EPISTEMIC REPRESENTATION UNDER PREDICTIVE COMPRESSION: SEPARATION, EXCESS RATE, AND COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a compressed predictive representation preserve the information needed to evaluate a future measurement? We study settings in which an epistemic functional of a parameter-indexed observation experiment is recoverable from the complete predictive marginal before compression. We characterize residual sensing-context information at the predictive rate–distortion boundary. For a coordinate-weighted Gaussian model, we prove an all-encoder stability bound and derive the exact additional rate required for prescribed log-Fisher fidelity, including achievability. We then connect this mechanism to the native predictive KL: marginalizing an uncertain target suppresses predictive sensitivity to measurement quality while leaving log-Fisher sensitivity unchanged. For Gaussian predictive-family decoders, a global natural-KL theorem gives a strictly positive epistemic excess-rate bound at positive primary predictive rate. We test the mechanism with stochastic neural bottlenecks under the same natural-KL loss. A validation-locked sweep followed by ten-seed replication yields 40/40 predictive-selected models meeting predictive fidelity but 0/40 reaching the epistemic tolerance under the prespecified probe, whereas 40/40 epistemic-aware models reach both empirical tolerances. Epistemic-aware representations also reduce sensing-action error from approximately chance to 0.1370.140. These results quantify how information present in a predictive distribution can remain costly to preserve under predictive compression.

1 INTRODUCTION

Predictive representation learning is a central principle for sequential decision-making. Predictive state representations characterize state through future observable tests, latent world models compress histories for forecasting and planning, and predictive information bottlenecks retain future-relevant information under compression (15; 20; 4; 11; 12; 21; 3; 7). Across these formulations, prediction implicitly determines which distinctions deserve representational capacity.

In embodied inference, however, an agent must also ask: *what will a future observation make distinguishable?* Two sensing contexts may induce similar predictive marginals while making a candidate measurement sharply different in its ability to identify a hidden quantity. Fisher information, Bayes error, and related experimental-design quantities formalize this epistemic role (14; 5).

We study a stronger regime than simple predictive insufficiency. Let K_Y denote the complete predictive marginal and E an epistemic consequence of the future observation experiment. In the Type-II setting,

$$E = \phi_Y(K_Y),$$

so the full predictive object already contains everything required to recover E. Yet after compression, a representation optimized for predictive fidelity may assign little or no rate to directions that weakly affect prediction but strongly determine future measurement quality. Thus predictive sufficiency before compression need not imply epistemic sufficiency after predictive-optimal compression.

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.

Geometrically, predictive and epistemic objectives need not induce the same sensitivity over representation variables. A direction strongly contracted by predictive geometry may remain important for parameter distinguishability. The failure therefore arises from *rate allocation at the predictive optimum*, not absence of relevant information from the predictive distribution.

We ask whether

$$D_P^*(Z) \le D_P^{\mathrm{req}} \quad \Longrightarrow ? \quad D_E^*(Z) \le D_E^{\mathrm{req}}.$$

When this implication fails, we define the additional rate needed for epistemic fidelity as

$$\Delta R_E(D_P, D_E) = R_{P,E}(D_P, D_E) - R_P(D_P),$$

the *epistemic excess rate*. Within classical rate–distortion theory (10; 23; 8; 9; 18; 19; 2; 6), we study how much epistemic information remains available near the primary predictive rate–distortion boundary.

Our results characterize both collapse and separation. We give a condition under which predictive fidelity guarantees epistemic fidelity, and a separation criterion based on residual conditional information at the predictive rate–distortion boundary. For a coordinate-weighted Gaussian model, we derive an architecture-independent stability bound, the exact positive epistemic excess rate with achievability, and a sharp activation threshold for the sensing-quality variable. Under native predictive KL, marginalizing an uncertain target suppresses predictive sensitivity to sensing quality while leaving log-Fisher sensitivity unchanged, yielding a global finite-M excess-rate theorem for Gaussian predictive-family decoders. We further connect the separation to sensing decisions and validate it with stochastic neural bottlenecks under validation-only selection and ten-seed replication.

Task-oriented compression exhibits a related reconstruction–estimation mismatch. Weiss (22) shows that squared-error rate–distortion compression can discard spectral content weak in reconstruction energy but important for localization Fisher information. Our setting instead concerns a *predictive–epistemic* mismatch: the primary object is already predictive, the secondary quantity evaluates an action-conditioned future observation, and under Type-II factorization it is recoverable from the complete predictive object. The central phenomenon is

$$\boxed{V \in K_Y, \qquad R_{V,\mathrm{opt}} = 0, \qquad R_{E|U}(D_E) > 0,}$$

i.e., epistemically relevant context is present in the predictive law, receives zero or vanishing rate at the predictive optimum, yet remains necessary for future measurement evaluation.

The broader implication is that accurate prediction alone does not guarantee preservation of the distinctions needed to decide *what should be observed next*.

2 Setup: Observation Experiment and Compression

We consider a history H, a future action sequence a, and a latent quantity Θ to be estimated or distinguished in the future.

Definition 1 (Full observation experiment)**.**

$$\mathfrak{E}_Y(h, a) = \{p(y^+ \mid \theta, h, a)\}_{\theta \in \Theta}.$$

Fix a prior kernel $\pi_{h,a}(d\theta)$ as part of the evaluation law. In the Gaussian models below this is the stated Gaussian prior, independent of (h, a). The predictive marginalization operator therefore includes this prior:

Definition 2 (Ordinary predictive marginal)**.**

$$K_Y(h, a) = \int p(y^+ \mid \theta, h, a)\, \pi_{h,a}(d\theta) = T_{P,\pi}(\mathfrak{E}_Y)(h, a).$$

Definition 3 (Epistemic consequence)**.**

$$E(h, a) = \Phi(\mathfrak{E}_Y(h, a)).$$

Assumption 4 (Type-II factorization)**.** *We focus on the Type-II regime in which the epistemic functional factors through the predictive marginal:*

$$\Phi = \phi_Y \circ T_{P,\pi}.$$

Under Assumption 4, exact recovery of K_Y implies exact recovery of E; separation therefore concerns finite-rate compression rather than a mismatch between the predictive object and the epistemic task. We call Z predictively (epistemically) adequate when $D_P^*(Z) \le D_P^{\mathrm{req}}$ ($D_E^*(Z) \le D_E^{\mathrm{req}}$). The distinction is important: the same uncompressed predictive distribution can contain both kinds of information, while a rate-limited representation may allocate its capacity very differently across them. Type-I projection loss and factorization details are in Appendix A. Rates are total nats.

3 Predictive–Epistemic Rate Region

We adopt the mutual-information formulation of rate-distortion theory (19; 2; 6) and distinguish predictive, epistemic, and jointly constrained operational regions.

Let $Z \sim p(z \mid H)$ be stochastic. Under a fixed evaluation law, A is fixed or sampled conditional on H, $Z \perp A \mid H$, and both decoders observe A. The downstream selector in Section 5.5 is instead a post-representation map $Z \mapsto a$. Assume $R_P(D_P) < \infty$ where needed and predictive reconstructions lie in the domain of ϕ_Y. Define

$$D_P^*(Z) = \inf_{\widehat{K}_P} \mathbb{E}\big[d_P(K_Y(H,A), \widehat{K}_P(Z,A))\big],$$

$$D_E^*(Z) = \inf_{\widehat{E}} \mathbb{E}\big[d_E(E(H,A), \widehat{E}(Z,A))\big].$$

The predictive rate-distortion function is

$$R_P(D_P) = \inf\{I(Z;H) : D_P^*(Z) \le D_P\},$$

and the joint predictive–epistemic rate-distortion function is

$$R_{P,E}(D_P, D_E) = \inf\{I(Z;H) : D_P^*(Z) \le D_P,\ D_E^*(Z) \le D_E\}.$$

The central quantity is the epistemic excess rate

$$\Delta R_E(D_P, D_E) = R_{P,E}(D_P, D_E) - R_P(D_P).$$

For the conditional analysis, now fix a decomposition $H = (U, V)$ and use

$$R_{E|U}(D_E) = \inf_{p(z|u,v)} I(Z; V \mid U),$$

subject to epistemic distortion D_E, with decoder access to (Z, U) (Appendix C.2). $R_{E|U}(D_E)$ uses U as decoder side information only to quantify the irreducible conditional epistemic requirement; it does not posit a Wyner–Ziv communication architecture (23; 6).

4 Collapse and General Separation

4.1 Collapse: when no extra rate is needed

Define the expected distortion modulus

$$\bar{\omega}(\delta) = \sup_{\mathbb{E}[d_P(P,Q)] \le \delta} \mathbb{E}\big[d_E(\phi_Y(P), \phi_Y(Q))\big],$$

and its right-regularized version

$$\bar{\omega}^+(\delta) = \inf_{\eta > 0} \bar{\omega}(\delta + \eta).$$

Theorem 5 (Distortion-compatible collapse)**.** *Assume Assumption 4. If $D_E \ge \bar{\omega}^+(D_P)$, then*

$$R_{P,E}(D_P, D_E) = R_P(D_P), \qquad \Delta R_E = 0.$$

A near-optimal predictive decoder composed with ϕ_Y attains epistemic distortion at most $\bar{\omega}^+(D_P)$; see Appendix B. Hence the epistemic constraint is free whenever $D_E \ge \bar{\omega}^+(D_P)$.

4.2 Free epistemic rate

Fix a decomposition $H = (U, V)$, where U is predictive-relevant and V is epistemic-relevant.

Definition 6 (Free epistemic rate)**.**

$$J_E(R, D_P) = \sup_{\substack{I(Z;H)\le R \\ D_P^*(Z)\le D_P}} I(Z; V \mid U).$$

Definition 7 (Near-optimal free epistemic rate)**.**

$$J_E^+(D_P) = \lim_{\epsilon \downarrow 0} J_E(R_P(D_P) + \epsilon, D_P).$$

The $\epsilon \downarrow 0$ relaxation is required because the predictive RDF need not be attained (Appendix C.1).

The boundary quantity J_E^+ measures residual context information available as total rate approaches the primary predictive optimum. It is therefore a property of the geometry of the primary rate–distortion boundary, rather than of a second communication link. If $J_E^+ = 0$, an asymptotically predictive-optimal sequence cannot retain a positive amount of conditional context information "for free"; any epistemic requirement demanding such information must move the operating point away from the primary boundary. Classical coding connections are discussed in Appendix F.

4.3 General separation

Theorem 8 (General separation)**.** *Assume Assumption 4. Suppose there exists a decomposition $H = (U, V)$ and predictive distortion d_P such that*

$$J_E^+(D_P) < R_{E|U}(D_E).$$

Then $R_{P,E}(D_P, D_E) > R_P(D_P)$ and $\Delta R_E > 0$.

If joint fidelity were achievable at predictive-optimal rate, a near-optimal predictive sequence would carry at least $R_{E|U}(D_E)$ nats about V given U, contradicting J_E^+ (Appendix C.3). Thus the theorem compares what the primary-optimal boundary can retain with what epistemic reconstruction necessarily requires. The condition is sufficient, not necessary: $I(Z; V \mid U)$ summarizes a required information quantity, but by itself does not characterize every possible epistemic decoder.

5 Gaussian Phase Separation

5.1 Model

Let

$$U \sim \mathcal{N}(0, \sigma_U^2), \quad V \sim \mathcal{N}(0, \sigma_V^2), \quad U \perp V, \quad H = (U, V).$$

Let $\Theta \sim \mathcal{N}(0, \sigma_\Theta^2)$ be independent of (U, V). Future observations are

$$Y^+ = (Y_P^+, Y_E^+),$$

with

$$Y_P^+ = U + N_P, \qquad Y_E^+ = \Theta + \sigma(V) Z_E,$$

where $N_P \sim \mathcal{N}(0, \tau^2)$ and $Z_E \sim \mathcal{N}(0, 1)$ are independent of each other and of (U, V, Θ), and

$$\sigma^2(V) = \sigma_0^2 e^{cV}, \qquad c \neq 0.$$

The sensing context V controls the measurement quality for Θ:

$$V \;\to\; \sigma(V) \;\to\; \Gamma_\Theta(V) = \sigma_0^{-2} e^{-cV}.$$

The Fisher functional Γ_Θ is interpreted in the classical experimental-design sense of parameter-indexed observation laws (14; 5).

Since $\mathrm{Var}(Y_E^+ \mid V) = \sigma_\Theta^2 + \sigma_0^2 e^{cV}$ is strictly monotone in V when $c \neq 0$, both V and $\Gamma_\Theta(V)$ are recoverable from the ordinary predictive marginal K_Y. The Gaussian example is therefore a Type-II compression loss, not a projection loss (full proof in Appendix D.1).

5.2 PREDICTIVE ALLOCATION

Define the coordinate-form predictive distortion

$$d_P = (U - \hat{U})^2 + \lambda(V - \hat{V})^2,$$

the coordinate form of a distortion on the identifiable Gaussian predictive family.

Proposition 9 (Predictive allocation)**.** *Assume $0 < D < \sigma_U^2$ and $0 < \lambda < \lambda^\star = D/(2\sigma_V^2)$. The epistemic component receives zero rate under the predictive-optimal allocation, $R_{V,\mathrm{opt}}(\lambda) = 0$, and*

$$R_P(D) = \tfrac{1}{2} \log \frac{\sigma_U^2}{D - \lambda\sigma_V^2}.$$

For independent Gaussian components under weighted squared-error distortion, the predictive rate-distortion optimum follows the classical reverse water-filling solution (2; 6). The activation boundary is $D - \lambda^\star\sigma_V^2 = \lambda^\star\sigma_V^2$, giving $\lambda^\star = D/(2\sigma_V^2)$; below it, V is inactive and the remaining distortion budget $D - \lambda\sigma_V^2$ is allocated to the active U-component (Appendix D.3).

5.3 PHASE-SEPARATION THEOREM

Let $r = I(Z; V \mid U)$.

Lemma 10 (Free epistemic rate bound)**.** *Fix $0 < D < \sigma_U^2$ and $0 < \lambda < \lambda^\star$. For any representation Z with $D_P^*(Z) \le D$,*

$$I(Z; H) \ge F(r) = r + \tfrac{1}{2} \log \frac{\sigma_U^2}{D - \lambda\sigma_V^2 e^{-2r}}.$$

F is strictly convex for $r \ge 0$, hence $F(r) \ge R_P(D) + \kappa_\lambda r$ with

$$\kappa_\lambda = 1 - \frac{\lambda\sigma_V^2}{D - \lambda\sigma_V^2} > 0.$$

Hence if $I(Z; H) \le R_P(D) + \epsilon$, then $I(Z; V \mid U) \le \epsilon/\kappa_\lambda$, and therefore $J_E^+(D) = 0$.

The Gaussian conditional rate-distortion converse gives $D_V \ge \sigma_V^2 e^{-2r}$; substituting into the predictive constraint and coding U at distortion D_U yields $I(Z; H) \ge F(r)$ (full proof in Appendix D.5). For the separable Gaussian reproduction family with decoders $\hat{U}(Z)$ and $\hat{V}(Z)$, $D_P^*(Z) = D_U^*(Z) + \lambda D_V^*(Z)$; see Appendix D.4.

Theorem 11 (Gaussian phase separation)**.** *Assume $0 < D < \sigma_U^2$ and $0 < \lambda < \lambda^\star = D/(2\sigma_V^2)$. For the log-Fisher distortion*

$$d_E = (\log \Gamma_\Theta - \log \widehat{\Gamma}_\Theta)^2 = c^2(V - \hat{V})^2, \qquad 0 < D_E < c^2\sigma_V^2,$$

we have $J_E^+(D) = 0$ while

$$R_{E|U}(D_E) = \tfrac{1}{2} \log \frac{c^2\sigma_V^2}{D_E} > 0.$$

Hence $\Delta R_E(D, D_E) > 0$. Moreover,

$$\Delta R_E(D, D_E) \ge \frac{\kappa_\lambda}{2} \log \frac{c^2\sigma_V^2}{D_E}, \qquad \kappa_\lambda = 1 - \frac{\lambda\sigma_V^2}{D - \lambda\sigma_V^2} > 0.$$

Because V is recoverable from K_Y, the loss is caused by finite-rate allocation rather than predictive unobservability:

$$\boxed{V \in K_Y, \quad R_{V,\mathrm{opt}} = 0, \quad R_{E|U}(D_E) > 0.}$$

The three terms respectively rule out projection loss, identify the predictive compression bottleneck, and certify residual epistemic demand. Separation is therefore induced by finite-rate allocation, not absence of V from the predictive law (Appendix D.6).

5.4 EXACT EXCESS RATE AND ACHIEVABILITY

Theorem 12 (Exact Gaussian excess rate)**.** *Under the assumptions of Theorem 11, put* $t = D_E/c^2 \in (0, \sigma_V^2)$*. Then*

$$R_{P,E}(D, D_E) = \frac{1}{2}\log\frac{\sigma_U^2}{D - \lambda t} + \frac{1}{2}\log\frac{\sigma_V^2}{t},$$

$$\boxed{\Delta R_E(D, D_E) = \frac{1}{2}\log\left[\frac{\sigma_V^2(D - \lambda\sigma_V^2)}{t(D - \lambda t)}\right] > 0.}$$

Independent Gaussian test channels achieve this rate.

Proof and construction details are in Appendix I.1.

Corollary 13 (Quantitative fidelity degradation)**.** *If* $D_P^*(Z) \le D$ *and* $I(H; Z) \le R_P(D) + \epsilon$*, then*

$$D_E^*(Z) \ge c^2\sigma_V^2 \exp(-2\epsilon/\kappa_\lambda).$$

Proof and construction details are in Appendix I.2. This all-encoder statement is architecture-independent: every representation within an ϵ-neighborhood of the predictive optimum obeys the same degradation bound.

5.5 A DECISION CONSEQUENCE WITH AN EXPLICIT MARGIN

Consider two future scalar measurements of Θ. Conditional on V, action 1 has noise variance s_L^2 when $V \ge 0$ and s_H^2 otherwise; action 2 reverses these assignments, where $0 < s_L^2 < s_H^2$. Their information gains are $u_L = \frac{1}{2}\log(1 + \sigma_\Theta^2/s_L^2)$ and $u_H = \frac{1}{2}\log(1 + \sigma_\Theta^2/s_H^2)$. Put $m = u_L - u_H > 0$. This is a separate controlled experiment; it does not assert that arbitrary smooth utilities have constant margins. Its complete predictive object includes the earlier Gaussian predictive family and therefore identifies V.

Proposition 14 (Near-optimal compression forces sensing regret)**.** *For any action selector observing* Z *under the Gaussian constraints* $D_P^*(Z) \le D$ *and* $I(H; Z) \le R_P(D) + \epsilon$*, with* $\epsilon \ge 0$*,*

$$\mathrm{Regret}(Z) \ge m\left[\frac{1}{2} - \sqrt{\frac{\epsilon}{2\kappa_\lambda}}\right]_+.$$

The displayed lower bound is nontrivial exactly when $\epsilon < \kappa_\lambda/2$. Proof and construction details are in Appendix I.3. At the predictive-optimal boundary the regret approaches $m/2$. Under the achieving Gaussian context channel of Theorem 12, selecting from the sign of $\hat{V}$ instead has regret $m \arccos\sqrt{1 - t/\sigma_V^2}/\pi < m/2$. Decision gain here follows from an explicit nonzero utility margin; it is not implied by epistemic fidelity alone.

5.6 ORDER PARAMETER

Corollary 15 (Closed-form order parameter)**.** *Let* $R_{V,\mathrm{opt}}(\lambda)$ *be the rate allocated to* V *by the Gaussian predictive-optimal test channel. Then*

$$R_{V,\mathrm{opt}}(\lambda) = \begin{cases} 0, & 0 < \lambda \le D/(2\sigma_V^2), \\ \frac{1}{2}\log\dfrac{2\lambda\sigma_V^2}{D}, & \lambda > D/(2\sigma_V^2). \end{cases}$$

At equality the context mode is exactly at activation and still carries zero rate: $R_{V,\mathrm{opt}}(\lambda^\star) = 0$*.*

At $\lambda^\star = D/(2\sigma_V^2)$, the sensing-quality mode activates sharply: below it V receives zero predictive rate; above it the rate grows logarithmically. Related Gaussian bottleneck transitions exist (7), but here the activated variable controls future Fisher information (Figure 2).

6 SEPARATION UNDER NATURAL PREDICTIVE KL

The coordinate-weighted theorem alone does not establish separation under a predictive likelihood loss. We now derive the natural geometry and a separate global result. Write $n(v) = \sigma_0^2 e^{cv}$, $M = \sigma_\Theta^2 > 0$, and $q_M(v) = M + n(v)$. Take $N_P \sim \mathcal{N}(0, \tau^2)$, $\tau^2 > 0$, independent of all other variables. Then

$$K_{u,v} = \mathcal{N}(u, \tau^2) \otimes \mathcal{N}(0, q_M(v)), \qquad \Gamma_\Theta(v) = 1/n(v).$$

For every finite M the predictive variance identifies v, so this remains Type-II. Here M is target uncertainty in the predictive marginal, not a predictive-distortion weight.

6.1 NATURAL MARGINAL GEOMETRY

Proposition 16 (Predictive and epistemic sensitivity)**.** *For fixed u, v, the forward predictive KL satisfies*

$$\mathrm{D}_{\mathrm{KL}}(K_{u,v} \| K_{u+\delta_u, v+\delta_v}) = \frac{\delta_u^2}{2\tau^2} + \frac{1}{4}\left(\frac{c\, n(v)}{M + n(v)}\right)^2 \delta_v^2 + o(\delta_v^2).$$

Its Fisher metric in the v direction is

$$g_V(v) = \frac{1}{2}\left(\frac{c\, n(v)}{M + n(v)}\right)^2, \qquad g_U = \tau^{-2},$$

whereas squared log-Fisher distortion has exact quadratic coefficient c^2. Thus the ratio of the two quadratic coefficients is $4(1 + M/n(v))^2$. In locally normalized coordinates the predictive weight is

$$\lambda_{\mathrm{nat}}(v) = g_V(v)/g_U = \frac{\tau^2 c^2}{2}\left(\frac{n(v)}{M + n(v)}\right)^2.$$

Proof and construction details are in Appendix I.4. When $M \gg n(v)$, predictive KL discounts context variation while log-Fisher sensitivity remains fixed. Intuitively, uncertainty in the latent target dominates the marginal predictive variance, so changing measurement noise has little effect on the marginal forecast even though it still changes the informativeness of the measurement about Θ. This supplies the likelihood-based counterpart of the weighted Gaussian allocation mechanism. The local expansion is not used as a global reverse-water-filling argument; the next theorem instead uses a global expected-KL bound.

6.2 GLOBAL SEPARATION AT POSITIVE PREDICTIVE RATE

Predictive decoders here are product Gaussians $K_{\hat{u},\hat{v}}$ under exact forward KL; denote the corresponding RDFs by R_P^{KL} and $R_{P,E}^{\mathrm{KL}}$. The finite-M excess-rate theorem does not cover arbitrary mixture decoders, although the suppression mechanism persists: for the unrestricted Bayes decoder the minimum expected KL is $I(Y^+; H \mid Z)$, while

$$I(V; Y_E^+) \le \frac{\mathrm{Var}(n(V))}{4M^2} = \delta_M \to 0.$$

Thus the predictive marginal itself becomes asymptotically insensitive to sensing context as $M \to \infty$, although the present closed-form finite-M excess-rate bounds use the Gaussian predictive family. Let

$$\bar{n} = \mathbb{E}\, n(V) = \sigma_0^2 e^{c^2\sigma_V^2/2}, \qquad C = \mathrm{Var}(n(V)) = \sigma_0^4\left(e^{2c^2\sigma_V^2} - e^{c^2\sigma_V^2}\right),$$

$$\delta_M = \frac{C}{4M^2}, \qquad G_M(D) = \frac{1}{2}\log\frac{D}{D - \delta_M}.$$

Theorem 17 (Global natural-KL excess rate)**.** *Suppose $0 < \delta_M < D < \sigma_U^2/(2\tau^2)$ and $0 < D_E < c^2\sigma_V^2$. Put $r_E = \frac{1}{2}\log(c^2\sigma_V^2/D_E)$. Then*

$$\frac{1}{2}\log\frac{\sigma_U^2}{2\tau^2 D} \le R_P^{\mathrm{KL}}(D) \le \frac{1}{2}\log\frac{\sigma_U^2}{2\tau^2(D - \delta_M)},$$

M	D_P^{pred}	D_P^{epi}	D_E^{pred}	D_E^{epi}	ΔR_{KL}	$\Delta \hat{I}$	Err_{pred}	Err_{epi}
2	.0788	.0489	.6200	.1226	1.099	1.053	.499	.137
8	.0867	.0459	.6200	.1276	1.185	1.159	.499	.140
32	.0856	.0456	.6200	.1283	1.182	1.156	.499	.140
128	.0856	.0455	.6200	.1284	1.182	1.156	.499	.140

Table 1: Ten-seed test replication after validation-only locking. ΔR_{KL} is the posterior-to-prior KL gap and $\Delta \hat{I}$ the empirical-mixture mutual-information gap; errors are binary sensing-action errors. Rates are total nats.

and every feasible encoder with $I(H;Z) \le R_P^{\text{KL}}(D) + \epsilon$ obeys

$$I(V;Z \mid U) \le G_M(D) + \epsilon, \qquad D_E^*(Z) \ge c^2 \sigma_V^2 e^{-2(G_M(D)+\epsilon)}.$$

Moreover,

$$\boxed{[r_E - G_M(D)]_+ \le \Delta R_E^{\text{KL}}(D, D_E) \le r_E + G_M(D).}$$

In particular, strict separation at positive primary predictive rate is guaranteed when

$$M > \max\left\{ \frac{\sqrt{C}}{2\sqrt{D}}, \frac{\sqrt{C}}{2\sqrt{D(1 - D_E/(c^2\sigma_V^2))}} \right\}.$$

At fixed $D, D_E, \tau^2, \sigma_U^2, \sigma_V^2, c, \sigma_0^2$, as $M \to \infty$,

$$\Delta R_E^{\text{KL}}(D, D_E) \to r_E, \qquad J_{E,\text{KL}}^+(D) \le G_M(D) \to 0.$$

Proof and construction details are in Appendix I.5.

What this establishes. The weighted model gives exact zero residual context rate; the natural-KL theorem gives a finite-M bound under the native likelihood loss without treating its local quadratic expansion as globally exact. The fixed-noise experiment retains log-Fisher sensitivity as M varies; other epistemic functionals may behave differently.

7 Numerical and Neural Validation

Analytical checks. Monte Carlo channels match the exact rates within sampling error, and quadrature verifies the natural-KL bounds for $M \in \{2, 8, 32, 128\}$ (Appendix G).

Mechanism-isolation benchmark and locked selection. The neural benchmark is designed to isolate the rate-allocation mechanism predicted by the theory, not to claim end-to-end world-model performance. It uses the same natural-KL source as Section 6, with $D_P^{\text{req}} = 0.1$ and $D_E^{\text{req}} = 0.16$. For each M, a common (μ, β) discovery pool is evaluated over three training seeds using validation data only. For configuration $c = (\mu, \beta)$, let $\overline{R}_{\text{KL}}(c)$ denote mean validation rate. We lock

$$\widehat{c}_P = \arg\min_c \overline{R}_{\text{KL}}(c) \text{ s.t. } \max_s D_P(c,s) \le D_P^{\text{req}},$$

and

$$\widehat{c}_{P,E} = \arg\min_c \overline{R}_{\text{KL}}(c) \text{ s.t. } \max_s D_P(c,s) \le D_P^{\text{req}}, \ \max_s D_E(c,s) \le D_E^{\text{req}}.$$

The finite-candidate statistic $\Delta \widehat{R}_{\text{pool}} = \overline{R}_{\text{KL}}(\widehat{c}_{P,E}) - \overline{R}_{\text{KL}}(\widehat{c}_P)$ is not identified with the unrestricted theoretical ΔR_E.

After locking, ten-seed retraining gives $40/40$ predictive-selected runs with $D_P \le 0.1$ but $0/40$ with $D_E \le 0.16$; all $40/40$ epistemic-aware runs meet both. This is a validation-locked finite-family separation, not a lower bound on $D_E^*(Z)$ or an estimate of unrestricted ΔR_E.

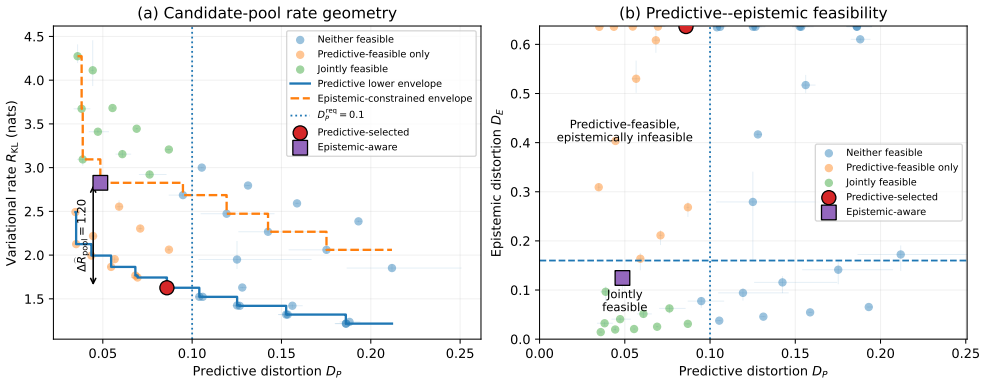

Figure 1: **Candidate-pool geometry for $M = 32$.** Three-seed discovery means with seed ranges. **(a)** Predictive distortion versus variational rate; highlighted points are the predictive and joint minima. **(b)** Predictive versus epistemic distortion with prespecified feasibility thresholds. Ten-seed replication appears in Table 1.

Rate accounting and sensing consequence. Since $R_{\text{KL}} = I(H;Z) + \mathrm{D}_{\text{KL}}(q(z)\|p_0(z))$, the raw variational-KL gap is not the unrestricted ΔR_E. Empirical-mixture estimates attribute 1.05–1.16 nats to mutual information, while aggregate-prior mismatch changes by at most 0.023 nats. Because epistemic-aware models also have lower predictive distortion, $\Delta\hat{I}$ should not be compared one-to-one with $r_E = \log 2 \simeq 0.693$: some rate can support predictive overfidelity. We therefore interpret it as mechanism evidence, not an estimator of unrestricted excess rate. Predictive-selected sensing decisions are near chance (0.499 error), whereas epistemic-aware models achieve 0.137–0.140. Architecture, objective, confidence intervals, regret values, and Monte Carlo sensitivity are reported in Appendix E.

8 Related Work and Discussion

Predictive representations and world models compress history for prediction or planning (15; 20; 4; 13; 16; 17; 11; 12), while classical RD work studies multiple fidelity constraints and side information (19; 2; 6; 9; 18; 10; 23; 8). Experimental design instead values observations for inference (14; 5; 1). The closest comparison is Weiss (22): squared-error RD compression can discard high-frequency content that is weak in reconstruction energy but strong in localization Fisher information, establishing a reconstruction–estimation mismatch. We instead study a *predictive–epistemic* mismatch: J_E^+ is evaluated at the predictive RD boundary, the secondary target is action-conditioned future measurement quality, and

$$V \in K_Y, \qquad R_{V,\text{opt}} = 0, \qquad R_{E|U}(D_E) > 0.$$

Thus the relevant context is present in the full predictive law but receives zero or vanishing rate at the predictive optimum. The secondary target is a parameter-indexed future experiment, not an arbitrary auxiliary label.

Limitations. The exact Gaussian result assumes independent sources; the finite-M natural-KL excess-rate theorem uses product-Gaussian predictive reconstructions. The neural benchmark is static and controlled, testing the predicted allocation mechanism rather than end-to-end sequential control. Unrestricted finite-M decoders, dynamic environments, and policy-coupled sensing remain open.

9 Conclusion

Even when epistemic information is identifiable from the full predictive law, finite-rate predictive optimality can suppress distinctions needed to decide what to observe next. J_E^+, exact Gaussian excess rates, global natural-KL bounds, and locked neural experiments characterize this predictive–epistemic separation.

ACKNOWLEDGMENTS

Omitted for anonymous review.

AI USE STATEMENT

Generative tools were used only for language editing, consistency checking, and code assistance. All claims, derivations, experiments, and reported numerical results were selected and verified by the authors.

REFERENCES

[1] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 1988.

[2] Toby Berger. *Rate Distortion Theory*. Prentice-Hall, 1971.

[3] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 2001.

[4] Byron Boots, Sajid M. Siddiqi, and Geoffrey J. Gordon. Closing the learning–planning loop with predictive state representations. *IJRR*, 2011.

[5] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 1995.

[6] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

[7] Felix Creutzig, Amir Globerson, and Naftali Tishby. Past–future information bottleneck in dynamical systems. *Physical Review E*, 2009.

[8] Abbas El Gamal and Thomas M. Cover. Achievable rates for multiple descriptions. *IEEE Transactions on Information Theory*, 1982.

[9] William H. R. Equitz and Thomas M. Cover. Successive refinement of information. *IEEE Transactions on Information Theory*, 1991.

[10] Robert M. Gray and Aaron D. Wyner. Source coding for a simple network. *Bell System Technical Journal*, 1974.

[11] David Ha and Juergen Schmidhuber. World models. *arXiv:1803.10122*, 2018.

[12] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, 2019.

[13] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998.

[14] Dennis V. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 1956.

[15] Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *NeurIPS*, 2002.

[16] Pascal Poupart and Craig Boutilier. Value-directed belief state approximation for pomdps. In *UAI*, 2000.

[17] Pascal Poupart and Craig Boutilier. Value-directed compression of pomdps. In *NeurIPS*, 2002.

[18] Bixio Rimoldi. Successive refinement of information: Characterization of the achievable rates. *IEEE Transactions on Information Theory*, 1994.

[19] Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record*, 1959.

[20] Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations. In *UAI*, 2004.

[21] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conference*, 1999.

[22] Amir Weiss. On the suboptimality of rate–distortion-optimal compression: Fundamental accuracy limits for distributed localization. *arXiv preprint arXiv:2603.23006*, 2026.

[23] Aaron D. Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 1976.

A Observation Experiment Foundations

A.1 Measurable factorization criterion

Assume the experiment space, the predictive-distribution space, and the epistemic consequence space are standard Borel, and that $T_{P,\pi}$ is a measurable quotient map. Then there exists a measurable ϕ_Y with $\Phi = \phi_Y \circ T_{P,\pi}$ if and only if

$$T_{P,\pi}(\mathfrak{E}_1) = T_{P,\pi}(\mathfrak{E}_2) \Rightarrow \Phi(\mathfrak{E}_1) = \Phi(\mathfrak{E}_2)$$

for all $\mathfrak{E}_1, \mathfrak{E}_2$ in the domain of Φ.

Proof. If $\Phi = \phi_Y \circ T_{P,\pi}$, the implication is immediate. Conversely, if the implication holds, Φ is constant on the fibers of $T_{P,\pi}$. By the universal property of the quotient, Φ factors as $\Phi = \phi_Y \circ T_{P,\pi}$ for a unique set-theoretic map ϕ_Y. Measurability of ϕ_Y follows from the standard Borel and measurable-quotient assumptions. □

A.2 Type-I vs Type-II examples

A Type-I example is any setting where Φ depends on the full likelihood family in a way that the marginal does not determine, e.g., Fisher information about a parameter whose identification requires the conditional family structure. A Type-II example is the Gaussian construction of Section 5.

A.3 Regularity assumptions

Standard-Borel spaces, measurable quotient $T_{P,\pi}$, and existence of regular conditional distributions are assumed throughout.

B Collapse Proofs

B.1 Concave modulus

Definition 18. *ϕ_Y is distortion-compatible with concave modulus ω if ω is nondecreasing and concave and $d_E(\phi_Y(P), \phi_Y(Q)) \le \omega(d_P(P, Q))$ for all P, Q.*

B.2 Expected modulus and regularization

Define

$$\bar{\omega}(\delta) = \sup_{\mathbb{E}[d_P(P,Q)] \le \delta} \mathbb{E}[d_E(\phi_Y(P), \phi_Y(Q))], \qquad \bar{\omega}^+(\delta) = \inf_{\eta > 0} \bar{\omega}(\delta + \eta).$$

Both are nondecreasing. We distinguish two zero-distortion conditions:

- *Exact zero-distortion compatibility*: $d_P(P, Q) = 0 \Rightarrow d_E(\phi_Y(P), \phi_Y(Q)) = 0$, equivalently $\bar{\omega}(0) = 0$.
- *Robust zero-distortion compatibility*: $\bar{\omega}^+(0) = 0$.

The collapse theorem with near-optimal decoders requires the robust version.

B.3 Full proof of Theorem 5

Proof. Let Z satisfy $D_P^*(Z) \le \delta$, and for any $\eta > 0$ let $\widehat{K}_P$ be η-optimal, i.e.,

$$\mathbb{E}[d_P(K_Y(H, A), \widehat{K}_P(Z, A))] \le D_P^*(Z) + \eta \le \delta + \eta.$$

Define $\widehat{E} = \phi_Y \circ \widehat{K}_P$. Then

$$\mathbb{E}[d_E(E(H, A), \widehat{E}(Z, A))] \le \mathbb{E}[d_E(\phi_Y(K_Y(H, A)), \phi_Y(\widehat{K}_P(Z, A)))] \le \bar{\omega}(\delta + \eta).$$

Since D_E^* is the infimum over all epistemic decoders, $D_E^*(Z) \le \bar{\omega}(\delta + \eta)$. Taking $\eta \downarrow 0$, $D_E^*(Z) \le \bar{\omega}^+(\delta)$. If $D_E^{\mathrm{req}} \ge \bar{\omega}^+(D_P^{\mathrm{req}})$, then any representation feasible for $D_P^*(Z) \le D_P^{\mathrm{req}}$ is feasible for $D_E^*(Z) \le D_E^{\mathrm{req}}$. Hence $R_{P,E}(D_P^{\mathrm{req}}, D_E^{\mathrm{req}}) = R_P(D_P^{\mathrm{req}})$. At $D_P^{\mathrm{req}} = 0$, robust zero-distortion compatibility gives $R_{P,E}(0, 0) = R_P(0)$. □

C General Separation Proofs

C.1 Non-attainment and J_E^+

For a predictively feasible sequence Z_n with $D_P^*(Z_n) \le D_P$ and $I(Z_n; H) \to R_P(D_P)$ from above,

$$\limsup_{n\to\infty} I(Z_n; V \mid U) \le J_E^+(D_P).$$

The finite-n bound $I(Z_n; V \mid U) \le J_E^+(D_P)$ need not hold.

C.2 Conditional epistemic RDF

Define

$$D_E^*(Z, U) = \inf_{\widehat{E}} \mathbb{E}[d_E(E(H, A), \widehat{E}(Z, U, A))]$$

and $R_{E|U}(D_E) = \inf\{I(Z; V \mid U) : D_E^*(Z, U) \le D_E\}$. The decoder $\widehat{E}(Z, U, A)$ may depend on U.

C.3 Full proof of Theorem 8

Proof. Suppose $R_{P,E}(D_P, D_E) = R_P(D_P)$. Then there exist Z_n with $I(Z_n; H) \to R_P(D_P)$, $D_P^*(Z_n) \le D_P$, $D_E^*(Z_n) \le D_E$. For all large n, $I(Z_n; H) \le R_P(D_P) + \epsilon$, so

$$I(Z_n; V \mid U) \le J_E(R_P(D_P) + \epsilon, D_P).$$

Taking $\epsilon \to 0$,

$$\limsup_{n} I(Z_n; V \mid U) \le J_E^+(D_P) < R_{E|U}(D_E).$$

On the other hand, $D_E^*(Z_n) \le D_E$ implies $D_E^*(Z_n, U) \le D_E$, since a decoder depending only on Z_n is admissible when the decoder may also use U. Therefore, by definition of $R_{E|U}(D_E)$,

$$I(Z_n; V \mid U) \ge R_{E|U}(D_E).$$

Contradiction. Hence $\Delta R_E > 0$. □

C.4 Exact discrete sanity example

Let U, V be independent uniform bits with $H = (U, V)$, $Y_P^+ = U \oplus N_P$, $Y_E^+ = V \oplus N_E$, d_P dependent only on Y_P^+, $\Phi(\mathfrak{E}_Y) = \mathbb{P}(Y_E^+ \mid V)$. Then

$$R_P(0) = \log 2, \quad R_{E|U}(0) = \log 2, \quad J_E^+(0) = 0, \quad R_{P,E}(0, 0) = 2 \log 2, \quad \Delta R_E = \log 2.$$

D Gaussian Theory

D.1 Full proof of the Type-II validity

Proof. From $Y_E^+ = \Theta + \sigma(V) Z_E$, with $\Theta \perp (U, V)$,

$$Y_E^+ \mid V \sim \mathcal{N}(0, \sigma_\Theta^2 + \sigma_0^2 e^{cV}).$$

The map $v \mapsto \sigma_\Theta^2 + \sigma_0^2 e^{cv}$ is strictly monotone when $c \neq 0$, so it is invertible. The ordinary predictive marginal for the pair $(U, V) = (u, v)$ contains the conditional law of Y_E^+ given $(U, V) = (u, v)$, which is exactly this Gaussian. Therefore

$$V = \frac{1}{c} \log \frac{\mathrm{Var}_{K_Y}(Y_E^+) - \sigma_\Theta^2}{\sigma_0^2},$$

and consequently

$$\Gamma_\Theta(V) = \sigma_0^{-2} e^{-cV} = \frac{1}{\mathrm{Var}_{K_Y}(Y_E^+) - \sigma_\Theta^2} = \phi_Y(K_Y).$$

□

D.2 Predictive-family reproduction alphabet

The coordinate form $d_P = (U - \hat{U})^2 + \lambda(V - \hat{V})^2$ is induced by the identifiable parameterization of the Gaussian predictive family $K_Y^{(U,V)}$.

D.3 Reverse water-filling derivation

Derivation of Proposition 9. For active dimensions in the weighted Gaussian source with distortion weights $(1, \lambda)$, reverse water-filling assigns

$$D_U^\star = \min(\sigma_U^2, \theta), \qquad D_V^\star = \min(\sigma_V^2, \theta/\lambda),$$

with θ chosen so that $D_U^\star + \lambda D_V^\star = D$. The V-component is inactive when $\theta/\lambda \geq \sigma_V^2$, i.e., $\theta \geq \lambda\sigma_V^2$. At the activation boundary, the constraint gives

$$D - \lambda\sigma_V^2 = \lambda\sigma_V^2,$$

hence $\lambda^\star = D/(2\sigma_V^2)$ and $\theta^\star = D/2$. For $\lambda < \lambda^\star$, $D_V^\star = \sigma_V^2$ and $D_U^\star = D - \lambda\sigma_V^2$, giving

$$R_P(D) = \tfrac{1}{2}\log\frac{\sigma_U^2}{D - \lambda\sigma_V^2}.$$

□

D.4 D_U^*, D_V^* separability and conditional MMSE

Definition 19.

$$D_U^*(Z) = \inf_g \mathbb{E}[(U - g(Z))^2], \qquad D_V^*(Z) = \inf_h \mathbb{E}[(V - h(Z))^2],$$

$$D_{V|U}^*(Z) = \inf_h \mathbb{E}[(V - h(Z, U))^2].$$

Both $D_U^*(Z)$ and $D_V^*(Z)$ use decoders depending only on Z; $D_{V|U}^*(Z)$ allows the decoder to depend on (Z, U).

Separability and conditional MMSE. The separability $D_P^*(Z) = D_U^*(Z) + \lambda D_V^*(Z)$ follows because the two terms in

$$\mathbb{E}[(U - g(Z))^2 + \lambda(V - h(Z))^2]$$

depend on g and h separately. Since $h(Z)$ is also a legal conditional decoder $h'(Z, U) = h(Z)$, we have $D_{V|U}^*(Z) \leq D_V^*(Z)$. The Gaussian conditional rate-distortion converse gives

$$r = I(Z; V \mid U) \geq \tfrac{1}{2}\log\frac{\sigma_V^2}{D_{V|U}^*(Z)},$$

hence $D_{V|U}^*(Z) \geq \sigma_V^2 e^{-2r}$. Combined with $D_V^*(Z) \geq D_{V|U}^*(Z)$,

$$D_V^*(Z) \geq \sigma_V^2 e^{-2r}.$$

This is the inequality used in Appendix D.5. □

D.5 Full proof of Lemma 10

Proof. By Appendix D.4, $D_V^*(Z) \geq \sigma_V^2 e^{-2r}$. The predictive constraint then gives

$$D_U^*(Z) \leq D - \lambda D_V^*(Z) \leq D - \lambda\sigma_V^2 e^{-2r}.$$

Coding U at distortion $D_U^*(Z)$ requires

$$I(Z; U) \geq \tfrac{1}{2}\log\frac{\sigma_U^2}{D_U^*(Z)} \geq \tfrac{1}{2}\log\frac{\sigma_U^2}{D - \lambda\sigma_V^2 e^{-2r}}.$$

Since $U \perp V$,

$$I(Z; H) = I(Z; U) + I(Z; V \mid U) \geq F(r).$$

Compute

$$F'(r) = 1 - \frac{\lambda\sigma_V^2 e^{-2r}}{D - \lambda\sigma_V^2 e^{-2r}}, \qquad F''(r) = \frac{2D\lambda\sigma_V^2 e^{-2r}}{(D - \lambda\sigma_V^2 e^{-2r})^2} > 0.$$

Thus F is convex and

$$F(r) \geq F(0) + F'(0)r = R_P(D) + \kappa_\lambda r,$$

with $\kappa_\lambda = F'(0) = 1 - \lambda\sigma_V^2/(D - \lambda\sigma_V^2) > 0$. If $I(Z;H) \leq R_P(D) + \epsilon$, then $R_P(D) + \kappa_\lambda r \leq R_P(D) + \epsilon$, so $r \leq \epsilon/\kappa_\lambda$. Taking $\epsilon \to 0$ gives $J_E^+(D) = 0$. □

D.6 Full proof of Theorem 11

Proof. By Appendix D.1, $\Gamma_\Theta = \phi_Y(K_Y)$, so Assumption 4 holds. By Lemma 10, $J_E^+(D) = 0$. By the Gaussian RDF for the source cV with variance $c^2\sigma_V^2$,

$$R_{E|U}(D_E) = \tfrac{1}{2}\log\frac{c^2\sigma_V^2}{D_E} > 0$$

whenever $0 < D_E < c^2\sigma_V^2$. Hence $J_E^+(D) = 0 < R_{E|U}(D_E)$, so by Theorem 8, $\Delta R_E > 0$. For the lower bound, any representation satisfying $D_E^*(Z) \leq D_E$ requires $r = I(Z;V \mid U) \geq R_{E|U}(D_E)$. By Lemma 10, transmitting r nats about V in addition to the predictive-optimal allocation costs at least $\kappa_\lambda r$. Therefore

$$R_{P,E}(D, D_E) \geq R_P(D) + \kappa_\lambda R_{E|U}(D_E),$$

giving

$$\Delta R_E(D, D_E) \geq \kappa_\lambda R_{E|U}(D_E) = \frac{\kappa_\lambda}{2}\log\frac{c^2\sigma_V^2}{D_E}.$$

□

D.7 Proof of Corollary 15

Proof. For $0 < \lambda \leq \lambda^\star$, Proposition 9 gives $R_{V,\mathrm{opt}}(\lambda) = 0$. For $\lambda > \lambda^\star$, both dimensions are active, and the reverse water-filling water level is $\theta = D/2$. The V-component distortion is $D_V^\star = \theta/\lambda = D/(2\lambda)$, and the rate assigned to V is

$$\tfrac{1}{2}\log\frac{\sigma_V^2}{D_V^\star} = \tfrac{1}{2}\log\frac{2\lambda\sigma_V^2}{D}.$$

Continuity at $\lambda = \lambda^\star$ gives $R_{V,\mathrm{opt}}(\lambda^\star) = 0$. □

D.8 Phase-transition visualization

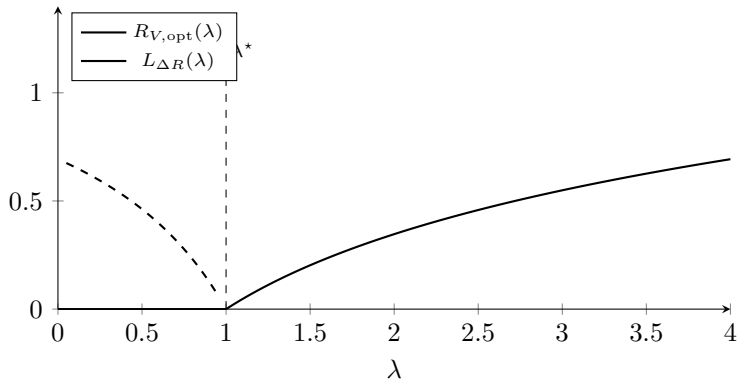

Figure 2: Normalized example with $\sigma_U^2 = 4$, $\sigma_V^2 = 1$, $D = 2$, so $\lambda^\star = 1$ and $D < \sigma_U^2$. Below $\lambda^\star$, $R_{V,\mathrm{opt}}(\lambda) = 0$; above, it is $\frac{1}{2}\log\lambda$. The dashed curve is the theorem-guaranteed excess-rate lower bound $L_{\Delta R}(\lambda) = C_E \cdot \frac{2(1-\lambda)}{2-\lambda}$ with $C_E = \frac{1}{2}\log 4 = \log 2 \approx 0.6931$.

D.9 Predictive-divergence interpretation

Proposition 16 derives the local marginal-KL geometry with its exact factor of $1/4$. Theorem 17 gives a separate global bound for the same natural loss. Neither result assumes that a local quadratic approximation is globally equivalent to KL. Figure 2 contains analytical curves only.

E Stochastic Neural Experimental Protocol and Results

E.1 Primary natural-KL benchmark

The executed neural benchmark uses the scalar source from Section 6 with

$$\sigma_U^2 = 4, \qquad \sigma_V^2 = 1, \qquad \tau^2 = 1, \qquad c = 0.8, \qquad \sigma_0^2 = 0.5,$$

and $M \in \{2, 8, 32, 128\}$. We fix $D_P^{\mathrm{req}} = 0.1$ and $D_E^{\mathrm{req}} = 0.16$. A Gaussian-family predictive decoder outputs $(\hat{u}, \hat{v})$ and is trained with the exact forward observation KL

$$d_P = \frac{(u-\hat{u})^2}{2\tau^2} + \frac{1}{2}\left(\frac{M+n(v)}{M+n(\hat{v})} - 1 - \log\frac{M+n(v)}{M+n(\hat{v})}\right).$$

The epistemic target is $\log \Gamma_\Theta(v) = -\log\sigma_0^2 - cv$. Predictive observation KL and posterior-to-prior regularization KL are kept separate throughout.

E.2 Stochastic representation, probes, and action readout

For the executed binary sensing readout we use $s_L^2 = 0.5$ and $s_H^2 = 2.0$; the optimal action is determined by the sign of V. The encoder is a diagonal-Gaussian stochastic bottleneck, $q_\theta(z \mid H)$, with $d_Z = 4$. The encoder and each decoder use two 64-unit ELU hidden layers; the encoder outputs latent means and log variances, the predictive decoder outputs $(\hat{u}, \hat{v})$, and the epistemic head outputs a scalar. The reference prior is $p_0 = \mathcal{N}(0, I_4)$. For epistemic weight μ and rate weight β, the complete training loss is

$$\mathcal{L}_{\mu,\beta} = \mathbb{E}\Big[d_P(K_{U,V}, K_{\hat{U}(Z),\hat{V}(Z)}) + \mu\big(\log\Gamma_\Theta(V) - g_\psi(Z)\big)^2\Big] + \beta\ \mathbb{E}_H\, \mathrm{D}_{\mathrm{KL}}(q_\theta(Z \mid H)\|p_0).$$

We optimize with AdamW (learning rate 10^{-3}, weight decay 10^{-2}), minibatches of 512, and early stopping on validation loss with patience 15. Training and validation objectives use posterior samples rather than posterior means. After encoder training, each encoder is frozen and a fresh epistemic probe is trained from newly sampled latents. Test fidelity is evaluated with 32 posterior draws per example; a 64-draw rerun is used for Monte Carlo sensitivity. A separate fresh binary action classifier is trained on frozen stochastic latents to predict the sign-dependent optimal sensing action. Its information-gain regret equals action error times the exact action margin.

A posterior-mean probe is not used as a rate–distortion point. Probe error is an achievable decoder error rather than a proof of the optimal $D_E^*(Z)$.

E.3 Discovery pool and locked replication

The discovery pool uses

$$\mu \in \{0, 0.1, 0.3, 1, 3\},$$

and the dense grid

$$\beta \in \{0.06, 0.07746, 0.10, 0.12649, 0.16, 0.19596, 0.24, 0.29394, 0.36\}.$$

For all four M values and three discovery seeds (0,1,2), this gives

$$4 \times 5 \times 9 \times 3 = 540$$

trained candidate models. Locked settings are chosen using validation data only and must meet the relevant empirical predictive and prespecified-probe criteria on all three discovery seeds. The final configurations are

M	predictive-selected (μ, β)	epistemic-aware (μ, β)
2	$(0, 0.12649)$	$(0.3, 0.07746)$
8	$(0, 0.16)$	$(0.3, 0.07746)$
32	$(0, 0.16)$	$(0.3, 0.07746)$
128	$(0, 0.16)$	$(0.3, 0.07746)$.

These settings are then frozen and retrained on ten independent replication seeds, 10–19. Each seed contains one predictive-selected and one epistemic-aware model for each M, for 80 final models. Including discovery, the reported neural study trains 620 models.

E.4 Ten-seed replication statistics

Table 1 reports the main means. Across all 40 predictive-selected replication runs, predictive fidelity is met, but the prespecified epistemic probe never reaches its tolerance. Across all 40 epistemic-aware runs, the prescribed decoders reach both empirical tolerances. The predictive-selected probe MSE is effectively invariant across M at approximately 0.620, near the no-context baseline 0.64. Epistemic-aware probe MSE is 0.123–0.128. Failure of the prescribed probe does not prove $D_E^*(Z) > D_E^{\text{req}}$; it establishes a locked finite-family decoder separation rather than an unrestricted decoder impossibility result.

Paired bootstrap intervals use 20,000 resamples of the ten seed pairs. For the posterior-KL gap, the 95% intervals are

M	95% CI for ΔR_{KL}
2	$[1.054, 1.159]$
8	$[1.152, 1.243]$
32	$[1.149, 1.242]$
128	$[1.148, 1.241]$.

For the empirical-mixture mutual-information gap they are

$$[1.019, 1.089], \quad [1.133, 1.189], \quad [1.131, 1.186], \quad [1.130, 1.186],$$

for $M = 2, 8, 32, 128$, respectively.

E.5 Rate decomposition

For reference prior p_0,

$$R_{\mathrm{KL}} = \mathbb{E}_H \, \mathrm{D}_{\mathrm{KL}}(q(z \mid H) \| p_0(z)) = I(H; Z) + \mathrm{D}_{\mathrm{KL}}(q(z) \| p_0(z)).$$

On an independent reference split we estimate the aggregate posterior as an empirical mixture of encoder conditionals and report

$$\hat{I} = \mathbb{E}_{H,Z}[\log q(Z \mid H) - \log \hat{q}(Z)], \qquad \hat{M} = \mathbb{E}_Z[\log \hat{q}(Z) - \log p_0(Z)].$$

The reference split contains 4096 histories and is disjoint from training, validation, and test data; log mixture densities use all 4096 Gaussian components. Expectations use one posterior draw per reference history and are repeated over the ten locked replication seeds. The mean $\Delta\hat{I}$ values are 1.053, 1.159, 1.156, and 1.156 nats for $M = 2, 8, 32, 128$, while mean aggregate-prior mismatch differences are 0.0234, 0.0067, 0.0071, and 0.0072 nats. The observed posterior-KL increase is therefore dominated by additional encoded information rather than aggregate-prior mismatch.

These quantities remain density-based finite-sample diagnostics. The finite-family neural gap is not identified with the unrestricted theoretical ΔR_E.

E.6 Sensing consequence

For this executed decision experiment we use $s_L^2 = 0.5$ and $s_H^2 = 2.0$. Predictive-selected representations yield mean binary action errors 0.4987, 0.4988, 0.4993, and 0.4994 for increasing M, approximately chance. Epistemic-aware errors are 0.1367, 0.1398, 0.1402, and 0.1402. The corresponding mean information-gain regrets decrease from 0.228, 0.305, 0.335, and 0.343 to 0.0626, 0.0855, 0.0940, and 0.0963, respectively. These results test the explicit context-sensitive decision consequence and do not imply that every epistemic-fidelity improvement must improve arbitrary downstream decisions.

E.7 Monte Carlo sensitivity

Repeating final evaluation with 64 rather than 32 posterior draws changes predictive distortion by 2.17×10^{-4} on average and 1.10×10^{-3} at worst. The corresponding epistemic-MSE changes are 3.14×10^{-4} on average and 1.41×10^{-3} at worst. Thus the reported separation is much larger than posterior-sampling variation.

E.8 Unexecuted extensions

The high-dimensional identifiable linear-prediction construction and smooth multiaction covariance benchmark are retained as extensions rather than evidence for the main claim. They are not mixed with the executed scalar natural-KL neural results.

F Scope Relative to Classical Rate–Distortion Theory

Joint fidelity constraints, Gaussian mode allocation, and Gaussian test channels are established rate–distortion tools. Natural predictive KL connects the sensing interpretation to the marginal observation loss; the global theorem uses a specified Gaussian decoder family and does not assert superiority to classical coding theory. Our analysis specializes them to sensing-quality information recoverable from the complete predictive object and emphasizes all-encoder stability at its primary compression boundary. Successive refinement studies staged fidelity; Gray–Wyner studies common/private descriptions; Wyner–Ziv studies decoder side information. These distinctions do not establish that classical theory cannot express our problem. The substantive claims to assess are the stated boundary bound, its exact specialization, and its implications for future-measurement evaluation. No novelty claim is made for reverse water-filling or the mere definition of a joint RDF.

G Reproduction of Executed Scalar Checks

The following Python script uses NumPy and SciPy. Values shown in the main text were generated with this script; rates are total nats. The natural-KL table uses numerical quadrature only for $\bar{d}_M$, and an analytic moment formula for δ_M.

```
import numpy as np
from scipy.special import roots_hermitenorm
rng=np.random.default_rng(20260912)
x,w=roots_hermitenorm(96);w=w/np.sqrt(2*np.pi)
c=.8;s0=.5;sv=1.;n=s0*np.exp(c*x);nb=s0*np.exp(c*c/2)
C=s0*s0*(np.exp(2*c*c)-np.exp(c*c));D=.1
print('NATURAL', 'C=',C,'RE=',.5*np.log(4))
for M in [2.,8.,32.,128.]:
 r=(M+n)/(M+nb);ex=np.dot(w,.5*(r-1-np.log(r)));bound=C/(4*M*M);g=.5*np.log(D/(D-bound))
 assert ex <= bound
 print('M',M,'expected_KL',ex,'delta_bound',bound,'g',g,'gap_lower',.5*np.log(4)-g)
# Finite difference check for local predictive metric.
M=8.;v=.3;h=1e-4; nv=s0*np.exp(c*v);q=M+nv;qp=M+s0*np.exp(c*(v+h));ratio=q/qp
kl=.5*(ratio-1-np.log(ratio));co=.25*(c*nv/q)**2
print('LOCAL',kl/h**2,co,'relative_error',abs(kl/h**2-co)/co)
# Achieving scalar backward Gaussian channels and empirical information densities.
N=200000; su2=4.;Dp=2.;t=.25
print('CHANNEL N=',N)
for lam in [.1,.5,.9]:
 du=Dp-lam*t
 uh=rng.normal(0,np.sqrt(su2-du),N);nu=rng.normal(0,np.sqrt(du),N);u=uh+nu
 vh=rng.normal(0,np.sqrt(1-t),N);nv=rng.normal(0,np.sqrt(t),N);v=vh+nv
 dp=np.mean(nu**2+lam*nv**2);de=np.mean(nv**2)
 information=.5*np.log(su2/du)+.5*(u*u/su2-nu*nu/du)+.5*np.log(1/t)+.5*(v*v-nv*nv/t)
 rate=np.mean(information);truth=.5*np.log(su2/du)+.5*np.log(1/t)
 err=np.mean((v>=0)!=(vh>=0));etr=np.arccos(np.sqrt(1-t))/np.pi
 print('lambda',lam,'DP',dp,'DE',de,'rate_MC',rate,
       'rate_exact',truth,'action_error',err,'error_exact',etr,
       'rate_SE',information.std(ddof=1)/np.sqrt(N))
 assert abs(dp-Dp)<.03 and abs(de-t)<.005 and abs(rate-truth)<.02 and abs(err-etr)<.005
```

H Matrix-Valued Extension

Let $V \in \mathbb{R}^m$ and $\log \Sigma(V) = \log \Sigma_0 + \sum_i V_i B_i$. Locally,

$$\operatorname{vec}(\log \Gamma_\Theta(V)) \approx GV + b,$$

so epistemic distortion becomes $d_E = (V-\hat{V})^\top G^\top G (V-\hat{V})$. When the source covariance and predictive and epistemic quadratic forms are simultaneously diagonalizable, the problem decomposes into generalized eigen-directions and scalar reverse water-filling suggests a direction-wise extension. Without this structure the coordinates remain coupled; a general matrix-valued theorem is left for future work.

I Proofs of Exact Rate, Geometry, and Decision Results

I.1 Exact Gaussian excess rate

Proof. For any feasible Z, let $d_U = D_U^*(Z)$ and $d_V = D_V^*(Z)$. Epistemic fidelity implies $d_V \le t$, and $d_U + \lambda d_V \le D$. Independence and conditional entropy subadditivity give

$$I(H;Z) \ge I(U;Z) + I(V;Z) \ge \frac{1}{2}\log\frac{\sigma_U^2}{d_U} + \frac{1}{2}\log\frac{\sigma_V^2}{d_V}.$$

Both component distortions are below their source variances. The last expression is bounded below by $g(d_V) = \frac{1}{2}\log[\sigma_U^2/(D-\lambda d_V)] + \frac{1}{2}\log(\sigma_V^2/d_V)$. Its derivative is $(2\lambda d_V - D)/[2d_V(D-\lambda d_V)] < 0$ for $d_V \le t < \sigma_V^2 < D/(2\lambda)$. Thus $g(d_V) \ge g(t)$. For achievability use independent backward channels $U = \hat{U} + N_U$ and $V = \hat{V} + N_V$, with independent Gaussian residual variances $D - \lambda t$ and t, respectively, and reproduction variances $\sigma_U^2 - D + \lambda t$ and $\sigma_V^2 - t$. Their induced forward kernel is an admissible encoder $Z = (\hat{U}, \hat{V})$ and attains both constraints. Subtracting $R_P(D)$ yields the formula. Strict positivity follows from the preceding separation theorem. □

I.2 Quantitative fidelity degradation

Proof. The conditional Gaussian RD converse gives $D_V^*(Z) \ge \sigma_V^2 e^{-2I(V;Z|U)}$. Apply Lemma 10 and $D_E^*(Z) = c^2 D_V^*(Z)$. □

I.3 Sensing regret

Proof. Let $B = \mathbf{1}\{V \ge 0\}$, a balanced bit. If $p_z = \Pr(B = 1 \mid Z = z)$, optimal classification error is $1/2 - \mathbb{E}|p_Z - 1/2|$. Pinsker's inequality and Jensen's inequality imply $\mathbb{E}|p_Z - 1/2| \le \sqrt{I(B;Z)/2}$. Data processing gives $I(B;Z) \le I(V;Z)$. Moreover,

$$I(V;Z \mid U) - I(V;Z) = I(U;V \mid Z) - I(U;V) = I(U;V \mid Z) \ge 0,$$

because $U \perp V$. Hence $I(B;Z) \le I(V;Z \mid U) \le \epsilon/\kappa_\lambda$. Wrong action choice costs exactly m. □

I.4 Natural predictive geometry

Proof. The mean-channel KL is exactly $\delta_u^2/(2\tau^2)$. For the variance channel, let $r = q_M(v)/q_M(v + \delta_v)$. Its KL is $(r - 1 - \log r)/2$. Expanding $r = 1 - [q_M'(v)/q_M(v)]\delta_v + O(\delta_v^2)$ yields the stated coefficient. Log Fisher equals $-\log\sigma_0^2 - cv$, giving coefficient c^2 without approximation. □

I.5 Global natural-KL separation

Proof. Choose a constant variance reconstruction $\hat{v} = \bar{v} = (\log(\bar{n}/\sigma_0^2))/c = c\sigma_V^2/2$. Its expected variance-channel KL is

$$\bar{d}_M = \mathbb{E}\left[\frac{1}{2}\left(\frac{M + n(V)}{M + \bar{n}} - 1 - \log\frac{M + n(V)}{M + \bar{n}}\right)\right] \le \delta_M.$$

Indeed, for $f(r) = r - 1 - \log r$, Taylor's theorem around 1 gives $f(r) \le (r-1)^2/[2\min(r,1)^2]$. Since $r = (M+n)/(M+\bar{n})$, $(M+\bar{n})\min(r,1) = \min(M+n, M+\bar{n}) \ge M$, proving the bound after averaging. The KL mean term forces $D_U^*(Z) \le 2\tau^2 D$. Gaussian RD and the chain rule therefore give

$$I(H;Z) = I(U;Z) + I(V;Z \mid U) \ge L_D + I(V;Z \mid U), \quad L_D = \tfrac{1}{2}\log[\sigma_U^2/(2\tau^2 D)].$$

For the predictive upper bound encode only U through its optimal Gaussian channel at squared error $2\tau^2(D - \delta_M)$ and predict variance $M + \bar{n}$ constantly. This costs $L_D + G_M(D)$. Thus a near-optimal encoder has conditional context information at most $G_M(D) + \epsilon$. Conditional Gaussian RD yields the epistemic-distortion bound. Any jointly feasible encoder has $I(V;Z \mid U) \ge r_E$ and hence joint rate at least $L_D + r_E$. Subtracting the predictive upper bound gives the excess-rate lower bound. For an upper bound, independently encode U at squared error $2\tau^2(D - \delta_M)$ and V at squared error D_E/c^2. The predictive decoder may ignore the latter code and still meet its KL constraint using $\bar{v}$, while a separate epistemic decoder uses it. This joint rate is at most $L_D + G_M(D) + r_E$. Subtract the predictive lower bound. Finally $\delta_M, G_M(D) \to 0$, which gives the limiting statements. Each finite M still identifies V; the literal infinite-variance limit is not an observation distribution. □

J Extended Experimental Protocol and Controls

Experiment 1: exact mechanism and negative controls. Use the scalar Gaussian model and sample independent Gaussian test channels; extend the reported checks across a locked grid. Sweep λ across $\lambda^\star$ and t/σ_V^2 across several fidelity levels. Compare the measured distortions and analytically computed mutual information against Theorem 12, Lemma 10, and Corollary 13. Add $D_E \geq c^2 D/\lambda$ as a distortion-compatible collapse control, since $d_E \leq (c^2/\lambda) d_P$. Numerical agreement tests implementation consistency, not theoretical novelty.

Experiment 2: learned stochastic representations. The primary scalar natural-KL benchmark has been executed with a dense common discovery pool, validation-only locking, and ten independent replication seeds; see Appendix E. Remaining extensions include high-dimensional predictive-relevance sweeps, unrestricted predictive decoder families, and environment/data replications.

Experiment 3: sensing consequence. The explicit-margin binary sensing experiment has been executed on frozen stochastic neural representations; see Section 7 and Appendix E. The smooth multiaction covariance model remains an extension and should report oracle action-margin distributions and margin-stratified regret. A gain in utility prediction is not treated as evidence of a gain in sensing decisions.

Rate accounting and interpretation. For a reference prior p_0, report the total KL in nats:

$$R_{\mathrm{KL}} = \mathbb{E}_H \mathrm{KL}(q(z \mid H) \| p_0(z)) = I(H; Z) + \mathrm{KL}(q(z) \| p_0(z)).$$

Here KL is between encoder posteriors and a reference prior, distinct from the observation-distribution KL in Section 6. A variational KL difference is not a mutual-information difference. Supplement it with an aggregate-posterior estimate and a learned-prior sensitivity analysis. Finite-family gaps do not estimate differences between unrestricted RDFs, even when each individual rate is accurately estimated. Common-pool nonnegativity follows by construction; the experimental targets are its magnitude, robustness, and mechanism.

Selection and uncertainty. Select checkpoints and candidate operating points using validation data only; lock tolerance thresholds before final testing. Report validation feasibility and independent test feasibility separately. Use paired training seeds, publish every seed-level gap, and report sensitivity to the number of seeds and the grid. Nearby fidelity thresholds are correlated evaluations of the same pool, not independent replications.