}, \ldots, \frac{j_d}{sn_d} \right), \ \Phi_{sn} = \sqrt{MN}\, Q, \tag{71}$$

where $Q \in \mathbb{R}^{MN \times MN}$ is orthogonal.

Let $\theta = (\theta_t)_{t=1}^{MN} \sim \mathcal{N}(0, (MN)^{-1} I_{MN})$ and define

$$F(x) = \sum_{t=1}^{MN} \theta_t \varphi_t(x), \tag{72}$$

where θ is the Gaussian coefficient vector of F in the real trigonometric basis $(\varphi_t)_{t=1}^{MN}$. The vector of all fine-grid values is $\Phi_{sn}\theta$. Its covariance is

$$\Phi_{sn}\frac{1}{MN}I_{MN}\Phi_{sn}^{\top} = I_{MN}. \tag{73}$$

Because it is jointly Gaussian, its coordinates are independent standard normal variables.

Every fine-grid index $j \in \mathcal{G}_{sn}$ has a unique componentwise decomposition $j = su + c$ with $u \in \mathcal{J}_n$ and $c \in \mathcal{J}_s$. Thus the points in equation (67) enumerate the entire fine grid exactly once, and the variables in equation (68) are independent $\mathcal{N}(0,1)$. Adding the independent observation and response noises gives equation (69) and identifies the parameters in Proposition 15.

For this specialization,

$$e_{mh} = \frac{1}{M(1+\nu^2)},\ \kappa_h = \frac{1}{1+\nu^2},\ \eta_m = \frac{N}{M(1+\nu^2)},\ V_Y = N(1+\tau^2). \tag{74}$$

Substitution into Theorem 10 proves equations in Corollary 11. Equation $\mathcal{B}([M]) = N\left(\tau^2 + \frac{\nu^2}{1+\nu^2}\right)$ follows from equation (15). The joint regression matrix has N nonzero rows with disjoint supports, so it has rank N. Each modality block is a nonzero scalar multiple of I_N and has rank N. Theorem 9 therefore gives equation $D_{\text{joint}}^{\text{exact}} = N,\ D_{\text{sep}}^{\text{exact}} = MN = s^d N$.

H Finite-Codebook Bounds and Quantized Neural Realizations

We prove Theorem 12. All coding alphabets are finite, all encoder maps are Borel, and all budgets are integers measured in bits. For a nonzero positive semidefinite matrix C, let $\det_+(C)$ denote the product of its positive eigenvalues. We use H for discrete entropy, h for differential entropy, and I for mutual information, with natural logarithms. Differential entropies are computed on the linear support of the relevant Gaussian vector.

H.1 A Gaussian message lower bound

Let Z be a centered Gaussian vector of rank $d > 0$ with covariance C, and let W be independent of Z. A message H may depend on Z, W, and other random variables, but takes at most 2^b values. For any measurable reconstruction $\widehat{Z}(H, W)$ with mean squared error D, we claim that

$$D \geq d \det_+(C)^{1/d} 2^{-2b/d}. \tag{75}$$

Projecting the reconstruction onto $\mathrm{Ran}(C)$ cannot increase its error, so identify that subspace with $\mathbb{R}^d$. The case $D = \infty$ is immediate. A zero error would imply that, conditional on W, the nondegenerate Gaussian Z lies in a set of at most 2^b reconstruction points, which has probability zero; hence $D > 0$. Entropy maximization at a fixed second moment, conditional translation invariance, and Jensen's inequality give

$$\begin{aligned} b\log 2 &\geq \mathrm{H}(H \mid W) \geq I(Z; H \mid W) \\ &= h(Z) - h(Z \mid H, W) \\ &\geq \frac{1}{2}\log\big((2\pi e)^d \det_+(C)\big) - \frac{d}{2}\log\left(\frac{2\pi e D}{d}\right). \end{aligned} \tag{76}$$

The conditional-entropy upper bound follows by comparing each conditional error law with the centered Gaussian of covariance $(D/d)I_d$: nonnegativity of relative entropy gives an upper bound by its cross entropy, and averaging its quadratic term yields D. Finite-valued H implies $I(Z; H \mid W) \leq b\log 2$, so the conditional entropy identity has no indeterminate subtraction. Rearranging proves equation (75). This is the usual Gaussian entropy argument underlying squared-error quantization bounds (Gray & Neuhoff, 1998); it does not require the message to be a function of Z alone.

H.2 Joint and separate lower bounds

By Gaussian conditioning and equation (5), the excess risk of every finite-message predictor $\widehat{Y}$ equals $\mathbb{E}\left\|KX - \widehat{Y}\right\|_2^2$. Put

$$C_* = K\Sigma_X K^{\top} = TT^{\top},\ \mathcal{A} = \{m : d_m > 0\},\ t = |\mathcal{A}|,\ G_m = K_m\Omega_m K_m^{\top}\ (m \in \mathcal{A}).$$

Positive definiteness of Σ_X and Ω_m gives $\operatorname{rank}(C_*) = d_{\text{joint}}$ and $\operatorname{rank}(G_m) = d_m$. Applying equation (75) with $Z = KX$ and no side information yields

$$\mathcal{E}_{\text{joint}}(B) \geq c_{\text{joint}} 2^{-2B/d_{\text{joint}}},\ c_{\text{joint}} = d_{\text{joint}} \det_+(C_*)^{1/d_{\text{joint}}} > 0. \tag{77}$$

Now fix any separate code with local budgets $(b_m)_m$ and excess risk D. Fix $m \in \mathcal{A}$ and supply $W = X_{-m}$ to the decoder. Every other local message is determined by W. Moreover,

$$KX = \mathbb{E}[KX \mid W] + Z_m,\ Z_m = K_m\big(X_m - \mathbb{E}[X_m \mid W]\big),$$

where Z_m is independent of W and has covariance G_m. Subtracting the known conditional mean from the original prediction gives a reconstruction of Z_m from (H_m, W) with error D, where $H_m = e_m(X_m)$. Thus

$$D \geq a_m 2^{-2b_m/d_m},\ a_m = d_m \det_+(G_m)^{1/d_m} > 0.$$

No independence among the different Z_m is needed. Since $\sum_{m\in\mathcal{A}} d_m/d_{\text{sep}} = 1$, their weighted geometric mean gives

$$\begin{aligned} D &\geq \prod_{m\in\mathcal{A}} \left(a_m 2^{-2b_m/d_m}\right)^{d_m/d_{\text{sep}}} \\ &= c_{\text{sep}} 2^{-2\sum_{m\in\mathcal{A}} b_m/d_{\text{sep}}} \geq c_{\text{sep}} 2^{-2B/d_{\text{sep}}},\ c_{\text{sep}} = \prod_{m\in\mathcal{A}} a_m^{d_m/d_{\text{sep}}} > 0. \end{aligned} \tag{78}$$

Taking the infimum over all codes proves the separate lower bound, including codes that spend bits on blocks with $K_m = 0$. The argument does not require independence among the modality-specific innovations and therefore applies directly to correlated modalities.

H.3 An explicit scalar quantizer with quadratic decay

Let $G \sim \mathcal{N}(0,1)$, and set $A_0 = 4\mathbb{E}(1+|G|)^4 = 40 + 48\sqrt{2/\pi}$. Define $q_1(x) = 0$. For even $n = 2k \geq 2$ and $x \geq 0$, define

$$u = \frac{x}{1+x},\ j = \lfloor ku \rfloor \in \{0, \dots, k-1\},\ v = \frac{j+1/2}{k},\ q_n(x) = \frac{v}{1-v}. \tag{79}$$

For $x < 0$ use $q_n(x) = -q_n(-x)$. This quantizer has exactly n levels; its largest positive level is $n-1$. We show that

$$|x - q_n(x)| \leq \frac{2}{n}(1+|x|)^2,\ \mathbb{E}|G - q_n(G)|^2 \leq A_0 n^{-2}. \tag{80}$$

By symmetry it suffices to consider $x \geq 0$. Write $a = 1-u = (1+x)^{-1}$ and $b = 1-v$. If $a \geq 1/k$, then $|u-v| \leq 1/(2k)$ and $b \geq a/2$. Consequently,

$$|x - q_n(x)| = \frac{|u-v|}{ab} \leq \frac{1}{ka^2} = \frac{(1+x)^2}{k}.$$

If $a < 1/k$, then $j = k-1$ and $q_n(x) = 2k-1$. With $s = 1+x > k$, the inequality $|s-2k| \leq s^2/k$ gives the same bound. Squaring and integrating proves equation (80); for $n = 1$ the mean squared error is $1 \leq A_0$. This construction controls the entire Gaussian tail and therefore incurs no additional logarithmic factor.

For any centered rank-d Gaussian vector Z with positive covariance eigenvalues $\lambda_1, \dots, \lambda_d$ and orthonormal eigenvectors $v_1, \dots, v_d$, write $Z = \sum_{i=1}^d \sqrt{\lambda_i} G_i v_i$. Quantizing each G_i with q_n produces a Borel product code with at most n^d messages and reconstruction $\widehat{Z}$ satisfying

$$\mathbb{E}\left\|Z - \widehat{Z}\right\|_2^2 = \sum_{i=1}^d \lambda_i \mathbb{E}|G_i - q_n(G_i)|^2 \leq A_0 \operatorname{tr}(C) n^{-2}. \tag{81}$$

Only orthogonality and the standard Gaussian marginal laws are used in this calculation.

H.4 MATCHING UPPER BOUNDS AND BIT COMPLEXITY

For joint encoding apply equation (81) to KX with $n = 2^{\lfloor B/d_{\text{joint}} \rfloor}$. The code has at most 2^B messages. Since $\lfloor z \rfloor > z - 1$,

$$\mathcal{E}_{\text{joint}}(B) \le C_{\text{joint}} 2^{-2B/d_{\text{joint}}},\ C_{\text{joint}} = 4A_0 \operatorname{tr}(C_*). \tag{82}$$

For separate encoding put $C_m = K_m \Sigma_{mm} K_m^\top$ for $m \in \mathcal{A}$, and choose $n = 2^{\lfloor B/d_{\text{sep}} \rfloor}$. Modality m quantizes the d_m Gaussian coordinates of $K_m X_m$, using $b_m = d_m \lfloor B/d_{\text{sep}} \rfloor$ bits; blocks outside $\mathcal{A}$ use zero bits. The total is at most B. The decoder adds the reconstructed local vectors. Writing their errors as E_m, Cauchy–Schwarz gives $\|\sum_m E_m\|_2^2 \le t \sum_m \|E_m\|_2^2$, irrespective of cross-modal dependence. Hence

$$\mathcal{E}_{\text{sep}}(B) \le C_{\text{sep}} 2^{-2B/d_{\text{sep}}},\ C_{\text{sep}} = 4A_0 t \sum_{m \in \mathcal{A}} \operatorname{tr}(C_m). \tag{83}$$

Together with equations (77) and (78), this proves equation (16). The exhibited upper constants exceed the lower constants, as also follows by applying both bounds at $B = 0$. If $K = 0$, the constant zero predictor is Bayes optimal and proves the stated degenerate case.

For $K \neq 0$ and sufficiently small $\epsilon > 0$, equation (16) gives

$$\frac{d_a}{2} \log_2 \frac{c_a}{\epsilon} \le B_a(\epsilon) \le \left\lceil \frac{d_a}{2} \log_2 \frac{C_a}{\epsilon} \right\rceil. \tag{84}$$

The upper bound also shows that the defining set for $B_a(\epsilon)$ is nonempty. The two bounds imply equation (17), and

$$\lim_{\epsilon \downarrow 0} \frac{B_{\text{sep}}(\epsilon)}{B_{\text{joint}}(\epsilon)} = \frac{d_{\text{sep}}}{d_{\text{joint}}}.$$

The shifted-lattice assertion follows from $d_{\text{joint}} = N$ and $d_{\text{sep}} = MN$. These comparisons keep the observation law fixed. Neither the entropy constants based on conditional innovations nor the resulting $O(1)$ terms are asserted to be uniform over nearly singular covariance families.

H.5 SCALAR RELU ENCODERS WITH A FINITE-LEVEL QUANTIZER

We finally show that these rates are determined by the finite codebook constraint rather than by the use of arbitrary Borel encoder partitions. For $L \in \mathbb{N}$, fix

$$Q_L(t) = \min\{L, \max\{1, \lfloor t + 1/2 \rfloor\}\},\ t \in \mathbb{R}.$$

This is a nearest-integer quantizer with clipping and a fixed tie convention. Consider any finite blockwise Borel code with alphabets $[L_i]$, input blocks Z_i, messages $e_i(Z_i)$, and decoder table $(c_{j_1 \dots j_k})$ in $\mathbb{R}^q$. As in Appendix A, select disjoint compact cores inside the local message cells with joint failure probability tending to zero. On the core of message j, prescribe the scalar value j. Tietze extension and uniform ReLU approximation give a finite network ϕ_i with error less than $1/4$ on all its cores. Therefore $Q_{L_i}(\phi_i(Z_i)) = e_i(Z_i)$ on the corresponding cores. Empty cores impose no condition; if $L_i = 1$, a constant map suffices.

The decoder table has an exact finite ReLU realization on all integer message tuples. Indeed, with $|t| = \rho(t) + \rho(-t)$, the functions

$$T_{\boldsymbol{j}}(z) = \rho\left(1 - \sum_{i=1}^{k} |z_i - j_i|\right),\ \psi(z) = \sum_{\boldsymbol{j}} c_{\boldsymbol{j}} T_{\boldsymbol{j}}(z)$$

form a finite ReLU network satisfying $\psi(\boldsymbol{j}) = c_{\boldsymbol{j}}$ for each valid tuple. Both the original predictor g and its quantized neural approximation $\widetilde{g}$ take values in this same finite table. If $R = \max_{\boldsymbol{j}} \|c_{\boldsymbol{j}}\|_2$ and $\mathcal{F}$ is the event of any core failure, then

$$\mathbb{E}\,\|g - \widetilde{g}\|_2^2 \le 4R^2 \Pr(\mathcal{F}) \longrightarrow 0.$$

For $Y \in L^2$, the difference of their risks is at most

$$\|g - \widetilde{g}\|_{L^2} \left(2\,\|Y\|_{L^2} + \|g\|_{L^2} + \|\widetilde{g}\|_{L^2}\right),$$

Table 2: Experiment 1. Excess risk above the Bayes value 0.2 for the explicit scalar ReLU construction, reported as mean ± sample standard deviation over ten evaluation seeds. The prescribed coefficients are not learned; parameter counts use the stored architecture and exclude structurally absent entries of diagonal affine layers.

G	Excess risk	Stored parameters	Largest absolute parameter
4	0.336059 ± 0.002899	63	64.5
8	0.111095 ± 0.000664	135	768.5
16	0.033960 ± 0.000265	279	7,168.5
32	0.009898 ± 0.000075	567	61,440.5
64	0.002840 ± 0.000016	1,143	507,904.5

which tends to zero. Every Borel finite-code predictor can thus be approximated in risk by quantized scalar ReLU encoders and a finite ReLU decoder without increasing any alphabet size. Conversely, every such quantized neural predictor is a Borel finite-code predictor. Taking infima proves equality of the classes' risk infima, jointly and separately. This is an infimum statement and does not assert that every Borel partition is represented exactly by a finite ReLU encoder. The intervening Q_L is the explicitly allowed discontinuous operation, while the neural encoder and decoder remain continuous.

I Additional Experimental Details and Results

The experiments cover explicit neural constructions, Gaussian risk calculations, bounded-network optimization, Lipschitz-constrained reconstruction, and finite-bit quantization, with attained risks reported separately from the corresponding population-risk infima.

Evaluation protocol. All experiments used an NVIDIA GeForce RTX 4090 with PyTorch 2.5.1 and CUDA 12.4. Floating-point computations used FP64, deterministic operations were enabled, and TF32 was disabled. We used seeds 1000–1009 throughout. In the constructive experiments, the seeds change evaluation samples; in the correlated Gaussian experiment, they also change the problem instance; in the training experiment, they determine initialization and sampled data. Unless indicated otherwise, tabulated uncertainties are sample standard deviations across these ten replicates, not standard errors. Reported 95% confidence intervals for replicate means use the Student distribution with nine degrees of freedom. Because the conditional means are known, we estimate excess risk directly from squared Euclidean prediction errors relative to the appropriate conditional mean. Total-risk estimates add the corresponding analytic Bayes risk, as in equation (5); noisy-response losses provide a separate check.

I.1 Experiment 1: explicit scalar ReLU approximation

We used $X \sim \mathcal{N}(0, I_2)$ and $Y = X + \zeta$, where $\zeta \sim \mathcal{N}(0, 0.1 I_2)$ is independent of X, giving Bayes risk 0.2. The best one-dimensional linear representation has excess risk 1, whereas the two-dimensional identity representation has zero excess risk. We implemented a scalar ReLU factorization with $G \times G$ grid cells for $G \in \{4, 8, 16, 32, 64\}$. Each coordinate was partitioned on $[-R_G, R_G]$, with $R_G = \sqrt{2 \log G + 1}$ and transition half-width $R_G/(4G^3)$. Saturated affine–ReLU ramps form continuous staircases, and their weighted sum assigns distinct scalar codes in $[0, 1]$ to the cell cores. A finite ReLU decoder interpolates the cell centers. The neural forward pass contains no rounding or discrete lookup. The maximum encoder hidden width is $2(G - 1)$, and the decoder hidden width is G.

We evaluated all nine products of the one-dimensional core, transition, and tail regions, drawing 8,192 conditional samples per stratum and weighting their errors by the analytic Gaussian probabilities. Independent identically distributed evaluation used 65,536 samples per seed. This procedure includes low-probability transition regions and tails explicitly rather than relying on their occurrence in an ordinary test sample.

Table 2 shows that the excess risk decreased from 0.336059 to 0.002840, while the bottleneck remained one-dimensional. At $G = 64$, the total-risk estimate was 0.202840 and the 95% inter-

Table 3: Experiment 2. Independent Gaussian model. Each budget has an analytic risk-infimum reference and the mean evaluated risk over ten seeds; scalar codes use $G = 64$. Evaluated risks add the known full or subset Bayes risk to the measured approximation error. Coinciding entries for constant or exact affine predictors are therefore not estimates obtained by averaging noisy-response losses. Additional local coordinates replace scalar codes by exact two-dimensional local representations when possible.

r		Joint linear	Separate linear	Joint neural	Separate neural
0	Reference	6.127500	6.127500	6.127500	6.127500
	Evaluated	6.129539	6.129539	6.127500	6.127500
1	Reference	3.472715	3.994167	1.472387	3.488693
	Evaluated	3.468290	3.991416	1.479093	3.492737
2	Reference	1.472387	2.703500	1.472387	1.856693
	Evaluated	1.472387	2.704201	1.472387	1.862860
3	Reference	1.472387	2.198026	1.472387	1.472387
	Evaluated	1.472387	2.200177	1.472387	1.479145
4	Reference	1.472387	1.856693	1.472387	1.472387
	Evaluated	1.472387	1.857074	1.472387	1.475054
5	Reference	1.472387	1.652505	1.472387	1.472387
	Evaluated	1.472387	1.652410	1.472387	1.472927
6	Reference	1.472387	1.472387	1.472387	1.472387
	Evaluated	1.472387	1.472387	1.472387	1.472387

val for excess risk was $[0.002829, 0.002852]$. The probability-weighted core, transition, and tail contributions decreased from 0.273126, 0.043753, and 0.019180 at $G = 4$ to 0.002454, 0.000358, and 0.000028 at $G = 64$. The independent-sample excess estimate at the latter resolution was 0.002930. Across the tested resolutions, the largest cell-center reconstruction discrepancy was below 2.0×10^{-12}, and the largest discrepancy from odd symmetry was below 6.5×10^{-12}.

The error reduction therefore occurs alongside both a finer partition and smaller probability-weighted errors outside its cores, as in Lemma 1. Over the same range, the stored parameter count increased from 63 to $1,143$, and the largest absolute parameter increased from 64.5 to $507,904.5$. Thus the observed approximation gains are accompanied by increasing network size and parameter magnitude, consistent with the nonattainment result of Proposition 14.

I.2 Experiment 2: four risk curves and modality allocation

Independent Gaussian model. We instantiated Section 7 with $M = 3$ and $N = 2$. The rows of $[a_{mh}]$ were $(1.6, 0.7)$, $(0.8, 1.2)$, and $(0.45, 0.55)$; the corresponding rows of $[v_{mh}]$ were $(1.0, 1.4)$, $(0.8, 1.1)$, and $(1.3, 0.9)$; and those of $[\omega_{mh}]$ were $(0.2, 0.5)$, $(0.4, 0.25)$, and $(0.6, 0.3)$. Response-noise variances were $(0.1, 0.15)$. This gives $V_Y = 6.127500$, $\mathcal{B}([M]) = 1.472387$, and modality energies approximately $(2.638807, 1.632000, 0.384306)$.

We evaluated all four encoder classes at budgets $r = 0, \ldots, 6$. Linear encoders and decoders implement the analytic optima as actual affine maps. For a joint scalar encoder, the full standardized input first yields the standardized two-dimensional conditional mean, which is passed through the scalar ReLU construction; the decoder receives only that scalar. Separate scalar encoders read only their own standardized modality observations. Additional local coordinates, when assigned, allow exact two-dimensional local representations. Neural constructions used $G \in \{8, 16, 32, 64\}$. Evaluation used 65,536 independent samples and 4,096 conditional samples per stratum. For separate encoders, disjoint strata were defined by the first scalar-coded modality outside its core, with later modalities sampled unconditionally. This avoids an exponential product enumeration while retaining the full probability law.

The linear estimates in Table 3 agree with the closed-form references up to sampling error. Joint linear encoding reaches the Bayes value at $r = 2$, while separate linear encoding requires $r = 6$. Below these exact dimensions, the finite neural constructions approach the different infima predicted

Table 4: Experiment 2. Approximation error above the corresponding unrestricted neural risk infimum in the independent model, reported as mean ± standard deviation. Both references equal the full-observation Bayes risk, but their bottleneck dimensions differ.

G	Joint neural, $r = 1$	Separate neural, $r = 3$
8	0.260729 ± 0.002143	0.263130 ± 0.002702
16	0.080228 ± 0.000793	0.079991 ± 0.000842
32	0.023439 ± 0.000174	0.023521 ± 0.000192
64	0.006706 ± 0.000058	0.006758 ± 0.000107

Table 5: Experiment 2. Enumeration of all modality subsets in the independent model at $G = 64$, with one scalar code per retained modality. The last column gives mean ± standard deviation. These independently evaluated subset predictors are distinct evaluation rows from the frontier and allocation controls.

Retained subset J	$\mathcal{B}(J)$	Evaluated risk
$\varnothing$	6.127500	6.127500 ± 0.000000
$\{1\}$	3.488693	3.492745 ± 0.000047
$\{2\}$	4.495500	4.497651 ± 0.000025
$\{3\}$	5.743194	5.743734 ± 0.000005
$\{1, 2\}$	1.856693	1.862914 ± 0.000062
$\{1, 3\}$	3.104387	3.108966 ± 0.000053
$\{2, 3\}$	4.111194	4.113895 ± 0.000026
$\{1, 2, 3\}$	1.472387	1.479123 ± 0.000095

by Theorem 10. At $r = 1$, joint neural excess above its reference decreased from 0.260729 to 0.006706 as G increased from 8 to 64; separate neural encoding at $r = 3$ decreased from 0.263130 to 0.006758 (Table 4). Thus the neural risk infima reach the Bayes value at dimensions 1 and 3, while the finite scalar constructions retain nonzero approximation error; the corresponding exact continuous factorization dimensions are 2 and 6.

Allocation and subset controls. At total budget $r = 2$, allocation $(2, 0, 0)$ gave risk 3.488693, whereas allocation $(1, 1, 0)$ gave 1.862859 ± 0.000073 at $G = 64$. Allocation $(1, 0, 0)$ gave 3.492733 ± 0.000040. A second coordinate for the first modality removes its finite coding error but does not change the available subset of observations; assigning that coordinate to the second modality instead changes the subset Bayes risk from 3.488693 to 1.856693. Table 5 includes all eight subsets. Their evaluated ordering agrees with the modality-energy ordering, with the smallest risk for each subset cardinality obtained by retaining the highest-energy modalities. The allocation comparison reflects the subset-risk characterization in equation (4): reallocating one coordinate can change the retained modality set and hence the subset Bayes risk.

Shifted-lattice realization. We used the real Fourier construction of Proposition 15, with $d = 2$, $s = 2$, coarse shape $(1, 2)$, observation-noise variance $\nu^2 = 0.4$, and response-noise variance $\tau^2 = 0.1$. There are $M = 4$ modalities, $N = 2$ response coordinates, and eight fine-grid values. We evaluated all four encoder classes for $r = 0, \ldots, 8$, using $G = 16$ and $G = 64$ for scalar codes, and enumerated all sixteen modality subsets at $G = 64$. The Bayes risk is 0.771429. Joint neural risk at $r = 1$ decreased from 0.795471 to 0.773450, and separate neural risk at $r = 4$ decreased from 0.795572 to 0.773462. At $G = 64$, the corresponding excess risks were 0.002022 ± 0.000014 and 0.002034 ± 0.000023. The joint and separate linear curves reach the Bayes value at $r = 2$ and $r = 8$, respectively, while Corollary 11 gives neural-infimum thresholds 1 and 4.

The scaled Fourier orthogonality discrepancy was 9.71×10^{-17}, and the fine-grid enumeration contained no duplicate indices. With 65,536 field samples per seed, the mean maximum-entry covariance discrepancy from I_8 was 0.0139, and the mean maximum absolute sample mean was 0.00653. These diagnostics verify the implemented Fourier realization for the tested configuration, while the construction itself remains specific to the finite periodic model.

Table 6: Experiment 3. Rank-one Gaussian encoding across ten problem instances per correlation level. The joint excess risk is 3.0125 throughout. The architecture gap is reported as mean $\pm$ standard deviation; other columns are means. The last column is the absolute error in the separate excess risk obtained by treating raw modality blocks as whitened feasible subspaces.

α	Separate excess risk	Architecture gap	Geometric lower bound	Unwhitened absolute error
0.00	7.2796	4.2671 ± 0.9728	3.3109	0.0000
0.25	7.1011	4.0886 ± 1.0389	3.1678	0.3570
0.50	7.0418	4.0293 ± 1.1015	3.1240	0.6653
0.75	7.1372	4.1247 ± 1.1802	3.2019	0.9287
0.90	7.3547	4.3422 ± 1.2658	3.3473	1.1822
0.98	7.6595	4.6470 ± 1.3551	3.5501	1.4765

I.3 Experiment 3: correlated Gaussian encoding

We used three four-dimensional modalities and a four-dimensional response. For each seed, the input covariance was obtained by a block-diagonal transformation of a Gaussian covariance with modality correlation matrix $(1-\alpha)I_3 + \alpha \mathbf{1}\mathbf{1}^\top$. Each transformation block was $Q_m \operatorname{diag}(0.5, 0.9, 1.5, 2.0)$, with Q_m a random orthogonal matrix. We varied $\alpha \in \{0, 0.25, 0.5, 0.75, 0.9, 0.98\}$ and held the whitened operator T fixed within a seed, with singular values $(3, 1.5, 0.8, 0.35)$. The regression matrix was therefore $K = T\Sigma_X^{-1/2}$ and changed with the covariance. Response-noise variances were $(0.1, 0.15, 0.2, 0.25)$, giving Bayes risk 0.7. This design changes the feasible separate subspaces without changing the joint spectral reference; it is not a fixed-regression comparison.

Joint encoders were checked at every budget from 0 to 12. The separate optimum was computed globally for $r = 1$ by solving the best rank-one encoding problem for each modality and selecting the best of the three candidates. Predictions were evaluated on 65,536 independent samples per instance. Across the 60 instances, the largest discrepancy between the joint spectral and covariance excess-risk calculations was 4.44×10^{-13}. The predicted joint excess risks at $r = 0, 1, 2, 3, 4$ were 12.0125, 3.0125, 0.7625, 0.1225, and 0; at $\alpha = 0.98$, the corresponding sampled means were 12.0069, 3.0096, 0.7628, 0.1226, and approximately 1.01×10^{-27}.

At $r = 1$, every tested instance had a positive joint–separate architecture gap. The mean gap ranged from 4.0293 to 4.6470, and equation $\mathcal{R}^{\mathrm{lin}}_{\mathrm{sep}}(r) - \mathcal{R}^{\mathrm{lin}}_{\mathrm{joint}}(r) \geq \left[\sigma_r^2(T) - \sigma_{r+1}^2(T)\right]\delta^2_{\mathrm{sep}}(r)$ held for every instance. Its mean lower bound was approximately 76–78% of the corresponding mean gap (Table 6). Treating raw modality blocks as whitened subspaces instead produced mean absolute separate-risk errors increasing from numerical zero at $\alpha = 0$ to 1.4765 at $\alpha = 0.98$. The covariance transformation in equation (10) is therefore numerically consequential; across the tested correlations, the architecture gap itself is nonmonotone.

For Proposition 8, we enumerated all $5^3 = 125$ local-rank allocations per instance and sampled two feasible encoders for each allocation. All 15,000 encoder-instance checks satisfied the lower bound within the prescribed 10^{-8} tolerance. The smallest slack was -6.84×10^{-14}, and the largest discrepancy in $\Delta_m = K_m\Omega_m$ was 6.14×10^{-15}. Because the feasible encoders were sampled rather than optimized, these calculations test the lower bound rather than the higher-rank separate optimum. The covariance condition number reached $2,368$, while the maximum rank-one risk-identity discrepancy remained below 5.05×10^{-13}. The measured ranks were $\operatorname{rank}(K) = 4$ and $\sum_m \operatorname{rank}(K_m) = 12$ in every instance, matching the rank inputs used in Theorem 9; the topological rank bound remains an analytic result.

I.4 Experiment 4: bounded networks and quantized representations

Training protocol. We returned to the two-dimensional Gaussian regression problem of Experiment 1. Both encoder and decoder were fully connected ReLU networks with an affine output layer. The initial scalar-bottleneck sweep used width and hidden-layer-count pairs $(128, 2)$, $(256, 2)$, $(512, 2)$, and $(256, 4)$, each with parameter bounds $C \in \{0.125, 0.5, 2, 8\}$. Every weight and bias was projected onto $[-C, C]$ after each optimizer update. We used Adam with fresh Gaussian batches of

Table 7: Experiment 4. Initial 2,500-step scalar-bottleneck runs, reported as mean $\pm$ standard deviation of excess risk. Width and hidden-layer count refer separately to the encoder and decoder. Bounds apply to every weight and bias. The $C = 2$ and $C = 8$ runs have identical results because neither bound activates along these shorter trajectories.

Width $\times$ layers	$C = 0.125$	$C = 0.5$	$C = 2$	$C = 8$
128×2	0.4693 ± 0.0058	0.2897 ± 0.0258	0.2614 ± 0.0144	0.2614 ± 0.0144
256×2	0.3711 ± 0.0113	0.2754 ± 0.0273	0.2472 ± 0.0183	0.2472 ± 0.0183
512×2	0.2939 ± 0.0240	0.2516 ± 0.0238	0.2532 ± 0.0280	0.2532 ± 0.0280
256×4	0.2736 ± 0.0438	0.2481 ± 0.0486	0.2419 ± 0.0348	0.2419 ± 0.0348

Table 8: Experiment 4. Training-budget comparison for width 256, four hidden layers in each network, and a scalar bottleneck. Entries are mean $\pm$ standard deviation of test excess risk for validation-selected checkpoints. Each column uses a cosine schedule defined over its own full step budget; the longer protocols are not checkpoint continuations of the shorter ones.

C	2,500 steps	10,000 steps	20,000 steps
0.125	0.2736 ± 0.0438	0.2534 ± 0.0356	Not run
0.5	0.2481 ± 0.0486	0.2052 ± 0.0502	Not run
2	0.2419 ± 0.0348	0.1844 ± 0.0357	0.1892 ± 0.0303

size 256, initial learning rate 0.002, cosine decay to one tenth of that rate, and gradient-norm clipping at 10. Training targets were the known conditional means X. Checkpoints were selected using 8,192 fixed validation samples, evaluated every 100 steps, without consulting the test set.

Test evaluation used 131,072 independent samples and 8,192 conditional samples in each radial interval $[0, 2)$, $[2, 4)$, $[4, 6)$, and $[6, \infty)$ for $\|X\|_2$, weighted by their analytic probabilities. The training protocols used 2,500-step runs. Additional protocols used 10,000 steps for width 256, four hidden layers in each network, and $C \in \{0.125, 0.5, 2\}$, followed by a separate 20,000-step protocol for $C = 2$. Each protocol was initialized afresh and defined its cosine schedule over its own step budget, so the longer runs are not continuations of shorter checkpoints.

Bounded-network results. The initial sweep shows lower attained risks with weaker bounds and, particularly under $C = 0.125$, with larger networks (Table 7). The dependence on width is not monotone for every bound. At 2,500 steps, $C = 2$ and $C = 8$ never triggered a projection and produced identical paired results, whereas $C = 0.125$ and $C = 0.5$ triggered projections at every update. The 10,000-step protocols attained mean excess risks 0.253394, 0.205206, and 0.184418 for their three bounds (Table 8).

The 20,000-step $C = 2$ protocol gave excess risk 0.189237 ± 0.030276, with a 95% interval of $[0.167579, 0.210895]$, corresponding to total risk 0.389237. The independent-sample excess estimate was 0.189751. The selected checkpoints had mean maximum absolute parameter 1.999787, and projection occurred in $9,762.5$ of the 20,000 updates on average. The radial contributions to excess risk were 0.119169, 0.068459, 0.001609, and 2.59×10^{-7}, respectively, so approximately 99.15% of the estimated error arose within $\|X\|_2 < 4$. The positive error is therefore not attributable mainly to rare Gaussian tails.

For an optimization diagnostic, we fitted a linear slope to each seed's validation errors over the 41 checkpoints from steps 16,000 to 20,000. The mean slope was -0.002448 per 1,000 steps, with a 95% interval of $[-0.009469, 0.004574]$; seven slopes were negative and three were positive. The mean final-checkpoint validation error was 0.214148, compared with 0.185322 for the validation-selected checkpoints. The interval includes zero, so these runs do not resolve a common late-stage trend across seeds.

A separately trained two-dimensional control with width 256, two hidden layers in each network, and $C = 8$ attained excess risk $(2.927 \pm 0.167) \times 10^{-6}$ after 2,500 steps. The same training and evaluation pipeline therefore closely approximates the identity when exact factorization is available. For the scalar bottleneck, Proposition 14 establishes strict positivity of the bounded model-class

Table 9: Experiment 4. Quantization of the fixed scalar code, reported as mean ± standard deviation of excess risk over ten matched evaluation seeds. The unquantized rows are separately evaluated controls for this experiment, not new training runs.

Bits	$G = 16$	$G = 32$	$G = 64$
Unquantized	0.033963 ± 0.000330	0.009938 ± 0.000083	0.002839 ± 0.000027
4	1.997234 ± 0.034561	6.243732 ± 0.114568	5.483095 ± 0.120094
6	0.282869 ± 0.009506	1.401166 ± 0.021528	1.972959 ± 0.051152
8	0.034423 ± 0.000339	0.082034 ± 0.005133	0.325812 ± 0.006846
10	0.035028 ± 0.000358	0.010006 ± 0.000084	0.023006 ± 0.000283
12	0.033973 ± 0.000323	0.010332 ± 0.000123	0.002848 ± 0.000027
16	0.033963 ± 0.000330	0.009947 ± 0.000083	0.002849 ± 0.000028
24	0.033963 ± 0.000330	0.009938 ± 0.000083	0.002839 ± 0.000027
32	0.033963 ± 0.000330	0.009938 ± 0.000083	0.002839 ± 0.000027

minimum, whereas the reported trained risks depend on the optimization procedure and need not equal the model-class minima. The 10,000- and 20,000-step protocols were initialized separately and are therefore treated as distinct training runs.

Representation quantization. We separately inserted a uniform b-bit quantizer between the fixed scalar encoder and decoder from Experiment 1, without retraining either network. For $h \in [0, 1]$, the quantizer was $h \mapsto \mathrm{round}((2^b - 1)h)/(2^b - 1)$. We used $G \in \{16, 32, 64\}$ and $b \in \{4, 6, 8, 10, 12, 16, 24, 32\}$, together with an unquantized control. Evaluation used 4,096 conditional samples per product stratum, with matched samples across bit depths. The arithmetic remained FP64; the bit budget restricts the transmitted representation, not the arithmetic used to evaluate the networks.

At $G = 64$, unquantized excess risk was 0.002839, compared with 0.325812, 0.023006, and 0.002848 at 8, 10, and 12 bits (Table 9). Near-unquantized performance was also observed at 8 bits for $G = 16$ and 10 bits for $G = 32$. In this construction, the G^2 cell-core codes are equally spaced on $[0, 1]$, so the quantizer levels coincide with them when $2^b = G^2$. This accounts for the observed alignment at $b = 2 \log_2 G$. Because the quantizer grids are not nested across all bit depths, the measured risk need not be monotone. The experiment characterizes quantization of this fixed scalar code rather than the optimal risk over all b-bit encoders.

I.5 Experiment 5: scalar reconstruction under a Lipschitz constraint

We evaluated the reflected-grid construction in equations (40) and (42) for a uniform input on Q_d. The grids were $n \in \{4, 6, 8, 12, 16, 24, 32\}$ for $d = 2$ and $n \in \{4, 6, 8, 10, 12, 16\}$ for $d = 3$. We set $w = 1/(8n^3)$, giving the analytic encoder bound $L_{n,w} = 16dn^{d+1}$ and decoder bound one. Thus the reported budget $\Lambda = 16dn^{d+1}$ is an analytic upper bound on the Lipschitz product, not an empirical estimate of its smallest possible value.

Evaluation used $16{,}384$ conditional samples per stratum and seed. The strata consisted of the joint core and the d disjoint events identified by the first coordinate outside its one-dimensional core. Earlier coordinates were sampled conditionally inside their cores, and later coordinates were sampled unconditionally. Conditional errors were weighted by the exact stratum probabilities. This includes the increasingly narrow transition regions without relying on their occurrence in an ordinary random sample. All computations used FP64; seeds changed evaluation samples, not network parameters.

The implementation evaluated neighboring candidate cells in the encoder minimum. For each seed and each tested $n \in \{4, 6, 8\}$, we compared this evaluation with the minimum over all grid cells on 512 independent uniform inputs. The largest recorded discrepancy was zero in both dimensions. The zero discrepancy provides a numerical check of the neighboring-cell implementation on the sampled inputs.

For each seed, we fitted ordinary least-squares slopes of log risk against log budget. Fits over all tested grids gave -0.7036 ± 0.0008 for $d = 2$ and -0.5304 ± 0.0006 for $d = 3$. Fits over the four largest budgets used $n \in \{12, 16, 24, 32\}$ and $n \in \{8, 10, 12, 16\}$, respectively, giving the slopes

Table 10: Experiment 5. Reconstruction risk of the prescribed scalar construction, reported as mean $\pm$ sample standard deviation over ten evaluation seeds. The last column is mean risk multiplied by $\Lambda^{2/(d+1)}$. Budgets are analytic Lipschitz product bounds.

d	n	Λ	Reconstruction risk	Rescaled risk
2	4	2,048	$(2.0073 \pm 0.0066) \times 10^{-2}$	3.237
2	6	6,912	$(8.4417 \pm 0.0357) \times 10^{-3}$	3.063
2	8	16,384	$(4.5254 \pm 0.0097) \times 10^{-3}$	2.919
2	12	55,296	$(1.9020 \pm 0.0086) \times 10^{-3}$	2.761
2	16	131,072	$(1.0386 \pm 0.0033) \times 10^{-3}$	2.680
2	24	442,368	$(4.5177 \pm 0.0161) \times 10^{-4}$	2.623
2	32	1,048,576	$(2.5033 \pm 0.0064) \times 10^{-4}$	2.584
3	4	12,288	$(4.1112 \pm 0.0089) \times 10^{-2}$	4.557
3	6	62,208	$(1.8462 \pm 0.0049) \times 10^{-2}$	4.605
3	8	196,608	$(9.9145 \pm 0.0156) \times 10^{-3}$	4.396
3	10	480,000	$(6.0874 \pm 0.0112) \times 10^{-3}$	4.217
3	12	995,328	$(4.0861 \pm 0.0116) \times 10^{-3}$	4.077
3	16	3,145,728	$(2.2046 \pm 0.0076) \times 10^{-3}$	3.910

reported in the main text. Their variation across seeds measures evaluation noise, not uncertainty from extrapolating a finite budget range to an asymptotic exponent.

At the largest grids, the core and transition contributions were 1.6212×10^{-4} and 8.8209×10^{-5} for $d = 2$, and 9.7414×10^{-4} and 1.2305×10^{-3} for $d = 3$. Transition error therefore remains a substantial part of the measured risk.

I.6 Experiment 6: fusion under matched Lipschitz budgets

We used the model of Theorem 6, with independent X_1, X_2 uniform on Q_3 and $Y = (X_1 + X_2)/\sqrt{2}$. Both architectures had total latent dimension two. Budgets were $\Lambda \in \{2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}, 2^{22}\}$. For each scalar construction in dimension d, we chose $n = \lfloor \Lambda^{1/(d+1)} \rfloor$ and $w = 2dn^{d-2}/\Lambda$. Separate encoding used $d = 3$ for each modality; joint encoding formed $S = (X_1 + X_2)/2$, retained its first coordinate, and used the $d = 2$ construction for the remaining coordinates.

The separate encoder and decoder have Lipschitz bounds Λ and one. For the joint construction, averaging and output rescaling give encoder and decoder bounds $\Lambda/\sqrt{2}$ and $\sqrt{2}$, respectively. Both products are therefore at most the same budget Λ. Grid sizes were $\{8, 11, 16, 22, 32, 45\}$ for the separate construction and $\{16, 25, 40, 64, 101, 161\}$ for the joint construction.

Each reconstruction-moment calculation used 32,768 conditional samples per stratum and seed, with the first-transition partition used in Experiment 5. Separate inputs were sampled from the uniform law. Joint inputs were sampled from the actual product triangular law: each coordinate of S has density $4t$ for $0 \le t \le 1/2$ and $4(1 - t)$ for $1/2 < t \le 1$. Conditional samples and stratum probabilities used this density rather than a uniform approximation.

For separate encoding, let $e_m = X_m - \psi(\phi(X_m))$ denote the error of the common local construction. Independence and identical local maps give $\frac{1}{2}\mathbb{E}\,\|e_1 + e_2\|_2^2 = \mathbb{E}\,\|e_1\|_2^2 + \|\mathbb{E}e_1\|_2^2$. We estimated the first and second moments by stratification and substituted them into this identity, retaining the squared mean-error term. Joint risk was twice the two-dimensional reconstruction error under the triangular law; its first output coordinate was recovered exactly.

Log–log fits over the four largest budgets used $\Lambda \in \{2^{16}, 2^{18}, 2^{20}, 2^{22}\}$. The risk-ratio slope was -0.1417 ± 0.0010, compared with separate and joint risk slopes -0.5295 ± 0.0005 and -0.6713 ± 0.0009. The ratio decreases throughout the tested range. Because the experiment evaluates prescribed constructions rather than architecture-level optima, its fitted slope need not equal the $-1/6$ exponent governing the bound on the optimal-risk ratio.

Table 11: Experiment 6. Attained risks under common latent dimension and Lipschitz budgets, reported as mean ± sample standard deviation over ten seeds. The last column summarizes the joint-to-separate risk ratio computed within each seed, not the ratio of the two mean risks.

Λ	Separate risk	Joint risk	Risk ratio
2^{12}	$(2.7843 \pm 0.0072) \times 10^{-1}$	$(1.9596 \pm 0.0060) \times 10^{-2}$	0.07038 ± 0.00020
2^{14}	$(1.2089 \pm 0.0034) \times 10^{-1}$	$(7.0730 \pm 0.0206) \times 10^{-3}$	0.05851 ± 0.00025
2^{16}	$(5.8166 \pm 0.0115) \times 10^{-2}$	$(2.7186 \pm 0.0075) \times 10^{-3}$	0.04674 ± 0.00019
2^{18}	$(2.6187 \pm 0.0076) \times 10^{-2}$	$(1.0690 \pm 0.0034) \times 10^{-3}$	0.04082 ± 0.00018
2^{20}	$(1.3167 \pm 0.0029) \times 10^{-2}$	$(4.2000 \pm 0.0136) \times 10^{-4}$	0.03190 ± 0.00010
2^{22}	$(6.3314 \pm 0.0104) \times 10^{-3}$	$(1.6689 \pm 0.0053) \times 10^{-4}$	0.02636 ± 0.00010

Table 12: Experiment 7. Excess risk above Bayes risk 0.2, reported as mean ± sample standard deviation over ten matched evaluation seeds. The prescribed product quantizer changes with the bit budget. All tested budgets are shown, including the nonmonotone low-bit separate results.

Total bits B	Joint excess risk	Separate excess risk
0	6.3045 ± 0.0076	6.3045 ± 0.0076
4	5.1664 ± 0.0113	1.8122 ± 0.0037
8	$(1.6927 \pm 0.0058) \times 10^{-1}$	3.9767 ± 0.0070
12	$(1.0093 \pm 0.0044) \times 10^{-2}$	$(4.7130 \pm 0.0350) \times 10^{-1}$
16	$(6.2695 \pm 0.0167) \times 10^{-4}$	$(1.0477 \pm 0.0035) \times 10^{-1}$
20	$(3.9227 \pm 0.0143) \times 10^{-5}$	$(2.5118 \pm 0.0098) \times 10^{-2}$
24	$(2.4498 \pm 0.0043) \times 10^{-6}$	$(6.2054 \pm 0.0261) \times 10^{-3}$
28	$(1.5313 \pm 0.0038) \times 10^{-7}$	$(1.5481 \pm 0.0050) \times 10^{-3}$
32	$(9.5758 \pm 0.0324) \times 10^{-9}$	$(3.8587 \pm 0.0109) \times 10^{-4}$

I.7 Experiment 7: finite-bit risk exponents

We fixed a centered Gaussian observation with two two-dimensional modalities. The marginal covariances were $\Sigma_{11} = \Sigma_{22} = I_2$, and the cross-covariance and regression blocks were

$$\Sigma_{12} = 0.7 \begin{pmatrix} \cos(0.4) & -\sin(0.4) \\ \sin(0.4) & \cos(0.4) \end{pmatrix}, \; K_1 = \begin{pmatrix} 1.0 & 0.3 \\ -0.2 & 0.8 \end{pmatrix}, \; K_2 = \begin{pmatrix} 0.7 & -0.4 \\ 0.5 & 1.1 \end{pmatrix}.$$

We used $\Sigma_{21} = \Sigma_{12}^\top$ and conditional mean $KX = K_1X_1 + K_2X_2$. Independent Gaussian response noise had covariance $0.1I_2$, giving Bayes risk 0.2. The input covariance condition number was 5.667, and the ranks were $\mathrm{rank}(K) = 2$ and $\mathrm{rank}(K_1) = \mathrm{rank}(K_2) = 2$. The model was fixed across all seeds and bit budgets.

For $B \in \{0, 4, 8, 12, 16, 20, 24, 28, 32\}$, we evaluated the explicit scalar quantizer in equation (79) on standardized Gaussian coordinates. Joint encoding quantized the covariance eigen-coordinates of KX, using $2^{\lfloor B/2 \rfloor}$ levels per coordinate. Separate encoding quantized those of each K_mX_m, using $2^{\lfloor B/4 \rfloor}$ levels per coordinate, and added the reconstructed local vectors. Each tested budget was used in full. The product quantizer was reconstructed independently at each budget; no fitted neural encoder or fixed neural code was used.

Evaluation used $262{,}144$ independent Gaussian observations per seed in FP64, with the same observations shared across both architectures and every bit budget. Excess risk was measured directly against KX; total risk adds the analytic Bayes value 0.2. In particular, the reported small high-bit errors were not obtained by subtracting two noisy total-risk estimates.

For each seed, ordinary least-squares fits of log-base-two excess risk against B over $B \in \{16, 20, 24, 28, 32\}$ gave slopes -0.99995 ± 0.00032 and -0.50475 ± 0.00033 for joint and separate encoding. Over this range, joint excess risk multiplied by 2^B stayed between 41.088 and 41.133, while separate excess risk multiplied by $2^{B/2}$ decreased from 26.821 to 25.288. Fits including $B = 8$

and $B = 12$ gave -1.00216 ± 0.00021 and -0.54081 ± 0.00018, respectively, showing the larger finite-bit effect for the separate construction.

The separate risk increased from 1.8122 at $B = 4$ to 3.9767 at $B = 8$ because the prescribed reconstruction levels are redesigned independently at each resolution and are neither nested nor optimized for monotonicity. This finite-budget behavior is specific to the prescribed quantizer; the optimal finite-codebook risk remains nonincreasing in B. At high bit rates, the measured decay agrees with the attainable exponents in Theorem 12, whose matching lower bounds are established analytically.

J Further Relation to Existing Theory

Multimodal learning has long studied where information from different modalities should be integrated, with classical formulations distinguishing, among other strategies, feature-level early fusion from decision-level late fusion (Baltrusaitis et al., 2019). We consider a more specific representation constraint within this broader setting: under a common total representation-dimension budget, information may either be compressed separately within each modality before fusion or compressed jointly after cross-modal interaction. In the linear setting, joint encoding is closely related to population reduced-rank regression (Izenman, 1975), whereas sufficient dimension reduction seeks low-dimensional subspaces that preserve a conditional distribution or conditional mean (Cook & Li, 2002). Distributed linear estimation and the distributed Karhunen–Loève transform further study local transforms, dimension allocation, and fusion under linear compression (Zhu et al., 2005; Gastpar et al., 2006; Schizas et al., 2005; Fang & Li, 2008; 2010; Zhang et al., 2019). More generally, multiterminal source coding and the CEO problem also consider multiple local encoders that process observations separately before a central decoder performs joint estimation, but their fundamental resource constraint is communication rate rather than the number of unconstrained, infinite-precision real-valued representation coordinates (Berger et al., 1996). Our linear results adopt the same population perspective but emphasize the geometry induced by a common total rank budget and use this structure as a benchmark for the dimension collapse that arises under unrestricted neural encoding.

Recent work has also compared intermediate and late multimodal fusion under Lipschitz assumptions (Li et al., 2026), deriving generalization-error bounds for multimodal representation learning. Our setting instead fixes the total latent dimension and constrains the encoder–decoder Lipschitz product, with population approximation risk as the objective. The resulting separation therefore concerns compression risk as the regularity budget grows rather than fusion-dependent generalization error.

Classical universal approximation establishes density of finite networks in continuous functions on compact sets (Hornik, 1991; Leshno et al., 1993), but does not by itself provide a continuous scalar factorization. Narrow-network proofs supply the more direct antecedent: Park et al. (2021) and Kim et al. (2024) explicitly use scalar coding before vector decoding. Their role in the present paper is therefore stronger than a comparison of whole-network width with designated bottleneck width. The half-space operation in Kim et al. (2024, Lemma 7) is $P_{a,b,c}(x) = x + \rho(-a^\top x - b)c/(a^\top c)$ for $a^\top c > 0$: it fixes one half-space and projects its complement onto the boundary. Their Appendix F, Lemma 27, also applies a scalar encoder tokenwise. Our proof replaces the Lebesgue-measure coding argument by finite measurable product partitions and high-probability compact cores under the given law. This yields the distribution-general blockwise approximation lemma and the exact modality-selection formula in Theorem 2; it does not make the underlying scalar-coding principle new.

Topological obstructions to low-dimensional continuous representations provide a more direct comparison with our nonlinear results. Batson et al. (2021) study how the topology of the data can obstruct faithful autoencoder reconstruction, while Kvalheim & Sontag (2024) show that, for data supported on manifolds of bounded intrinsic dimension, topological obstructions to exact continuous reconstruction can coexist with arbitrarily small reconstruction error outside exceptional sets of arbitrarily small measure, leading in particular to arbitrarily small global L^2 reconstruction loss under their assumptions. Our setting differs in several respects: we allow arbitrary finite-dimensional joint laws and arbitrary square-integrable prediction targets, impose no manifold-support assumption, and study both joint and block-separable encoding. Moreover, the scalar approximation phenomenon is not specific to prediction. Taking the identity map as the target in our approximation lemma yields scalar L^2 autoencoding for any square-integrable finite-dimensional distribution. The distinction from

the topological autoencoding literature therefore lies not simply in prediction versus reconstruction, but in the distributional assumptions, the representation architecture, and the notion of approximation being studied.

Exact continuous factorization exhibits a fundamentally different dimensional behavior. Our lower bounds rely on invariance of domain (Brouwer, 1912): when the target map retains sufficiently many independent directions on a high-dimensional subspace, a continuous encoder cannot factor it exactly through a lower-dimensional Euclidean space. By contrast, this topological obstruction may disappear if the representation is required to be only measurable; the Borel isomorphism theorem for uncountable standard Borel spaces provides the classical basis for this phenomenon (Kuratowski, 1934). In statistical learning, Pestov (2013) similarly observes that the standard learning formalism can be insensitive to ambient dimension and uses a Borel isomorphism to obtain a deterministic reduction to a one-dimensional domain. Our setting lies between these two extremes. The encoder and decoder are finite ReLU networks and are therefore continuous, but the object of interest is the population-risk infimum under a fixed probability law rather than pointwise exact factorization. This distinction explains why a one-dimensional neural bottleneck can collapse in L^2 risk while exact continuous factorization continues to obey a rank barrier.

The information bottleneck provides a different notion of representation complexity. Rather than treating the number of Euclidean coordinates as the sole resource, the classical information-bottleneck formulation characterizes compression through mutual information between the representation, the input, and the prediction target (Tishby et al., 2000; Goldfeld & Polyanskiy, 2020). Subsequent work on distributed information bottlenecks, task-oriented communication, and related representation-learning problems likewise constrains representation capacity through mutual information, communication rate, or stochastic encoding (Vera et al., 2019; Aguerri & Zaidi, 2019; Shao et al., 2022). These formulations remain nontrivial precisely because a real-valued coordinate is not treated as a cost-free channel of unlimited precision. In a related information-theoretic direction, Wu & Verdú (2010) show that the fundamental limits of almost-lossless analog compression are governed by Rényi information dimension under specified regularity conditions on the compressor or decompressor, including linear or Lipschitz restrictions. Our dimension-only neural bottleneck intentionally removes such rate, noise, and regularity constraints and thereby exposes the degeneracy that arises when only Euclidean output dimension is controlled. This comparison also clarifies that Lipschitz stability, bounded network complexity, quantization, representation noise, and communication rate are not alternative descriptions of representation dimension, but additional resource constraints capable of restoring a nontrivial nonlinear bottleneck theory.

Finally, our shifted-lattice construction is structurally related to multichannel and generalized sampling (Papoulis, 1977; Izen, 2005). Classical sampling theory studies the recovery of signals from multiple interlaced or heterogeneous sampling channels, whereas our purpose is not to establish a new sampling theorem. We use a finite periodic field and interlaced lattices only to provide a concrete finite-dimensional realization in which the dimensional separation among joint encoding, block-separable encoding, neural population-risk infima, and exact continuous factorization can be exhibited within a single multiresolution sensing architecture.