INTRINSIC INFORMATION RATES OF SUMMARIES IN NEURAL POSTERIOR ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural posterior estimation (NPE) often encodes observations Y into a fixed-dimensional summary S, with the neural posterior conditioning only on S. Any such summary imposes a finite information budget on inference, whereas exact preservation requires infinite information whenever posteriors vary continuously with the data. The budget typically emerges from architecture and training rather than an explicit measurement or choice. Better performance may therefore reflect better encoding or simply a larger budget. In this paper, we make the budget explicit and ask two questions: how large must it be, and how best is it spent? To answer them, we model the summary as a (possibly stochastic) channel from Y to S. For the first question, we show that this budget, measured by the mutual information $I(Y;S)$, grows unboundedly as the posterior concentrates: in regular d-dimensional models, bounded posterior distortion requires $\frac{d}{2}\log n - O(1)$ nats, even when a fixed-dimensional sufficient statistic exists. More generally, the required budget is set by the distribution of the random posterior, and we bound it from above by quantizing posteriors and from below by converses that weight each posterior state by its probability. For the second question, we show that a finite budget is best spent stochastically. Under a common-support condition on the simulator, which covers the continuous observation spaces typical of simulators, a deterministic summary necessarily loses something: at every positive budget too small for exact preservation, some stochastic summary achieves *strictly* lower distortion than every deterministic summary and every time-shared mixture of deterministic summaries within the same budget. Motivated by these results, we insert a binary bottleneck into NPE, measure each trained summary's rate, and compare summaries at matched rate, which removes the budget as a confounder in architecture comparisons. Across four simulator benchmarks and several encoder families, stochastic bottlenecks achieve favorable rate–loss tradeoffs, and architecture rankings change with the budget.

1 INTRODUCTION

Neural posterior estimation (NPE) learns an amortized posterior for parameter θ from simulated parameter–observation pairs (Papamakarios & Murray, 2016; Greenberg et al., 2019; Lueckmann et al., 2017; Radev et al., 2022; Chen et al., 2024). In typical NPE pipelines, a summary network first maps a data set Y to a fixed-dimensional representation S, and the posterior decoder conditions only on S. This makes the summary network a central design choice, in that a summary allows for accurate posterior reconstruction only insofar as it can correctly distinguish data sets that correspond to different posteriors (Blum et al., 2013; Fearnhead & Prangle, 2012; Jiang et al., 2017; Chan et al., 2018; Åkesson et al., 2022).

This paper begins from the premise that every such NPE pipeline already operates under an information budget, usually an implicit one. A trained summary network passes only a limited amount of information about the data to the decoder: finite precision, injected noise, bottleneck layers, and training all limit how much gets through. This budget is not chosen explicitly; instead, it emerges as a byproduct of decisions made for other reasons. The budget is not often measured, either, and the quantities we do report, such as output dimension and parameter count, are poor proxies, since one real-valued coordinate can carry an enormous amount of information, while a wide vector passed

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.

through noise can carry almost none. This budget is also binding, in that whenever posteriors vary continuously with the data, preserving the posterior exactly requires infinite information, and no finite budget suffices (Proposition E.1). The upshot is that the information budget is typically a confounder for experiments on summary network architecture: when one summary network outperforms another, we cannot tell whether that architecture distinguishes the data better, or merely spends a larger information budget.

In this paper, we make the information budget explicit by modeling the summary as a possibly randomized channel $Y \to S$. The budget a summary spends is its rate, the mutual information $I(Y;S)$, and its distortion is the expected KL divergence from the full-data posterior Π_Y to the summary-based posterior Π_S. The intrinsic rate $R^\star(\varepsilon)$ is then the smallest budget sufficient for distortion at most ε. This quantity depends on the prior and the simulator, not on the architecture or optimization procedure used to learn a summary.

We address two questions about this budget:

1. At a given tolerance ε, what properties of a Bayesian experiment govern the intrinsic rate $R^\star(\varepsilon)$?
2. At a prescribed rate R, can a stochastic summary achieve *strictly* lower posterior distortion than every deterministic summary, including their time-sharing mixtures (i.e, a random choice among deterministic summaries, revealed to the decoder)?

For the first question, one part of the answer is that the budget must grow with sample size. This holds even when a fixed-dimensional sufficient statistic exists: the statistic does not grow, but the precision at which it must be transmitted does. In regular d-dimensional models, keeping the posterior error bounded requires at least $\frac{d}{2}\log n - O(1)$ nats (Lemma 2.3), which follows by combining the identity between posterior error and lost parameter information (Tishby et al., 2000; Harremoës & Tishby, 2007) with a classical expansion of Clarke & Barron (1994). Beyond parametric models, what sets the required budget is the collection of posteriors the experiment can produce, together with how often each occurs. The summary's job is, in effect, to tell the decoder which posterior to report. If the experiment produces many well-separated posteriors and visits them often, the summary needs many distinguishable messages. But if most posteriors are rarely visited, the summary can get by with less information. We make this intuition precise in both directions: quantizing the posteriors gives an achievable budget (Theorem 3.2), and converses that weight each posterior by its probability give lower bounds (Theorem 3.3).

For the second question, the answer is that a finite budget is best spent with a randomized summary. This is not obvious, and at first sight looks false: if we limited the number of values a summary may take, rather than its information, deterministic summaries would be optimal (Theorem 3.2). But an information budget behaves differently. The Gaussian model gives a first hint, since its optimal summary at every positive rate is the sufficient sample mean plus deliberate noise (Theorem 3.4). Our main theorem shows that the Gaussian hint is general, under the common-support condition that no observation can rule out any parameter value (Assumption 4.1), which typical noisy simulators satisfy. At every positive budget too small for exact posterior preservation, some stochastic summary achieves strictly lower posterior error than every deterministic summary and every time-shared mixture of them within the same budget (Theorem 4.2). The intuition behind the proof is that we can occasionally replace a code's label, without telling the decoder, by a random draw from the code's own output distribution. This hidden swap saves more rate than it costs in posterior error, and spending the savings on a finer code beats the original. The argument needs neither finite observation spaces nor the existence of optimal codes, and the gap is not a curvature accident, since the two frontiers can share their slope at zero rate and still separate at every positive rate (Proposition B.8). The support assumption is sufficient, rather than a characterization of all experiments with a stochastic advantage; Section 4 gives examples on both sides.

Finally, we test this empirically. Because the information budget cannot be set directly, we measure it: every summary is trained with a binary bottleneck, giving all architectures a common code alphabet and decoder, and each trained model's deployed rate is estimated from its code distribution. Comparisons are then made at matched rate, which removes the budget as a confounder. A conjugate count benchmark enumerates all deterministic count encoders and verifies the second answer exactly, with decoder error far smaller than the gap. Four simulator benchmarks compare stochastic deployment against annealed training with hard deployment, over several encoders and three training seeds. Stochastic bottlenecks trace better rate–loss curves throughout, and architecture rankings change with

the budget, so comparisons at unmatched budgets can pick different winners. These comparisons involve trained models, so optimization and decoder error contribute, whereas the count benchmark isolates the gap exactly.

The paper proceeds from posterior preservation and its parametric consequence (Section 2), through posterior geometry (Section 3), to strict stochastic separation (Section 4). Sections 5 and 5.2 develop and evaluate the NPE comparison protocol. Appendix L gives a detailed account of the related work, including connections to literature on summary selection in likelihood-free inference, approximate sufficiency, information theory and compression.

2 Posterior preservation as an information constraint

Let θ have prior π, let $Y \mid \theta$ follow a simulator, and let S be generated by a measurable channel $q(s \mid y)$. Thus $\theta \to Y \to S$ is a Markov chain. We work on standard Borel spaces, assume regular conditional distributions and measurable divergences, and write $\Pi_Y = P_{\theta|Y}$ and $\Pi_S = P_{\theta|S}$. Although $I(\theta; Y)$ may grow with sample size, we assume it is finite for each experiment else every finite-rate summary will yield infinite posterior distortion.

Definition 2.1 (Intrinsic information rate)**.** For a summary S, define its posterior distortion by $\Delta_{\mathrm{KL}}(S) := \mathbb{E}\,\mathrm{KL}(\Pi_Y \| \Pi_S)$ and information rate by. $R(S) := I(Y; S)$ The intrinsic rate required for tolerance $\varepsilon \geq 0$, and the minimum distortion under a rate budget $R \geq 0$ are

$$R^\star(\varepsilon) := \inf_{q(s|y):\, \Delta_{\mathrm{KL}}(S) \leq \varepsilon} I(Y; S), \qquad D^\star(R) := \inf_{q(s|y):\, I(Y;S) \leq R} \Delta_{\mathrm{KL}}(S). \tag{1}$$

Zero distortion corresponds to exact posterior sufficiency (Lindley, 1956), potentially requiring an unbounded information rate. A deterministic continuous statistic may supply that rate even in one dimension, however, quantization or stochastic encoding makes precision, and hence rate, an explicit modeling choice for approximate sufficiency (Joyce & Marjoram, 2008; Hayashi & Tan, 2018).

Lemma 2.2 (Posterior loss and compensation)**.** *For every summary channel,*

$$\Delta_{\mathrm{KL}}(S) = I(\theta; Y \mid S) = I(\theta; Y) - I(\theta; S), \qquad \Pi_S := \mathbb{E}[\Pi_Y \mid S]. \tag{2}$$

For any measurable reconstruction Q_S with finite expected divergence,

$$\mathbb{E}\,\mathrm{KL}(\Pi_Y \| Q_S) = \Delta_{\mathrm{KL}}(S) + \mathbb{E}\,\mathrm{KL}(\Pi_S \| Q_S). \tag{3}$$

Consequently $R^\star(\varepsilon) \geq [I(\theta; Y) - \varepsilon]_+$ and $D^\star(R) \geq [I(\theta; Y) - R]_+$.

Equation (2) identifies $R^\star$ with the information-bottleneck frontier (Tishby et al., 2000; Harremoës & Tishby, 2007) of the Bayesian experiment. When conditional entropies are well-defined and finite with respect to a common reference measure, the same identity reads $\Delta_{\mathrm{KL}}(S) = H(\theta \mid S) - H(\theta \mid Y)$, the expected-posterior-entropy selection criterion of Hoffmann & Onnela (2026). However, they minimize it within a pre-selected summary class. This class can restrict dimension, architecture, network size, or the number of selected statistics, but it places no direct constraint on the rate $I(Y; S)$. For example, if $Y = (U, V)$ with U sufficient and V carrying no information about θ, then U and (U, V) induce identical posteriors and identical distortion, so no fidelity criterion separates them, while $I(Y; U) < I(Y; (U, V))$. Only a criterion that charges for rate prefers the smaller summary. Our work is therefore complementary to theirs.

Furthermore, the KL and mutual-information definitions above remain applicable even when such an entropy difference is unavailable. Equation (3) further separates information discarded by the summary from error introduced by its decoder. This separation also guides our interpretation of the empirical NPE rate-loss frontiers shown in Section 5.2.

Lemma 2.2 further motivates the reason to make information rate explicit. The Bayes-information expansion of Clarke & Barron (1994) for regular parametric models combined with Equation (2) makes it immediate that as the data accumulates, a summary must resolve increasingly narrow and concentrated posterior distributions. Therefore, even when a fixed-dimensional sufficient statistic exists, it must be transmitted with an unbounded increasing precision, hence the information rate, to achieve bounded distortion. The following lemma states the result formally.

Lemma 2.3 (Summary-rate consequence in parametric models). *For a regular d-dimensional i.i.d. model satisfying the prior-averaged Clarke–Barron conditions, every summary with $\Delta_{\mathrm{KL}}(S_n) \le \varepsilon$ obeys*

$$I(Y^n; S_n) \ge \frac{d}{2} \log \frac{n}{2\pi e} + h(\pi) + \frac{1}{2}\mathbb{E}_\pi \log\det \mathcal{I}(\theta) - \varepsilon + o(1), \tag{4}$$

where $\mathcal{I}$ is the per-observation Fisher information. Thus, for fixed d and ε, $R_n^\star(\varepsilon) \ge \frac{d}{2}\log n - O(1)$.

Appendix I gives an alternative lower bound from Bayes mean-square error. We further show via a matching upper bound that the leading rate of $O(\frac{d}{2}\log n)$ is indeed achievable under a quadratic posterior chart condition, which is verified for a non-Gaussian exponential-family class.

3 Posterior geometry and the information rate

In this section, we develop upper and lower bounds on the intrinsic rate. We start by identifying that $R^\star$ depends on the *law* of Π_Y, not just its support. We state the posterior-source reduction (Harremoës & Tishby, 2007; Armstrong, 2026) in the following lemma and then make the connection to the relevant geometry precise; proof and endpoint properties are provided in Appendix E.

Proposition 3.1 (Posterior-source reduction). *Let $Z := \Pi_Y$. Then $P(\theta \in \cdot \mid Z) = Z$ almost surely, and $R^\star(\varepsilon) = \inf_{P_{\widehat{Z}|Z}:\, \mathbb{E}\mathrm{KL}(Z\|\widehat{Z}) \le \varepsilon} I(Z;\widehat{Z})$, where $\widehat{Z}$ ranges over probability-measure-valued reconstructions.*

We use posterior-source reduction to show that achievable rates and converses depend on two features of the random posterior: how accurately a finite codebook represents it on average, and how much predictive probability one accurate reconstruction captures. We develop these bounds, then use the exact Gaussian frontier to motivate the distinction between deterministic and stochastic compression.

3.1 Achievability by posterior quantization

For an integer $M \ge 1$, let

$$D_M^{\mathrm{av}} := \inf_{Q_1,\dots,Q_M} \mathbb{E} \min_{j \le M} \mathrm{KL}(\Pi_Y \| Q_j), \qquad N^{\mathrm{av}}(\varepsilon) := \inf\{M : D_M^{\mathrm{av}} < \varepsilon\}. \tag{5}$$

The centers are probability measures, and the expectation weights each posterior by its occurrence under the simulator and prior.

Theorem 3.2 (Average posterior quantization). *The optimal distortion with at most M summary values satisfies*

$$D_M^{\mathrm{av}} = \inf_{g:\mathcal{Y}\to[M]} \Delta_{\mathrm{KL}}(g(Y)) = \inf_{q(s|y):\, |\mathcal{S}|\le M} \Delta_{\mathrm{KL}}(S). \tag{6}$$

Moreover $D_1^{\mathrm{av}} = I(\theta;Y)$, and $R^\star(\varepsilon) \le \log N^{\mathrm{av}}(\varepsilon)$.

We therefore identify the optimal M-symbol posterior-distortion infimum with quantization of the random posterior. For the deterministic quantizer, $R(S) = H(S) \le \log M$, hence it also yields an upper bound on the information frontier. The result holds in our standard-Borel Bayesian setup and is proved by directly replacing any stochastic summary with a deterministic nearest-center rule, followed by posterior reconstruction; see Appendix F. The equality further distinguishes *cardinality* from *information*: every stochastic M-symbol summary admits a deterministic M-symbol replacement with no greater posterior distortion. This replacement preserves the alphabet-size constraint, but need not preserve mutual information. However, as we show in the Section 4, when comparison happens at matched information rates rather than cardinality, randomized summaries achieve strictly lower distortion.

3.2 Converses that retain predictive mass

A packing of possible posteriors alone cannot lower-bound the rate of the observed experiment. For example, let $\theta = Y$ have probability $1-\delta$ at 0 and probability δ/M at each of $1,\dots,M$. Its posterior family always contains $M+1$ mutually singular distributions, but the constant summary has distortion $h_2(\delta) + \delta \log M \to 0$. A converse therefore must retain the masses of those posterior states. Our next theorem states the probability-weighted analogue of a packing lower bound that accounts for the mass represented by a single reconstruction.

Theorem 3.3 (Posterior small-ball bound; specialization of Xu & Raginsky (2017)). *For $t > \varepsilon$, define the largest predictive mass captured by a forward-KL ball as $\mathcal{Q}_Z(t) := \sup_Q P\{\mathrm{KL}(Z\|Q) \le t\} > 0$. Then, the intrinsic rate is lower bounded as $R^\star(\varepsilon) \ge \left[\left(1 - \frac{\varepsilon}{t}\right) \log \frac{1}{\mathcal{Q}_Z(t)} - \log 2\right]_+$.*

Hence, applying general small-ball information bounds to the random posterior yields a converse governed by predictive mass. Take, for example, a uniform law on M sufficiently separated posterior states, $\mathcal{Q}_Z(t) = 1/M$ and the bound has the familiar $\log M$ scale. Appendix H gives a complementary cluster bound that retains the entropy of unequal cluster weights, and a KL-Laplace bound (Riegler et al., 2018). Together, these results identify posterior approximation error and predictive concentration as complementary controls on the information budget, and clarify why cardinality and mutual information lead to different conclusions about encoder randomization.

We now discuss the Gaussian location model where the exact frontier is analytically available. Its sufficient sample mean has a continuous distribution, so transmitting it exactly costs infinite information. Under a positive finite rate budget, the classical Gaussian information-bottleneck solution (Chechik et al., 2005; Asoodeh & Calmon, 2020) instead adds noise to the mean.

Proposition 3.4 (Gaussian frontier). *Let $\theta \sim N(0, \tau^2 I_d)$ and $Y_i \mid \theta \sim N(\theta, \sigma^2 I_d)$ independently, and put $a = n\tau^2/\sigma^2$. Then, for $R \ge 0$ and $0 < \varepsilon < \frac{d}{2}\log(1+a)$*

$$D_n^\star(R) = \frac{d}{2}\log(1 + ae^{-2R/d}), \qquad R_n^\star(\varepsilon) = \frac{d}{2}\log\frac{a}{e^{2\varepsilon/d} - 1} \tag{7}$$

Independent Gaussian noise with variance chosen to meet the rate budget therefore attains this frontier; a constant summary attains the zero-rate endpoint. A conditional entropy-power argument proves the converse against arbitrary channels (Appendix I). In the next section, we show that the choice of randomizing summaries is neither arbitrary nor specific to the Gaussian model. We show that for general Bayesian experiments with a common sampling support, any summary that approaches the optimal frontier must be stochastic.

4 When stochastic summaries improve the frontier

To formalize the comparison suggested by the Gaussian example, let $\mathcal{G}$ contain all measurable $g : \mathcal{Y} \to \mathbb{N}$ with $H(g(Y)) < \infty$, including constants, and set

$$D^{\mathrm{det}}(R) := \inf_{g \in \mathcal{G}:\, H(g(Y)) \le R} \Delta_{\mathrm{KL}}(g(Y)), \qquad \mathcal{A}_{\mathrm{det}} := \{(H(g(Y)), \Delta_{\mathrm{KL}}(g(Y))) : g \in \mathcal{G}\}. \tag{8}$$

The information cost of a deterministic code is its entropy and therefore restricting to countable outputs loses no finite-rate deterministic summaries. One can further draw a finite selector J independently of the experiment and output $(J, g_J(Y))$, revealing which code was used. Rate and distortion then average over the selected codes, giving the deterministic time-sharing frontier

$$D^{\mathrm{ts}}(R) := \inf\{d : (r, d) \in \mathrm{conv}(\mathcal{A}_{\mathrm{det}}),\ r \le R\}. \tag{9}$$

Here conv denotes finite convex combinations (Wagner & Ballé, 2021; Witsenhausen & Wyner, 1975). Appendix A establishes the time-sharing identities and frontier properties. For the theorem below, we assume that the sampling laws share a common support. Thus, the experiment is homogeneous in the sense of Le Cam (n.d., Section 3, p. 14) except on sets of measure zero under the prior.

Assumption 4.1 (Common support). The sampling distributions $\{P_{Y|\theta} : \theta \in \Theta_0\}$ are mutually absolutely continuous on a measurable set Θ_0 with $\pi(\Theta_0) = 1$.

Equivalently, under the standing finite-information assumption, $\pi \otimes P_Y \sim P_{\theta Y}$, or $\Pi_y \sim \pi$ for P_Y-almost every y, where $\sim$ denotes mutual absolute continuity. Proposition B.1 proves these equivalences. Define the zero-distortion rate $R_0 := \inf\{R \ge 0 : D^{\mathrm{ts}}(R) = 0\}$ with $\inf \varnothing = \infty$. We are now ready to state our main result.

Theorem 4.2 (Strict stochastic advantage). *Let $\theta \to Y$ be a Bayesian experiment on standard Borel spaces with $0 < I(\theta; Y) < \infty$ satisfying Assumption 4.1. Then*

$$D^\star(R) < D^{\mathrm{ts}}(R) \qquad \text{for every } 0 < R < R_0. \tag{10}$$

For each such R, a finite-alphabet summary channel has information rate exactly R and distortion strictly below $D^{\mathrm{ts}}(R)$. Both frontiers equal $I(\theta; Y)$ at zero; if $R_0 < \infty$, both vanish for $R \ge R_0$.

Theorem 4.2 establishes a limitation of the entire deterministic encoder class, rather than of any particular construction or training procedure: each interior rate, any sequence of rate-feasible deterministic time-sharing mixtures cannot approach the stochastic frontier. This conclusion applies to general Bayesian experiments, including continuous observations, and requires neither frontier to be attained. Moreover, the interval of separation need not be bounded: whenever $R^\star(\epsilon) \to \infty$ as $\epsilon \downarrow 0$, we have $R_0 = \infty$, so stochastic summaries retain a strict advantage at every positive finite information budget. The Gaussian experiment of Proposition 3.4 provides an explicit example: its noisy Gaussian summary outperforms the full deterministic time-sharing envelope on the entire positive half-line.

Proof sketch. Fix a deterministic code T. With probability $1 - t$, retain its label; otherwise replace it by an independent draw from its own marginal distribution, without revealing the replacement. The resulting code V_t has the same output distribution but saves at least $\kappa(T)t\log(1/t)$, where $\kappa(T) := 1-\sum_c P(T=c)^2$,in information rate. Its added posterior distortion has an exact expression involving a weighted Jensen–Shannon divergence between the code-level joint and product laws. Data processing bounds this cost by a single experiment-level quantity, uniformly over all deterministic codes; under Assumption 4.1, that bound is $o(t\log(1/t))$.

To turn this improvement mechanism into a strict gap between frontiers, the rate saving must also be controlled uniformly. At a fixed interior rate, nearly optimal deterministic mixtures retain a positive amount of parameter information. Finite Bayes information prevents codes with relevance bounded away from zero from becoming arbitrarily concentrated on one label, giving a positive lower bound on the mixture's average collision gap. Relabeling therefore saves a *uniformly* positive multiple of $t\log(1/t)$ at a smaller-order posterior cost. Revealed time-sharing with a fixed finite code of lower distortion then spends this saving and restores the original rate while strictly reducing distortion. Thus hidden mixing frees the budget and revealed mixing spends it. Appendix B fixes all comparison constants before choosing the nearly optimal mixture, avoiding any assumption of an attained deterministic optimum or a finite-rate sufficient code.

Finite observation spaces give a complementary setting in which both the separation interval and the required output alphabet can be made explicit. If Y takes m values with positive probabilities and its posteriors are pairwise equivalent, then $R_0 = H(\Pi_Y)$, and separation holds throughout $0 < R < H(\Pi_Y)$. Moreover, an optimum is attained with at most $m + 1$ output values, including when the parameter space is general; this is the finite-observation IB cardinality bound (Witsenhausen & Wyner, 1975; Benger et al., 2023). In particular, Beta–Binomial experiments with positive prior parameters satisfy these conditions, with $H(\Pi_C) = H(C)$. They provide a controlled illustration of the theorem because the deterministic count envelope can be enumerated independently of neural optimization. The theorem also applies separately to the ordered Bernoulli sample, although enumeration of count encoders does not enumerate all deterministic functions of that sample. Appendix C gives the complete statements, the attainment argument, and an explicit stochastic count-channel witness.

The common-support condition identifies a broad regime of strict advantage, but is not a characterization of all such experiments. Three examples distinguish it from weaker or stronger claims. If the target itself is determined by the observation, $\theta = f(Y)$, and has finite entropy, then

$$D^{\mathrm{ts}}(R) = D^\star(R) = [H(\theta) - R]_+. \tag{11}$$

Revealed time-sharing between the constant code and $f(Y)$ attains this line (Kolchinsky et al., 2019), making randomization beyond time-sharing unnecessary. Nor does introducing posterior overlap necessarily create an advantage. Let Y be a fair bit and let θ equal Y with probability $\alpha \in (0, 1)$ and an erasure symbol otherwise, using an independent erasure coin. Here $I(\theta; S) = \alpha I(Y; S)$ for every summary, so both frontiers equal $\alpha[\log 2 - R]_+$ despite the overlapping posteriors. Conversely, a fair-bit observation with posteriors $(0.1, 0.9, 0)$ and $(0, 0.1, 0.9)$ has strict separation at every $0 < R < \log 2$, even though the sampling laws are not equivalent. Together, these examples show that posterior overlap alone is insufficient and common support is not necessary. The directional singular-mass criterion in Appendix C explains the last example; Appendix D gives all three calculations and further boundary cases.

Discussion. The advantage is also not a consequence of a particular frontier shape. The scalar Gaussian frontiers have the same initial slope (Proposition B.8), while full-support finite experiments can have an initial linear segment (Benger et al., 2023). Neither feature prevents strict separation at every interior rate. What drives the proof is instead the difference in scale between the information saved by hidden relabeling and its posterior cost. Under the additional assumption that the *lautum*

information $\mathrm{KL}(\pi \otimes P_Y \| P_{\theta Y})$ (Palomar & Verdú, 2008) is finite, the cost admits a linear upper bound in t, giving a quantitative gap (Corollary B.7). This, however, is not required for the qualitative separation.

The choice of constraint is therefore essential. Relabeling preserves the output distribution, so it reduces mutual information without reducing output entropy. Together with Theorem 3.2, the separation result establishes a genuine contrast: deterministic summaries suffice for the optimal fixed-cardinality distortion infimum, but are strictly suboptimal under an interior information budget satisfying the theorem's hypotheses. The entropy-penalized deterministic information bottleneck of Strouse & Schwab (2017) studies a different objective; neither its hard-assignment conclusions nor our cardinality result imply deterministic optimality for every output-entropy-constrained problem.

Finally, the result concerns a summary applied to one dataset at a time. It goes beyond removing irrelevant variation within posterior fibers by sufficient-statistic averaging (Armstrong, 2026): strict improvement remains possible even when there is no such variation to remove. It does not, however, compare against deterministic block codes that jointly encode independent Bayesian experiments, which can attain the information-bottleneck tradeoff asymptotically in classical finite-alphabet logarithmic-loss settings (Courtade & Weissman, 2014). The single-dataset setting is precisely the one implemented by the NPE protocol below, where the theorem motivates testing stochastic summaries at matched information rates.

5 Implications for NPE: Comparing Summaries at Measured Rate

The preceding sections distinguish nominal summary dimension, code cardinality, and information rate as different rate axes. In this section, we develop investigate the empirical value of encoder randomization, and motivate comparing trained summaries at their measured rate. We examine:

- **(H1)** Stochastic bottlenecks can achieve lower evaluation posterior loss than deterministic bottlenecks at matched estimated rate. Where an exhaustive deterministic comparison is available, we also test improvement beyond deterministic time-sharing.
- **(H2)** Comparisons between summary architectures depend on the information budget; one no-bottleneck score need not predict their ranking under compression.

5.1 Binary-bottleneck NPE

To make information rates explicitly measurable, we insert a binary-bottleneck between the encoder and decoder networks of the standard NPE architecture. A summary encoder maps a dataset to $s = e_\psi(y) \in \mathbb{R}^{d_s}$. A binary encoder produces code probabilities $p_\eta(b_k = 1 \mid s) = \sigma(a_{\eta,k}(s))$ for $k \in [K]$, and a binary decoder maps a code $b \in \{0,1\}^K$ back to $\widetilde{s} = d_\eta(b) \in \mathbb{R}^{d_s}$. The posterior network receives only $\widetilde{s}$, with no skip connection from the original summary. The stochastic channel uses the conditionally independent Bernoulli channel and samples bits b from $q_\eta(b \mid y) = \prod_{k=1}^{K} \mathrm{Bern}(b_k; p_{\eta,k}(e_\psi(y)))$. The deterministic channel uses a hard threshold of the same probabilities $b_k = \mathbf{1}\{p_{\eta,k}(e_\psi(y)) > 1/2\}$.

Training and evaluation. The summary network, binary autoencoder, and posterior decoder are trained jointly with a sparse Bernoulli penalty which regularizes the information rate of the channel:

$$\mathcal{L}(\phi, \psi, \eta) = \mathbb{E}[-\log q_\phi(\theta \mid d_\eta(B))] + \beta\, \mathbb{E}_Y \sum_{k=1}^{K} \mathrm{KL}(q_\eta(b_k \mid y) \,\|\, \mathrm{Bern}(p_0))\,, \tag{12}$$

where $p_0 = 0.01$. Let $r_0(b) := \prod_k \mathrm{Bern}(b_k; p_0)$, then the variational identity (Alemi et al., 2017) $\mathbb{E}_Y \mathrm{KL}(q_\eta(B \mid Y) \| r_0) = I(Y; B) + \mathrm{KL}(q_\eta(B) \| r_0)$ shows that the penalty includes aggregate-marginal mismatch. We train both channels using forward codes and use straight-through gradients during the backward pass (Bengio et al., 2013). The deterministic channel's forward pass is annealed with a decreasing temperature: we start with the same stochastic training and slowly decrease the temperature of the channel followed by a hard threshold training phase to ensure convergence (Jang et al., 2017; Agustsson et al., 2017).

We vary β to obtain operating points and measure the deployed rate after training. Let $q_i(b) = q_\eta(b \mid y_i)$ and $\bar{q}_M = M^{-1} \sum_i q_i$ for a reference set of M simulator observations. We estimate

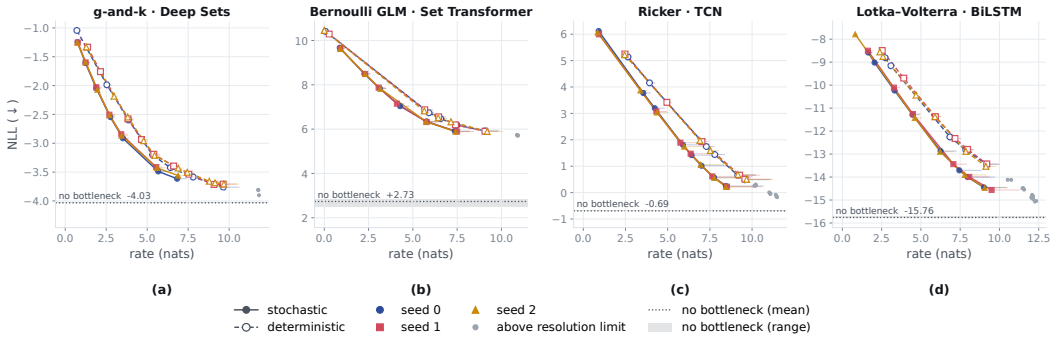


Figure 2: **Rate–NLL comparisons on four simulators.** All trained stochastic NPEs achieve lower NLL than the deterministic ones at matched information rate. Horizontal segments connect the MC and marginal-based rate estimates. Gray points exceed the 10-nat MC rate threshold and are omitted from averaged curves. Dotted lines are standard NPE references trained without a bottleneck.

$R_M = \mathbb{E}_{B\sim\bar{q}_M}[-\log \bar{q}_M(B)] - M^{-1}\sum_i H(q_i)$ by Monte Carlo. The exact empirical-mixture information obeys $R_M \le \log M$ and has downward bias relative to population information after averaging over random reference sets. Hence, we also fit marginal models r_ω on the code distribution of the trained bottleneck NPE, giving a variational upper bound on the estimated rate. The two estimates are displayed as horizontal segments for experiments where exact enumeration is not feasible.

The vertical axis is evaluation NLL rather than a direct estimate of posterior KL. Write $q_\phi(\theta \mid B)$ for the decoder including its deterministic binary reconstruction. When the entropy and expected losses are finite, Lemma 2.2 gives $L = H(\theta \mid Y) + \Delta_{\mathrm{KL}}(B) + \mathbb{E}_B\,\mathrm{KL}(p(\theta \mid B)\|q_\phi(\theta \mid B))$. Within one simulator and parameterization, the first term is common. NLL differences still combine information loss and decoder error. The resulting curves describe the trained family. We verify on the analytic benchmarks that our trained models have multiple orders of magnitude less error than the measured separation in NLL.

5.2 Experiments

We use a finite conjugate experiment and four simulator benchmarks to examine the hypotheses of Section 5. The count experiment isolates information loss from decoder approximation and compares achievable stochastic channels with an exhaustive deterministic time-sharing hull. On the simulators, we compare evaluation NLL at estimated information rate within trained model families. The interpolation and comparator conventions are specified in Section 5.

An enumerated posterior-compression example. Let $\theta \sim \mathrm{Beta}(5,5)$ and $Y_1,\dots,Y_8 \mid \theta$ be independent Bernoulli observations. We encode their sufficient count $C = \sum_i Y_i$ using four bits. The nine count values have 21,147 partitions, permitting enumeration of all deterministic count encoders up to relabeling. Figure 1 compares these pairs and their lower convex hull with achievable stochastic channels from information-bottleneck iterations. The largest observed reduction in information loss beyond this time-sharing hull is approximately 0.036 nats.

All BNPE models use an exact finite-sum mutual-information penalty in this example and training lands at the two plotted frontiers. The maximum mean decoder approximation gap is approximately 4×10^{-4} nats, substantially smaller than the separation.

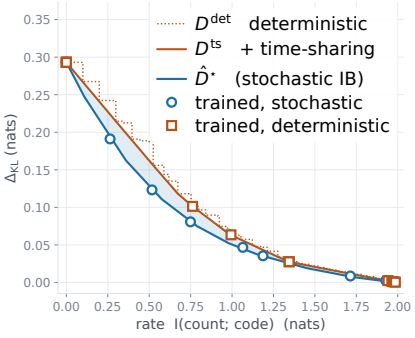


Figure 1: Randomization beyond deterministic time-sharing in the count benchmark.

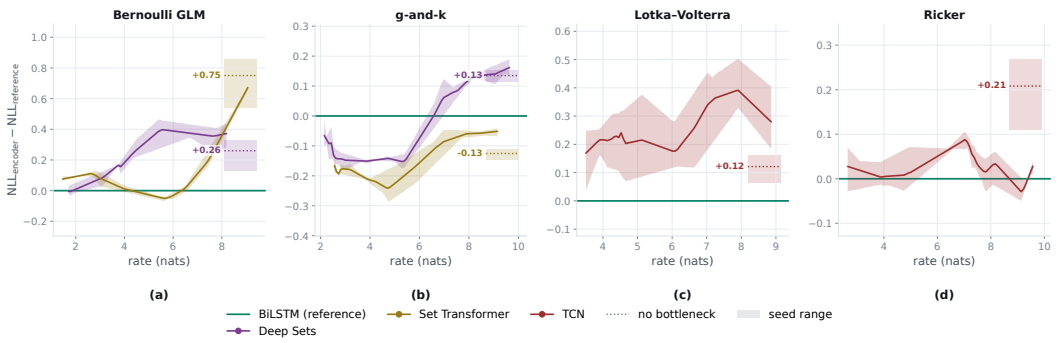


Figure 3: **Architecture differences at matched rate.** Each curve is encoder NLL minus BiLSTM NLL; negative values favor the named encoder. Shading shows the minimum-to-maximum range of the three paired differences. Dotted segments are corresponding mean no-bottleneck differences.

Simulator protocol and estimands. We study a Bernoulli GLM, g-and-k, Lotka–Volterra, and Ricker, with Deep Sets, Set Transformer, BiLSTM, or TCN encoders as appropriate. Each supplies a 128-bit bottleneck followed by the same conditional-flow family. Appendix K reports the configuration, estimators, and additional panels.

(H1) Stochastic versus deterministic codes. Figure 2 shows one encoder per simulator; Figure 5 contains the other six combinations. Curves are interpolated separately within each training seed. All evaluated stochastic BNPE models exhibit a favorable rate–NLL tradeoff across the displayed regimes: the mean matched-rate NLL differences remain negative in all ten simulator-encoder combinations, ranging from -0.168 to -0.617 nats.

(H2) Architecture comparisons depend on rate. Figure 3 forms differences from BiLSTM within each training seed, then averages the differences on a common rate grid. On g-and-k, the Deep Sets comparison changes sign, while Set Transformer remains favorable over the displayed shared range. On the GLM, the Set Transformer gap is smaller at intermediate rates than its no-bottleneck difference. The Lotka–Volterra TCN remains above BiLSTM in the displayed range; the Ricker difference is smaller and varies with rate and seed. They comparisons demonstrate why one no-bottleneck score does not characterize an encoder under compression.

6 Conclusion and Limitations

Posterior fidelity and retained information provide complementary coordinates for comparing summaries. A fixed representation dimension does not fix its information rate, and regular parametric concentration can require increasing rate even when a fixed-dimensional sufficient statistic exists. Posterior quantization and mass-sensitive converses connect this requirement to the law of the random posterior. At a prescribed cardinality, deterministic summaries achieve the same distortion infimum as stochastic ones, whereas, at a prescribed information rate, Theorem 4.2 establishes a strict stochastic advantage over the entire deterministic time-sharing envelope. The extension to general observation spaces follows from controlling the perturbation cost uniformly over candidate codes.

The experiments turn this distinction into a comparison of measured binary-code rates and posterior-decoder loss. The count benchmark resolves the deterministic envelope directly. The simulator comparisons show favorable stochastic rate–NLL tradeoffs and architecture rankings that vary with rate, while also reflecting encoder restrictions, optimization, and decoder approximation.

The separation theorem concerns a channel applied to one observation or simulated dataset. It does not compare deterministic block codes across independent experiments or guarantee that a neural optimizer reaches the better channel. Its support assumption is sufficient rather than necessary; characterizing separation when sampling supports differ remains an open problem. The results establish an explicit information budget as a useful basis for studying both the statistical requirements and the learned performance of NPE summaries.

AI USE STATEMENT

Generative AI tools assisted with literature searches, mathematical and experimental audits, code and figure preparation, and manuscript editing. The authors are responsible for the scientific claims, source attribution, and final presentation.

REFERENCES

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Mattias Åkesson, Prashant Singh, Fredrik Wrede, and Andreas Hellander. Convolutional neural networks as summary statistics for approximate bayesian computation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3353–3365, 2022. doi: 10.1109/TCBB.2021.3108695. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9847490/.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.00410.

Justin Alsing and Benjamin Wandelt. Generalized massive optimal data compression. *Monthly Notices of the Royal Astronomical Society: Letters*, 476(1):L60–L64, 2018. doi: 10.1093/mnrasl/sly029. URL https://academic.oup.com/mnrasl/article/476/1/L60/4909822.

Joss Armstrong. Source-side sufficiency for the information bottleneck: exact reduction and finite-block equivalence, 2026. URL https://arxiv.org/abs/2604.26744v2. arXiv:2604.26744v2.

Shahab Asoodeh and Flavio P. Calmon. Bottleneck problems: An information and estimation-theoretic view. *Entropy*, 22(11):1325, 2020. doi: 10.3390/e22111325. URL https://www.mdpi.com/1099-4300/22/11/1325.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL https://jmlr.org/papers/v6/banerjee05b.html.

Etam Benger, Shahab Asoodeh, and Jun Chen. The cardinality bound on the information bottleneck representations is tight. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 1478–1483, 2023. doi: 10.1109/ISIT54713.2023.10206791. URL https://arxiv.org/abs/2305.07000.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. URL https://arxiv.org/abs/1308.3432.

Michael G. B. Blum, Matthew A. Nunes, Dennis Prangle, and Scott A. Sisson. A comparative review of dimension reduction methods in approximate bayesian computation. *Statistical Science*, 28(2):189–208, 2013. doi: 10.1214/12-STS406. URL https://arxiv.org/abs/1202.3819.

Jeffrey Chan, Valerio Perrone, Jeffrey P. Spence, Paul A. Jenkins, Sara Mathieson, and Yun S. Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/2e9f978b222a956ba6bdf427efbd9ab3-Abstract.html.

Tom Charnock, Guilhem Lavaux, and Benjamin D Wandelt. Automatic physical inference with information maximizing neural networks. *Physical Review D*, 97(8):083004, 2018. doi: 10.1103/PhysRevD.97.083004. URL https://journals.aps.org/prd/abstract/10.1103/PhysRevD.97.083004.

Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6(6):165–188, 2005. URL https://jmlr.org/papers/v6/chechik05a.html.

Tianyu Chen, Vansh Bansal, and James G. Scott. Conditional diffusions for amortized neural posterior estimation. *arXiv preprint arXiv:2410.19105*, 2024. URL https://arxiv.org/abs/2410.19105.

Yanzhi Chen, Dinghuai Zhang, Michael U. Gutmann, Aaron Courville, and Zhanxing Zhu. Neural approximate sufficient statistics for implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.10079.

Yanzhi Chen, Michael U Gutmann, and Adrian Weller. Is learning summary statistics necessary for likelihood-free inference? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4529–4544. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23h.html.

Philip A. Chou, Tom Lookabaugh, and Robert M. Gray. Entropy-constrained vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(1):31–42, 1989. doi: 10.1109/29.17498. URL https://web.stanford.edu/class/ee398a/handouts/papers/Chou%20-%20EC%20VQ.pdf.

Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994. doi: 10.1016/0378-3758(94)90153-8. URL https://www.sciencedirect.com/science/article/pii/0378375894901538.

Thomas A. Courtade and Tsachy Weissman. Multiterminal source coding under logarithmic loss. *IEEE Transactions on Information Theory*, 60(1):740–761, 2014. doi: 10.1109/TIT.2013.2288257. URL https://people.eecs.berkeley.edu/~courtade/pdfs/CourtadeWeissman_TransITJan2014.pdf.

Rancy El Nmeir, Harald Luschgy, and Gilles Pagès. New approach to greedy vector quantization, 2020. URL https://arxiv.org/abs/2003.14145. arXiv:2003.14145.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(3):419–474, 2012. doi: 10.1111/j.1467-9868.2011.01010.x. URL https://academic.oup.com/jrsssb/article-abstract/74/3/419/7075312.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pp. 881–889. Pmlr, 2015.

Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In *Learning Theory and Kernel Machines*, volume 2777 of *Lecture Notes in Computer Science*, pp. 595–609. Springer, 2003. doi: 10.1007/978-3-540-45167-9_43. URL https://doi.org/10.1007/978-3-540-45167-9_43.

David Greenberg, Marcel Nonnenmacher, and Jakob H. Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2404–2414. PMLR, 2019. URL https://proceedings.mlr.press/v97/greenberg19a.html.

Peter Harremoës and Naftali Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 566–570, 2007. doi: 10.1109/ISIT.2007.4557285. URL https://doi.org/10.1109/ISIT.2007.4557285.

David Haussler and Manfred Opper. Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492, 1997. doi: 10.1214/aos/1030741081. URL https://doi.org/10.1214/aos/1030741081.

Masahito Hayashi and Vincent Y. F. Tan. Minimum rates of approximate sufficient statistics. *IEEE Transactions on Information Theory*, 64(2):875–888, 2018. doi: 10.1109/TIT.2017.2775612. URL https://arxiv.org/abs/1612.02542.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4239–4248. PMLR, 2020. URL https://proceedings.mlr.press/v119/hermans20a.html.

Till Hoffmann and Jukka-Pekka Onnela. Unifying summary statistic selection for approximate bayesian computation. *Statistics and Computing*, 36:70, 2026. doi: 10.1007/s11222-025-10808-2. URL https://link.springer.com/article/10.1007/s11222-025-10808-2.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, 27(4):1595–1618, 2017. doi: 10.5705/ss.202015.0340. URL https://www3.stat.sinica.edu.tw/statistica/oldpdf/A27n411.pdf.

Paul Joyce and Paul Marjoram. Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Artemy Kolchinsky, Brendan D. Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1808.07593. arXiv:1808.07593.

Lucien Le Cam. On some results of J. Hájek concerning asymptotic normality. Unpublished manuscript, University of California, Berkeley, n.d. URL https://www.stat.berkeley.edu/~rice/LeCam/papers/Hajek2a.pdf. Undated manuscript in the Le Cam archive; Section 3, p. 14.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Jan-Matthis Lueckmann, Pedro J. Gonçalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/file/addfa9b7e234254d26e9c7f2af1005cb-Paper.pdf.

Matthew A Nunes and David J Balding. On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics & Molecular Biology*, 9(1):Article 34, 2010. doi: 10.2202/1544-6115.1576. URL https://doi.org/10.2202/1544-6115.1576.

Daniel P. Palomar and Sergio Verdú. Lautum information. *IEEE Transactions on Information Theory*, 54(3):964–975, 2008. doi: 10.1109/TIT.2007.915715. URL https://palomar.home.ece.ust.hk/papers/2008/PalomarVerdu_TransIT2008_lautum.pdf.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/6aca97005c68f1206823815f66102863-Abstract.html.

Stefan T Radev, Ulf K Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. BayesFlow: Learning complex stochastic models with invertible neural networks. *IEEE transactions on neural networks and learning systems*, 33(4):1452–1466, 2022. doi: 10.1109/TNNLS.2020.3042395. URL https://arxiv.org/abs/2003.06281.

Erwin Riegler, Günther Koliander, and Helmut Bölcskei. Rate-distortion theory for general sets and measures. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 101–105, 2018. URL https://www.mins.ee.ethz.ch/pubs/p/rikobo18.

Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45–114, 2014. doi: 10.1214/14-SS107. URL https://projecteuclid.org/journals/statistics-surveys/volume-8/issue-none/Log-concavity-and-strong-log-concavity-A-review/10.1214/14-SS107.pdf.

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. URL https://ira.lib.polyu.edu.hk/handle/10397/107084.

Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems*, volume 12, pp. 617–623, 2000. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/be3e9d3f7d70537357c67bb3f4086846-Paper.pdf. Conference held in 1999.

DJ Strouse and David J. Schwab. The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630, 2017. doi: 10.1162/NECO_a_00961. URL https://direct.mit.edu/neco/article/29/6/1611/8273/The-Deterministic-Information-Bottleneck.

Ce Sui, Xiaosheng Zhao, Tao Jing, and Yi Mao. Evaluating summary statistics with mutual information for cosmological inference. In *ICML Workshop on Machine Learning for Astrophysics*, 2023. URL https://ml4astro.github.io/icml2023/assets/18.pdf.

Lucas Theis and Eirikur Agustsson. On the advantages of stochastic encoders. In *Neural Compression Workshop, International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2102.09270.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. URL https://arxiv.org/abs/physics/0004057.

Aaron B. Wagner and Johannes Ballé. Neural networks optimally compress the Sawbridge. In *2021 Data Compression Conference (DCC)*, pp. 143–152. IEEE, 2021. doi: 10.1109/DCC50243.2021.00022. URL https://arxiv.org/abs/2011.05065.

Hans S. Witsenhausen and Aaron D. Wyner. A conditional entropy bound for a pair of discrete random variables. *IEEE Transactions on Information Theory*, 21(5):493–501, 1975. doi: 10.1109/TIT.1975.1055437. URL https://doi.org/10.1109/TIT.1975.1055437.

Aaron D. Wyner and Jacob Ziv. A theorem on the entropy of certain binary sequences and applications: Part I. *IEEE Transactions on Information Theory*, 19(6):769–772, 1973. doi: 10.1109/TIT.1973.1055107. URL https://doi.org/10.1109/TIT.1973.1055107.

Aolin Xu and Maxim Raginsky. Information-theoretic lower bounds on bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, 2017. URL https://arxiv.org/abs/1607.00550.

APPENDIX GUIDE

The separation proof and its scope are developed in Appendices A–D. The remaining appendices give the posterior-geometry and parametric arguments, experimental details, and consolidated related work.

		Page
A	Foundations for the separation theorem	15
B	Strict separation for general observation spaces	21
C	Finite observations: improvement and attainment	29
D	Examples delimiting the support and rate assumptions	34
E	Posterior identities and source reduction	35
F	Posterior quantization and uniform covering	37
G	Hellinger covering and the local modulus	39
H	Further converses	40
I	Parametric rates: hypotheses and achievability	43
J	Parametric rate–distortion calculations	48
K	Experimental protocol and additional results	50
L	Related work and connections	52

A FOUNDATIONS FOR THE SEPARATION THEOREM

Throughout these appendices, Θ and $\mathcal{Y}$ are standard Borel spaces, $I := I(\theta;Y) \in (0,\infty)$, and $\theta \to Y \to S$ is the summary Markov chain. Auxiliary random variables are independent of the experiment unless specified otherwise. Write $\mathcal{G}$ for countable-valued deterministic codes $T = g(Y)$ with $H(T) < \infty$; finite codes and constants are included. The deterministic time-sharing frontier uses finite revealed mixtures of these codes and an average information-rate constraint, as in Proposition A.3. All statements below concern infima unless attainment is explicitly proved.

Notation for measures, code cells, and mixtures. In the general mixture statements below, P and Q denote probability measures on the same measurable space $(\mathcal{X},\mathcal{A})$; Q is the reference measure in $\mathrm{KL}(P\|Q)$. They need not be the joint and product laws of an experiment. For a code $T = g(Y)$, write

$$p_c := \mathbb{P}(T=c), \qquad P_T := \sum_{c:p_c>0} p_c\delta_c, \qquad \Pi_c := P_{\theta|T=c}, \qquad \pi = \sum_{c:p_c>0} p_c\Pi_c.$$

Thus p_c is a cell probability, P_T is the marginal law of the code, and Π_c is the parameter posterior in cell c; δ_c is unit mass at c. Thus Π_T is the random posterior Π_S of the main text with $S = T$. For $p_c > 0$, define its density relative to the prior by

$$f_c := \frac{d\Pi_c}{d\pi}, \qquad \Pi_c(d\theta) = f_c(\theta)\pi(d\theta).$$

Following the information-bottleneck literature, we call $I(\theta;T)$ the *relevance* of the code T: it is the information about the parameter that the code retains, and by Lemma 2.2 it equals $I - \Delta_{\mathrm{KL}}(T)$, so a code has high relevance exactly when it has low posterior distortion. The same term is used below for the relevance $I(\theta;S)$ of a general summary.

The general compensation lemma (Lemma A.6) uses arbitrary component laws P_j and positive weights λ_j summing to one, with mixture $\bar{P} := \sum_j \lambda_j P_j$. Appendix B specifies the joint and product laws used for the experiment.

For arbitrary P, Q, their Lebesgue decomposition is $P = \tilde{P}_Q + P_Q^{\perp}$, where $\tilde{P}_Q \ll Q$ and $P_Q^{\perp} \perp Q$. The *directed singular mass* is

$$\alpha(P \mid Q) := P_Q^{\perp}(\mathcal{X}) = \sup\{P(A) : A \in \mathcal{A},\ Q(A) = 0\}. \tag{13}$$

It measures the mass that P places on events impossible under Q. It lies in $[0,1]$, vanishes exactly when $P \ll Q$, and need not equal $\alpha(Q \mid P)$. For $w \in (0,1)$, define the weighted Jensen-Shannon

divergence as:

$$M_w := wP + (1-w)Q, \qquad \mathrm{JS}_w(P,Q) := w\mathrm{KL}(P\|M_w) + (1-w)\mathrm{KL}(Q\|M_w).$$

The first argument is the rare component as $w \downarrow 0$. We use $h_2(w) = -w\log w - (1-w)\log(1-w)$ for binary entropy.

Lemma A.1 (Finite-rate deterministic summaries)**.** *A deterministic summary with a standard Borel output space has finite information rate if and only if its law is supported on countably many atoms with finite entropy. In that case $I(Y; g(Y)) = H(g(Y))$. This restriction does not apply to stochastic summaries.*

Proof. Let $T = g(Y)$ and let μ be the law of T on its output space $\mathcal{X}$. We first express the rate of T as a supremum of finite entropies, and then evaluate that supremum separately for a law with a nonatomic part and for a purely atomic law.

Mutual information between general random variables is the supremum of the mutual informations between their finite quantizations, and it is enough to quantize one of the two arguments. Hence

$$I(Y;T) = \sup_a I(Y; a(T)),$$

where a ranges over measurable maps from the output space of T to a finite set. For a finite-valued $a(T)$, the mutual information is the entropy difference $H(a(T)) - H(a(T) \mid Y)$, and the conditional entropy vanishes because $a(T) = a(g(Y))$ is a function of Y. Therefore

$$I(Y;T) = \sup_a H(a(T)). \tag{14}$$

Since each $a(T)$ is a function of T, every term in the supremum is at most $H(T)$, where $H(T)$ is the (possibly infinite) entropy of the countable law when μ is atomic.

Suppose now that μ has a nonatomic component of total mass $c > 0$. Let N be the countable set of atoms of μ; the nonatomic part gives N zero mass, so μ restricted to $\mathcal{X} \setminus N$ is a nonatomic measure of mass c. A nonatomic finite measure on a standard Borel space takes every value between zero and its total mass on measurable subsets, so $\mathcal{X} \setminus N$ can be partitioned into measurable sets $A_1, \ldots, A_m$ with $\mu(A_i) = c/m$ for each i. Define the quantizer a to map A_i to i and N to the symbol 0. Dropping the nonnegative term of the symbol 0,

$$H(a(T)) \;\geq\; \sum_{i=1}^{m} \frac{c}{m} \log \frac{m}{c} \;=\; c\log m - c\log c,$$

which tends to infinity with m. By (14) the rate is infinite.

Suppose instead that μ is purely atomic, $\mu = \sum_k p_k \delta_{x_k}$ with $p_1 \geq p_2 \geq \cdots$. Let a_m retain the first m atoms as separate symbols and merge all remaining atoms into one symbol of mass $r_m := \sum_{k>m} p_k$. Then

$$H(a_m(T)) = \sum_{k \leq m} p_k \log \frac{1}{p_k} + r_m \log \frac{1}{r_m}.$$

As $m \to \infty$, the partial sums increase to $H(T)$ and $r_m \log(1/r_m) \to 0$ because $r_m \to 0$. Thus $\sup_a H(a(T)) \geq H(T)$, and the reverse inequality was noted above, so $I(Y;T) = H(T)$. The rate is finite exactly when $H(T) < \infty$, which proves both assertions. A stochastic summary can have a nonatomic law and finite rate, for instance the observation plus independent Gaussian noise in Theorem 3.4, so the restriction is specific to deterministic codes. □

Lemma A.2 (Approximation by finite posterior codes)**.** *There are finite-valued deterministic codes U_n such that $I(\theta; U_n) \uparrow I$ and $\Delta_{\mathrm{KL}}(U_n) \downarrow 0$. Consequently, for every $\eta > 0$ a finite code has distortion less than η, and $D^{\mathrm{det}}(R) \to 0$ as $R \to \infty$.*

Proof. The idea is to write $I(\theta; U_n)$ as the integral of a convex function of a martingale that converges to the joint density of the experiment, and to pass to the limit from both sides.

Because $\mathcal{Y}$ is standard Borel, its Borel sigma-field is generated by a countable family of sets $C_1, C_2, \ldots$. Let $\mathcal{P}_n$ be the finite partition of $\mathcal{Y}$ generated by $C_1, \ldots, C_n$, and let U_n be the label of

the cell of $\mathcal{P}_n$ that contains Y. The partitions refine one another, so U_n is a function of U_{n+1}, each U_n takes finitely many values, and $H(U_n) < \infty$.

Put $Q_{\theta Y} := \pi \otimes P_Y$. Since $I = \mathrm{KL}(P_{\theta Y} \| Q_{\theta Y}) < \infty$, the joint law is absolutely continuous with respect to the product law, and we write $f := dP_{\theta Y}/dQ_{\theta Y}$. Consider the increasing sigma-fields $\mathcal{F}_n := \mathcal{B}(\Theta) \otimes \sigma(\mathcal{P}_n)$ on $\Theta \times \mathcal{Y}$ and the conditional expectations

$$f_n := \mathbb{E}_{Q_{\theta Y}}[f \mid \mathcal{F}_n].$$

Each f_n is $\mathcal{F}_n$-measurable, hence of the form $f_n(\theta, y) = \sum_{C \in \mathcal{P}_n} f_{n,C}(\theta) \mathbf{1}_C(y)$. For $A \in \mathcal{B}(\Theta)$ and a cell C, the defining property of conditional expectation gives

$$\mathbb{P}(\theta \in A,\, U_n = C) = \int_{A \times C} f \, dQ_{\theta Y} = \int_{A \times C} f_n \, dQ_{\theta Y} = P_Y(C) \int_A f_{n,C} \, d\pi.$$

Thus $f_{n,C}$ is the density of the cell posterior $P_{\theta | U_n = C}$ relative to the prior whenever $P_Y(C) > 0$, and

$$\begin{aligned} \int f_n \log f_n \, dQ_{\theta Y} &= \sum_C P_Y(C) \int f_{n,C} \log f_{n,C} \, d\pi \\ &= \sum_C P(U_n = C)\, \mathrm{KL}(P_{\theta | U_n = C} \| \pi) = I(\theta; U_n). \end{aligned}$$

We now show that these integrals converge to I. The sigma-fields $\mathcal{F}_n$ increase to $\mathcal{B}(\Theta) \otimes \mathcal{B}(\mathcal{Y})$, so by Lévy's upward theorem the martingale (f_n) converges to f almost surely under $Q_{\theta Y}$. Write $\phi(u) := u \log u$, a convex function bounded below by $-1/e$. Conditional Jensen gives $\phi(f_n) \le \mathbb{E}_{Q_{\theta Y}}[\phi(f) \mid \mathcal{F}_n]$, and integrating,

$$I(\theta; U_n) = \int \phi(f_n) \, dQ_{\theta Y} \le \int \phi(f) \, dQ_{\theta Y} = I.$$

For the reverse direction, apply Fatou's lemma to the nonnegative functions $\phi(f_n) + 1/e$, using continuity of ϕ and the almost sure convergence $f_n \to f$:

$$I + \frac{1}{e} = \int \Big(\phi(f) + \frac{1}{e}\Big) dQ_{\theta Y} \le \liminf_n \int \Big(\phi(f_n) + \frac{1}{e}\Big) dQ_{\theta Y} = \liminf_n I(\theta; U_n) + \frac{1}{e}.$$

Together, $I(\theta; U_n) \to I$. The sequence is nondecreasing because U_n is a function of U_{n+1} and data processing gives $I(\theta; U_n) \le I(\theta; U_{n+1})$. Lemma 2.2 then gives $\Delta_{\mathrm{KL}}(U_n) = I - I(\theta; U_n) \downarrow 0$.

For the consequences, fix $\eta > 0$ and choose n with $\Delta_{\mathrm{KL}}(U_n) < \eta$; this is the required finite code. For every $R \ge H(U_n)$ the code U_n is feasible in the definition of $D^{\mathrm{det}}(R)$, so $D^{\mathrm{det}}(R) \le \Delta_{\mathrm{KL}}(U_n) < \eta$. Since D^{det} is nonincreasing and nonnegative, and η was arbitrary, $D^{\mathrm{det}}(R) \to 0$ as $R \to \infty$. □

Proposition A.3 (Revealed time-sharing). *Let $S_1, \dots, S_k$ be summary channels and let J be independent of the experiment and their auxiliary randomness, with probabilities λ_j. The revealed mixture $S = (J, S_J)$ satisfies*

$$I(Y; S) = \sum_j \lambda_j I(Y; S_j), \qquad \Delta_{\mathrm{KL}}(S) = \sum_j \lambda_j \Delta_{\mathrm{KL}}(S_j). \tag{15}$$

For deterministic components $S_j = g_j(Y)$, the component rate is $H(g_j(Y))$, and their finite revealed mixtures achieve exactly $\mathrm{conv}(\mathcal{A}_{\mathrm{det}})$. *The frontier D^{ts} in (9) is convex, nonincreasing, and continuous on $[0, \infty)$, with $D^{\mathrm{ts}}(0) = I$ and $D^{\mathrm{ts}}(R) \to 0$ as $R \to \infty$. Moreover $0 \le D^\star \le D^{\mathrm{ts}} \le D^{\mathrm{det}} \le I$. Its zero-distortion threshold satisfies $R_0 \ge I > 0$; $D^{\mathrm{ts}} > 0$ on $[0, R_0)$, and, if $R_0 < \infty$, $D^{\mathrm{ts}} = 0$ on $[R_0, \infty)$.*

Proof. We prove the two affine identities first, then identify the achieved pairs, and finally derive the shape properties of D^{ts} from them.

The revealed mixture is a valid summary channel: given Y, the pair (J, S_J) is generated from Y and randomness independent of θ, so $\theta \to Y \to (J, S_J)$ is a Markov chain. By the chain rule for mutual information,

$$I(Y; J, S_J) = I(Y; J) + I(Y; S_J \mid J) = 0 + \sum_j \lambda_j I(Y; S_J \mid J = j).$$

The first term vanishes because J is independent of Y. On the event $J = j$, the pair (Y, S_J) equals (Y, S_j), and the conditional law of (Y, S_j) given $J = j$ is its unconditional law because J is independent of the experiment and of the auxiliary randomness of S_j. Hence $I(Y; S_J \mid J = j) = I(Y; S_j)$, which is the first identity in (15). The same conditioning gives

$$I(\theta; Y \mid J, S_J) = \sum_j \lambda_j I(\theta; Y \mid S_J, J = j) = \sum_j \lambda_j I(\theta; Y \mid S_j),$$

and Lemma 2.2 identifies each conditional information with the corresponding posterior distortion, proving the second identity.

For deterministic components $S_j = g_j(Y)$, Lemma A.1 gives $I(Y; S_j) = H(g_j(Y))$, so the mixture $(J, g_J(Y))$ realizes the pair $\sum_j \lambda_j \big(H(g_j(Y)), \Delta_{\mathrm{KL}}(g_j(Y))\big)$. Conversely, every finite convex combination of points of $\mathcal{A}_{\mathrm{det}}$ has this form for a suitable finite law of J. The achieved set is therefore exactly $\mathrm{conv}(\mathcal{A}_{\mathrm{det}})$.

We turn to the shape of D^{ts}. The constant code gives $(0, I) \in \mathcal{A}_{\mathrm{det}}$, so $D^{\mathrm{ts}}(R) \le I < \infty$ for every $R \ge 0$, and D^{ts} is nonincreasing because the feasible set $\{(r, d) \in \mathrm{conv}(\mathcal{A}_{\mathrm{det}}) : r \le R\}$ grows with R. For convexity, fix $R_1, R_2 \ge 0$, $\mu \in (0, 1)$, and $\varepsilon > 0$, and choose feasible (r_i, d_i) with $r_i \le R_i$ and $d_i \le D^{\mathrm{ts}}(R_i) + \varepsilon$. The convex combination $\mu(r_1, d_1) + (1-\mu)(r_2, d_2)$ again lies in $\mathrm{conv}(\mathcal{A}_{\mathrm{det}})$ and has rate at most $\mu R_1 + (1-\mu)R_2$, so

$$D^{\mathrm{ts}}(\mu R_1 + (1-\mu)R_2) \le \mu d_1 + (1-\mu) d_2 \le \mu D^{\mathrm{ts}}(R_1) + (1-\mu) D^{\mathrm{ts}}(R_2) + \varepsilon,$$

and $\varepsilon \downarrow 0$ gives convexity.

The chain of inequalities follows from inclusions of feasible classes. A single deterministic code with $H(g(Y)) \le R$ is a one-component mixture, so $D^{\mathrm{ts}} \le D^{\mathrm{det}}$, and the constant code gives $D^{\mathrm{det}} \le I$. Every feasible point (r, d) of $\mathrm{conv}(\mathcal{A}_{\mathrm{det}})$ with $r \le R$ is realized by a revealed mixture, which is a summary channel of rate $r \le R$ and distortion d, so $D^\star(R) \le d$; taking the infimum over such points gives $D^\star \le D^{\mathrm{ts}}$. Finally $D^\star \ge 0$ because divergences are nonnegative.

For continuity, a finite convex function on $[0, \infty)$ is continuous on the open half-line $(0, \infty)$. At the origin, only rate-zero mixtures are feasible, that is, mixtures of constant codes, each with distortion $I(\theta; Y \mid \text{constant}) = I$; thus $D^{\mathrm{ts}}(0) = I$. The universal converse in Lemma 2.2 gives, for $0 \le R < I$,

$$I - R \le D^\star(R) \le D^{\mathrm{ts}}(R) \le I,$$

so $D^{\mathrm{ts}}(R) \to I$ as $R \downarrow 0$, which is continuity from the right at zero. The limit at infinity follows from $0 \le D^{\mathrm{ts}} \le D^{\mathrm{det}}$ and Lemma A.2.

For the threshold, the display above shows $D^{\mathrm{ts}}(R) \ge I - R > 0$ for $R < I$, so $R_0 \ge I > 0$. On $[0, R_0)$ the frontier is positive by the definition of R_0 as an infimum. If $R_0 < \infty$, then $D^{\mathrm{ts}}(R) = 0$ for some R arbitrarily close to R_0 from the right, monotonicity gives $D^{\mathrm{ts}} = 0$ on (R_0, ∞), and continuity gives $D^{\mathrm{ts}}(R_0) = 0$ as well. □

Lemma A.4 (Truncating countable codes)**.** *If $T \in \mathcal{G}$, define $T_n = T$ on $\{T \le n\}$ and $T_n = n + 1$ otherwise. Then $H(T_n) \le H(T)$ and $\Delta_{\mathrm{KL}}(T_n) \downarrow \Delta_{\mathrm{KL}}(T)$. Thus D^{ts} is unchanged if its component codes must be finite.*

Proof. Take the code values in $\mathbb{N} = \{1, 2, \ldots\}$. Since T_n is a function of T, both $H(T_n) \le H(T)$ and, by data processing, $I(\theta; T_n) \le I(\theta; T)$. Moreover T_n is a function of T_{n+1}, so $I(\theta; T_n)$ is nondecreasing in n, and by Lemma 2.2 the distortions $\Delta_{\mathrm{KL}}(T_n) = I - I(\theta; T_n)$ are nonincreasing in n and bounded below by $\Delta_{\mathrm{KL}}(T)$.

To bound the gap, apply the chain rule to $I(\theta; T, T_n)$ in both orders. Because T_n is a function of T,

$$I(\theta; T) = I(\theta; T, T_n) = I(\theta; T_n) + I(\theta; T \mid T_n),$$

so that

$$\Delta_{\mathrm{KL}}(T_n) - \Delta_{\mathrm{KL}}(T) = I(\theta; T \mid T_n) \le H(T \mid T_n).$$

Conditioning on $T_n = c$ with $c \le n$ determines $T = c$ and contributes no entropy. Conditioning on $T_n = n + 1$ leaves the tail law $(p_c / r_n)_{c > n}$ with $r_n := \sum_{c > n} p_c$, so

$$H(T \mid T_n) = \sum_{c > n} p_c \log \frac{r_n}{p_c} \le \sum_{c > n} p_c \log \frac{1}{p_c},$$

using $r_n \le 1$. The right side is a tail of the convergent series $H(T) = \sum_c p_c \log(1/p_c)$, hence tends to zero. This proves $\Delta_{\mathrm{KL}}(T_n) \downarrow \Delta_{\mathrm{KL}}(T)$.

For the final claim, let $(\lambda_j, T_j)_{j \le k}$ be any finite mixture of codes in $\mathcal{G}$ with average rate at most R and average distortion d. Truncating every component at level n gives a mixture of finite codes whose average rate is still at most R, by the first inequality, and whose average distortion $\sum_j \lambda_j \Delta_{\mathrm{KL}}(T_{j,n})$ decreases to d as $n \to \infty$. Hence the infimum over mixtures of finite codes is at most d, and taking the infimum over all mixtures shows it is at most $D^{\mathrm{ts}}(R)$. The reverse inequality holds because finite codes belong to $\mathcal{G}$. □

Lemma A.5 (Cost of exact preservation)**.** *If $\Delta_{\mathrm{KL}}(S) = 0$, then $\Pi_Y = \Pi_S$ almost surely, hence Π_Y is measurable from S. If Π_Y is countably valued, every such summary has rate at least $H(\Pi_Y)$, allowing infinite entropy.*

Proof. The random variable $\mathrm{KL}(\Pi_Y \| \Pi_S)$ is nonnegative, so $\mathbb{E}\,\mathrm{KL}(\Pi_Y \| \Pi_S) = 0$ forces $\mathrm{KL}(\Pi_Y \| \Pi_S) = 0$ almost surely, and a divergence vanishes only between equal measures. Thus $\Pi_Y = \Pi_S$ almost surely, and Π_S is a measurable function of S.

Suppose the posterior label $Z := \Pi_Y$ takes countably many values. Since Z is a function of Y and Π_S is a function of S, two applications of data processing give

$$I(Y;S) \;\ge\; I(Z;S) \;\ge\; I(Z;\Pi_S) \;=\; I(Z;Z),$$

the last equality because $\Pi_S = Z$ almost surely. For a countably valued Z, the identity $I(Z;Z) = H(Z)$ holds with both sides possibly infinite: it is the purely atomic case of Lemma A.1, where finite quantizations of Z have entropies increasing to $H(Z)$. □

Lemma A.6 (Countable mixture compensation)**.** *Let $\bar{P} = \sum_j \lambda_j P_j$, with $\lambda_j > 0$, and suppose $\mathrm{KL}(\bar{P} \| Q) < \infty$. Then every $\mathrm{KL}(P_j \| Q)$ is finite, $\mathrm{KL}(P_j \| \bar{P}) \le \log(1/\lambda_j)$, and*

$$\sum_j \lambda_j \mathrm{KL}(P_j \| Q) = \sum_j \lambda_j \mathrm{KL}(P_j \| \bar{P}) + \mathrm{KL}(\bar{P} \| Q),$$

with the sums allowed to be infinite.

Proof. We first collect the integrability facts, then prove the identity for a single component, and finally sum over components.

Since $\lambda_j P_j \le \bar{P}$, we have $P_j \ll \bar{P}$ with $dP_j/d\bar{P} \le 1/\lambda_j$, and therefore

$$\mathrm{KL}(P_j \| \bar{P}) = \int \log \frac{dP_j}{d\bar{P}}\, dP_j \le \log \frac{1}{\lambda_j}.$$

The hypothesis $\mathrm{KL}(\bar{P} \| Q) < \infty$ gives $\bar{P} \ll Q$; write $r := d\bar{P}/dQ$. The negative part of $\log r$ is $\bar{P}$-integrable without any assumption: since $u \log(1/u) \le 1/e$ for $u \in (0,1]$,

$$\int (\log r)_-\, d\bar{P} = \int_{\{r<1\}} r \log \frac{1}{r}\, dQ \le \frac{1}{e}.$$

The positive part is then integrable because $\int (\log r)_+\, d\bar{P} = \mathrm{KL}(\bar{P} \| Q) + \int (\log r)_-\, d\bar{P} < \infty$. Hence $\log r \in L^1(\bar{P})$, and by the density bound also $\log r \in L^1(P_j)$ with $\int |\log r|\, dP_j \le \lambda_j^{-1} \int |\log r|\, d\bar{P}$.

Since $P_j \ll \bar{P} \ll Q$, the chain rule for densities gives $dP_j/dQ = (dP_j/d\bar{P})\, r$ almost everywhere under Q, hence

$$\log \frac{dP_j}{dQ} = \log \frac{dP_j}{d\bar{P}} + \log r \qquad P_j\text{-almost surely.}$$

Both terms on the right are P_j-integrable, so integrating against P_j is legitimate and yields

$$\mathrm{KL}(P_j \| Q) = \mathrm{KL}(P_j \| \bar{P}) + \int \log r\, dP_j,$$

a finite quantity. This proves the first two claims.

Multiply the last display by λ_j and sum over j. The series $\sum_j \lambda_j \int |\log r|\, dP_j = \int |\log r|\, d\bar{P}$ converges, so the sum and the integral in the last term can be exchanged:

$$\sum_j \lambda_j \int \log r\, dP_j = \int \log r\, d\bar{P} = \mathrm{KL}(\bar{P}\|Q).$$

The remaining two series have nonnegative terms, and they differ by the finite number $\mathrm{KL}(\bar{P}\|Q)$, so they are either both finite or both infinite, and in the finite case the identity holds as stated. □

Lemma A.7 (Rare-mixture asymptotics)**.** *The divergence $\mathrm{JS}_w(P,Q)$ is at most $h_2(w)$ and decreases under measurable maps. Furthermore,*

$$\lim_{w\downarrow 0} \frac{\mathrm{KL}(P\|M_w)}{\log(1/w)} = \alpha(P \mid Q), \qquad \lim_{w\downarrow 0} \frac{\mathrm{JS}_w(P,Q)}{w\log(1/w)} = \alpha(P \mid Q).$$

In particular, the latter divergence is $o(w\log(1/w))$ exactly when $P \ll Q$. If $\mathrm{KL}(P\|Q) < \infty$, then

$$\mathrm{JS}_w(P,Q) = w\mathrm{KL}(P\|Q) - \mathrm{KL}(M_w\|Q) \le w\mathrm{KL}(P\|Q).$$

Proof. The first two properties follow from an information-theoretic reading of JS_w. Let $\xi \sim \mathrm{Bernoulli}(w)$ and, given ξ, draw X from P if $\xi = 1$ and from Q if $\xi = 0$. Then $X \sim M_w$, and the mutual information between the label and the draw is

$$I(\xi; X) = w\mathrm{KL}(P\|M_w) + (1-w)\mathrm{KL}(Q\|M_w) = \mathrm{JS}_w(P,Q).$$

Hence $\mathrm{JS}_w(P,Q) \le H(\xi) = h_2(w)$, and for a measurable map φ, data processing gives $\mathrm{JS}_w(P\circ\varphi^{-1}, Q\circ\varphi^{-1}) = I(\xi;\varphi(X)) \le I(\xi;X)$.

We now prove the first limit. Take densities $p := dP/d(P+Q)$ and $q := dQ/d(P+Q)$, so that M_w has density $wp + (1-w)q$. Since $M_w \ge wP$, the density $dP/dM_w = p/(wp+(1-w)q)$ is at most $1/w$ on $\{p > 0\}$, so $\mathrm{KL}(P\|M_w) \le \log(1/w)$ is finite, and

$$\mathrm{KL}(P\|M_w) = \int_{\{p>0\}} \log \frac{p}{wp+(1-w)q}\, dP.$$

Split the domain into $S_0 := \{p > 0, q = 0\}$ and $S_1 := \{p > 0, q > 0\}$. On S_0 the integrand equals $\log(1/w)$ exactly. The set S_0 has Q-measure zero, and P restricted to $\{q = 0\}$ is the singular part of P relative to Q, so $P(S_0) = \alpha(P \mid Q)$ and

$$\int_{S_0} \log \frac{p}{wp+(1-w)q}\, dP = \alpha(P \mid Q)\log\frac{1}{w}.$$

On S_1, write the integrand as $g_w = g_w^+ - g_w^-$. As $w \downarrow 0$ the ratio $p/(wp+(1-w)q)$ converges pointwise to $p/q \in (0,\infty)$, so g_w converges to the finite number $\log(p/q)$ and $g_w^+/\log(1/w) \to 0$ pointwise on S_1. The bound $g_w^+ \le \log(1/w)$ makes $g_w^+/\log(1/w) \le 1$, which is P-integrable, so dominated convergence gives $\int_{S_1} g_w^+\, dP = o(\log(1/w))$. For the negative part, put $m := wp + (1-w)q$ and note that on $\{m > p\}$,

$$p\log\frac{m}{p} = m\cdot\frac{p}{m}\log\frac{m}{p} \le \frac{m}{e},$$

again by $u\log(1/u) \le 1/e$. Integrating against $P+Q$ gives $\int_{S_1} g_w^-\, dP \le \frac{1}{e}\int m\, d(P+Q) = \frac{1}{e}$, uniformly in w. Dividing by $\log(1/w)$ and combining the three contributions,

$$\frac{\mathrm{KL}(P\|M_w)}{\log(1/w)} = \alpha(P \mid Q) + \frac{\int_{S_1} g_w^+\, dP}{\log(1/w)} - \frac{\int_{S_1} g_w^-\, dP}{\log(1/w)} \longrightarrow \alpha(P \mid Q).$$

For the second limit, the other component of JS_w is small: $dQ/dM_w \le 1/(1-w)$ gives $\mathrm{KL}(Q\|M_w) \le \log\frac{1}{1-w}$, and $-\log(1-w)/w \to 1$ while $\log(1/w) \to \infty$, so

$$\frac{(1-w)\mathrm{KL}(Q\|M_w)}{w\log(1/w)} \le \frac{-\log(1-w)}{w\log(1/w)} \longrightarrow 0.$$

Dividing $\mathrm{JS}_w(P,Q) = w\mathrm{KL}(P\|M_w) + (1-w)\mathrm{KL}(Q\|M_w)$ by $w\log(1/w)$ and using the first limit gives the second. Since $\alpha(P \mid Q) = 0$ exactly when $P \ll Q$, the divergence is $o(w\log(1/w))$ exactly in that case.

Finally assume $\mathrm{KL}(P\|Q) < \infty$. Apply Lemma A.6 to the two-component mixture $M_w = wP + (1-w)Q$ with reference Q. Its hypothesis holds because convexity of the divergence in its first argument gives $\mathrm{KL}(M_w\|Q) \le w\mathrm{KL}(P\|Q) + (1-w)\mathrm{KL}(Q\|Q) = w\mathrm{KL}(P\|Q) < \infty$. The lemma's identity reads

$$w\mathrm{KL}(P\|Q) + (1-w)\mathrm{KL}(Q\|Q) = w\mathrm{KL}(P\|M_w) + (1-w)\mathrm{KL}(Q\|M_w) + \mathrm{KL}(M_w\|Q),$$

that is, $w\mathrm{KL}(P\|Q) = \mathrm{JS}_w(P,Q) + \mathrm{KL}(M_w\|Q)$. Rearranging gives the stated identity, and dropping the nonnegative term $\mathrm{KL}(M_w\|Q)$ gives the inequality. □

B Strict separation for general observation spaces

Notation for the experiment and its codes. For $T = g(Y)$, write $P_{\theta Y}$ and $P_{\theta T}$ for the joint laws of (θ, Y) and (θ, T). The corresponding product laws are

$$Q_{\theta Y} := \pi \otimes P_Y, \qquad Q_{\theta T} := \pi \otimes P_T, \qquad I = \mathrm{KL}(P_{\theta Y}\|Q_{\theta Y}).$$

Here P_Y and P_T are the marginal laws of the observation and code. Use the cell probabilities $p_c = P_T(\{c\})$ and cell posteriors Π_c defined in Appendix A. For a measurable $A \subseteq \Theta$ and a positive-probability cell c,

$$P_{\theta T}(A \times \{c\}) = p_c \Pi_c(A), \qquad Q_{\theta T}(A \times \{c\}) = p_c \pi(A).$$

The laws $P_{\theta T}, Q_{\theta T}$ are the images of $P_{\theta Y}, Q_{\theta Y}$ under $(\theta, y) \mapsto (\theta, g(y))$: $P_{\theta T}$ retains the parameter–code dependence, while $Q_{\theta T}$ makes them independent with the same marginals.

Since $\mathrm{KL}(P_{\theta Y}\|Q_{\theta Y}) = I < \infty$, a joint density $f = dP_{\theta Y}/dQ_{\theta Y}$ exists. Versions of the conditional distributions satisfy

$$\Pi_y(d\theta) = f(\theta, y)\pi(d\theta), \qquad P_{Y|\theta}(dy) = f(\theta, y)P_Y(dy)$$

for almost every conditioning value. Indeed, Tonelli's theorem and the two marginal identities give $\int f(\theta, y)\pi(d\theta) = 1$ for P_Y-almost every y, and $\int f(\theta, y)P_Y(dy) = 1$ for π-almost every θ. Defining the kernels arbitrarily on the exceptional null sets gives regular conditional versions.

Proposition B.1 (Equivalent support formulations)**.** *Under $I(\theta; Y) < \infty$, Assumption 4.1 is equivalent to each of*

(i) $P_{\theta Y} \sim \pi \otimes P_Y$*;*

(ii) $\Pi_y \sim \pi$ *for P_Y-almost every y;*

(iii) $P_{Y|\theta} \sim P_Y$ *for π-almost every θ.*

For a dominated simulator with density $p(y \mid \theta)$, these conditions are equivalent to $p(y \mid \theta) > 0$ for $\pi \otimes P_Y$-almost every (θ, y).

Proof. All three conditions concern the zero set of the joint density f. We first show that each of (i), (ii), (iii) is equivalent to $Q_{\theta Y}\{f = 0\} = 0$, and then relate (iii) to the assumption as stated.

For (i): we already know $P_{\theta Y} \ll Q_{\theta Y}$, so (i) is the reverse relation $Q_{\theta Y} \ll P_{\theta Y}$. If $Q_{\theta Y}\{f = 0\} = 0$ and $P_{\theta Y}(A) = \int_A f\, dQ_{\theta Y} = 0$, then $f = 0$ almost everywhere on A under $Q_{\theta Y}$, so $Q_{\theta Y}(A) = Q_{\theta Y}(A \cap \{f = 0\}) = 0$. Conversely, the set $\{f = 0\}$ has $P_{\theta Y}$-measure $\int_{\{f=0\}} f\, dQ_{\theta Y} = 0$, so $Q_{\theta Y} \ll P_{\theta Y}$ forces $Q_{\theta Y}\{f = 0\} = 0$. Thus

$$\text{(i)} \iff Q_{\theta Y}\{f = 0\} = 0.$$

By Tonelli's theorem this common quantity can be computed in either order:

$$Q_{\theta Y}\{f = 0\} = \int \pi\{\theta : f(\theta, y) = 0\}\, P_Y(dy) = \int P_Y\{y : f(\theta, y) = 0\}\, \pi(d\theta).$$

For (ii): by the displayed version of the posterior, $\Pi_y \ll \pi$ holds for P_Y-almost every y with density $f(\cdot, y)$. The same one-variable argument as above shows that $\pi \ll \Pi_y$ holds exactly when $\pi\{\theta : f(\theta, y) = 0\} = 0$. Hence (ii) holds exactly when the integrand of the first integral vanishes for P_Y-almost every y, that is, exactly when $Q_{\theta Y}\{f = 0\} = 0$.

For (iii): likewise $P_{Y|\theta} \ll P_Y$ with density $f(\theta, \cdot)$ for π-almost every θ, and $P_Y \ll P_{Y|\theta}$ holds exactly when $P_Y\{y : f(\theta, y) = 0\} = 0$. Hence (iii) holds exactly when the integrand of the second integral vanishes for π-almost every θ, again exactly when $Q_{\theta Y}\{f = 0\} = 0$.

It remains to connect (iii) with Assumption 4.1, which asks for a measurable set Θ_0 with $\pi(\Theta_0) = 1$ on which all sampling laws are mutually absolutely continuous. Suppose first that the assumption holds and fix $\theta_0 \in \Theta_0$. Since $\pi(\Theta_0) = 1$, the predictive law is the mixture $P_Y = \int_{\Theta_0} P_{Y|\theta}\, \pi(d\theta)$. If $P_{Y|\theta_0}(A) = 0$, then $P_{Y|\theta}(A) = 0$ for every $\theta \in \Theta_0$ by equivalence, so $P_Y(A) = 0$. If $P_Y(A) = 0$, then $P_{Y|\theta}(A) = 0$ for π-almost every $\theta \in \Theta_0$, hence for at least one such θ, and equivalence gives $P_{Y|\theta_0}(A) = 0$. Thus $P_{Y|\theta_0} \sim P_Y$, and therefore $P_{Y|\theta} \sim P_{Y|\theta_0} \sim P_Y$ for every $\theta \in \Theta_0$, which is (iii). Conversely, suppose (iii) holds and let $\Theta_0 := \{\theta : P_Y\{y : f(\theta, y) = 0\} = 0\}$, a measurable set by Tonelli's theorem, intersected with the full-measure set on which the displayed version of $P_{Y|\theta}$ is a probability measure. Then $\pi(\Theta_0) = 1$ by (iii), and for $\theta, \theta' \in \Theta_0$ we have $P_{Y|\theta} \sim P_Y \sim P_{Y|\theta'}$, so the assumption holds with this Θ_0.

Finally, let the simulator be dominated by a sigma-finite measure ν with density $p(y \mid \theta)$ and predictive density $m(y) = \int p(y \mid \theta)\pi(d\theta)$, so that $P_Y(dy) = m(y)\nu(dy)$. Then $P_{\theta Y}(d\theta, dy) = p(y \mid \theta)\pi(d\theta)\nu(dy)$ and $Q_{\theta Y}(d\theta, dy) = m(y)\pi(d\theta)\nu(dy)$, so on the set $\{m > 0\}$, which carries all of P_Y, the joint density is $f(\theta, y) = p(y \mid \theta)/m(y)$. The condition $Q_{\theta Y}\{f = 0\} = 0$ is therefore exactly $p(y \mid \theta) > 0$ for $\pi \otimes P_Y$-almost every (θ, y). □

Scope of the support condition. The condition is mutual absolute continuity, not equality of topological supports. A sufficient statistic preserves it: conditional distributions of the raw observation given the statistic are parameter-independent, so equivalence of the statistic laws lifts through their shared kernel. It is also preserved by finite i.i.d. replication. A parameter-independent observation kernel having strictly positive densities on a common support gives equivalent sampling laws after integrating any latent dynamics. Finiteness of I must still be checked separately; equivalence alone does not bound mutual information.

For a code $T = g(Y)$ and a cell with $p_c > 0$, the cell posterior is the average of the observation posteriors over the cell, so its density relative to the prior is

$$f_c(\theta) = p_c^{-1} \int_{\{g=c\}} f(\theta, y) P_Y(dy) \le p_c^{-1} \int f(\theta, y) P_Y(dy) = p_c^{-1}$$

for π-almost every θ. Hence $\mathrm{KL}(\Pi_c \| \pi) = \int \log f_c \, d\Pi_c \le \log(1/p_c)$, and $I(\theta; T) = \sum_c p_c \mathrm{KL}(\Pi_c \| \pi) \le I$ by data processing. Define the integrable posterior-information function $\varphi(y) = \mathrm{KL}(\Pi_y \| \pi)$, whose integral is $\int \varphi \, dP_Y = I < \infty$, and, for $0 \le \varepsilon < 1$, set

$$\omega(\varepsilon) = \sup_{A : P_Y(A) \le \varepsilon} \int_A \varphi \, dP_Y, \qquad \psi(\varepsilon) = \log \frac{1}{1 - \varepsilon} + \omega(\varepsilon).$$

Absolute continuity of the integral gives $\omega(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$. Explicitly, $\int_A \varphi \, dP_Y \le M\varepsilon + \int (\varphi - M)_+ \, dP_Y$, and the second term tends to zero as $M \to \infty$. Both moduli are nondecreasing.

Lemma B.2 (Informative codes cannot be nearly constant)**.** *For $T \in \mathcal{G}$, let $p_{\max} = \max_c p_c$ and $\kappa(T) = 1 - \sum_c p_c^2$. Then*

$$1 - p_{\max} \le \kappa(T) \le 2(1 - p_{\max}), \qquad I(\theta; T) \le \psi(1 - p_{\max}).$$

Consequently, if $\psi(\delta) < u$, every code with relevance at least u has $\kappa(T) > \delta$.

Proof. The collision bounds come from comparing $\sum_c p_c^2$ with $p_{\max}$ in both directions. Since every $p_c \le p_{\max}$ and the p_c sum to one, $\sum_c p_c^2 \le p_{\max} \sum_c p_c = p_{\max}$, so $\kappa(T) \ge 1 - p_{\max}$. Since the largest cell alone contributes $p_{\max}^2$,

$$\kappa(T) \le 1 - p_{\max}^2 = (1 - p_{\max})(1 + p_{\max}) \le 2(1 - p_{\max}).$$

For the relevance bound, let b be a cell with $p_b = p_{\max}$ and split $I(\theta;T) = \sum_c p_c \mathrm{KL}(\Pi_c\|\pi)$ into the largest cell and the rest. For the largest cell, the density bound $f_b \le 1/p_b$ gives

$$p_b \mathrm{KL}(\Pi_b\|\pi) \le p_b \log \frac{1}{p_b} \le \log \frac{1}{p_b}.$$

For every other cell, the cell posterior is the mixture $\Pi_c = p_c^{-1} \int_{\{g=c\}} \Pi_y \, P_Y(dy)$. The divergence $P \mapsto \mathrm{KL}(P\|\pi)$ is convex, being a supremum of affine functionals of P by the Donsker–Varadhan representation, so the integral form of Jensen's inequality gives

$$p_c \mathrm{KL}(\Pi_c\|\pi) \le \int_{\{g=c\}} \mathrm{KL}(\Pi_y\|\pi)\, P_Y(dy) = \int_{\{g=c\}} \varphi \, dP_Y.$$

Summing over $c \ne b$, the cells $\{g = c\}$ with $c \ne b$ are disjoint with union $\{g \ne b\}$, a set of P_Y-probability $1 - p_b$. By the definition of ω,

$$\sum_{c \ne b} p_c \mathrm{KL}(\Pi_c\|\pi) \le \int_{\{g \ne b\}} \varphi \, dP_Y \le \omega(1 - p_b).$$

Adding the two parts, $I(\theta;T) \le \log(1/p_b) + \omega(1-p_b) = \psi(1-p_{\max})$.

For the last assertion, suppose $\psi(\delta) < u \le I(\theta;T)$. Then $\psi(1-p_{\max}) \ge I(\theta;T) > \psi(\delta)$, and since ψ is nondecreasing this forces $1 - p_{\max} > \delta$. The lower collision bound gives $\kappa(T) \ge 1 - p_{\max} > \delta$. □

Definition B.3 (Lautum information (Palomar & Verdú, 2008))**.** The lautum information of the experiment is the reverse product/joint divergence

$$L := \mathrm{KL}(Q_{\theta Y}\|P_{\theta Y}) = \int \mathrm{KL}(\pi\|\Pi_y)\, P_Y(dy),$$

and its code-level version is

$$L_T := \mathrm{KL}(Q_{\theta T}\|P_{\theta T}) = \sum_{c:p_c>0} p_c \mathrm{KL}(\pi\|\Pi_c).$$

Either quantity may be infinite. Mutual information uses the opposite direction, $I = \mathrm{KL}(P_{\theta Y}\|Q_{\theta Y})$; its finiteness does not imply $L < \infty$.

Since $P_{\theta T}, Q_{\theta T}$ are images of $P_{\theta Y}, Q_{\theta Y}$ under the same map, data processing gives $L_T \le L$. Finite lautum information will yield a quantitative bound; the qualitative separation needs only the support condition.

Lemma B.4 (Random relabeling: channel law and rate saving)**.** *For $T \in \mathcal{G}$ and $t \in (0,1)$, independently draw $\xi \sim \mathrm{Bernoulli}(t)$ and $T' \sim P_T$, and set $V_t = T$ if $\xi = 0$ and $V_t = T'$ otherwise.*

(i) Relabeling preserves the code marginal and mixes the joint law.

$$V_t \sim P_T, \qquad P_{\theta V_t} = (1-t)P_{\theta T} + tQ_{\theta T}.$$

(ii) Relabeling gives a $t\log(1/t)$ lower bound on the rate saving. *We have $I(Y;V_t) = I(T;V_t)$, and*

$$\begin{aligned} B_t(T) &:= H(T) - I(Y;V_t) \\ &= H(T \mid V_t) = H(V_t \mid T) \ge \kappa(T) t \log(1/t). \end{aligned} \tag{16}$$

(iii) A nonconstant code retains positive information after relabeling.

$$0 < B_t(T) < H(T) \qquad \text{if } H(T) > 0.$$

Proof. The pair (ξ, T') is independent of (θ, Y), so V_t is obtained from T by a channel that does not look at θ or Y; in particular $\theta \to Y \to T \to V_t$ is a Markov chain.

For (i), condition on the coin. On $\{\xi = 0\}$ the pair (θ, V_t) equals (θ, T), whose law is $P_{\theta T}$. On $\{\xi = 1\}$ it equals (θ, T'), and T' is independent of θ with law P_T, so its law is $\pi \otimes P_T = Q_{\theta T}$. Averaging over the coin,

$$P_{\theta V_t} = (1-t)P_{\theta T} + tQ_{\theta T}.$$

Both components have θ-marginal π and code marginal P_T, so $V_t \sim P_T$.

For (ii), the Markov chain $Y \to T \to V_t$ gives $I(Y;V_t) \le I(T;V_t)$, and T being a function of Y gives $I(T;V_t) \le I(Y;V_t)$; hence $I(Y;V_t) = I(T;V_t)$. All entropies below are finite because $H(T) < \infty$ and $H(V_t) = H(T)$. Writing $I(T;V_t) = H(T) - H(T \mid V_t)$,

$$B_t(T) = H(T) - I(T;V_t) = H(T \mid V_t).$$

The joint law of (T, V_t) is

$$\mathbb{P}(T = c, V_t = c') = (1-t)p_c \mathbf{1}\{c = c'\} + t p_c p_{c'},$$

which is symmetric in (c, c') and has both marginals equal to P_T. Therefore $H(T \mid V_t) = H(T, V_t) - H(V_t) = H(T, V_t) - H(T) = H(V_t \mid T)$. To bound $H(V_t \mid T)$ from below, condition on $T = c$: the label V_t equals c with probability $(1-t) + tp_c$ and equals each $c' \neq c$ with probability $tp_{c'}$. Dropping the nonnegative diagonal term and using $p_{c'} \le 1$,

$$H(V_t \mid T = c) \;\ge\; \sum_{c' \neq c} t p_{c'} \log \frac{1}{t p_{c'}} \;\ge\; t \log \frac{1}{t} \sum_{c' \neq c} p_{c'} \;=\; t \log \frac{1}{t} (1 - p_c).$$

Averaging over c with weights p_c,

$$B_t(T) = \sum_c p_c H(V_t \mid T = c) \;\ge\; t \log \frac{1}{t} \sum_c p_c (1 - p_c) = \kappa(T)\, t \log \frac{1}{t}.$$

For (iii), let T be nonconstant, so that $0 < p_c < 1$ for at least one cell and $\kappa(T) = \sum_c p_c(1-p_c) > 0$. The lower bound in (ii) then gives $B_t(T) > 0$. For the upper bound, $B_t(T) < H(T)$ is the same as $I(T;V_t) > 0$, that is, T and V_t are not independent. If they were, the diagonal of the joint law would satisfy $(1-t)p_c + tp_c^2 = p_c^2$, that is, $(1-t)p_c(1-p_c) = 0$ for every c, which forces every $p_c \in \{0, 1\}$ and contradicts nonconstancy. □

Lemma B.5 (Posterior cost of relabeling and uniform bounds)**.** *For the channel in Lemma B.4, let $L_t(T) := \Delta_{\mathrm{KL}}(V_t) - \Delta_{\mathrm{KL}}(T)$ be its added posterior distortion. The following bounds separate the exact cost from the conditions used to control it.*

(i) The exact cost is a relevance term plus a mixture divergence.

$$L_t(T) = tI(\theta;T) + \mathrm{JS}_t(Q_{\theta T}, P_{\theta T}). \tag{17}$$

(ii) One experiment-level quantity bounds the cost for every code.

$$0 \le L_t(T) \le A(t) := tI + \mathrm{JS}_t(Q_{\theta Y}, P_{\theta Y}) \le tI + h_2(t). \tag{18}$$

(iii) Common support makes this uniform cost smaller than the saved-rate scale. *Under Assumption 4.1,*

$$A(t) = o(t \log(1/t)) \qquad (t \downarrow 0).$$

(iv) Finite experiment-level lautum information gives a linear bound.

$$L < \infty \quad \Longrightarrow \quad A(t) \le t(I + L).$$

(v) Finite code-level lautum information gives a second exact identity. *If $L_T < \infty$, then*

$$L_t(T) = t\{I(\theta;T) + L_T\} - \mathrm{KL}((1-t)P_{\theta T} + tQ_{\theta T} \| P_{\theta T}).$$

Proof. By Lemma 2.2, $\Delta_{\mathrm{KL}}(V_t) - \Delta_{\mathrm{KL}}(T) = I(\theta;T) - I(\theta;V_t)$, and this is nonnegative by data processing along $\theta \to T \to V_t$. The exact identity in (i) comes from expanding $I(\theta;V_t)$ with the compensation lemma.

For (i), Lemma B.4(i) gives $P_{\theta V_t} = \bar{P} := (1-t)P_{\theta T} + tQ_{\theta T}$ with marginals π and P_T, so

$$I(\theta;V_t) = \mathrm{KL}(\bar{P} \| \pi \otimes P_{V_t}) = \mathrm{KL}(\bar{P} \| Q_{\theta T}).$$

Apply Lemma A.6 to the two-component mixture $\bar{P}$ with reference $Q_{\theta T}$. Its finite-divergence hypothesis holds because convexity of the divergence gives $\mathrm{KL}(\bar{P}\|Q_{\theta T}) \le (1-t)\mathrm{KL}(P_{\theta T}\|Q_{\theta T}) + t\cdot 0 = (1-t)I(\theta;T) \le I < \infty$. The compensation identity reads

$$(1-t)\mathrm{KL}(P_{\theta T}\|Q_{\theta T}) + t\mathrm{KL}(Q_{\theta T}\|Q_{\theta T}) = \big[(1-t)\mathrm{KL}(P_{\theta T}\|\bar{P}) + t\mathrm{KL}(Q_{\theta T}\|\bar{P})\big] + \mathrm{KL}(\bar{P}\|Q_{\theta T}).$$

The left side is $(1-t)I(\theta;T)$. The bracket on the right is $\mathrm{JS}_t(Q_{\theta T}, P_{\theta T})$, since $\bar{P} = tQ_{\theta T} + (1-t)P_{\theta T}$ is the weighted mixture with the rare component $Q_{\theta T}$ carrying weight t. The last term is $I(\theta; V_t)$. Hence

$$(1-t)I(\theta;T) = \mathrm{JS}_t(Q_{\theta T}, P_{\theta T}) + I(\theta; V_t),$$

and therefore

$$\begin{aligned} L_t(T) &= I(\theta;T) - I(\theta;V_t) \\ &= I(\theta;T) - (1-t)I(\theta;T) + \mathrm{JS}_t(Q_{\theta T}, P_{\theta T}) \\ &= tI(\theta;T) + \mathrm{JS}_t(Q_{\theta T}, P_{\theta T}). \end{aligned}$$

For (ii), bound the two terms of (17) separately. First, $I(\theta;T) \le I$. Second, the map $(\theta, y) \mapsto (\theta, g(y))$ sends $Q_{\theta Y}$ to $Q_{\theta T}$ and $P_{\theta Y}$ to $P_{\theta T}$, so the data-processing property of Lemma A.7 gives $\mathrm{JS}_t(Q_{\theta T}, P_{\theta T}) \le \mathrm{JS}_t(Q_{\theta Y}, P_{\theta Y})$. Together, $L_t(T) \le tI + \mathrm{JS}_t(Q_{\theta Y}, P_{\theta Y}) = A(t)$, and the entropy bound $\mathrm{JS}_t \le h_2(t)$ from the same lemma gives the last inequality.

For (iii), Proposition B.1(i) turns the support assumption into $Q_{\theta Y} \ll P_{\theta Y}$, so the directed singular mass $\alpha(Q_{\theta Y} \mid P_{\theta Y})$ is zero. The second limit in Lemma A.7, applied with the rare component $Q_{\theta Y}$, gives $\mathrm{JS}_t(Q_{\theta Y}, P_{\theta Y}) = o(t\log(1/t))$. The other term satisfies $tI/(t\log(1/t)) = I/\log(1/t) \to 0$. Hence $A(t) = o(t\log(1/t))$.

For (iv), if $L = \mathrm{KL}(Q_{\theta Y}\|P_{\theta Y}) < \infty$, the final inequality of Lemma A.7 with $P = Q_{\theta Y}$ and $Q = P_{\theta Y}$ gives $\mathrm{JS}_t(Q_{\theta Y}, P_{\theta Y}) \le tL$, so $A(t) \le tI + tL$.

For (v), if $L_T = \mathrm{KL}(Q_{\theta T}\|P_{\theta T}) < \infty$, the final identity of Lemma A.7 with $P = Q_{\theta T}$ and $Q = P_{\theta T}$ reads

$$\mathrm{JS}_t(Q_{\theta T}, P_{\theta T}) = tL_T - \mathrm{KL}\big(tQ_{\theta T} + (1-t)P_{\theta T} \,\big\|\, P_{\theta T}\big).$$

Substituting this into (17) gives the stated formula. □

B.1 Proof of Theorem 4.2

Proof. Fix $R \in (0, R_0)$ and write $D = D^{\mathrm{ts}}(R)$. The argument has three stages. We first fix constants that depend only on the experiment and on R; then we take a deterministic mixture that is nearly optimal at rate R and relabel each of its components, which saves rate at a $t\log(1/t)$ scale while costing distortion at a smaller scale; finally we spend the saved rate on a fixed finite code with distortion below D. Fixing the constants before choosing the mixture is what makes the argument independent of attainment.

We begin by showing $0 < D < I$. Positivity holds because $R < R_0$ and $D^{\mathrm{ts}} > 0$ on $[0, R_0)$ by Proposition A.3. For the upper bound, Lemma A.2 supplies a finite code U' with $I(\theta;U') > 0$; such a code is nonconstant, so $H(U') > 0$. Time-sharing U' with the constant code at weight $\lambda := \min\{1, R/H(U')\}$ on U' has rate $\lambda H(U') \le R$ and, by (15) and Lemma 2.2, distortion

$$\lambda\Delta_{\mathrm{KL}}(U') + (1-\lambda)I = I - \lambda I(\theta;U') < I.$$

Hence $D < I$, and we set $a := I - D > 0$.

Next, Lemma A.2 also supplies a finite code U with $d_U := \Delta_{\mathrm{KL}}(U) < D$. Its rate $r_U := H(U)$ must exceed R: otherwise U alone would be a feasible one-component mixture at rate R with distortion below $D^{\mathrm{ts}}(R)$.

We now fix the remaining constants. Since $\psi(\varepsilon) \to 0$ as $\varepsilon \downarrow 0$, choose $\delta \in (0,1)$ with $\psi(\delta) < a/4$, and set

$$K := \frac{a\delta}{4I} > 0, \qquad c := \frac{K(D - d_U)}{r_U} > 0.$$

By Lemma B.5(iii), $A(t)/(t\log(1/t)) \to 0$, so we can choose $t \in (0, 1/2]$ with

$$A(t) \le \frac{c}{2}\, t\log\frac{1}{t}.$$

Finally choose η with

$$0 < \eta < \min\Big\{\frac{a}{2},\ \frac{c}{4}\, t \log\frac{1}{t}\Big\}.$$

Hence, every constant so far depends only on the experiment, R, and the choices of U and δ.

We now take a nearly optimal mixture. By the definition of $D^{\mathrm{ts}}(R)$ as an infimum and Lemma A.4, there is a finite revealed mixture of finite codes $(\lambda_j, T_j)_{j\le k}$ with average rate $r := \sum_j \lambda_j H(T_j) \le R$ and average distortion $d := \sum_j \lambda_j \Delta_{\mathrm{KL}}(T_j)$ satisfying $D \le d \le D + \eta$. By Lemma 2.2, its average relevance is

$$\sum_j \lambda_j I(\theta; T_j) = I - d \ \ge\ I - D - \eta = a - \eta > \frac{a}{2}.$$

Let v be the total weight of the components with $I(\theta; T_j) \ge a/4$. Bounding the relevance of the other components by $a/4$ and that of all components by I,

$$\frac{a}{2} < \sum_j \lambda_j I(\theta; T_j) \le \frac{a}{4}(1 - v) + Iv \le \frac{a}{4} + Iv,$$

so $v > a/(4I)$. Each component counted in v has relevance at least $a/4 > \psi(\delta)$, so Lemma B.2 gives $\kappa(T_j) > \delta$ for it. Therefore

$$\sum_j \lambda_j \kappa(T_j) \ \ge\ \delta v \ >\ \frac{a\delta}{4I} = K.$$

We now relabel. Apply the channel of Lemma B.4 to every component T_j with the same t, using independent coins and draws for different components, and keep the original selector J visible. The result is the revealed mixture $S' := (J, V_{t,J})$ of the relabeled codes $V_{t,j}$, which is a summary channel because all added randomness is independent of the experiment. By (15) and Lemma B.4(ii), its rate is

$$r' = \sum_j \lambda_j I(Y; V_{t,j}) = \sum_j \lambda_j \{H(T_j) - B_t(T_j)\} = r - B,$$

where the total saving satisfies

$$B := \sum_j \lambda_j B_t(T_j) \ \ge\ \sum_j \lambda_j \kappa(T_j)\, t \log\frac{1}{t} \ \ge\ Kt \log\frac{1}{t} > 0.$$

By (15) and Lemma B.5(ii), its distortion is

$$\Delta_{\mathrm{KL}}(S') = \sum_j \lambda_j \{\Delta_{\mathrm{KL}}(T_j) + L_t(T_j)\} = d + L', \qquad 0 \le L' := \sum_j \lambda_j L_t(T_j) \le A(t).$$

In particular $0 \le r' < r \le R < r_U$.

We now spend the saved rate. Use a fresh revealed selector, independent of everything else, to transmit U with probability

$$u := \frac{R - r'}{r_U - r'} \in (0, 1)$$

and S' otherwise. By (15) the final summary S has rate $(1 - u)r' + ur_U = r' + u(r_U - r') = R$ exactly, and distortion

$$\Delta_{\mathrm{KL}}(S) = (1 - u)(d + L') + ud_U = (d + L') - u\{(d + L') - d_U\}.$$

Two facts control the subtracted term. First, $d \ge D > d_U$ and $L' \ge 0$ give $(d + L') - d_U \ge D - d_U > 0$. Second, using $R \ge r$ and $r_U - r' \le r_U$,

$$u = \frac{R - r'}{r_U - r'} \ \ge\ \frac{r - r'}{r_U} = \frac{B}{r_U} \ \ge\ \frac{Kt \log(1/t)}{r_U}.$$

Therefore

$$\begin{aligned}\Delta_{\mathrm{KL}}(S) &\le d + L' - \frac{Kt\log(1/t)}{r_U}(D - d_U)\\ &\le D + \eta + A(t) - c\,t\log\frac{1}{t}\\ &\le D + \frac{c}{4}\,t\log\frac{1}{t} + \frac{c}{2}\,t\log\frac{1}{t} - c\,t\log\frac{1}{t}\\ &= D - \frac{c}{4}\,t\log\frac{1}{t} \;<\; D,\end{aligned}$$

where the second line uses $d \le D + \eta$, $L' \le A(t)$, and the definition of c, and the third line uses the choices of η and t.

Every component code and every selector takes finitely many values, so the alphabet of S is finite, and all randomization is generated independently of the experiment, so S is a valid summary channel of rate exactly R. Hence $D^\star(R) \le \Delta_{\mathrm{KL}}(S) < D^{\mathrm{ts}}(R)$, which is (10); no attainment of either frontier was used.

For the endpoints, a summary with $I(Y;S) = 0$ is independent of Y, hence of θ, so $\Delta_{\mathrm{KL}}(S) = I - I(\theta;S) = I$ and $D^\star(0) = I = D^{\mathrm{ts}}(0)$. If $R_0 < \infty$, Proposition A.3 gives $D^{\mathrm{ts}} = 0$ on $[R_0, \infty)$, and $0 \le D^\star \le D^{\mathrm{ts}}$ gives the same for $D^\star$. □

Corollary B.6 (Pointwise exclusion of deterministic near-optimizers)**.** *For each fixed $R \in (0, R_0)$, put $\gamma_R = D^{\mathrm{ts}}(R) - D^\star(R) > 0$. A channel with $I(Y;S) \le R$ and $\Delta_{\mathrm{KL}}(S) < D^\star(R) + \gamma_R$ cannot be a finite revealed mixture of deterministic codes, including a deterministic code itself. Every rate-feasible minimizing sequence eventually lies outside this class. The gap is pointwise in R, not asserted uniform over rates, experiments, or neural model classes.*

Proof. Theorem 4.2 gives $\gamma_R > 0$. A finite revealed mixture of deterministic codes with average rate at most R realizes a point $(r, d) \in \mathrm{conv}(\mathcal{A}_{\mathrm{det}})$ with $r \le R$, so by the definition of the time-sharing frontier $d \ge D^{\mathrm{ts}}(R) = D^\star(R) + \gamma_R$; a single deterministic code is the one-component case. If S_n are channels with $I(Y;S_n) \le R$ and $\Delta_{\mathrm{KL}}(S_n) \to D^\star(R)$, then $\Delta_{\mathrm{KL}}(S_n) < D^\star(R) + \gamma_R$ for all large n, so these S_n are not deterministic mixtures. □

Corollary B.7 (A quantitative gap under finite lautum information)**.** *If $L < \infty$, retain a, U, δ, K, c from the proof above. For every $t \in (0, 1/2]$ with $\log(1/t) \ge 2(I + L)/c$,*

$$D^\star(R) \le D^{\mathrm{ts}}(R) - \frac{c}{2}t\log(1/t).$$

Proof. By Lemma B.5(iv), $A(t) \le t(I + L)$, and the condition $\log(1/t) \ge 2(I + L)/c$ gives

$$A(t) \le t(I + L) \le \frac{c}{2}\,t\log\frac{1}{t},$$

so this t satisfies the requirement imposed on t in the proof of Theorem 4.2. Run the construction of that proof with this t and with any $\eta \in (0, a/2)$; the restriction $\eta < a/2$ is all that was used to obtain $v > a/(4I)$, and the final chain of inequalities becomes

$$\Delta_{\mathrm{KL}}(S) \le D + \eta + A(t) - c\,t\log\frac{1}{t} \le D + \eta - \frac{c}{2}\,t\log\frac{1}{t}.$$

Since S has rate exactly R, $D^\star(R) \le D + \eta - \frac{c}{2}t\log(1/t)$ for every such η, and letting $\eta \downarrow 0$ gives the claim. □

B.2 Equal initial slopes do not rule out a strict gap

The separation theorem concerns every positive interior rate, whereas an initial-slope comparison describes only the first-order behavior at zero. In information-bottleneck theory, the best relevance-to-rate ratio is the sharp constant in a mutual-information strong data-processing inequality; it also determines the frontier's initial slope (Asoodeh & Calmon, 2020). For the Gaussian experiment this constant is ρ^2, consistently with the classical Gaussian IB solution (Chechik et al., 2005).

Proposition B.8 shows that deterministic threshold codes approach the same constant as stochastic channels. Nevertheless, the deterministic time-sharing frontier remains strictly above the stochastic frontier at every positive rate. Thus a strict inequality between initial slopes is unnecessary for our theorem: the Gaussian example has the same first-order information efficiency in both classes while retaining a strict gap at each fixed positive rate.

Proposition B.8 (Gaussian initial slopes)**.** *Let (θ, X) be standard jointly Gaussian with correlation $\rho \in (0,1)$, with X the observation. Define*

$$\beta_{\mathrm{det}} = \sup_{g:\, 0<H(g(X))<\infty} \frac{I(\theta; g(X))}{H(g(X))}, \qquad s^\star = \sup_{S:\, 0<I(X;S)<\infty} \frac{I(\theta; S)}{I(X;S)}.$$

Then $\beta_{\mathrm{det}} = s^\star = \rho^2$, the deterministic supremum is not attained, and the right derivatives of D^{ts} and $D^\star$ at zero both equal $-\rho^2$. Nevertheless $D^\star(R) < D^{\mathrm{ts}}(R)$ for every $R > 0$.

Proof. We first identify the stochastic frontier and read off $s^\star$ from its tangent at the origin; then we show that threshold codes approach the same ratio, which gives β_{det}; then we convert the two ratios into initial slopes; and finally we verify the hypotheses of Theorem 4.2.

Write $X = \rho\theta + \sqrt{1-\rho^2}\, W$ with W standard normal and independent of θ. Rescaling the observation to X/ρ changes no mutual information and puts the model in the form of Theorem 3.4 with $d = 1$, $n = 1$, $\tau^2 = 1$, and $\sigma^2 = (1-\rho^2)/\rho^2$, so that $a := n\tau^2/\sigma^2 = \rho^2/(1-\rho^2)$. The theorem gives

$$D^\star(R) = \tfrac{1}{2}\log(1 + ae^{-2R}), \qquad I = D^\star(0) = \tfrac{1}{2}\log(1+a) = \tfrac{1}{2}\log\frac{1}{1-\rho^2}.$$

Differentiating,

$$\frac{d}{dR}D^\star(R) = -\frac{ae^{-2R}}{1+ae^{-2R}}, \qquad \frac{d^2}{dR^2}D^\star(R) = \frac{2ae^{-2R}}{(1+ae^{-2R})^2} > 0,$$

so $D^\star$ is strictly convex, positive at every finite R, and has slope $-a/(1+a) = -\rho^2$ at zero. Strict convexity places the graph above its tangent at the origin,

$$D^\star(R) \;\geq\; I - \rho^2 R \qquad \text{with strict inequality for } R > 0.$$

For any channel with $0 < I(X;S) < \infty$, Lemma 2.2 gives $I(\theta;S) = I - \Delta_{\mathrm{KL}}(S) \leq I - D^\star(I(X;S)) \leq \rho^2 I(X;S)$, so $s^\star \leq \rho^2$. The Gaussian channels in the achievability part of Theorem 3.4 attain $\Delta_{\mathrm{KL}} = D^\star(R)$ at every rate $R > 0$, and for them $I(\theta;S)/I(X;S) = \{I - D^\star(R)\}/R \to -\frac{d}{dR}D^\star(0) = \rho^2$ as $R \downarrow 0$. Hence $s^\star = \rho^2$.

We now show that threshold codes approach the same ratio. For $b > 0$ let $T_b = \mathbf{1}\{X > b\}$, $p_b = \mathbb{P}(X > b)$, and $\mu_b = \mathbb{E}[X \mid X > b]$. The rate of T_b is $H(T_b) = h_2(p_b)$. For its relevance, keep only the cell $\{X > b\}$ in $I(\theta;T_b) = \sum_c p_c \mathrm{KL}(\Pi_c \| \pi)$:

$$I(\theta;T_b) \;\geq\; p_b\, \mathrm{KL}\big(P_{\theta|X>b} \,\big\|\, N(0,1)\big).$$

The conditional mean of θ in this cell is $\mathbb{E}[\theta \mid X > b] = \mathbb{E}[\mathbb{E}[\theta \mid X] \mid X > b] = \rho\mu_b$. A law P on $\mathbb{R}$ with finite mean m satisfies $\mathrm{KL}(P\|N(0,1)) \geq m^2/2$: the Donsker–Varadhan inequality $\mathrm{KL}(P\|N(0,1)) \geq \int g\, dP - \log \int e^g\, dN(0,1)$ with $g(x) = sx$ gives $sm - s^2/2$, and the choice $s = m$ gives $m^2/2$. Therefore

$$I(\theta;T_b) \;\geq\; \frac{p_b \rho^2 \mu_b^2}{2}, \qquad H(T_b) = h_2(p_b).$$

As $b \to \infty$, the Gaussian tail estimates give $p_b \sim \phi(b)/b$ and $\mu_b = \phi(b)/p_b \sim b$, where ϕ is the standard normal density, hence $\log(1/p_b) \sim b^2/2$; also $h_2(p_b) = p_b \log(1/p_b) + (1-p_b)\log\frac{1}{1-p_b}$, whose second term is of order p_b and negligible against the first, so $h_2(p_b) \sim p_b \log(1/p_b) \sim p_b b^2/2$. Consequently

$$\liminf_{b\to\infty} \frac{I(\theta;T_b)}{H(T_b)} \;\geq\; \lim_{b\to\infty} \frac{p_b\rho^2 b^2/2}{p_b b^2/2} = \rho^2.$$

A deterministic code is a channel with $I(X;g(X)) = H(g(X))$, so $\beta_{\mathrm{det}} \leq s^\star = \rho^2$, and the threshold codes show $\beta_{\mathrm{det}} \geq \rho^2$. Thus $\beta_{\mathrm{det}} = s^\star = \rho^2$. The deterministic supremum is not attained:

a code with $0 < H(g(X)) = R < \infty$ has $I(\theta; g(X)) = I - \Delta_{\mathrm{KL}}(g(X)) \le I - D^\star(R) < \rho^2 R$ by the strict tangent inequality at $R > 0$.

Next we convert the ratios into initial slopes. For the time-sharing frontier, every deterministic component satisfies $I(\theta; g_j(X)) \le \beta_{\mathrm{det}} H(g_j(X))$, so a mixture with average rate at most R has distortion $I - \sum_j \lambda_j I(\theta; g_j(X)) \ge I - \beta_{\mathrm{det}} R$; hence $D^{\mathrm{ts}}(R) \ge I - \beta_{\mathrm{det}} R$ and the right derivative at zero, which exists by convexity, is at least $-\beta_{\mathrm{det}}$. Conversely, fix a code g with $0 < H(g(X)) =: r < \infty$ and relevance $\iota := I(\theta; g(X))$. Time-sharing it with the constant code at weight R/r for $R \le r$ gives rate R and distortion $I - (\iota/r)R$, so $D^{\mathrm{ts}}(R) \le I - (\iota/r)R$ near zero and the right derivative is at most $-\iota/r$. Taking the supremum over g gives at most $-\beta_{\mathrm{det}}$. The right derivative of D^{ts} at zero is therefore $-\beta_{\mathrm{det}} = -\rho^2$. For the unrestricted frontier the same two-sided argument with $s^\star$ in place of β_{det} applies, and agrees with the direct computation $\frac{d}{dR} D^\star(0) = -\rho^2$ above.

Finally we check the hypotheses of Theorem 4.2. The joint law of (θ, X) is a nondegenerate bivariate Gaussian and the product law $\pi \otimes P_X$ is a standard bivariate Gaussian; both have strictly positive densities on $\mathbb{R}^2$, so they are mutually absolutely continuous, which is condition (i) of Proposition B.1. The Bayes information $I = \frac{1}{2}\log(1+a)$ lies in $(0, \infty)$, and the lautum information is also finite: for the two zero-mean Gaussians with covariances $\Sigma_1 = I_2$ and $\Sigma_0 = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$,

$$\begin{aligned} L = \mathrm{KL}(N(0,\Sigma_1)\|N(0,\Sigma_0)) &= \tfrac{1}{2}\big\{ \operatorname{tr}(\Sigma_0^{-1}\Sigma_1) - 2 - \log\det(\Sigma_0^{-1}\Sigma_1)\big\} \\ &= \frac{\rho^2}{1-\rho^2} - \tfrac{1}{2}\log\frac{1}{1-\rho^2} = a - \tfrac{1}{2}\log(1+a), \end{aligned}$$

using $\operatorname{tr}\Sigma_0^{-1} = 2/(1-\rho^2)$ and $\det\Sigma_0^{-1} = 1/(1-\rho^2)$. Since $D^{\mathrm{ts}}(R) \ge D^\star(R) > 0$ at every finite R, the threshold is $R_0 = \infty$, and Theorem 4.2 gives $D^\star(R) < D^{\mathrm{ts}}(R)$ for every $R > 0$. □

C Finite observations: strict improvement and attainment

Finite observation spaces reveal both additional conclusions and a concrete mechanism behind the general theorem. The finite set of deterministic partitions makes its time-sharing envelope attainable and identifies the zero-distortion threshold exactly. A perturbation that confuses just two cells then expresses the distortion cost per saved nat through the directed singular mass of their posteriors. This gives a criterion that can prove strict improvement even when common support fails, as in the unequal-support example of Appendix D.

Finite observations also permit an existence result beyond the general separation statement: the stochastic optimum is attained with at most $|\mathcal{Y}| + 1$ outputs, including when the parameter space is infinite. The final construction connects these results to the count benchmark by specifying a feasible one-bit channel and the exhaustive deterministic comparison. Thus the appendix supplies a support-sensitive extension, an attainment result, and an explicit example, while also giving an elementary finite-cell explanation of the separation mechanism.

In this section, we assume that Y has finite support with m positive-probability atoms. The parameter space may be any standard Borel space. Every deterministic code is a partition of the observation support, so there are finitely many distinct deterministic rate–distortion pairs and their convex hull is compact.

Lemma C.1 (Finite-experiment envelope)**.** *Every $D^{\mathrm{ts}}(R)$ is attained by a finite mixture. Its zero-distortion threshold is $R_0 = H(\Pi_Y)$, which is positive when $I > 0$. At each $0 < R < R_0$, every optimal mixture has positive distortion and uses rate exactly R.*

Proof. Two deterministic codes that induce the same partition of the m atoms have the same entropy and the same cell posteriors, hence the same rate–distortion pair. So $\mathcal{A}_{\mathrm{det}}$ is a finite set and $\mathrm{conv}(\mathcal{A}_{\mathrm{det}})$ is a compact convex polygon in $\mathbb{R}^2$.

For attainment, fix $R \ge 0$. The feasible set $\{(r,d) \in \mathrm{conv}(\mathcal{A}_{\mathrm{det}}) : r \le R\}$ is a closed subset of a compact set, and it is nonempty because it contains the constant code's point $(0, I)$. The continuous coordinate $(r,d) \mapsto d$ therefore attains its minimum $D^{\mathrm{ts}}(R)$ at some feasible point, and every point of the convex hull is a finite convex combination of points of $\mathcal{A}_{\mathrm{det}}$, which by Proposition A.3 is realized by a finite revealed mixture.

For the threshold, consider the posterior label code $g(y) := \Pi_y$, which takes at most m values. Since $g(Y)$ determines Π_Y, the barycenter formula gives $\Pi_{g(Y)} = \mathbb{E}[\Pi_Y \mid g(Y)] = \Pi_Y$, so $\Delta_{\mathrm{KL}}(g(Y)) = 0$, and its rate is $H(g(Y)) = H(\Pi_Y)$. Thus $D^{\mathrm{ts}}(H(\Pi_Y)) = 0$ and $R_0 \le H(\Pi_Y)$. Conversely, suppose $D^{\mathrm{ts}}(R) = 0$ for some R, and let (λ_j, T_j) be an optimal mixture at rate R, which exists by the first part. Since $\sum_j \lambda_j \Delta_{\mathrm{KL}}(T_j) = 0$ with all $\lambda_j > 0$, every component has $\Delta_{\mathrm{KL}}(T_j) = 0$, and Lemma A.5 gives $H(T_j) \ge H(\Pi_Y)$ for each j. The average rate is then at least $H(\Pi_Y)$, so feasibility forces $R \ge H(\Pi_Y)$. Hence $R_0 \ge H(\Pi_Y)$, and the two bounds give $R_0 = H(\Pi_Y)$.

The threshold is positive when $I > 0$: $H(\Pi_Y) = 0$ means that Π_Y is almost surely constant, and since $\mathbb{E}\,\Pi_Y = \pi$ this constant is π, which gives $I = \mathbb{E}\,\mathrm{KL}(\Pi_Y \| \pi) = 0$.

Finally fix $0 < R < R_0$. Every optimal mixture has distortion $D^{\mathrm{ts}}(R) > 0$ by Proposition A.3. Suppose an optimal mixture had average rate $r < R$. Mix it, through a fresh revealed selector, with the posterior label code, which has rate R_0 and distortion zero, giving the label code weight $u \in (0, 1)$. The new average rate is $(1-u)r + uR_0$, which is at most R when $u \le (R-r)/(R_0-r)$, a positive number because $r < R < R_0$. The new distortion is $(1-u)D^{\mathrm{ts}}(R) < D^{\mathrm{ts}}(R)$. This feasible mixture is strictly better, contradicting optimality; so every optimal mixture uses rate exactly R. □

Lemma C.2 (Rare confusion between two cells)**.** *Choose two cells $a \ne b$ of a deterministic code T, with masses $p_a, p_b > 0$ and posteriors Π_a, Π_b. Change label a to b with probability $t \in (0, 1)$, using a hidden independent coin, and leave all other labels unchanged. Put $w_t = p_a t/(p_b + p_a t)$ and $M_t = w_t \Pi_a + (1 - w_t)\Pi_b$. The resulting summary V_t has rate saving and added distortion*

$$B_t = H(T) - I(Y; V_t) = (p_b + p_a t) h_2(w_t) > 0, \tag{19}$$

$$L_t = \Delta_{\mathrm{KL}}(V_t) - \Delta_{\mathrm{KL}}(T) = p_a t \mathrm{KL}(\Pi_a \| M_t) + p_b \mathrm{KL}(\Pi_b \| M_t). \tag{20}$$

Moreover $B_t \sim p_a t \log(1/t)$ and $L_t / B_t \to \alpha(\Pi_a \mid \Pi_b)$. If $\Pi_a \ll \Pi_b$, then $L_t = o(B_t)$; if $\mathrm{KL}(\Pi_a \| \Pi_b) < \infty$, then $L_t \le p_a t \mathrm{KL}(\Pi_a \| \Pi_b)$.

Proof. The channel from T to V_t uses only T and an independent coin, so $\theta \to Y \to T \to V_t$ is a Markov chain, and as in Lemma B.4 we have $I(Y; V_t) = I(T; V_t)$. Both the rate saving and the added distortion are conditional quantities given V_t, and only one output value of V_t leaves anything uncertain.

For the rate saving, $B_t = H(T) - I(T; V_t) = H(T \mid V_t)$. If $V_t = c$ with $c \ne b$, then $T = c$ with certainty, because a label is changed only into b. If $V_t = b$, then either $T = b$, or $T = a$ and the coin came up. The output b has probability $p_b + p_a t$, and conditionally on it,

$$\mathbb{P}(T = a \mid V_t = b) = \frac{p_a t}{p_b + p_a t} = w_t, \qquad \mathbb{P}(T = b \mid V_t = b) = 1 - w_t.$$

Hence $H(T \mid V_t) = (p_b + p_a t) h_2(w_t)$, which is positive because $w_t \in (0, 1)$.

For the added distortion, Lemma 2.2 and the chain rule give, as in Lemma A.4,

$$L_t = I(\theta; T) - I(\theta; V_t) = I(\theta; T \mid V_t) = \sum_c \mathbb{P}(V_t = c)\, I(\theta; T \mid V_t = c).$$

For $c \ne b$ the label T is determined given $V_t = c$, and the corresponding term vanishes. Given $V_t = b$, the label is a with probability w_t and b with probability $1 - w_t$, and the conditional law of θ given (T, V_t) is Π_T because the coin is independent of θ given T. So the conditional law of θ given $V_t = b$ is the mixture M_t, and

$$I(\theta; T \mid V_t = b) = w_t \mathrm{KL}(\Pi_a \| M_t) + (1 - w_t)\mathrm{KL}(\Pi_b \| M_t) = \mathrm{JS}_{w_t}(\Pi_a, \Pi_b).$$

Multiplying by $\mathbb{P}(V_t = b) = p_b + p_a t$ and using $(p_b + p_a t) w_t = p_a t$ and $(p_b + p_a t)(1 - w_t) = p_b$ gives the stated formula for L_t.

For the asymptotics as $t \downarrow 0$, we have $w_t \sim (p_a/p_b)t$, so $\log(1/w_t) = \log(1/t) + O(1) \sim \log(1/t)$, and $h_2(w) \sim w \log(1/w)$ as $w \downarrow 0$. Therefore

$$B_t = (p_b + p_a t) h_2(w_t) \sim p_b \cdot \frac{p_a}{p_b} t \log \frac{1}{t} = p_a t \log \frac{1}{t}.$$

For $L_t = (p_b + p_a t)\,\mathrm{JS}_{w_t}(\Pi_a, \Pi_b)$, the second limit in Lemma A.7 gives $\mathrm{JS}_{w_t}(\Pi_a, \Pi_b) \sim \alpha(\Pi_a \mid \Pi_b)\, w_t \log(1/w_t)$ when the singular mass is positive, and $\mathrm{JS}_{w_t}(\Pi_a, \Pi_b) = o(w_t \log(1/w_t))$ when it is zero. In either case

$$\frac{L_t}{B_t} = \frac{\mathrm{JS}_{w_t}(\Pi_a, \Pi_b)}{h_2(w_t)} = \frac{\mathrm{JS}_{w_t}(\Pi_a, \Pi_b)}{w_t \log(1/w_t)} \cdot \frac{w_t \log(1/w_t)}{h_2(w_t)} \longrightarrow \alpha(\Pi_a \mid \Pi_b) \cdot 1.$$

If $\Pi_a \ll \Pi_b$ the limit is zero, so $L_t = o(B_t)$. If $\mathrm{KL}(\Pi_a \| \Pi_b) < \infty$, the last inequality of Lemma A.7 gives $\mathrm{JS}_{w_t}(\Pi_a, \Pi_b) \le w_t \mathrm{KL}(\Pi_a \| \Pi_b)$, and multiplying by $p_b + p_a t$ gives $L_t \le p_a t \mathrm{KL}(\Pi_a \| \Pi_b)$. □

Proposition C.3 (A directional support criterion)**.** *Fix $0 < R < R_0$ and $D = D^{\mathrm{ts}}(R)$. Suppose an optimal deterministic mixture contains a positive-weight component having two positive-mass cells with*

$$\alpha(\Pi_a \mid \Pi_b) < \frac{D}{R_0 - R}.$$

Then a finite stochastic channel of rate exactly R has distortion strictly below D. The comparison threshold satisfies $D/(R_0 - R) \le I/R_0 \le 1$.

Proof. The construction perturbs one component of the optimal mixture with the two-cell confusion of Lemma C.2, and then restores the rate by mixing in the exact posterior label. The perturbation trades rate for distortion at the ratio $\alpha(\Pi_a \mid \Pi_b)$, while the label code buys distortion back at the ratio $D/(R_0 - R)$; the hypothesis says the first ratio is the cheaper one.

Let (λ_j, T_j) be an optimal mixture with the stated component $T = T_{j_0}$ of weight $v := \lambda_{j_0} > 0$. By Lemma C.1 it has rate exactly R and distortion $D > 0$. Replace T_{j_0} by the channel V_t of Lemma C.2, keeping the visible selector. By the affine identities (15), the rate and distortion become

$$R - \delta r, \qquad D + \delta d, \qquad \delta r := v B_t > 0, \quad \delta d := v L_t \ge 0,$$

and Lemma C.2 gives $\delta d / \delta r = L_t / B_t \to \alpha(\Pi_a \mid \Pi_b)$ as $t \downarrow 0$.

Now use a fresh revealed selector to transmit the exact posterior label code, of rate $R_0 = H(\Pi_Y)$ and distortion zero, with probability

$$u := \frac{\delta r}{R_0 - R + \delta r},$$

and the perturbed mixture otherwise. The rate returns to

$$(1 - u)(R - \delta r) + u R_0 = R - \delta r + u\,(R_0 - R + \delta r) = R,$$

and the distortion becomes

$$(1 - u)(D + \delta d) = \frac{R_0 - R}{R_0 - R + \delta r}\,(D + \delta d).$$

This is strictly less than D exactly when $(R_0 - R)(D + \delta d) < D(R_0 - R + \delta r)$, that is, when

$$\frac{\delta d}{\delta r} < \frac{D}{R_0 - R}.$$

Since $\delta d / \delta r \to \alpha(\Pi_a \mid \Pi_b) < D/(R_0 - R)$, this holds for all sufficiently small t. The resulting channel has finitely many outputs and rate exactly R.

For the threshold, D^{ts} is convex with $D^{\mathrm{ts}}(0) = I$ and $D^{\mathrm{ts}}(R_0) = 0$, so its graph lies below the chord between these points and $D \le I(1 - R/R_0)$, which rearranges to $D/(R_0 - R) \le I/R_0$. Finally, the posterior label is a sufficient statistic, so $I = I(\theta; \Pi_Y) \le H(\Pi_Y) = R_0$, giving $I/R_0 \le 1$. □

Corollary C.4 (Finite observations)**.** *If Y has finite support with positive probabilities, $I(\theta; Y) > 0$, and the posteriors Π_y are pairwise equivalent, then $R_0 = H(\Pi_Y)$ and (10) holds for every $0 < R < H(\Pi_Y)$.*

Proof. We verify the hypotheses of Theorem 4.2, using the finite structure to identify R_0.

For Corollary C.4, the prior is the finite mixture $\pi = \sum_y \mathbb{P}(Y = y)\Pi_y$ with positive weights. Each $\Pi_y \ll \pi$ because $\mathbb{P}(Y = y)\Pi_y \le \pi$. Conversely, if $\Pi_y(A) = 0$ for some y, pairwise

equivalence gives $\Pi_{y'}(A) = 0$ for every y', hence $\pi(A) = 0$. Thus $\Pi_y \sim \pi$ for every y, which is condition (ii) of Proposition B.1, so Assumption 4.1 holds. The Bayes information satisfies $0 < I \le H(Y) < \infty$, the spaces are standard Borel, and Theorem 4.2 applies on $(0, R_0)$, where $R_0 = H(\Pi_Y)$ by Lemma C.1. Alternatively, one can use Proposition C.3: every cell posterior of every deterministic code is a positive mixture of observation posteriors, hence equivalent to π, so any two cells have $\alpha(\Pi_a \mid \Pi_b) = 0 < D/(R_0 - R)$; and an optimal mixture at $0 < R < R_0$ has rate exactly $R > 0$, so it contains a nonconstant component, which has at least two cells. □

Corollary C.5 (Beta–Binomial experiments)**.** *For $\theta \sim \mathrm{Beta}(\alpha, \beta)$ and $C \mid \theta \sim \mathrm{Binomial}(n, \theta)$ with $\alpha, \beta > 0$ and $n \ge 1$, separation holds throughout $0 < R < H(C)$. It also holds when the observation is the full ordered Bernoulli sample.*

Proof. For Corollary C.5, every count $c \in \{0, \dots, n\}$ has positive predictive probability, and the posterior given $C = c$ is $\mathrm{Beta}(\alpha + c, \beta + n - c)$. These $n + 1$ laws have strictly positive densities on $(0, 1)$, so they are pairwise equivalent, and they are distinct because their parameters differ. Distinctness makes $c \mapsto \Pi_c$ injective, so $H(\Pi_C) = H(C)$, which is positive because C is nonconstant for $n \ge 1$; and $I > 0$ because no posterior equals the prior. Corollary C.4 gives separation on $0 < R < H(C)$. For the ordered Bernoulli sample $Y \in \{0, 1\}^n$, all 2^n outcomes have positive probability and the posterior given $Y = y$ is $\mathrm{Beta}(\alpha + \sum_i y_i, \beta + n - \sum_i y_i)$, which depends on y only through the count. These posteriors are again pairwise equivalent, and $H(\Pi_Y) = H(C)$. Corollary C.4 applies directly to this finite observation space with the same threshold. This is a separate application, not a transfer of a deterministic restriction through a sufficient statistic. □

Proposition C.6 (Finite-alphabet attainment with a general parameter)**.** *If Y has m positive-probability atoms, $D^\star(R)$ is attained, for every $R \ge 0$, by a channel with at most $m + 1$ outputs, even when θ is not finite-valued.*

Proof. The proof rewrites the optimization over channels as an optimization over probability laws on the simplex, shows that this second problem has a compact feasible set and a continuous objective, and then reduces the support of an optimizer by Carathéodory's theorem.

Write $p_i > 0$ for the observation probabilities and Π_i for the posteriors, $i = 1, \dots, m$, so that $\pi = \sum_i p_i \Pi_i$. For a summary S, define the conditional observation-probability vector $w(S) := (\mathbb{P}(Y = i \mid S))_{i=1}^m$, which takes values in the compact simplex $\mathcal{D}_m$, and let ν be its law. Taking expectations, the barycenter of ν is $\int w\, \nu(dw) = (\mathbb{P}(Y = i))_i = p$. Conversely, let ν be any probability law on $\mathcal{D}_m$ with barycenter p, and define the channel

$$q(dw \mid Y = i) := \frac{w_i}{p_i}\, \nu(dw).$$

This is a probability kernel because $\int (w_i/p_i)\nu(dw) = p_i/p_i = 1$. Under it, $\mathbb{P}(Y = i, S \in dw) = p_i(w_i/p_i)\nu(dw) = w_i\nu(dw)$, so S has law ν and $\mathbb{P}(Y = i \mid S = w) = w_i$; that is, $w(S) = S$ and the law of $w(S)$ is ν. The laws on the simplex with barycenter p are therefore exactly the laws of $w(S)$ over all summaries.

We express rate and distortion through ν. With $h(w) := -\sum_i w_i \log w_i$,

$$I(Y; S) = H(Y) - H(Y \mid S) = H(p) - \mathbb{E}\, h(w(S)) = H(p) - \int h\, d\nu.$$

By the barycenter formula, $\Pi_S = \sum_i \mathbb{P}(Y = i \mid S)\Pi_i = \sum_i w_i(S)\Pi_i$, so with $F(w) := \mathrm{KL}(\sum_i w_i \Pi_i \| \pi)$,

$$I(\theta; S) = \mathbb{E}\, \mathrm{KL}(\Pi_S \| \pi) = \int F\, d\nu, \qquad \Delta_{\mathrm{KL}}(S) = I - \int F\, d\nu.$$

Since a summary's distortion depends on S only through the joint law of (Y, S), and any ν is realized by the explicit channel above,

$$D^\star(R) = I - \sup\Big\{ \int F\, d\nu : \ \nu \in \mathcal{P}(\mathcal{D}_m),\ \int w\, \nu(dw) = p,\ \int h\, d\nu \ge H(p) - R \Big\}.$$

Both integrands are continuous on $\mathcal{D}_m$. The entropy h is continuous with the convention $0 \log 0 = 0$. For F, note that $p_i \Pi_i \le \pi$ gives $d\Pi_i/d\pi \le 1/p_i$, so the density $f_w := d(\sum_i w_i \Pi_i)/d\pi =$

$\sum_i w_i \, d\Pi_i/d\pi$ takes values in $[0, 1/\min_i p_i]$ and depends linearly, hence continuously, on w at every point of Θ. The function $u \mapsto u \log u$ is bounded on this interval, so dominated convergence under π gives $F(w) = \int f_w \log f_w \, d\pi \to F(w')$ as $w \to w'$.

The feasible set is nonempty and compact. It contains δ_p, the law of the constant summary, since $\int h \, d\delta_p = H(p) \geq H(p) - R$. The space $\mathcal{P}(\mathcal{D}_m)$ is compact in the weak topology because $\mathcal{D}_m$ is a compact metric space. The barycenter constraints $\int w_i \, d\nu = p_i$ and the rate constraint $\int h \, d\nu \geq H(p) - R$ are closed, since w_i and h are continuous and bounded. The objective $\nu \mapsto \int F \, d\nu$ is weakly continuous for the same reason. Therefore the supremum is attained by some $\nu^\star$.

We reduce the support. Consider the continuous map

$$\Phi(w) := \big(w_1, \ldots, w_{m-1}, h(w), F(w)\big) \in \mathbb{R}^{m+1}.$$

The moment vector $\int \Phi \, d\nu^\star$ lies in the convex hull of the compact set $\Phi(\mathcal{D}_m)$. By Carathéodory's theorem in $\mathbb{R}^{m+1}$, it is a convex combination of at most $m+2$ points $\Phi(w^{(1)}), \ldots, \Phi(w^{(m+2)})$ with weights μ_k. The discrete law $\nu' := \sum_k \mu_k \delta_{w^{(k)}}$ has the same first $m-1$ barycenter coordinates as $\nu^\star$, hence the same barycenter since $w_m = 1 - \sum_{i<m} w_i$, the same entropy moment, and the same F-moment. So ν' is feasible and optimal.

Suppose ν' has exactly $m+2$ atoms with positive weights. The linear map $\varepsilon \mapsto \big(\sum_k \varepsilon_k, \ \sum_k \varepsilon_k w_i^{(k)} \ (i < m), \ \sum_k \varepsilon_k h(w^{(k)})\big)$ from $\mathbb{R}^{m+2}$ to $\mathbb{R}^{m+1}$ has a nonzero kernel element ε. For $|s|$ small, the weights $\mu + s\varepsilon$ are positive, sum to one, and give a law with the same barycenter and the same entropy moment, hence a feasible law, whose F-moment is $\int F \, d\nu' + s \sum_k \varepsilon_k F(w^{(k)})$. Optimality of ν' for both signs of s forces $\sum_k \varepsilon_k F(w^{(k)}) = 0$, so the F-moment is constant along the whole segment. Since $\sum_k \varepsilon_k = 0$ and $\varepsilon \neq 0$, some coordinate of ε is negative; increase s until the first weight reaches zero. The resulting law is still feasible and optimal and has at most $m+1$ atoms.

Finally, the explicit channel $q(dw \mid Y = i) = (w_i/p_i)\nu(dw)$ built from this finitely supported optimal ν has at most $m+1$ outputs, rate at most R, and distortion $I - \int F \, d\nu = D^\star(R)$. □

Under the separation hypotheses, a finite-observation optimum at an interior rate cannot be deterministic or a revealed deterministic mixture. The attainment proposition does not supply a global numerical optimizer; it establishes existence and a cardinality bound.

The explicit count witness and enumeration. For $\alpha = \beta = 5$, $n = 8$, take the observation to be C, write $w_c = P(C = c)$ and $\Pi_c = P_{\theta|C=c} = \mathrm{Beta}(5 + c, 13 - c)$. For any finite channel $q(b \mid c)$, its induced posterior is

$$\Pi_b = \frac{\sum_c w_c q(b \mid c) \Pi_c}{\sum_c w_c q(b \mid c)},$$

on positive-probability outputs, with

$$R = \sum_{c,b} w_c q(b \mid c) \log \frac{q(b \mid c)}{\sum_j w_j q(b \mid j)}, \qquad \Delta_{\mathrm{KL}} = \sum_{c,b} w_c q(b \mid c) \mathrm{KL}(\Pi_c \| \Pi_b).$$

All $\mathrm{Bell}(9) = 21{,}147$ count partitions are enumerated; posterior mixtures are integrated by Gauss–Legendre quadrature and the deterministic envelope is their lower convex hull. The displayed quadrature values are numerical approximations, not certified integration enclosures.

Put $w_A = P(C \leq 3)$ and define $\varrho_\star$ as the unique solution in $(0, 1)$ of

$$h_2(w_A \varrho) - w_A h_2(\varrho) = 0.5.$$

The left side is the information rate of $q(1 \mid c) = \varrho \mathbf{1}\{c \leq 3\}$: the output B has $\mathbb{P}(B = 1) = w_A \varrho$, so $H(B) = h_2(w_A \varrho)$, while $H(B \mid C) = w_A h_2(\varrho)$ because B is random only on the counts $c \leq 3$. It is strictly increasing in ϱ: its derivative is $w_A \log\{(1 - w_A\varrho)/(w_A(1 - \varrho))\}$, which is positive because $1 - w_A\varrho > w_A(1 - \varrho)$ reduces to $w_A < 1$. The solution is approximately 0.8782923057. Its distortion is approximately 0.154529, below the time-sharing envelope 0.162253 and the best individual deterministic count code 0.187462 at rate 0.5. The stochastic point is feasible, not claimed globally optimal. Neither the nine-count enumeration nor its numerical envelope enumerates all deterministic functions of the ordered raw sample.

D EXAMPLES DELIMITING THE SUPPORT AND RATE ASSUMPTIONS

A deterministic target. Suppose $\theta = f(Y)$ is discrete with $H(\theta) < \infty$. For every summary, $\Delta_{\mathrm{KL}}(S) = H(\theta) - I(\theta; S) \geq H(\theta) - I(Y; S)$. Time-sharing the constant summary and $f(Y)$ attains the bound up to rate $H(\theta)$; transmitting $f(Y)$ handles larger budgets, proving (11). Not every deterministic code is efficient: if $Y = (\theta, U)$ with independent fair bits, the code $T = U$ has positive rate and zero relevance, lying above the frontier. Channels driven by θ alone satisfy $I(Y; S) = I(\theta; S)$ and attain the line at their achieved rates.

Overlapping posteriors without an advantage. Let Y be a fair bit and independently draw $E \sim \mathrm{Bernoulli}(\alpha)$, where $0 < \alpha < 1$. Set $\theta = Y$ when $E = 1$ and $\theta = e$ otherwise, with e a distinct erasure symbol. For every channel from Y, the indicator E is independent of (Y, S) and is determined by θ. The chain rule therefore gives $I(\theta; S) = \alpha I(Y; S)$ and $I(\theta; Y) = \alpha \log 2$. It follows that

$$D^\star(R) = D^{\mathrm{ts}}(R) = \alpha[\log 2 - R]_+,$$

attained by visible time-sharing of the constant and identity codes. The target is not a function of Y, and the posteriors overlap at e with mass $1 - \alpha$: neither property ensures strict separation.

Strict advantage despite unequal supports. Let Y be a fair bit with posteriors $\Pi_0 = (0.1, 0.9, 0)$ and $\Pi_1 = (0, 0.1, 0.9)$. There are only two deterministic partition types, so $R_0 = \log 2$ and $D^{\mathrm{ts}}(R) = I(1 - R/\log 2)$ on $[0, \log 2]$. The prior is $(0.05, 0.5, 0.45)$; only its middle atom leaves Y uncertain, with conditional binary entropy $h_2(0.1)$. Hence $I = \log 2 - \frac{1}{2} h_2(0.1)$ and $I/\log 2 \geq 1/2 > 0.1 = \alpha(\Pi_0 \mid \Pi_1)$. At every interior rate, the optimal mixture uses the identity code with positive weight. Proposition C.3 therefore gives strict separation at every such rate, despite failure of equivalence.

Topological support is insufficient. For a probability law on a metric observation space, its topological support is

$$\mathrm{supp}_{\mathrm{top}}(P) := \{y : P(U) > 0 \text{ for every open neighborhood } U \text{ of } y\}.$$

This definition tests neighborhoods. Mutual absolute continuity instead requires agreement on all measurable null events: $P(A) = 0$ if and only if $Q(A) = 0$ for every measurable A. Equal topological supports need not ensure this stronger property. Let θ be a fair bit. Given $\theta = 0$, let Y be uniform on $[0, 1]$; given $\theta = 1$, give positive masses to an enumeration of the rationals in $[0, 1]$. Both sampling laws have topological support $[0, 1]$, but they are mutually singular and $\theta = \mathbf{1}\{Y \in \mathbb{Q}\}$ almost surely. Thus (11) gives no advantage despite equal topological supports, unlike the measure equivalence in Assumption 4.1. The separation proof uses the latter property to make the directed singular mass $\alpha(Q_{\theta Y} \mid P_{\theta Y})$ vanish for its joint and product laws. No choice of topology is needed for that argument. Common support in this measure-theoretic sense is sufficient for the theorem, but the preceding unequal-support example shows that it is not necessary for separation.

An output-entropy constraint is a different problem. For $\theta = Y$ a fair bit and $0 < R < \log 2$, every deterministic code with entropy at most R is constant. Choose $t \in (0, 1)$ with $h_2(t/2) = R$ and use $q(B = 1 \mid Y = 1) = t$, $q(B = 1 \mid Y = 0) = 0$. Then $H(B) = R$ but

$$I(\theta; B) = h_2(t/2) - \tfrac{1}{2} h_2(t) > 0.$$

This improves on every feasible deterministic code under $H(S) \leq R$; the cardinality result and hard-assignment claims for an entropy-penalized Lagrangian do not extend to this constraint. Relabeling preserves output entropy and saves no cost under that measure.

Infinite mutual information need not mean a noiseless component. The finite-information consequences in Section 2 do not characterize which noisy models satisfy that assumption, nor do they imply finite exact-preservation rate (Proposition E.1). For a counterexample, let $\theta = N$ have a countable full-support distribution p with $H(N) = \infty$, for example a normalized $p_n \propto 1/[n(\log n)^2]$ on $n \geq 2$. Let $Y = N + G$ with $G \sim N(0, 1)$ independent. All sampling laws have strictly positive densities on $\mathbb{R}$, so they are equivalent, and no nonconstant function of N is determined by Y. For

$Z = \mathrm{round}(Y)$ and $E = \mathrm{round}(G)$, translation by the integer N gives $Z = N + E$ almost surely. The discretized Gaussian E has finite entropy, and hence

$$H(N \mid Y) \le H(N \mid Z) = H(E \mid Z) \le H(E) < \infty.$$

For finite quantizations U_m of N with $H(U_m) \to \infty$, $I(N;Y) \ge I(U_m;Y) \ge H(U_m) - H(E)$, proving $I(N;Y) = \infty$ despite additive Gaussian observation noise. Thus an exactly observed continuous parameter coordinate is only one possible cause of infinite mutual information.

E Proofs for Section 2

Proof of Lemma 2.2. We prove the three parts of (2) in turn, then the compensation identity (3) by applying the chain rule for divergences in two different orders, and finally the two converses.

Since $\theta \to Y \to S$ is a Markov chain, θ is conditionally independent of S given Y, so the conditional law of θ given (Y, S) is the conditional law given Y alone: $P_{\theta|Y,S} = \Pi_Y$ almost surely. On standard Borel spaces the conditional mutual information is the expected divergence between the two conditional laws of θ,

$$I(\theta; Y \mid S) = \mathbb{E}\,\mathrm{KL}(P_{\theta|Y,S} \| P_{\theta|S}) = \mathbb{E}\,\mathrm{KL}(\Pi_Y \| \Pi_S) = \Delta_{\mathrm{KL}}(S).$$

For the second equality in (2), expand $I(\theta; Y, S)$ with the chain rule in both orders:

$$I(\theta; Y, S) = I(\theta; Y) + I(\theta; S \mid Y) = I(\theta; Y), \qquad I(\theta; Y, S) = I(\theta; S) + I(\theta; Y \mid S),$$

where the first line uses $I(\theta; S \mid Y) = 0$ from the Markov property. Equating the two expansions gives $I(\theta; Y \mid S) = I(\theta; Y) - I(\theta; S)$, and all terms are finite because $0 \le I(\theta; S) \le I(\theta; Y) < \infty$. For the barycenter formula, fix a measurable $A \subseteq \Theta$ and apply the tower property together with the Markov property:

$$\mathbb{P}(\theta \in A \mid S) = \mathbb{E}\big[\mathbb{P}(\theta \in A \mid Y, S) \,\big|\, S\big] = \mathbb{E}\big[\Pi_Y(A) \,\big|\, S\big].$$

Since Θ is standard Borel, a countable generating algebra of sets A suffices to identify the two kernels, giving $\Pi_S = \mathbb{E}[\Pi_Y \mid S]$ almost surely.

For (3), let Q_S be a measurable reconstruction with $\mathbb{E}\,\mathrm{KL}(\Pi_Y \| Q_S) < \infty$, and fix a value s of S. On $\Theta \times \mathcal{Y}$ compare the conditional joint law $P_{\theta Y|S=s}$ with the product $Q_s \otimes P_{Y|S=s}$, whose $\mathcal{Y}$-marginal agrees with that of the joint. The chain rule for divergences holds in $[0, \infty]$ without integrability assumptions. Disintegrating along Y first,

$$\begin{aligned}\mathrm{KL}\big(P_{\theta Y|S=s} \,\big\|\, Q_s \otimes P_{Y|S=s}\big) &= \mathrm{KL}(P_{Y|S=s} \| P_{Y|S=s}) + \int \mathrm{KL}(P_{\theta|Y=y,S=s} \| Q_s)\, P_{Y|S=s}(dy) \\ &= \int \mathrm{KL}(\Pi_y \| Q_s)\, P_{Y|S=s}(dy).\end{aligned}$$

Disintegrating along θ first,

$$\begin{aligned}\mathrm{KL}\big(P_{\theta Y|S=s} \,\big\|\, Q_s \otimes P_{Y|S=s}\big) &= \mathrm{KL}(P_{\theta|S=s} \| Q_s) \\ &\quad + \int \mathrm{KL}(P_{Y|\theta,S=s} \| P_{Y|S=s})\, P_{\theta|S=s}(d\theta).\end{aligned}$$

The last integral is the mutual information between θ and Y under the conditional law given $S = s$, which by symmetry of mutual information equals $\int \mathrm{KL}(\Pi_y \| \Pi_s)\, P_{Y|S=s}(dy)$. Equating the two expansions and averaging over s with respect to the law of S,

$$\mathbb{E}\,\mathrm{KL}(\Pi_Y \| Q_S) = \mathbb{E}\,\mathrm{KL}(\Pi_Y \| \Pi_S) + \mathbb{E}\,\mathrm{KL}(\Pi_S \| Q_S) = \Delta_{\mathrm{KL}}(S) + \mathbb{E}\,\mathrm{KL}(\Pi_S \| Q_S).$$

The left side is finite by hypothesis, so both terms on the right are finite and no difference of infinities occurs. When all measures have densities p_Y, p_S, q_S with respect to a common reference measure, the same identity is the expectation of the pointwise decomposition $\log(p_Y/q_S) = \log(p_Y/p_S) + \log(p_S/q_S)$, using that $\log(p_S/q_S)$ is $\sigma(S)$-measurable and $\mathbb{E}[p_Y \mid S] = p_S$; the argument above avoids the choice of a reference measure.

For the converses, data processing along $\theta \to Y \to S$ gives $I(\theta; S) \le I(Y; S)$, and with the identity just proved,

$$\Delta_{\mathrm{KL}}(S) = I(\theta; Y) - I(\theta; S) \;\ge\; I(\theta; Y) - I(Y; S).$$

If $\Delta_{\mathrm{KL}}(S) \le \varepsilon$ then $I(Y; S) \ge I(\theta; Y) - \varepsilon$, and rates are nonnegative, so $R^\star(\varepsilon) \ge [I(\theta; Y) - \varepsilon]_+$. If $I(Y; S) \le R$ then $\Delta_{\mathrm{KL}}(S) \ge I(\theta; Y) - R$, and distortions are nonnegative, so $D^\star(R) \ge [I(\theta; Y) - R]_+$. □

Proof of Proposition 3.1. We first verify that $Z = \Pi_Y$ is its own posterior, then prove the two inequalities behind the reduction identity $R^\star(\varepsilon) = \inf_{P_{\widehat{Z}|Z}:\, \mathbb{E}\mathrm{KL}(Z\|\widehat{Z})\le\varepsilon} I(Z;\widehat{Z})$, and finally give the explicit reduction of an arbitrary kernel, which also proves Remark E.2.

For a measurable $A \subseteq \Theta$, the tower property gives

$$\mathbb{P}(\theta \in A \mid Z) = \mathbb{E}\big[\mathbb{P}(\theta \in A \mid Y) \,\big|\, Z\big] = \mathbb{E}[Z(A) \mid Z] = Z(A),$$

since $Z(A)$ is $\sigma(Z)$-measurable and Z is a function of Y. Consequently the law of Z determines the joint law of (θ, Z): for measurable A and C, $\mathbb{P}(\theta \in A, Z \in C) = \mathbb{E}[Z(A)\mathbf{1}\{Z \in C\}]$.

We first show that the right side of the reduction identity is at most the left. Let S be a summary channel with $\Delta_{\mathrm{KL}}(S) \le \varepsilon$ and set $\widehat{Z} := \Pi_S$, a probability-measure-valued function of S. The joint law of $(Z, \widehat{Z})$ defines a channel from Z to reconstructions, with distortion $\mathbb{E}\,\mathrm{KL}(Z\|\widehat{Z}) = \mathbb{E}\,\mathrm{KL}(\Pi_Y\|\Pi_S) = \Delta_{\mathrm{KL}}(S) \le \varepsilon$. Since Z is a function of Y and $\widehat{Z}$ a function of S, data processing gives $I(Z;\widehat{Z}) \le I(Y;S)$. Thus every feasible summary yields a feasible reconstruction channel of no larger rate, and taking the infimum over summaries shows that the right side is at most $R^\star(\varepsilon)$.

Now, we show the reverse inequality. Let $P_{\widehat{Z}|Z}$ be a channel with $\mathbb{E}\,\mathrm{KL}(Z\|\widehat{Z}) \le \varepsilon$. Build a summary from Y in two stages: compute $Z = \Pi_Y$, then draw S from $P_{\widehat{Z}|Z}(\cdot \mid Z)$ using fresh randomness. Then $\theta \to Y \to S$ is a Markov chain, and the joint law of (Z, S) equals that of $(Z, \widehat{Z})$. Because S depends on Y only through Z, and Z is a function of Y, the chain rule gives

$$I(Y;S) = I(Y,Z;S) = I(Z;S) + I(Y;S \mid Z) = I(Z;S) = I(Z;\widehat{Z}).$$

For the distortion, use the value of S itself as a reconstruction, $Q_S := S$. Its expected divergence is $\mathbb{E}\,\mathrm{KL}(\Pi_Y\|Q_S) = \mathbb{E}\,\mathrm{KL}(Z\|\widehat{Z}) \le \varepsilon < \infty$, so (3) applies and gives

$$\Delta_{\mathrm{KL}}(S) = \mathbb{E}\,\mathrm{KL}(\Pi_Y\|Q_S) - \mathbb{E}\,\mathrm{KL}(\Pi_S\|Q_S) \le \mathbb{E}\,\mathrm{KL}(Z\|\widehat{Z}) \le \varepsilon.$$

Thus S is feasible for $R^\star(\varepsilon)$ with rate $I(Z;\widehat{Z})$, and taking the infimum over reconstruction channels gives $R^\star(\varepsilon) \le$ the right side of the posterior source reduction.

Finally we reduce an arbitrary kernel explicitly. Let S be generated from Y by a kernel $q(ds \mid y)$, define $q'(ds \mid z) := \mathbb{P}(S \in ds \mid Z = z)$, and generate S' from Z through q' with fresh randomness. For measurable A and C, using $\mathbb{P}(\theta \in A \mid Y, S) = Z(A)$ and $\mathbb{P}(S \in C \mid Y) = q(C \mid Y)$,

$$\begin{aligned}\mathbb{P}(\theta \in A, S \in C) &= \mathbb{E}\big[Z(A)\, q(C \mid Y)\big] = \mathbb{E}\big[Z(A)\, \mathbb{E}[q(C \mid Y) \mid Z]\big] \\ &= \mathbb{E}\big[Z(A)\, q'(C \mid Z)\big] = \mathbb{P}(\theta \in A, S' \in C),\end{aligned}$$

where the second equality is the tower property with the $\sigma(Z)$-measurable factor $Z(A)$, the third uses $\mathbb{E}[q(C \mid Y) \mid Z] = \mathbb{P}(S \in C \mid Z) = q'(C \mid Z)$, and the last uses that S' is conditionally independent of θ given Y with $\mathbb{P}(S' \in C \mid Y) = q'(C \mid Z)$. Hence (θ, S) and (θ, S') have the same joint law, so they induce the same reconstruction Π_S and, by (2), the same distortion. Their rates compare as $I(Y;S') = I(Z;S') = I(Z;S) \le I(Y;S)$, by the same chain-rule computation as above and data processing. This proves Remark E.2: even if q is degenerate, so that $S = g(Y)$ is deterministic, the averaged kernel $q'(\cdot \mid z) = \mathbb{P}(g(Y) \in \cdot \mid Z = z)$ is degenerate only when $g(Y)$ is a function of Z. □

Proposition E.1 (Shape and endpoints). *$R^\star$ is nonincreasing and convex, $R^\star(\varepsilon) = 0$ for $\varepsilon \ge I(\theta;Y)$, and $R^\star(0) = \infty$ whenever the law of $Z = \Pi_Y$ is non-atomic.*

Remark E.2 (Deterministic restrictions and sufficient statistics). Replacing $q(\cdot \mid Y)$ by its conditional average given Z preserves the joint law of (θ, S) and can only reduce rate. Even if $S = g(Y)$, this averaged kernel need not be deterministic. A deterministic restriction on a sufficient statistic is therefore distinct from a deterministic restriction on the raw observation. The deterministic comparison in Section 4 refers to functions of the original observation.

Proof of Proposition E.1. Monotonicity holds because the feasible set $\{q : \Delta_{\mathrm{KL}}(S) \le \varepsilon\}$ grows with ε.

For convexity, fix $\varepsilon_1, \varepsilon_2 \geq 0$, $\mu \in (0,1)$, and $\eta > 0$, and choose channels S_1, S_2 with $\Delta_{\mathrm{KL}}(S_i) \leq \varepsilon_i$ and $I(Y;S_i) \leq R^\star(\varepsilon_i) + \eta$; if some $R^\star(\varepsilon_i)$ is infinite there is nothing to prove. Let $S = (J, S_J)$ be the revealed mixture with $\mathbb{P}(J=1) = \mu$. By (15),

$$\Delta_{\mathrm{KL}}(S) = \mu\Delta_{\mathrm{KL}}(S_1) + (1-\mu)\Delta_{\mathrm{KL}}(S_2) \leq \mu\varepsilon_1 + (1-\mu)\varepsilon_2,$$

$$I(Y;S) = \mu I(Y;S_1) + (1-\mu)I(Y;S_2) \leq \mu R^\star(\varepsilon_1) + (1-\mu)R^\star(\varepsilon_2) + \eta.$$

Hence $R^\star(\mu\varepsilon_1 + (1-\mu)\varepsilon_2) \leq \mu R^\star(\varepsilon_1) + (1-\mu)R^\star(\varepsilon_2) + \eta$, and $\eta \downarrow 0$ gives convexity.

For the zero-rate endpoint, the constant summary has $I(Y;S) = 0$ and, by (2), distortion $I(\theta;Y) - I(\theta;S) = I(\theta;Y)$. It is feasible whenever $\varepsilon \geq I(\theta;Y)$, so $R^\star(\varepsilon) = 0$ there.

For exact preservation, let $\Delta_{\mathrm{KL}}(S) = 0$. Lemma A.5 gives $Z = \Pi_S$ almost surely, so Z coincides with a measurable function of S. Data processing then gives

$$I(Y;S) \;\geq\; I(Z;S) \;\geq\; I(Z;\Pi_S) \;=\; I(Z;Z).$$

The space of probability measures on Θ is standard Borel, so Lemma A.1 applies to the deterministic map $Z \mapsto Z$: when the law of Z is non-atomic, $I(Z;Z) = \infty$. Hence every summary with zero distortion has infinite rate, and $R^\star(0) = \infty$. □

F Proofs for Section 3

Proof of Theorem 3.2. Write

$$D_M^{\mathrm{card}} := \inf_{g:\mathcal{Y}\to[M]} \Delta_{\mathrm{KL}}(g(Y)), \qquad D_M^{\mathrm{stoch}} := \inf_{q(s|y):\,|\mathcal{S}|\leq M} \Delta_{\mathrm{KL}}(S)$$

for the deterministic and stochastic M-symbol infima. We show $D_M^{\mathrm{av}} \leq D_M^{\mathrm{stoch}} \leq D_M^{\mathrm{card}} \leq D_M^{\mathrm{av}}$, then evaluate the case $M = 1$, and finally derive the rate bound.

We first show $D_M^{\mathrm{av}} \leq D_M^{\mathrm{stoch}}$. Let S be any summary with values in $[M]$, deterministic or not. For each output j with $\mathbb{P}(S=j) > 0$ put $Q_j := \Pi_{S=j} = P_{\theta|S=j}$, and choose the remaining Q_j arbitrarily. Pointwise, the nearest center is at least as close as the center actually selected by S:

$$\min_{j\leq M} \mathrm{KL}(\Pi_Y\|Q_j) \;\leq\; \mathrm{KL}(\Pi_Y\|Q_S) = \mathrm{KL}(\Pi_Y\|\Pi_S).$$

Taking expectations gives $\mathbb{E}\min_j \mathrm{KL}(\Pi_Y\|Q_j) \leq \Delta_{\mathrm{KL}}(S)$, so $D_M^{\mathrm{av}} \leq \Delta_{\mathrm{KL}}(S)$; the infimum over S gives the claim. The inequality $D_M^{\mathrm{stoch}} \leq D_M^{\mathrm{card}}$ holds because a deterministic $g : \mathcal{Y} \to [M]$ is a particular M-symbol channel.

We now show $D_M^{\mathrm{card}} \leq D_M^{\mathrm{av}}$ by the nearest-center construction. Let $Q_1, \ldots, Q_M$ be a codebook with finite average distortion $\mathbb{E}\min_j \mathrm{KL}(\Pi_Y\|Q_j) < \infty$; such codebooks exist because $Q_1 = \cdots = Q_M = \pi$ gives $\mathbb{E}\,\mathrm{KL}(\Pi_Y\|\pi) = I(\theta;Y) < \infty$. Define the nearest-center quantizer with the smallest index breaking ties,

$$g(y) := \min \arg\min_{j\leq M} \mathrm{KL}(\Pi_y\|Q_j).$$

It is measurable because each $y \mapsto \mathrm{KL}(\Pi_y\|Q_j)$ is measurable by the standing assumption and $\{g = j\} = \bigcap_{i<j}\{\mathrm{KL}(\Pi_y\|Q_j) < \mathrm{KL}(\Pi_y\|Q_i)\} \cap \bigcap_{i>j}\{\mathrm{KL}(\Pi_y\|Q_j) \leq \mathrm{KL}(\Pi_y\|Q_i)\}$. Put $S := g(Y)$ and use Q_S as a reconstruction. By construction $\mathrm{KL}(\Pi_Y\|Q_S) = \min_j \mathrm{KL}(\Pi_Y\|Q_j)$, so $\mathbb{E}\,\mathrm{KL}(\Pi_Y\|Q_S)$ is finite and the compensation identity (3) applies:

$$\mathbb{E}\min_j \mathrm{KL}(\Pi_Y\|Q_j) = \mathbb{E}\,\mathrm{KL}(\Pi_Y\|Q_S) = \Delta_{\mathrm{KL}}(S) + \mathbb{E}\,\mathrm{KL}(\Pi_S\|Q_S) \;\geq\; \Delta_{\mathrm{KL}}(S).$$

Conditioning on each positive-mass cell, this is the statement that replacing the center Q_j by the cell barycenter $\Pi_{g=j}$ can only reduce the cell's expected divergence. Hence $D_M^{\mathrm{card}} \leq \Delta_{\mathrm{KL}}(g(Y)) \leq \mathbb{E}\min_j \mathrm{KL}(\Pi_Y\|Q_j)$ for every codebook, and taking the infimum over codebooks gives $D_M^{\mathrm{card}} \leq D_M^{\mathrm{av}}$. The three infima are therefore equal, which is (6).

For $M = 1$ the only deterministic code is the constant, whose distortion is $I(\theta;Y \mid \text{constant}) = I(\theta;Y)$, so $D_1^{\mathrm{av}} = I(\theta;Y)$. The same value is visible from the compensation identity with a constant S and a single center Q, which reads $\mathbb{E}\,\mathrm{KL}(\Pi_Y\|Q) = I(\theta;Y) + \mathrm{KL}(\pi\|Q)$ and is minimized at $Q = \pi$.

For the rate bound, suppose $D_M^{\mathrm{av}} < \varepsilon$. By the definition of the infimum there is a codebook with $\mathbb{E}\min_j \mathrm{KL}(\Pi_Y \| Q_j) < \varepsilon$, and by the display above its nearest-center quantizer $S = g(Y)$ has $\Delta_{\mathrm{KL}}(S) < \varepsilon$. Its rate is $I(Y;S) = H(S) \le \log M$ because S is a deterministic code with at most M values. Hence $R^\star(\varepsilon) \le \log M$ whenever $D_M^{\mathrm{av}} < \varepsilon$, and taking $M = N^{\mathrm{av}}(\varepsilon)$ gives the bound. The strict inequality in the definition of N^{av} is what guarantees a feasible codebook without an attainment assumption; if only $D_M^{\mathrm{av}} \le \varepsilon$ is known, the same argument gives $R^\star(\varepsilon + \eta) \le \log M$ for every $\eta > 0$. □

Corollary F.1 (Uniform posterior covering). *Let $N_{\mathrm{KL}}^{\mathrm{unif}}(\varepsilon)$ be the least number of probability-measure centers Q_j with $\min_j \mathrm{KL}(\Pi_y \| Q_j) \le \varepsilon$ for P_Y-almost every y. If this number is finite, a deterministic summary achieves distortion at most ε and rate at most $\log N_{\mathrm{KL}}^{\mathrm{unif}}(\varepsilon)$.*

Proof of Corollary F.1. Let $Q_1, \dots, Q_N$ be a uniform cover with $N = N_{\mathrm{KL}}^{\mathrm{unif}}(\varepsilon)$, and let $S = g(Y)$ be its nearest-center quantizer as in the proof above. The pointwise bound $\min_j \mathrm{KL}(\Pi_y \| Q_j) \le \varepsilon$ for P_Y-almost every y gives $\mathbb{E}\min_j \mathrm{KL}(\Pi_Y \| Q_j) \le \varepsilon$, and the compensation step gives $\Delta_{\mathrm{KL}}(S) \le \mathbb{E}\min_j \mathrm{KL}(\Pi_Y \| Q_j) \le \varepsilon$. The rate is $H(S) \le \log N$. Since the codebook is given, no infimum has to be approximated. □

Proposition F.2 (Unbounded Gaussian posterior means). *For $\mathcal{P} = \{N(m, vI_d) : m \in \mathbb{R}^d\}$, fixed $v > 0$, and a full-support predictive law for m, every finite codebook has infinite essential-uniform forward-KL radius. With normalized squared Hellinger distance $\mathrm{H}^2(P,Q) = \frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2$, no finite uniform cover of radius $r < 1$ exists.*

Proof of Proposition F.2. Both claims rest on the same picture: a finite codebook is tight, so it places little mass far from the origin, while the Gaussian $N(m, vI_d)$ places half its mass near m; for large $\|m\|$ these two regions are disjoint. Write $P_m := N(m, vI_d)$.

Fix centers $Q_1, \dots, Q_M$ and $\delta \in (0,1)$. By translation invariance, choose $A > 0$ with $P_m(B(m,A)) = N(0, vI_d)(B(0,A)) \ge 1/2$ for every m. Each Q_j is tight, and a finite union of compact sets is compact, so there is a compact K with $Q_j(K^{\mathrm{c}}) < \delta$ for all j. Let $\rho_K := \sup_{x \in K} \|x\|$. If $\|m\| > A + \rho_K$, the ball $B(m,A)$ does not meet K, so $B(m,A) \subseteq K^{\mathrm{c}}$ and

$$P_m(B(m,A)) \ge \tfrac{1}{2}, \qquad Q_j(B(m,A)) \le Q_j(K^{\mathrm{c}}) < \delta.$$

Data processing for the two-cell partition $\{B(m,A), B(m,A)^{\mathrm{c}}\}$ bounds the divergence by the binary divergence of the two cell probabilities, and with $p := P_m(B(m,A)) \ge 1/2$, $q := Q_j(B(m,A)) < \delta$,

$$\mathrm{KL}(P_m \| Q_j) \;\ge\; \mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \;\ge\; p\log\frac{1}{q} - h_2(p) \;\ge\; \tfrac{1}{2}\log\frac{1}{\delta} - \log 2,$$

where the middle step drops the nonnegative term $(1-p)\log\frac{1}{1-q}$ and the last uses $p \ge 1/2$, $q < \delta$, and $h_2 \le \log 2$. This bound holds for every j and every m with $\|m\| > A + \rho_K$. That set of means is open and nonempty, hence has positive probability under the full-support predictive law, so the essential supremum of $\min_j \mathrm{KL}(P_m \| Q_j)$ is at least $\frac{1}{2}\log(1/\delta) - \log 2$. Letting $\delta \downarrow 0$ shows the essential-uniform radius is infinite.

For the Hellinger claim, write the affinity as $\mathcal{A}(P,Q) := \int \sqrt{pq}$, so that $\mathrm{H}^2(P,Q) = 1 - \mathcal{A}(P,Q)$. Fix $\delta \in (0, 1/4)$ and again a compact K with $Q_j(K^{\mathrm{c}}) < \delta$ for all j. Since $P_m(K) = N(0, vI_d)(K - m)$ and $K - m$ moves away from the origin as $\|m\| \to \infty$, there is A' with $P_m(K) < \delta$ whenever $\|m\| > A' + \rho_K$. Splitting the affinity over K and K^{c} and applying the Cauchy–Schwarz inequality on each piece,

$$\mathcal{A}(P_m, Q_j) \le \sqrt{P_m(K)\, Q_j(K)} + \sqrt{P_m(K^{\mathrm{c}})\, Q_j(K^{\mathrm{c}})} \le \sqrt{\delta} + \sqrt{\delta} = 2\sqrt{\delta}.$$

Hence $\mathrm{H}^2(P_m, Q_j) \ge 1 - 2\sqrt{\delta}$ for all j on the positive-probability set $\{\|m\| > A' + \rho_K\}$. Given any radius $r < 1$, choose δ so small that $1 - 2\sqrt{\delta} > r^2$; then $\min_j \mathrm{H}(P_m, Q_j) > r$ on a set of positive predictive probability, so the codebook is not a uniform cover of radius r. □

G HELLINGER COVERING AND THE LOCAL MODULUS

This extension makes the quantization argument usable when posterior geometry is controlled in Hellinger distance. An explicit comparison between the two divergences then converts the geometric bound into a posterior KL guarantee.

We use normalized squared Hellinger distance $\mathrm{H}^2(P,Q) = \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2\,d\mu$. Let $\mathcal{P}$ be the essential posterior family, $\mathrm{Mix}(\mathcal{P})$ its measurable mixtures, and

$$\omega(r) = \sup\big\{\mathrm{KL}(P\|Q) : P \in \mathcal{P},\ Q \in \mathrm{Mix}(\mathcal{P}),\ \mathrm{H}(P,Q) \le r\big\}.$$

Proposition G.1. *A uniform Hellinger cover of radius r with M centres yields a deterministic summary with $I(Y;S) \le \log M$, $\Delta_{\mathrm{H}}(S) := \mathbb{E}\,\mathrm{H}^2(\Pi_Y,\Pi_S) \le 4r^2$ and $\Delta_{\mathrm{KL}}(S) \le \omega(2r)$. For an arbitrary M-centre codebook with expected squared nearest-centre error e, the nearest-centre quantizer satisfies $\Delta_{\mathrm{H}}(S) \le 4e$, and if $\mathrm{KL}(P\|Q) \le C\,\mathrm{H}^2(P,Q)$ holds for every relevant posterior–mixture pair then $\Delta_{\mathrm{KL}}(S) \le 4Ce$.*

Proof. Let $Q_1,\dots,Q_M$ be the codebook and let $S = g(Y)$ be its nearest-centre quantizer, $g(y) = \min\arg\min_j \mathrm{H}(\Pi_y, Q_j)$, which is measurable as in the proof of Theorem 3.2. Its rate is $H(S) \le \log M$. Write $\Pi_j := \Pi_{S=j} = \mathbb{E}[\Pi_Y \mid S=j]$ for the cell posterior, which is a mixture of posteriors in cell j and hence an element of $\mathrm{Mix}(\mathcal{P})$. The argument compares Π_Y with Π_S through the centre Q_S, using two facts about the Hellinger distance: it is a metric, and its square is convex in each argument.

The convexity is seen from the affinity: $\mathrm{H}^2(P,Q) = 1 - \int\sqrt{pq}\,d\mu$, and for fixed q the map $p \mapsto \sqrt{pq}$ is concave pointwise, so $P \mapsto \mathrm{H}^2(P,Q)$ is convex. The integral form of Jensen's inequality applied to the mixture Π_j gives, for each positive-mass cell,

$$\mathrm{H}^2(Q_j,\Pi_j) \ \le\ \mathbb{E}\big[\mathrm{H}^2(Q_j,\Pi_Y)\,\big|\,S=j\big]. \tag{21}$$

We first treat the uniform cover. Its defining property is $\mathrm{H}(\Pi_y, Q_{g(y)}) \le r$ for P_Y-almost every y, so the right side of (21) is at most r^2 and $\mathrm{H}(Q_j,\Pi_j) \le r$ for every cell. The triangle inequality then gives, almost surely,

$$\mathrm{H}(\Pi_Y,\Pi_S) \le \mathrm{H}(\Pi_Y,Q_S) + \mathrm{H}(Q_S,\Pi_S) \le r + r = 2r.$$

Squaring and taking expectations gives $\Delta_{\mathrm{H}}(S) \le 4r^2$. For the divergence, $\Pi_Y \in \mathcal{P}$ and $\Pi_S \in \mathrm{Mix}(\mathcal{P})$ are at Hellinger distance at most $2r$, so by the definition of the modulus $\mathrm{KL}(\Pi_Y\|\Pi_S) \le \omega(2r)$ almost surely, and $\Delta_{\mathrm{KL}}(S) \le \omega(2r)$.

We now treat an arbitrary codebook with expected squared nearest-centre error $e = \mathbb{E}\min_j \mathrm{H}^2(\Pi_Y,Q_j) = \mathbb{E}\,\mathrm{H}^2(\Pi_Y,Q_S)$. The triangle inequality and $(a+b)^2 \le 2a^2+2b^2$ give, pointwise,

$$\mathrm{H}^2(\Pi_Y,\Pi_S) \le 2\mathrm{H}^2(\Pi_Y,Q_S) + 2\mathrm{H}^2(Q_S,\Pi_S).$$

Take conditional expectations given $S=j$ and use (21) on the second term:

$$\mathbb{E}\big[\mathrm{H}^2(\Pi_Y,\Pi_S)\,\big|\,S=j\big] \le 2\,\mathbb{E}\big[\mathrm{H}^2(\Pi_Y,Q_j)\,\big|\,S=j\big] + 2\mathrm{H}^2(Q_j,\Pi_j) \le 4\,\mathbb{E}\big[\mathrm{H}^2(\Pi_Y,Q_j)\,\big|\,S=j\big].$$

Averaging over the cells gives $\Delta_{\mathrm{H}}(S) \le 4e$. If $\mathrm{KL}(P\|Q) \le C\,\mathrm{H}^2(P,Q)$ for every $P \in \mathcal{P}$ and $Q \in \mathrm{Mix}(\mathcal{P})$, apply it pointwise to the pair (Π_Y,Π_S) and take expectations: $\Delta_{\mathrm{KL}}(S) \le C\,\Delta_{\mathrm{H}}(S) \le 4Ce$. □

A local quadratic modulus is only usable when individual distances are controlled on its range of validity, which a uniform cover supplies and an average Hellinger bound does not. A checkable sufficient condition for the global comparison is that the density ratios $t = p/q$ of every relevant pair lie in a compact subinterval of $(0,\infty)$: the function $t \mapsto (t\log t - t + 1)/(\sqrt{t}-1)^2$, extended by the value 2 at $t=1$, is then bounded, and integrating gives $\mathrm{KL} \le C\,\mathrm{H}^2$. Conversely, unbounded likelihood ratios alone do not establish that every local comparison fails. We record this route because Hellinger entropy is the standard currency in nonparametric Bayesian theory, but we do not import covering rates for sampling models into posterior geometry: the two are related only through an explicit map and tail control, and Proposition F.2 shows the translation can fail.

H FURTHER CONVERSES

H.1 PROOF OF THE POSTERIOR SMALL-BALL BOUND

Proof of Theorem 3.3. The bound compares the joint law of the random posterior and its reconstruction with the corresponding product law on a single event, the ball of radius t. Under the joint law the ball is likely, because the average distortion is small; under the product law it is unlikely, because no single reconstruction captures more than $\mathcal{Q}_Z(t)$ of the predictive mass.

Let S be any summary with $\Delta_{\mathrm{KL}}(S) \le \varepsilon$, put $\widehat{Z} := \Pi_S$, and define the event $E_t := \{(z, Q) : \mathrm{KL}(z\|Q) \le t\}$ in the product of two copies of the space of probability measures on Θ. Write

$$p := P_{Z,\widehat{Z}}(E_t) = \mathbb{P}\big(\mathrm{KL}(Z\|\widehat{Z}) \le t\big), \qquad q := (P_Z \otimes P_{\widehat{Z}})(E_t).$$

For the product probability, Tonelli's theorem integrates the ball probability of a fixed reconstruction against the law of $\widehat{Z}$:

$$q = \int \mathbb{P}\{\mathrm{KL}(Z\|Q) \le t\}\, P_{\widehat{Z}}(dQ) \;\le\; \sup_Q \mathbb{P}\{\mathrm{KL}(Z\|Q) \le t\} = \mathcal{Q}_Z(t).$$

For the joint probability, Markov's inequality applied to the nonnegative variable $\mathrm{KL}(Z\|\widehat{Z})$ gives

$$1 - p = \mathbb{P}\big(\mathrm{KL}(Z\|\widehat{Z}) > t\big) \le \frac{\mathbb{E}\,\mathrm{KL}(Z\|\widehat{Z})}{t} = \frac{\Delta_{\mathrm{KL}}(S)}{t} \le \frac{\varepsilon}{t},$$

so $p \ge a := 1 - \varepsilon/t > 0$.

Now compare the two laws. Since Z is a function of Y and $\widehat{Z}$ a function of S, data processing gives $I(Y;S) \ge I(Z;\widehat{Z})$. The mutual information $I(Z;\widehat{Z}) = \mathrm{KL}(P_{Z,\widehat{Z}}\|P_Z \otimes P_{\widehat{Z}})$ can only decrease when both laws are reduced to the two-cell partition $\{E_t, E_t^c\}$, where it becomes the binary divergence of the cell probabilities:

$$I(Z;\widehat{Z}) \;\ge\; \mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

Dropping the nonnegative term $(1-p)\log\frac{1}{1-q}$ and rearranging,

$$\mathrm{kl}(p,q) \;\ge\; p\log\frac{1}{q} - h_2(p) \;\ge\; p\log\frac{1}{\mathcal{Q}_Z(t)} - \log 2 \;\ge\; a\log\frac{1}{\mathcal{Q}_Z(t)} - \log 2,$$

using $q \le \mathcal{Q}_Z(t) < 1$, $h_2 \le \log 2$, and $p \ge a$ together with $\log(1/\mathcal{Q}_Z(t)) > 0$. Since S was an arbitrary feasible summary and rates are nonnegative, this yeilds the bound.

For the strengthening, recall that $\mathrm{kl}(p,q)$ is nondecreasing in p on $p \ge q$ and nonincreasing in q on $q \le p$. If $a > \mathcal{Q}_Z(t)$, then $p \ge a > \mathcal{Q}_Z(t) \ge q$, so both monotonicity properties apply and $\mathrm{kl}(p,q) \ge \mathrm{kl}(a,q) \ge \mathrm{kl}(a, \mathcal{Q}_Z(t))$.

No metric or density structure was used: only the measurability of the event E_t and the nonnegativity of the distortion, so the argument applies verbatim to any jointly measurable nonnegative distortion. □

H.2 A LAPLACE FORM

For $s > 0$ let $\mathcal{A}_Z(s) := \sup_Q \mathbb{E}\exp\{-s\mathrm{KL}(Z\|Q)\}$.

Proposition H.1. *$R^\star(\varepsilon) \;\ge\; \big[\sup_{s>0}\{-s\varepsilon - \log\mathcal{A}_Z(s)\}\big]_+$, and the statement holds for any measurable nonnegative distortion and arbitrary reconstructions.*

Proof. The tool is the Gibbs variational inequality: for probability measures $P \ll \mu$ and a measurable $f \le 0$,

$$\mathrm{KL}(P\|\mu) \;\ge\; \int f\,dP - \log\int e^f\,d\mu. \tag{22}$$

For bounded f this follows by writing $\mu_f(dz) := e^{f(z)}\mu(dz)/\int e^f\,d\mu$ and computing

$$\mathrm{KL}(P\|\mu) - \Big[\int f\,dP - \log\int e^f\,d\mu\Big] = \int \log\Big(\frac{dP}{d\mu}\,e^{-f}\int e^f\,d\mu\Big)dP = \mathrm{KL}(P\|\mu_f) \;\geq\; 0.$$

For unbounded $f \leq 0$, apply this to the truncations $f_N := \max\{f, -N\}$; as $N \to \infty$, $\int f_N\,dP \downarrow \int f\,dP$ by monotone convergence and $\int e^{f_N}d\mu \to \int e^f\,d\mu$ by dominated convergence with the bound $e^{f_N} \leq 1$, so the right side of (22) for f_N converges to that for f.

Fix $s > 0$, a summary S, and a measurable reconstruction Q_S. Apply (22) with $\mu := P_Z$, $P := P_{Z|S=s_0}$, and $f(z) := -s\mathrm{KL}(z\|Q_{s_0}) \leq 0$:

$$\begin{aligned}\mathrm{KL}(P_{Z|S=s_0}\|P_Z) \;&\geq\; -s\,\mathbb{E}\big[\mathrm{KL}(Z\|Q_S)\,\big|\,S=s_0\big] - \log\mathbb{E}_{Z\sim P_Z}\exp\{-s\mathrm{KL}(Z\|Q_{s_0})\}\\ &\geq\; -s\,\mathbb{E}\big[\mathrm{KL}(Z\|Q_S)\,\big|\,S=s_0\big] - \log\mathcal{A}_Z(s),\end{aligned}$$

where the last step bounds the Laplace transform of the fixed reconstruction Q_{s_0} by its supremum over reconstructions. Averaging over s_0 with respect to the law of S turns the left side into $I(Z;S) = \mathbb{E}\,\mathrm{KL}(P_{Z|S}\|P_Z)$, and data processing gives

$$I(Y;S) \;\geq\; I(Z;S) \;\geq\; -s\,\mathbb{E}\,\mathrm{KL}(Z\|Q_S) - \log\mathcal{A}_Z(s).$$

Now take $Q_S := \Pi_S$ and a feasible summary, so that $\mathbb{E}\,\mathrm{KL}(Z\|Q_S) = \Delta_{\mathrm{KL}}(S) \leq \varepsilon$; then $I(Y;S) \geq -s\varepsilon - \log\mathcal{A}_Z(s)$. Taking the supremum over $s > 0$ and the infimum over feasible summaries gives the claim, and the bracket $[\cdot]_+$ records that rates are nonnegative. The quantity $\mathcal{A}_Z(s)$ is strictly positive, since the choice $Q = \pi$ gives $\mathrm{KL}(Z\|\pi) < \infty$ almost surely because $\mathbb{E}\,\mathrm{KL}(Z\|\pi) = I(\theta;Y) < \infty$; and it is at most one, so $-\log\mathcal{A}_Z(s) \geq 0$. □

This is the general Shannon lower bound of Riegler et al. (2018) with the prior-predictive law of the random posterior as reference measure, in which case the generalized entropy term vanishes. It can be stronger than the small ball bound of Theorem 3.3 because it integrates the whole distortion profile rather than thresholding it.

H.3 Matching exponents under small-ball regularity

Suppose $\mathcal{Q}_Z(t) \leq Ct^\alpha$ for all $t > 0$. For any reconstruction Q, put $X := \mathrm{KL}(Z\|Q) \geq 0$ and write the Laplace transform as a layer-cake integral, substituting $u = e^{-st}$:

$$\begin{aligned}\mathbb{E}\,e^{-sX} = \int_0^1 \mathbb{P}\big(e^{-sX} \geq u\big)\,du &= \int_0^\infty \mathbb{P}(X \leq t)\,se^{-st}\,dt\\ &\leq \int_0^\infty Ct^\alpha se^{-st}\,dt = C\,\Gamma(\alpha+1)\,s^{-\alpha}.\end{aligned}$$

Taking the supremum over Q gives $\mathcal{A}_Z(s) \leq C\Gamma(\alpha+1)s^{-\alpha}$, so Proposition H.1 yields $R^\star(\varepsilon) \geq -s\varepsilon + \alpha\log s - \log\{C\Gamma(\alpha+1)\}$ for every $s > 0$. The right side is maximized at $s = \alpha/\varepsilon$, where it equals

$$R^\star(\varepsilon) \;\geq\; \alpha\log\frac{\alpha}{\varepsilon} - \alpha - \log\{C\Gamma(\alpha+1)\}.$$

If in addition $D_M^{\mathrm{av}} \leq KM^{-1/\alpha}$, Theorem 3.2 gives the matching upper bound and $R^\star(\varepsilon) = \alpha\log(1/\varepsilon) + O(1)$ as $\varepsilon \downarrow 0$. These are explicit sufficient conditions for a matching exponent for a fixed law, not a general theorem; for a sequence of experiments indexed by n all dependence of C, K, α on n must be tracked, and the conditions exclude a finite posterior source at resolutions below its atom scale.

H.4 A mass-sensitive cluster converse

Proposition H.2. *Let $A_1, \dots, A_M$ be disjoint measurable data sets with $M \geq 2$, $A = \bigcup_j A_j$, $m = P_Y(A) > 0$ and $q_j = P_Y(A_j)/m$. Suppose $\mathrm{H}(\Pi_y, P_j) \leq \rho$ for $y \in A_j$ and $\mathrm{H}(P_i, P_j) \geq 2(\rho+\eta)$ for $i \neq j$, with $\eta > 0$. Then every summary satisfies*

$$\Delta_{\mathrm{H}}(S) \;\geq\; m\eta^2\left[\frac{H(q) - I(Y;S)/m - \log 2}{\log M}\right]_+, \qquad \Delta_{\mathrm{KL}}(S) \;\geq\; 2\Delta_{\mathrm{H}}(S).$$

Proof. The argument turns a summary into a decoder of the cluster index and applies Fano's inequality. An accurate summary posterior identifies the cluster; so if the cluster is hard to identify from S, the summary posterior is often inaccurate.

On the event A, let J be the cluster index, $J = j$ when $Y \in A_j$, and set $J := 0$ off A. Define a decoder $\widehat{J}(S)$ as the unique j with $\mathrm{H}(P_j, \Pi_S) < \rho + \eta$ when such a j exists, and arbitrarily otherwise. Uniqueness holds because two such indices $j \neq j'$ would give $\mathrm{H}(P_j, P_{j'}) \leq \mathrm{H}(P_j, \Pi_S) + \mathrm{H}(\Pi_S, P_{j'}) < 2(\rho + \eta)$, contradicting the separation hypothesis. If $Y \in A_j$ and $\mathrm{H}(\Pi_Y, \Pi_S) < \eta$, then

$$\mathrm{H}(P_j, \Pi_S) \leq \mathrm{H}(P_j, \Pi_Y) + \mathrm{H}(\Pi_Y, \Pi_S) < \rho + \eta,$$

so the decoder outputs j correctly. Hence the conditional error probability $e := \mathbb{P}(\widehat{J} \neq J \mid A)$ satisfies

$$e \;\leq\; \mathbb{P}\big(\mathrm{H}(\Pi_Y, \Pi_S) \geq \eta \,\big|\, A\big).$$

Fano's inequality in its nonuniform form states that for a variable J with M possible values and any estimator $\widehat{J}$ with error probability e, $H(J \mid \widehat{J}) \leq h_2(e) + e \log(M-1) \leq \log 2 + e \log M$. Under the conditional law given A, and since $\widehat{J}$ is a function of S,

$$I(J; S \mid A) = H(J \mid A) - H(J \mid S, A) \;\geq\; H(q) - H(J \mid \widehat{J}, A) \;\geq\; H(q) - \log 2 - e \log M.$$

To relate this to the rate, note that $(\mathbf{1}_A, J)$ is a function of Y, so by data processing and the chain rule,

$$I(Y; S) \;\geq\; I(\mathbf{1}_A, J; S) \;\geq\; I(J; S \mid \mathbf{1}_A) = m\, I(J; S \mid A) + (1-m)\, I(J; S \mid A^{\mathrm{c}}) = m\, I(J; S \mid A),$$

because J is constant off A. Combining the last two displays and solving for e,

$$e \;\geq\; \frac{H(q) - I(Y; S)/m - \log 2}{\log M}.$$

Finally, restrict the Hellinger distortion to A and apply Markov's inequality to the event $\{\mathrm{H}(\Pi_Y, \Pi_S) \geq \eta\}$:

$$\Delta_{\mathrm{H}}(S) \;\geq\; m\, \mathbb{E}\big[\mathrm{H}^2(\Pi_Y, \Pi_S) \,\big|\, A\big] \;\geq\; m\eta^2\, \mathbb{P}\big(\mathrm{H}(\Pi_Y, \Pi_S) \geq \eta \,\big|\, A\big) \;\geq\; m\eta^2 e.$$

Inserting the lower bound on e, and noting that $\Delta_{\mathrm{H}} \geq 0$ makes the positive part harmless, gives the first display.

For the second display, compare the two divergences pointwise. By Jensen's inequality for the concave logarithm,

$$\begin{aligned}\mathrm{KL}(P \| Q) = -2\, \mathbb{E}_P \log \sqrt{q/p} \;&\geq\; -2 \log \mathbb{E}_P \sqrt{q/p} = -2 \log \int \sqrt{pq} \\ &= -2 \log\big(1 - \mathrm{H}^2(P, Q)\big) \;\geq\; 2\mathrm{H}^2(P, Q),\end{aligned}$$

the last step using $-\log(1-x) \geq x$. Taking expectations with $(P, Q) = (\Pi_Y, \Pi_S)$ gives $\Delta_{\mathrm{KL}}(S) \geq 2\Delta_{\mathrm{H}}(S)$. □

Proposition H.2 is not dominated by Theorem 3.3: it depends on the entropy $H(q)$ of the cluster weights, whereas the small-ball quantity can be controlled by the single largest cluster. It is the appropriate statement when the experiment concentrates on a few well-separated regions of unequal probability. If the law is genuinely uniform on M separated posterior atoms, setting $m = 1$, $\rho = 0$ and $q_j = 1/M$ recovers a finite Fano converse for that experiment.

H.5 Evaluating the converses

Both $\mathcal{Q}_Z(t)$ and $\mathcal{A}_Z(s)$ are suprema over reconstructions, and the converses require *upper* bounds on them. Maximizing over a restricted family, or reporting the best value returned by a local optimizer, bounds the supremum from below and therefore yields an optimistic quantity that is not a valid converse. Repeated restarts do not repair this. Any numerical value should either certify the supremum or be reported as a candidate.

I PARAMETRIC RATES: HYPOTHESES AND ACHIEVABILITY

The prior-averaged expansion. Lemma 2.3 uses the prior-averaged entropy-risk theorem of Clarke & Barron (1994, Theorem 1, Eq. (2.3)). Its hypotheses include an interior compact prior support with the stated continuity and integrability conditions, likelihood smoothness and score/Hessian moment bounds, positive-definite Fisher information, and identifiability. Under those conditions,

$$I(\theta; Y^n) = \frac{d}{2} \log \frac{n}{2\pi e} + h(\pi) + \frac{1}{2} \mathbb{E}_\pi \log \det \mathcal{I}(\theta) + o(1). \tag{23}$$

Combining this equality with the posterior-loss identity proves the corollary: for a summary with $\Delta_{\mathrm{KL}}(S_n) \le \varepsilon$, Lemma 2.2 and data processing give $I(Y^n; S_n) \ge I(\theta; S_n) = I(\theta; Y^n) - \Delta_{\mathrm{KL}}(S_n) \ge I(\theta; Y^n) - \varepsilon$, and substituting (23) gives (4). Pointwise asymptotic normality alone does not justify averaging a remainder under the prior.

I.1 A MEAN-SQUARE ROUTE TO THE LOWER BOUND

Proposition I.1. *Suppose $h(\pi)$ is finite and some estimator $\widehat{\theta}_n(Y^n)$ has $0 < D_n = \mathbb{E}\|\theta - \widehat{\theta}_n\|^2 < \infty$. Then $I(\theta; Y^n) \ge h(\pi) - \frac{d}{2} \log \frac{2\pi e D_n}{d}$. If $\limsup_n n D_n \le K \in (0, \infty)$, every summary with $\Delta_{\mathrm{KL}}(S) \le \varepsilon$ satisfies $I(Y^n; S) \ge \frac{d}{2} \log n + h(\pi) - \frac{d}{2} \log \frac{2\pi e K}{d} - \varepsilon - o(1)$.*

Proof. The first claim bounds the conditional entropy of θ given the data by comparing each posterior with a Gaussian centred at the estimator. If $I(\theta; Y^n) = \infty$ there is nothing to prove, so assume it is finite; then, with $h(\pi)$ finite, $I(\theta; Y^n) = h(\pi) - h(\theta \mid Y^n)$, where $h(\theta \mid Y^n) = \mathbb{E}\, h(\Pi_Y)$ and a posterior without a Lebesgue density is assigned entropy $-\infty$.

Fix a data value y and let $G_y := N\big(\widehat{\theta}_n(y), (D_n/d) I_d\big)$ with density g_y. If Π_y has a density, nonnegativity of $\mathrm{KL}(\Pi_y \| G_y) = -h(\Pi_y) - \mathbb{E}_{\Pi_y} \log g_y$ gives

$$h(\Pi_y) \;\le\; -\mathbb{E}_{\Pi_y} \log g_y = \frac{d}{2} \log \frac{2\pi D_n}{d} + \frac{d}{2 D_n} \mathbb{E}_{\Pi_y} \|\theta - \widehat{\theta}_n(y)\|^2,$$

and the inequality is trivial otherwise. Averaging over y and using $\mathbb{E}\|\theta - \widehat{\theta}_n\|^2 = D_n$,

$$h(\theta \mid Y^n) \;\le\; \frac{d}{2} \log \frac{2\pi D_n}{d} + \frac{d}{2 D_n} D_n = \frac{d}{2} \log \frac{2\pi e D_n}{d}.$$

Subtracting from $h(\pi)$ gives $I(\theta; Y^n) \ge h(\pi) - \frac{d}{2} \log \frac{2\pi e D_n}{d}$.

For the second claim, let $\Delta_{\mathrm{KL}}(S) \le \varepsilon$. By Lemma 2.2 and data processing,

$$I(Y^n; S) \;\ge\; I(\theta; S) = I(\theta; Y^n) - \Delta_{\mathrm{KL}}(S) \;\ge\; h(\pi) - \frac{d}{2} \log \frac{2\pi e D_n}{d} - \varepsilon.$$

The hypothesis $\limsup_n n D_n \le K$ gives $D_n \le (K + o(1))/n$, hence

$$-\frac{d}{2} \log \frac{2\pi e D_n}{d} \;\ge\; -\frac{d}{2} \log \frac{2\pi e (K + o(1))}{nd} = \frac{d}{2} \log n - \frac{d}{2} \log \frac{2\pi e K}{d} - o(1),$$

which gives the stated bound. □

If $D_n = 0$, then θ is determined by the data and, for a continuous prior, $I(\theta; Y^n) = \infty$. For positive error, the proposition replaces the Clarke–Barron conditions by a posterior localization hypothesis at scale $n^{-1/2}$, which is implied by regular Bernstein–von Mises theory together with uniform integrability. Weak posterior convergence alone does not imply it.

I.2 ACHIEVABILITY UNDER A QUADRATIC POSTERIOR CHART

Assumption I.2 (Quadratic posterior chart)**.** There is a measurable statistic $T_n(Y^n) \in \mathbb{R}^d$ and a constant $L_n < \infty$ with $\mathrm{KL}(\Pi_y \| \Pi_{y'}) \le L_n \|T_n(y) - T_n(y')\|^2$ for all y, y' in a set of full probability. Write $V_n = \mathbb{E}\|T_n - \mathbb{E}T_n\|^2 < \infty$.

The hypothesis implies that observations with equal chart coordinates have equal posteriors, and is therefore considerably stronger than the existence of an asymptotically normal estimator. It should be verified, not inferred from regularity.

Lemma I.3. *Under Assumption I.2, every deterministic quantizer $S = g(T_n)$ satisfies $\Delta_{\mathrm{KL}}(S) \le 2L_n\,\mathbb{E}\|T_n - \mathbb{E}[T_n \mid S]\|^2 \le 2L_n\,\mathbb{E}\|T_n - t_S\|^2$ for any cell representatives t_S.*

Proof. Write Y for Y^n. The cell posterior Π_S is the average of the observation posteriors over the cell, and the chart assumption controls the divergence between two observation posteriors; the proof compares Π_Y with a random second observation from the same cell.

Given (Y, S), let Y' be drawn from the conditional law $P_{Y|S}$, independently of Y. Then the barycenter formula reads $\Pi_S = \mathbb{E}[\Pi_Y \mid S] = \mathbb{E}[\Pi_{Y'} \mid Y, S]$. The divergence $Q \mapsto \mathrm{KL}(P\|Q)$ is convex in its second argument, so the integral form of Jensen's inequality gives, almost surely,

$$\begin{aligned}\mathrm{KL}(\Pi_Y\|\Pi_S) = \mathrm{KL}\big(\Pi_Y \,\big\|\, \mathbb{E}[\Pi_{Y'} \mid Y, S]\big) \;&\le\; \mathbb{E}\big[\mathrm{KL}(\Pi_Y\|\Pi_{Y'}) \,\big|\, Y, S\big] \\ &\le\; L_n\,\mathbb{E}\big[\|T_n(Y) - T_n(Y')\|^2 \,\big|\, Y, S\big],\end{aligned}$$

where the last step is Assumption I.2 applied to the pair (Y, Y'). Taking expectations,

$$\Delta_{\mathrm{KL}}(S) \;\le\; L_n\,\mathbb{E}\|T_n - T_n'\|^2, \qquad T_n' := T_n(Y').$$

Given S, the vectors T_n and T_n' are independent with the same conditional law, so the expected squared difference is twice the trace of the conditional covariance:

$$\mathbb{E}\big[\|T_n - T_n'\|^2 \,\big|\, S\big] = 2\,\mathbb{E}\big[\|T_n - \mathbb{E}[T_n \mid S]\|^2 \,\big|\, S\big].$$

This gives the first inequality. For the second, the conditional mean minimizes conditional expected squared error among $\sigma(S)$-measurable vectors: $\mathbb{E}[\|T_n - t_S\|^2 \mid S] = \mathbb{E}[\|T_n - \mathbb{E}[T_n \mid S]\|^2 \mid S] + \|\mathbb{E}[T_n \mid S] - t_S\|^2$. □

Theorem I.4. *Under Assumption I.2 with $L_n > 0$, for every $\varepsilon > 0$ there is a countable-valued deterministic summary with $\Delta_{\mathrm{KL}}(S) \le \varepsilon$ and*

$$I(Y^n; S) \;\le\; \frac{d}{2}\log\left\{2\pi e\left[\left(\sqrt{\frac{L_n V_n}{2\varepsilon}} + \frac{1}{2}\right)^2 + \frac{1}{12}\right]\right\}.$$

In particular, if $L_n = O(n)$ and $V_n = O(1)$ then $R_n^\star(\varepsilon) \le \frac{d}{2}\log n + O_{d,\varepsilon}(1)$, which with Lemma 2.3 gives the rate $\frac{d}{2}\log n + O(1)$.

Proof. The summary is a cubic lattice quantizer of the chart statistic. We first choose the cell width so that the distortion bound of Lemma I.3 equals ε, and then bound the entropy of the lattice label by the entropy of a continuous variable with the same covariance up to the dithering term.

Set $\delta^2 := 2\varepsilon/(L_n d)$ and $S := \mathrm{round}(T_n/\delta) \in \mathbb{Z}^d$, rounding each coordinate to the nearest integer. Each coordinate of $T_n - \delta S$ has absolute value at most $\delta/2$, so $\|T_n - \delta S\|^2 \le d\delta^2/4$. Lemma I.3 with the cell representatives $t_S = \delta S$ gives

$$\Delta_{\mathrm{KL}}(S) \;\le\; 2L_n \cdot \frac{d\delta^2}{4} = \frac{L_n d\delta^2}{2} = \varepsilon.$$

For the rate, $I(Y^n; S) = H(S)$ because S is a deterministic countable-valued code (Lemma A.1), so it suffices to bound $H(S)$. Let U be uniform on $[-\frac{1}{2}, \frac{1}{2}]^d$ and independent of everything else. The dithered variable $W := S + U$ has a density equal to $\mathbb{P}(S = k)$ on each unit cube $k + [-\frac{1}{2}, \frac{1}{2})^d$, so its differential entropy is

$$h(W) = -\sum_{k \in \mathbb{Z}^d} \mathbb{P}(S = k)\log \mathbb{P}(S = k) = H(S).$$

Among $\mathbb{R}^d$-valued variables with covariance trace v, the isotropic Gaussian maximizes differential entropy, which gives $h(W) \le \frac{d}{2}\log\frac{2\pi e v}{d}$ with $v = \mathrm{tr}\,\mathrm{Cov}(W)$. By independence, $\mathrm{tr}\,\mathrm{Cov}(W) = \mathrm{tr}\,\mathrm{Cov}(S) + \mathrm{tr}\,\mathrm{Cov}(U) = \mathrm{tr}\,\mathrm{Cov}(S) + d/12$, since each coordinate of U has variance $1/12$.

It remains to bound $\operatorname{tr}\operatorname{Cov}(S) = \mathbb{E}\|S - \mathbb{E}S\|^2$. Centring at $\mathbb{E}[T_n]/\delta$ instead of $\mathbb{E}S$ can only increase the mean squared norm, and the triangle inequality in L^2 gives

$$\begin{aligned}\sqrt{\operatorname{tr}\operatorname{Cov}(S)} \;&\le\; \sqrt{\mathbb{E}\|S - \mathbb{E}[T_n]/\delta\|^2} \\ &\le\; \sqrt{\mathbb{E}\|S - T_n/\delta\|^2} + \sqrt{\mathbb{E}\|T_n/\delta - \mathbb{E}[T_n]/\delta\|^2} \;\le\; \frac{\sqrt{d}}{2} + \frac{\sqrt{V_n}}{\delta}.\end{aligned}$$

Substituting $\delta = \sqrt{2\varepsilon/(L_n d)}$ gives $\sqrt{V_n}/\delta = \sqrt{d}\,\sqrt{L_n V_n/(2\varepsilon)}$, hence

$$\operatorname{tr}\operatorname{Cov}(S) \;\le\; d\Big(\sqrt{\frac{L_n V_n}{2\varepsilon}} + \frac{1}{2}\Big)^2, \qquad \frac{\operatorname{tr}\operatorname{Cov}(W)}{d} \;\le\; \Big(\sqrt{\frac{L_n V_n}{2\varepsilon}} + \frac{1}{2}\Big)^2 + \frac{1}{12}.$$

Inserting this into the maximum-entropy bound gives the display; in particular the finite second moment of T_n makes $H(S)$ finite. If $L_n = O(n)$ and $V_n = O(1)$, the bracket is $O(n/\varepsilon)$ and the bound reads $\frac{d}{2}\log n + O_{d,\varepsilon}(1)$. □

A fixed-cardinality variant follows from the extended Pierce lemma (El Nmeir et al., 2020, Thm. 1.1(b)): if $\mathbb{E}\|T_n\|^{2+\gamma} < \infty$ for some $\gamma > 0$ then $D_M^{\mathrm{av}} \le 2L_n\kappa_{d,2,\gamma}^2(\mathbb{E}\|T_n\|^{2+\gamma})^{2/(2+\gamma)}M^{-2/d}$, whence $\log N^{\mathrm{av}}(\varepsilon) \le \frac{d}{2}\log L_n + \frac{d}{2}\log(1/\varepsilon) + O(1)$. The displayed $O(1)$ is uniform over a sequence of experiments only when the required $(2+\gamma)$ moments are uniformly bounded. The entropy-constrained bound above needs only $V_n < \infty$.

Remark I.5. The deterministic quantizer of Theorem I.4 attains the leading order $\frac{d}{2}\log n$ but not the optimal frontier: by Theorem 4.2, its rate exceeds $R_n^\star(\varepsilon)$ by a strictly positive amount, which is bounded in n and therefore absorbed in the $O(1)$.

I.3 A VERIFIED NON-GAUSSIAN CLASS

Proposition I.6. *In one dimension let $p(y \mid x) = h_0(y)\exp\{xy - A(x)\}$ with A finite and twice differentiable on $\mathbb{R}$ and $0 < \kappa \le A''(x) \le K < \infty$ for all x. Let $\theta \sim N(0,\tau^2)$ and take n conditionally independent observations. Then $T_n = \bar{Y}$ satisfies Assumption I.2 with $L_n = n^2/\{2(\tau^{-2} + n\kappa)\}$ and $V_n \le K^2\tau^2 + K/n$, and $\widehat{\theta}_n = (A')^{-1}(\bar{Y})$ has Bayes mean-square error at most $K/(\kappa^2 n)$. Consequently $R_n^\star(\varepsilon) = \frac{1}{2}\log n + O_\varepsilon(1)$ for each fixed $\varepsilon > 0$, and $\frac{d}{2}\log n + O_{d,\varepsilon}(1)$ for d independent such coordinates.*

Proof. We verify the chart assumption through the log-partition function of the posterior, then bound the chart variance and the estimator's error, check that the Bayes information is finite, and finally combine the lower and upper bounds.

Write $k_n := \tau^{-2} + n\kappa$. The likelihood of n observations depends on the data only through $t = \bar{Y}$, and the posterior P_t given $T_n = t$ has density proportional to $\exp\{-\psi_t(x)\}$ with

$$\psi_t(x) := \frac{x^2}{2\tau^2} - ntx + nA(x), \qquad \psi_t''(x) = \tau^{-2} + nA''(x) \;\ge\; k_n.$$

Thus P_t is strongly log-concave with parameter k_n. The scalar Brascamp–Lieb inequality states that a density proportional to $e^{-\psi}$ with $\psi'' \ge k > 0$ satisfies $\operatorname{Var}(g(X)) \le \mathbb{E}[g'(X)^2/\psi''(X)]$ for smooth g (Saumard & Wellner, 2014, Prop. 10.1); with $g(x) = x$ it gives $\operatorname{Var}_{P_t}(\theta) \le \mathbb{E}_{P_t}[1/\psi_t''(\theta)] \le k_n^{-1}$.

Define the posterior log-partition function

$$F_n(t) := \log \int_{\mathbb{R}} \exp\{-\psi_t(x)\}\,dx.$$

Since $A(x) \ge A(0) + A'(0)x + \kappa x^2/2$, the integrand is dominated by a Gaussian-type function, so F_n is finite for every t and can be differentiated under the integral sign; the derivatives multiply the integrand by polynomials in x, which the same quadratic bound keeps integrable. Differentiating twice,

$$F_n'(t) = n\,\mathbb{E}_{P_t}[\theta], \qquad F_n''(t) = n^2\operatorname{Var}_{P_t}(\theta) \;\le\; \frac{n^2}{k_n}.$$

The posterior densities are $p_t(x) = \exp\{-\psi_t(x) - F_n(t)\}$, so $\log\{p_t(x)/p_u(x)\} = n(t-u)x - F_n(t) + F_n(u)$, and integrating against P_t,

$$\mathrm{KL}(P_t\|P_u) = n(t-u)\,\mathbb{E}_{P_t}[\theta] - F_n(t) + F_n(u) = F_n(u) - F_n(t) - F_n'(t)(u-t).$$

This is the second-order Taylor remainder of F_n, so by the mean value form of the remainder and the bound on F_n'',

$$\mathrm{KL}(P_t\|P_u) = \tfrac{1}{2}F_n''(\xi)(u-t)^2 \;\le\; \frac{n^2}{2k_n}(u-t)^2$$

for some ξ between t and u. This verifies Assumption I.2 with $T_n = \bar{Y}$ and $L_n = n^2/(2k_n)$.

For the chart variance, the exponential-family identities give $\mathbb{E}[T_n \mid \theta] = A'(\theta)$ and $\mathrm{Var}(T_n \mid \theta) = A''(\theta)/n \le K/n$. By the law of total variance,

$$V_n = \mathrm{Var}(T_n) = \mathbb{E}\,\mathrm{Var}(T_n \mid \theta) + \mathrm{Var}(A'(\theta)) \;\le\; \frac{K}{n} + \mathrm{Var}(A'(\theta)).$$

Since $|A''| \le K$, the function A' is K-Lipschitz, and for an independent copy θ' of θ,

$$\mathrm{Var}(A'(\theta)) = \tfrac{1}{2}\mathbb{E}\big[(A'(\theta) - A'(\theta'))^2\big] \;\le\; \tfrac{1}{2}K^2\,\mathbb{E}\big[(\theta-\theta')^2\big] = K^2\tau^2.$$

Hence $V_n \le K^2\tau^2 + K/n$.

For the estimator, $A'' \ge \kappa > 0$ makes A' strictly increasing with $A'(x) \to \pm\infty$ as $x \to \pm\infty$, so $A' : \mathbb{R} \to \mathbb{R}$ is a bijection, and its inverse has derivative $1/A'' \le 1/\kappa$, hence is κ^{-1}-Lipschitz. With $\widehat{\theta}_n = (A')^{-1}(T_n)$ and $\theta = (A')^{-1}(A'(\theta))$,

$$\mathbb{E}\big[(\widehat{\theta}_n - \theta)^2 \,\big|\, \theta\big] \;\le\; \kappa^{-2}\,\mathbb{E}\big[(T_n - A'(\theta))^2 \,\big|\, \theta\big] = \frac{A''(\theta)}{\kappa^2 n} \;\le\; \frac{K}{\kappa^2 n}.$$

Averaging over θ gives the Bayes mean-square error bound. This error is positive, because the posterior has positive variance, so Proposition I.1 applies with $\limsup_n nD_n \le K/\kappa^2$.

The experiment has finite Bayes information. For any fixed law Q on the data space, $\mathbb{E}\,\mathrm{KL}(P_{Y^n|\theta}\|Q) = I(\theta;Y^n) + \mathrm{KL}(P_{Y^n}\|Q)$, so $I(\theta;Y^n) \le \mathbb{E}\,\mathrm{KL}(P_{Y^n|\theta}\|Q)$. Take Q to be the sampling law at $x = 0$. For one observation, $\mathrm{KL}(p(\cdot \mid x)\|p(\cdot \mid 0)) = \mathbb{E}_x[xY - A(x) + A(0)] = xA'(x) - A(x) + A(0)$, and Taylor's theorem with $A'' \le K$ gives $xA'(x) - A(x) + A(0) \le Kx^2/2$. The n observations are conditionally independent, so

$$I(\theta;Y^n) \;\le\; n\,\mathbb{E}\{\theta A'(\theta) - A(\theta) + A(0)\} \;\le\; \frac{nK}{2}\mathbb{E}[\theta^2] = \frac{nK\tau^2}{2} < \infty.$$

Also $h(\pi) = \frac{1}{2}\log(2\pi e\tau^2)$ is finite.

We combine the bounds. Proposition I.1 with $d = 1$ gives, for every summary with $\Delta_{\mathrm{KL}}(S) \le \varepsilon$, $I(Y^n;S) \ge \frac{1}{2}\log n + h(\pi) - \frac{1}{2}\log(2\pi eK/\kappa^2) - \varepsilon - o(1)$, so $R_n^\star(\varepsilon) \ge \frac{1}{2}\log n - O_\varepsilon(1)$. Theorem I.4 with $L_n = n^2/(2k_n) \le n/(2\kappa) = O(n)$ and $V_n = O(1)$ gives $R_n^\star(\varepsilon) \le \frac{1}{2}\log n + O_\varepsilon(1)$. Together, $R_n^\star(\varepsilon) = \frac{1}{2}\log n + O_\varepsilon(1)$.

For d independent coordinates (θ_j, Y_j^n), the posterior is the product of the coordinate posteriors, and the divergence between two product measures is the sum of the coordinate divergences. Hence Assumption I.2 holds for $T_n = (\bar{Y}_1, \dots, \bar{Y}_d)$ with the same L_n, since $\mathrm{KL}(\Pi_y\|\Pi_{y'}) = \sum_j \mathrm{KL}(P_{t_j}\|P_{t_j'}) \le L_n\sum_j (t_j - t_j')^2 = L_n\|T_n(y) - T_n(y')\|^2$. The chart variances, the squared estimation errors, the prior entropies, and the Bayes informations all add over coordinates, so the same two bounds give $R_n^\star(\varepsilon) = \frac{d}{2}\log n + O_{d,\varepsilon}(1)$. □

The condition is not vacuous outside Gaussian data: taking h_0 to be an equal mixture of $N(\pm b, \sigma^2)$ gives $A(x) = \frac{1}{2}\sigma^2x^2 + \log\cosh(bx)$ and hence $\kappa = \sigma^2$, $K = \sigma^2 + b^2$, with a genuinely non-Gaussian two-component observation law. Nor is it automatic: Bernoulli curvature degenerates and Poisson curvature is unbounded.

I.4 Proof of Theorem 3.4

Proof. The proof has a converse part, which lower-bounds the distortion of every channel of rate R by a conditional entropy-power argument, and an achievability part, which shows that adding Gaussian noise to the posterior mean makes every inequality of the converse an equality.

We first record the conjugate computation. Given Y^n, the posterior is $\theta \mid Y^n \sim N(m, vI_d)$ with

$$m = \rho\bar{Y}, \qquad \rho = \frac{n\tau^2}{\sigma^2 + n\tau^2}, \qquad v = \frac{\tau^2\sigma^2}{\sigma^2 + n\tau^2},$$

since the posterior precision is $\tau^{-2} + n\sigma^{-2} = 1/v$ and the posterior mean is $v \cdot n\bar{Y}/\sigma^2$. Marginally $\bar{Y} \sim N(0, (\tau^2 + \sigma^2/n)I_d)$, so $m \sim N(0, s^2 I_d)$ with

$$s^2 = \rho^2\Big(\tau^2 + \frac{\sigma^2}{n}\Big) = \frac{n\tau^4}{\sigma^2 + n\tau^2}, \qquad \frac{s^2}{v} = \frac{n\tau^2}{\sigma^2} = a.$$

Define $\zeta := \theta - m$. Since the conditional law of ζ given Y^n is $N(0, vI_d)$ whatever the value of Y^n, the vector ζ is independent of Y^n; and because a summary's auxiliary randomness is independent of (θ, Y^n), the same conditional law holds given (Y^n, S), so ζ is independent of (Y^n, S) as well. Thus $\theta = m + \zeta$ with m a function of Y^n and $\zeta \sim N(0, vI_d)$ independent of (m, S).

For the converse, let S be any summary with $I(Y^n; S) = R < \infty$. Fix a value s_0 of S. Given $S = s_0$, the vectors m and ζ are independent, so the entropy power inequality $e^{2h(X_1+X_2)/d} \geq e^{2h(X_1)/d} + e^{2h(X_2)/d}$ for independent X_1, X_2 in $\mathbb{R}^d$, with the convention $e^{2h(X_1)/d} = 0$ when X_1 has no density, gives

$$h(\theta \mid S = s_0) \;\geq\; \frac{d}{2}\log\big\{e^{2h(m|S=s_0)/d} + e^{2h(\zeta)/d}\big\} = \frac{d}{2}\log\big\{e^{2h(m|S=s_0)/d} + 2\pi ev\big\},$$

using $h(\zeta) = \frac{d}{2}\log(2\pi ev)$. The function $x \mapsto \log(e^{2x/d} + 2\pi ev)$ is convex and nondecreasing, so averaging over s_0 and applying Jensen's inequality,

$$h(\theta \mid S) = \mathbb{E}\, h(\theta \mid S = s_0) \;\geq\; \frac{d}{2}\log\big\{e^{2h(m|S)/d} + 2\pi ev\big\}, \qquad h(m \mid S) := \mathbb{E}\, h(m \mid S = s_0).$$

Next, $h(m \mid S) = h(m) - I(m; S)$, and m is a function of Y^n, so $I(m; S) \leq I(Y^n; S) = R$ and

$$h(m \mid S) \;\geq\; h(m) - R = \frac{d}{2}\log(2\pi e s^2) - R, \qquad \text{hence} \qquad e^{2h(m|S)/d} \;\geq\; 2\pi e s^2 e^{-2R/d}.$$

Substituting into the previous display,

$$h(\theta \mid S) \;\geq\; \frac{d}{2}\log\big\{2\pi e s^2 e^{-2R/d} + 2\pi ev\big\} = \frac{d}{2}\log(2\pi ev) + \frac{d}{2}\log\Big\{1 + \frac{s^2}{v}e^{-2R/d}\Big\}.$$

The first term is $h(\theta \mid Y^n)$, the entropy of $N(m, vI_d)$. By Lemma 2.2, and since all entropies here are finite,

$$\Delta_{\mathrm{KL}}(S) = I(\theta; Y^n) - I(\theta; S) = h(\theta \mid S) - h(\theta \mid Y^n) \;\geq\; \frac{d}{2}\log\{1 + ae^{-2R/d}\}.$$

The right side is decreasing in R, so every channel with $I(Y^n; S) \leq R$ satisfies the same bound, and $D_n^\star(R) \geq \frac{d}{2}\log(1 + ae^{-2R/d})$.

For achievability at $R > 0$, let $S := m + N$ with $N \sim N(0, \lambda I_d)$ independent of everything else and

$$\lambda := \frac{s^2}{e^{2R/d} - 1}.$$

Since $Y^n \to m \to S$ is a Markov chain and m is a function of Y^n, $I(Y^n; S) = I(m; S)$, and the Gaussian channel formula gives

$$I(m; S) = \frac{d}{2}\log\Big(1 + \frac{s^2}{\lambda}\Big) = \frac{d}{2}\log e^{2R/d} = R.$$

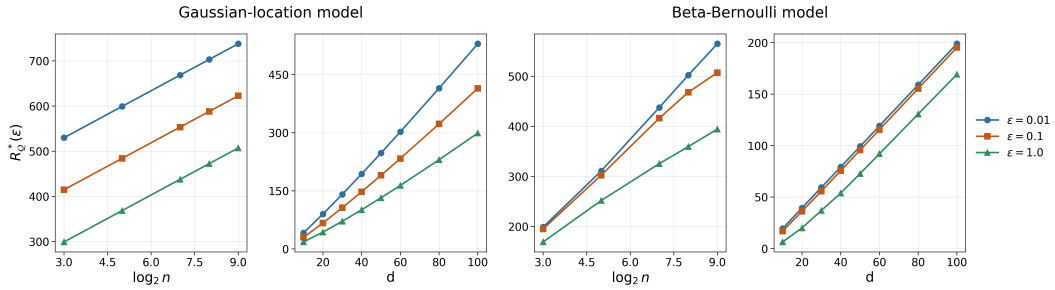


Figure 4: Analytic rate calibration in parametric models. Gaussian panels evaluate the exact frontier in Theorem 3.4; Beta–Bernoulli panels report achievable rates within the specified noisy-count channel grid. Distortion budgets $\varepsilon \in \{0.01, 0.1, 1\}$ are total budgets. Left to right: Gaussian rate versus n at $d = 100$, Gaussian rate versus d at $n = 8$, and the corresponding Beta–Bernoulli quantities. The Beta–Bernoulli curves are not optimized over all channels.

Given S, the vector m is Gaussian with covariance $\frac{s^2\lambda}{s^2+\lambda} I_d = s^2\big(1 - \frac{s^2}{s^2+\lambda}\big) I_d = s^2 e^{-2R/d} I_d$, because $s^2/(s^2+\lambda) = 1 - e^{-2R/d}$ for this λ. Hence $\theta = m + \zeta$ given S is Gaussian with covariance $(s^2 e^{-2R/d} + v) I_d$, and

$$\Delta_{\mathrm{KL}}(S) = h(\theta \mid S) - h(\theta \mid Y^n) = \frac{d}{2} \log \frac{s^2 e^{-2R/d} + v}{v} = \frac{d}{2} \log\{1 + a e^{-2R/d}\}.$$

So the converse bound is attained, and $D_n^\star(R) = \frac{d}{2}\log(1 + a e^{-2R/d})$ for $R > 0$; at $R = 0$ the constant summary has distortion $I(\theta; Y^n) = \frac{d}{2}\log(1+a)$, which agrees with the formula.

For the rate function, $R \mapsto \frac{d}{2}\log(1 + a e^{-2R/d})$ is a continuous strictly decreasing bijection from $[0, \infty)$ onto $(0, \frac{d}{2}\log(1+a)]$. For $0 < \varepsilon < \frac{d}{2}\log(1+a)$, the converse shows that every channel with distortion at most ε has rate at least the solution R of $\frac{d}{2}\log(1 + a e^{-2R/d}) = \varepsilon$, and the noisy channel at that rate has distortion exactly ε. Solving,

$$a e^{-2R/d} = e^{2\varepsilon/d} - 1 \quad \Longleftrightarrow \quad R = \frac{d}{2} \log \frac{a}{e^{2\varepsilon/d} - 1},$$

which is the second formula in (7). Finally, the law of the random posterior $N(m, v I_d)$ is non-atomic because the law of m is, so Proposition E.1 gives $R_n^\star(0) = \infty$. □

Under Assumption I.2 the same model admits $T_n = m$ with $L_n = 1/(2v) = O(n)$, since $\mathrm{KL}(N(m, vI_d) \| N(m', vI_d)) = \|m - m'\|^2/(2v)$, so average quantization attains the same leading order even though by Proposition F.2 no finite uniform cover exists.

J Details of the parametric rate–distortion calculations

Figure 4 uses the posterior distortion and information rate of Definition 2.1, in nats, with total distortion budgets rather than per-coordinate budgets.

Gaussian location model. For the Gaussian model in Theorem 3.4, the sufficient mean $T = n^{-1}\sum_i Y_i$ has $I(\theta; Y^n) = \frac{d}{2}\log(1 + n\tau^2/\sigma^2)$. We use $(\tau, \sigma) = (1, 1)$ in the figure and evaluate the optimal channel

$$S_\lambda = T + \eta, \qquad \eta \sim \mathcal{N}(0, \lambda I_d),$$

with η independent of (θ, Y^n). Its rate and distortion are

$$R_\lambda = I(Y^n; S_\lambda) = I(T; S_\lambda) = \frac{d}{2} \log\left(1 + \frac{\tau^2 + \sigma^2/n}{\lambda}\right), \tag{24}$$

and

$$\Delta_\lambda = I(\theta; Y^n \mid S_\lambda) = I(\theta; T \mid S_\lambda) = \frac{d}{2}\left[\log\left(1 + \frac{n\tau^2}{\sigma^2}\right) - \log\left(1 + \frac{\tau^2}{\sigma^2/n + \lambda}\right)\right].$$

For $0 < \varepsilon < I(\theta; Y^n)$, distortion increases and rate decreases with λ, so the optimal noise variance solves $\Delta_\lambda = \varepsilon$:

$$\lambda_\varepsilon = \frac{\tau^2}{\left(1 + \frac{n\tau^2}{\sigma^2}\right) e^{-2\varepsilon/d} - 1} - \frac{\sigma^2}{n}.$$

Theorem 3.4 and its proof in Appendix I establish optimality against arbitrary channels. At fixed $d, \varepsilon, \tau, \sigma$, (24) gives

$$R_{\lambda_\varepsilon} = \frac{d}{2} \log n + O(1) \qquad (n \to \infty).$$

The figure uses the exact rate, including its fixed-total-tolerance $\frac{d}{2} \log d$ growth as $d \to \infty$ at fixed n; per-coordinate tolerance instead gives linear growth in d.

Beta–Bernoulli model. For the second experiment, we consider independent coordinates

$$\theta_j \sim \mathrm{Beta}(a, b), \qquad Y_{ij} \mid \theta_j \sim \mathrm{Bernoulli}(\theta_j), \qquad j = 1, \dots, d, \quad i = 1, \dots, n,$$

with $(a, b) = (5, 5)$ in the figure. For each coordinate, the sufficient statistic is the count

$$C_j := \sum_{i=1}^{n} Y_{ij}.$$

We evaluate a noisy-count summary channel of the form

$$S_j \sim q_\rho(\cdot \mid C_j), \qquad q_\rho(s \mid c) \propto \exp\left(-\frac{(s-c)^2}{2\rho^2}\right), \qquad s \in \{0, \dots, n\},$$

applied independently across coordinates. Because the model factorizes across j, both rate and distortion factorize:

$$R_{\rho,d} = d\, R_\rho^{(1)}, \qquad \Delta_{\rho,d} = d\, \Delta_\rho^{(1)}.$$

In one dimension, the prior-predictive law of the count is Beta–Binomial:

$$p(c) = \binom{n}{c} \frac{B(a + c, b + n - c)}{B(a, b)}, \qquad c = 0, \dots, n.$$

The achieved rate is

$$R_\rho^{(1)} = I(C; S) = \sum_{c=0}^{n} \sum_{s=0}^{n} p(c)\, q_\rho(s \mid c) \log \frac{q_\rho(s \mid c)}{p_\rho(s)}, \qquad p_\rho(s) := \sum_{c=0}^{n} p(c) q_\rho(s \mid c).$$

The full posterior given the sufficient statistic is

$$\theta \mid C = c \sim \mathrm{Beta}(a + c, b + n - c),$$

whereas the posterior given the summary value $S = s$ is the mixture

$$p(\theta \mid s) = \sum_{c=0}^{n} p(c \mid s)\, \mathrm{Beta}(a + c, b + n - c).$$

We compute the one-dimensional distortion as

$$\Delta_\rho^{(1)} = \mathbb{E}\left[\mathrm{KL}\big(p(\theta \mid C) \,\|\, p(\theta \mid S)\big)\right] = H(\theta \mid S) - H(\theta \mid C),$$

where the entropies are evaluated by Gauss–Legendre quadrature on $[0, 1]$. In the experiments we use 300 quadrature nodes and the grid $\rho \in \mathcal{G} = \{0.25, 0.35, 0.5, 0.7, 1.0, 1.4, 2.0, 2.8, 4.0, 5.6, 8.0, 11.3, 16.0, 22.6, 32.0, 45.2, 64.0\}$. Since we do not invert $\rho \mapsto \Delta_\rho$ analytically, the plotted Beta–Bernoulli curves are grid-based achievable upper bounds:

$$R_{\mathrm{grid}}^{\mathrm{BB}}(\varepsilon; n, d) := \min_{\rho \in \mathcal{G}:\, \Delta_{\rho,d} \le \varepsilon} R_{\rho,d}.$$

Numerical details. The Gaussian curves require only closed-form evaluation. The Beta–Bernoulli curves require only finite sums over $c, s \in \{0, \dots, n\}$ together with one-dimensional quadrature for the mixture posterior entropy. No neural networks are trained for these figures. When varying n, we use the values shown on the horizontal axis of Figure 4; when varying d, we analogously use the displayed values of d.

K EXPERIMENTAL PROTOCOL AND ADDITIONAL RESULTS

K.1 TRAINING AND CHECKPOINT SELECTION

The simulator experiments use a 256-dimensional encoder summary, a 128-bit code with fixed prior activation probability 0.01, and a conditional normalizing flow with conditioner width 32 and 32 flow layers. Optimization uses Adam with learning rate 5×10^{-5} throughout g-and-k. On the other simulators the initial rate is 10^{-4} for $\beta < 0.1$ and 5×10^{-5} otherwise, with the training schedule's decay with 72,000 steps, each with fresh batch of 128 samples. All Ricker configurations share the same fitted $\log(1 + x)$ input normalizer.

The stochastic BNPE samples its code during training and deployment. The deterministic one samples Bernoulli codes with a temperature that decreases log-linearly from one to 0.1 between $0.1E$ and $0.8E$, and then uses thresholded codes for stabilization. Both models use a straight-through gradient during training, and evaluation of the deterministic BNPE always uses a hard threshold.

K.2 EVALUATION LOSS AND INFORMATION-RATE ESTIMATES

Each checkpoint is scored on five simulator sets, with $M = 2^{18} = 262{,}144$ samples. NLL averages eight deployed code draws per observation for stochastic deployment and uses one pass for hard deployment. The rate calculation uses 65,536 sampled codes and the empirical mixture over the full reference set.

For Bernoulli conditional probabilities $q_i(b) = q(b \mid y_i)$, define $\bar{q}_M(b) = M^{-1} \sum_i q_i(b)$. The empirical-mixture estimate is

$$\widehat{R}_M \approx \mathbb{E}_{b \sim \bar{q}_M}[-\log \bar{q}_M(b)] - \frac{1}{M} \sum_i H(q_i). \tag{25}$$

The finite empirical mixture's information is bounded by $\log M \simeq 12.477$. Evaluation clips Bernoulli probabilities to $[10^{-5}, 1 - 10^{-5}]$.

A a two-hidden-layer MADE (Germain et al., 2015) model of width 512 is fitted on an independently simulated dataset. For a fixed fitted marginal r, the population identity

$$\mathbb{E}_{Y,B}[-\log r(B)] - H(B \mid Y) = I(Y; B) + \mathrm{KL}(p_B \| r) \tag{26}$$

gives a variational upper bound.

K.3 CONTROLLED BENCHMARK COMPUTATION

The Beta–Bernoulli experiment enumerates count and code states and uses 512-point quadrature for the remaining integral over θ. The deterministic comparison enumerates all partitions of the nine count values. The convex hull includes independent time-sharing between deterministic count encoders. Iterative information-bottleneck updates provide attainable stochastic channels. Neural training directly penalizes finite-sum $I(C; B)$, unlike the simulator experiments' sparse-prior KL surrogate. Reported neural summaries use five training seeds per arm and penalty coefficient.

K.4 GAUSSIAN DIAGNOSTIC WITH A FIXED-COVARIANCE DECODER

This controlled benchmark measures how efficiently a binary code preserves the posterior mean when the decoder covariance is fixed. Let $\theta \sim N(0, I_4)$ and $Y_i \mid \theta \sim N(\theta, I_4)$ for $i = 1, \ldots, 4$. Writing $T = \bar{Y}$, the full posterior is $N(\mu(T), vI_4)$ with $\mu(T) = 0.8T$ and $v = 0.2$. The encoder has 16 binary outputs and the decoder is $q_\phi(\theta \mid B) = N(g_\phi(B), vI_4)$. Both networks use two hidden layers of width 256 with ReLU activations. The objective is NLL plus the sparse-prior Bernoulli KL penalty with prior activation probability 0.01. Training uses 4000 Adam updates, batch size 4096, initial learning rate 10^{-3} with cosine decay to 10^{-5}, and training seed zero.

Let $g^\star(B) = \mathbb{E}[\mu(T) \mid B]$. The population decomposition is

$$L_\phi - h(\theta \mid T) = D_{\mathrm{mean}}(B) + G_\phi, \tag{27}$$

$$D_{\mathrm{mean}}(B) := \frac{\mathbb{E}\|\mu(T) - g^\star(B)\|^2}{2v}, \qquad G_\phi := \frac{\mathbb{E}\|g^\star(B) - g_\phi(B)\|^2}{2v}. \tag{28}$$

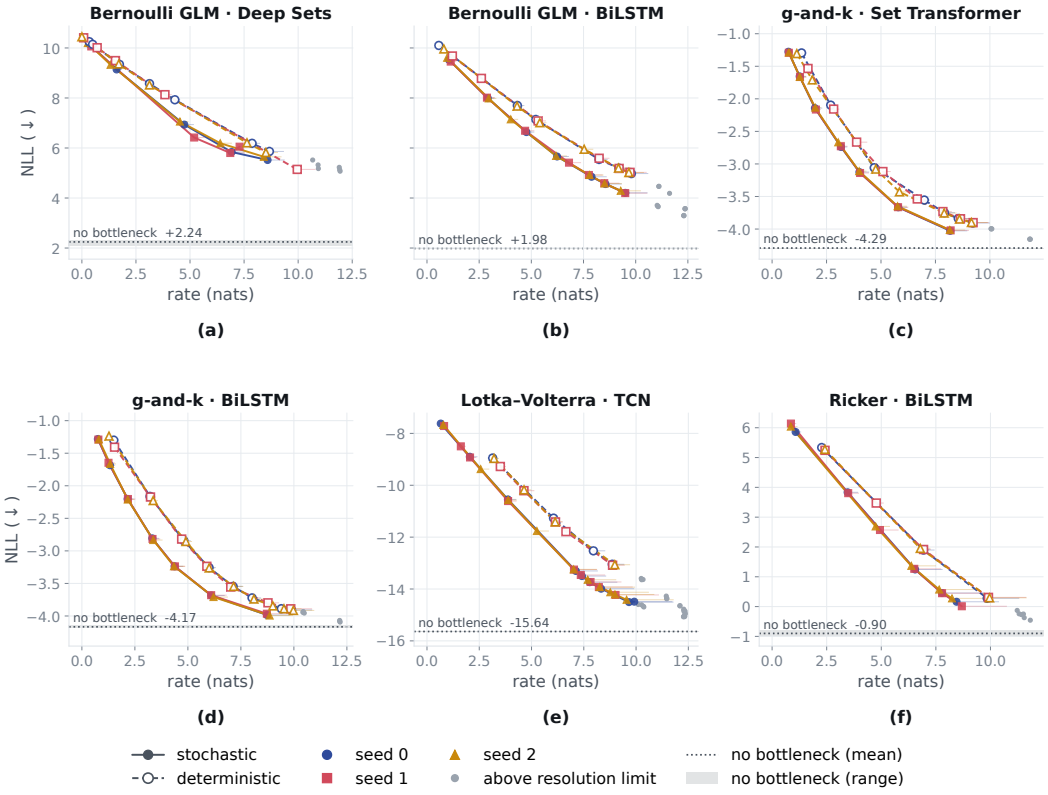


Figure 5: The six simulator–encoder combinations omitted from Figure 2.

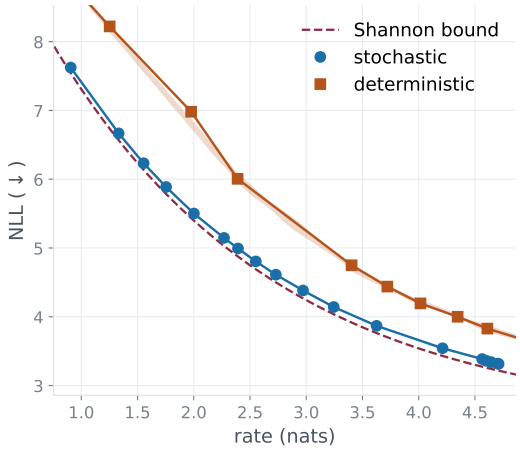


Figure 6: Gaussian posterior-mean diagnostic

Here $h(\theta \mid T) = 2.45688$ nats. This separates the code's posterior-mean error from the learned decoder's remaining mean error. The unrestricted posterior loss satisfies (Lemma 2.2):

$$\Delta_{\mathrm{KL}}(B) = D_{\mathrm{mean}}(B) - \mathbb{E}_B \mathrm{KL}(\Pi_B \,\|\, N(g^\star(B), vI_4)) \,. \tag{29}$$

The subtracted term includes covariance mismatch and any non-Gaussian shape of the coarsened posterior. At zero rate, $D_{\mathrm{mean}} = 8$, whereas $\Delta_{\mathrm{KL}} = 2\log 5 \simeq 3.2189$. Thus the diagnostic uses a restricted decoder family whose error remains meaningful even when its mean is fitted optimally.

Since $\mu(T) \sim N(0, 0.8I_4)$, Gaussian squared-error rate–distortion gives the exact unrestricted-channel reference $D_{\mathrm{mean}}(R) = 8e^{-R/2}$, which is a lower bound for the finite binary channels tested here. It is distinct from the posterior-KL frontier $D^\star(R) = 2\log(1 + 4e^{-R/2})$ in Theorem 3.4.

K.5 Additional posterior diagnostics

Neural ratio estimation offers a complementary route to measuring retained parameter information (Hermans et al., 2020). For the exact joint-to-product ratio $r(\theta, s) = p(\theta, s)/(\pi(\theta)p(s))$, its joint expected log ratio is $I(\theta; S)$. Comparing this quantity for full observations and summaries would target the information loss in Lemma 2.2. With learned ratios, however, the difference also includes both estimators' approximation errors. The conjugate experiment directly evaluates posterior loss and decoder mismatch; the simulator results report NLL and code rate.

L Related work and connections

Summary statistics for likelihood-free inference. Summary construction includes dimension reduction and subset selection for ABC (Blum et al., 2013; Nunes & Balding, 2010), regression-based statistics (Fearnhead & Prangle, 2012; Jiang et al., 2017), Fisher-information and score-based compression (Charnock et al., 2018; Alsing & Wandelt, 2018), and neural summary learning (Chen et al., 2021; 2023). Exchangeable and convolutional encoders adapt summaries to observation structure (Chan et al., 2018; Åkesson et al., 2022); amortized posterior estimators can also train a summary jointly with its density decoder (Radev et al., 2022). Closest to the fidelity criterion here, Hoffmann & Onnela (2026) connect expected posterior entropy, parameter–summary information, and expected posterior KL loss. Sui et al. (2023) evaluate cosmological summaries through parameter–summary information. We place this posterior fidelity alongside an explicit data–summary information budget, asking how much compression is compatible with posterior preservation and comparing learned encoders at matched estimated rates.

Posterior preservation and the information bottleneck. The information bottleneck (IB) trades information retained about a target against information retained about the observation (Tishby et al.,

2000; Gilad-Bachrach et al., 2003). Under $\theta \to Y \to S$, the identity $\Delta_{\mathrm{KL}}(S) = I(\theta;Y) - I(\theta;S)$ identifies our intrinsic rate with this classical frontier, expressed in posterior-error coordinates. Conditional-distribution KL distortion and posterior-mixture reconstruction are classical IB constructions (Harremoës & Tishby, 2007). Courtade & Weissman (2014) explicitly study rate-limited posterior compression under expected KL as a shifted logarithmic-loss coding problem. Their finite-alphabet operational results concern coding across independent experiments; our frontier concerns one use of a summary channel. The sufficient-statistic reduction of Armstrong (2026) allows the observation to be replaced by its random posterior without changing the unrestricted channel frontier. These foundations identify the object whose representation cost governs posterior-preserving summaries.

Posterior quantization and converses. Average posterior quantization uses the standard nearest-center and centroid arguments of entropy-constrained and Bregman quantization (Chou et al., 1989; Banerjee et al., 2005). The posterior mixture is the KL instance of the Bregman centroid, connecting deterministic summaries to average quantization of the random posterior. Our converses specialize arbitrary-prior small-ball information bounds (Xu & Raginsky, 2017) and general Shannon lower bounds (Riegler et al., 2018) to this source. They retain the predictive mass of distinguishable posterior states, which support covering and packing numbers alone do not capture. The resulting question is how finely the *law of the random posterior*, under a fixed Bayesian experiment, must be represented to meet a prescribed average distortion.

Haussler & Opper (1997) identify Bayes information with cumulative Bayesian predictive relative-entropy risk and bound it using prior-weighted geometry of sampling distributions. The compensation inequality transfers any such lower bound directly to $R_n^\star(\varepsilon) \ge I(\theta;Y_{1:n}) - \varepsilon$. In particular, suppose $Y_{1:n} \mid \theta \sim P_\theta^{\otimes n}$ satisfies their dominated-model setup. With this paper's normalization $\mathrm{H}^2(P,Q) = \frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2$, their Corollary 1(1) (p. 2466, with $\alpha = 1/2$) gives

$$R_n^\star(\varepsilon) \;\ge\; \left[-\int \pi(d\theta)\log\int \pi(d\vartheta)\exp\{-n\mathrm{H}^2(P_\theta,P_\vartheta)\} - \varepsilon\right]_+ .$$

This lower bound needs no uniform likelihood-ratio bound. Their corresponding Hellinger upper bounds require additional control (Corollary 1 or Theorem 3); mutual absolute continuity alone is insufficient. Their Section 7 takes a supremum over priors to obtain minimax bounds from sampling-model covering and packing numbers. Their Theorem 4(3), pp. 2482–2483, likewise concerns minimax predictive risk. These results control information acquired about a parameter from repeated sampling. Our posterior-space bounds control compression under a fixed predictive law, and the fixed-prior parametric converse uses the prior-averaged Clarke–Barron expansion below.

Parametric information and approximate sufficiency. Clarke & Barron (1994) give the regular, prior-averaged Bayes-information expansion underlying the $\frac{d}{2}\log n$ converse in Lemma 2.3. Hayashi & Tan (2018) obtain the same leading coefficient for finite-memory approximate sufficient-statistic coding under their reconstruction and alphabet-size criteria. Here distortion is posterior KL averaged under a fixed prior, and rate is data–summary mutual information; matching upper bounds follow under the stated posterior-chart conditions. The exact Gaussian benchmark specializes Gaussian IB (Chechik et al., 2005; Asoodeh & Calmon, 2020), making the dependence on sample size, dimension, and posterior tolerance explicit.

Stochastic summaries and deterministic time-sharing. Stochastic assignments are central to classical IB, and finite-alphabet entropy tradeoffs have long been studied through convex representations (Tishby et al., 2000; Witsenhausen & Wyner, 1975; Wyner & Ziv, 1973). Agglomerative IB expresses the relevance cost of merging cells through weighted Jensen–Shannon divergence (Slonim & Tishby, 2000). For deterministic targets, the IB frontier is piecewise linear and can be attained by visible time-sharing of deterministic summaries (Kolchinsky et al., 2019); full-support experiments can also have linear segments (Benger et al., 2023). Theorem 4.2 compares stochastic summaries with the entire visible time-sharing envelope of deterministic summaries. For finite observations, rare confusion between two cells yields a support-sensitive refinement of the weighted Jensen–Shannon merge cost. The general theorem replaces this pair-specific control by a uniform, experiment-level bound for random relabeling. Combined with finite-quantizer approximation and uniform integrability, it yields strict improvement at every interior rate for standard Borel experiments satisfying

the common-support condition, without requiring finite pairwise posterior KL or attainment of the deterministic envelope.

The compression cost and the role of randomness. The deterministic IB of Strouse & Schwab (2017) prices output entropy, changing the optimization problem. For ordinary additive coding-cost Lagrangians, encoder randomness can be fixed without increasing the objective; distribution-preserving reconstruction constraints can instead make shared randomness useful (Theis & Agustsson, 2021). Our separation theorem charges mutual information and permits private encoder randomness. Relabeling preserves the output distribution while reducing mutual information, so its gain depends on that choice of cost.

Learning and evaluating an information budget. The training penalty builds on neural variational IB (Alemi et al., 2017) and task-oriented inference compression (Shao et al., 2022). The binary-bottleneck protocol makes the summary channel explicit and permits comparisons across architectures and stochastic or deterministic modes. We measure deployed rate separately from the variational training penalty; the empirical rate estimates and decoder losses are also distinguished from the population posterior-information frontier.