THE REPRESENTATION COST OF CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Approximate inference in Bayesian neural networks requires choosing which weight dependencies to represent. We study shallow networks with a Gaussian prior, fixed data, Lipschitz activations and smooth convex likelihood energies of at most quadratic growth. Products over whole neurons allow dependence within each neuron; in Gaussian regression, their limiting optimum recovers the posterior mean but retains prior fluctuations. For mixtures of products over individual weights, the minimum log component count at width N is $\Theta(\sqrt{N})$ to approach a strictly better whole-neuron variational value and $\Theta(N)$ to approximate the full weight posterior in KL, at fixed tolerances below the respective positive gaps. The factors may be non-Gaussian. For ReLU, full Gaussian weight distributions with bounded prior KL can contract one direction of limiting uncertainty while missing mean directions available to whole-neuron products. The optimal ELBO ordering changes with the labels. In the width limit, the optimal full Gaussian can give a conditional forecast in the wrong direction even after both predictive marginals are made exact. Finite fits, a covariance intervention that preserves the mean function and a separate preference-head study illustrate consequences for prediction and evidence-based prior selection.

1 INTRODUCTION

Bayesian neural networks use a distribution over weights to describe predictions and uncertainty (Blundell et al., 2015). Approximate inference requires choosing which dependencies between weights to represent. A full-covariance Gaussian approximate posterior allows correlations between all weights; Gaussian conditional distributions can yield a jointly non-Gaussian approximate posterior (Ober & Aitchison, 2021). Recent comparisons with a fixed prior report tighter variational bounds for conditional approximations than for fitted full Gaussians (Rochussen & Fortuin, 2026). Explaining such differences requires understanding what each family can represent, beyond the outcome of a particular optimization run.

Learning a mean, contracting uncertainty, and predicting the effect of another observation place different demands on a weight distribution. Optimal diagonal-Gaussian approximations can revert to the prior in wide networks with odd activations (Coker et al., 2022). Even in conjugate regression, a correct variational mean can coexist with incorrect predictive variance (Odgers et al., 2026). Joint-prediction evaluations test dependence across inputs (Osband et al., 2022), while prediction-focused active learning uses dependence to value another observation (Bickford Smith et al., 2023; 2024). Invariance and permutation symmetry provide complementary explanations for variational failures (Kurle et al., 2022; Gelberg et al., 2024). We compare what variational families can preserve at their best solutions and quantify the component cost of expressing the needed dependence as a mixture.

We study shallow networks with a Gaussian prior and fixed data. A mixture of L products selects one component for the whole network, then draws individual weights independently in the original coordinates (Jaakkola & Jordan, 1998). A *whole-neuron* product instead permits dependence within neurons while keeping neurons independent. Both allow non-Gaussian factors. Comparing their best evidence lower bounds (ELBOs) separates family restrictions from optimization failure.

Consider the two observations $(x, y) = (1, 1)$ and $(-1, -1)$ studied by Coker et al. (2022), here with a bias-free ReLU network and Gaussian observation variance $1/2$. A product over scalar weights at bounded KL to the prior can shift the limiting mean network output only along $|x|$. It cannot fit the

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.

opposite labels; its optimal limiting mean at the two inputs remains $(0, 0)$. Whole-neuron dependence recovers the limiting posterior output mean $(1/2, -1/2)$ but retains output covariance $I_2/2$ instead of the posterior's $I_2/4$. Figure 1 shows this difference.

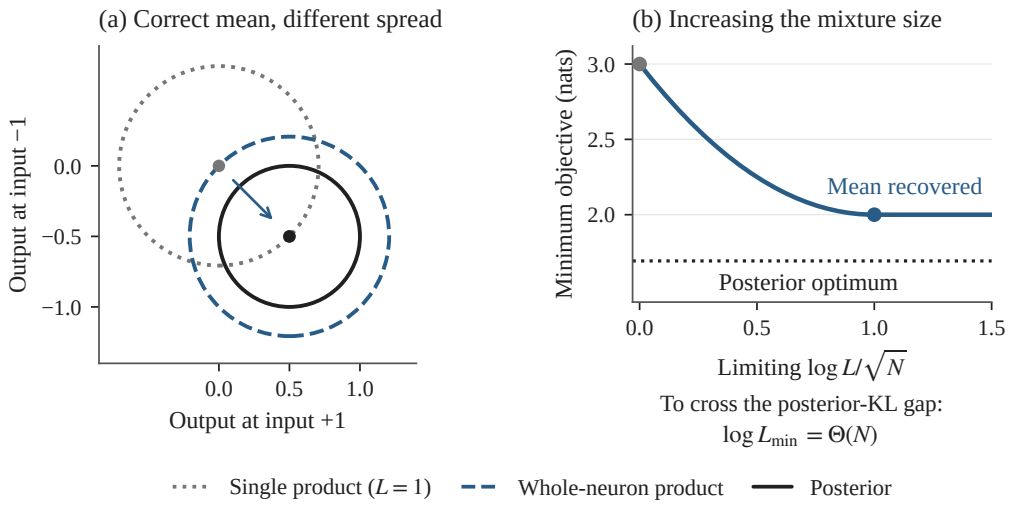


Figure 1: **Recovering the mean does not recover posterior spread.** Analytic width limits for the two-input example. (a) Means and one-standard-deviation contours of the best scalar-factorized product ($L = 1$), whole-neuron product and posterior. (b) Exact limiting minimum of prior KL plus expected squared error versus $\log L/\sqrt{N}$. Recovering the mean leaves a posterior-KL gap $1 - \log 2$; crossing it at fixed accuracy requires $\log L$ of order N (Appendix B.4; Theorem 3).

At width N, the minimum $\log L$ is $\Theta(\sqrt{N})$ to approach a strictly better whole-neuron value and $\Theta(N)$ to cross the remaining posterior-KL gap, each at a fixed tolerance below its positive gap. Coker et al. (2022, Section 6) leave the optimal ReLU mean-field limit open. We characterize it for the one-hidden-layer model of Section 2.

We also characterize the limiting outputs of full Gaussian weight distributions. For ReLU, this family can contract only the radial output direction while missing whole-neuron mean directions, and changing only the labels can reverse the families' optimal ELBO ordering. In the width limit, even after its individual predictive distributions are made exact, the optimal full Gaussian can move a conditional forecast in the wrong direction. An established conditional readout supplies the needed dependence (Hron et al., 2022). We isolate its covariance effect on marginal prediction by holding its mean function fixed.

Dependence also matters when the ELBO is used to select a prior. In a separate preference-head study on fixed language-model features, product-mixture restrictions reverse evidence-based prior selection, while structured covariance restores the evidence ordering.

2 Scalar-factorized mixtures and whole-neuron products

Consider a shallow network of width N. Neuron i has an outgoing weight $a_i \in \mathbb{R}$ and incoming weights $b_i \in \mathbb{R}^d$, collected in $\theta_i = (a_i, b_i)$. Under the prior U_N, all scalar weights are independent standard Gaussians. At input h, the neuron's contribution and the normalized network output are

$$z_h(a, b) = a\varphi(b^T h), \qquad F_N(h) = N^{-1/2} \sum_{i=1}^{N} z_h(\theta_i). \tag{1}$$

Here $F_N(h)$ is the latent network output; a future response is drawn from the likelihood conditional on this output. The incoming dimension d is fixed and φ is globally Lipschitz. Augmenting h with

a constant coordinate includes a Gaussian hidden bias; there is no separate output bias outside the $N^{-1/2}$ sum.

The observed inputs $H = (h_1, \ldots, h_s)$ remain fixed as width grows. The observed responses enter through the negative log likelihood $V(F_N(H))$ of the latent outputs. We assume that $V : \mathbb{R}^s \to [0, \infty)$ is convex and C^1, with $V(f) \le c(1 + \|f\|^2)$. Gaussian regression uses $V(f) = \frac{1}{2}(f - y)^T \Sigma^{-1}(f - y)$ with observation-noise covariance $\Sigma \succ 0$; logistic likelihoods also satisfy the assumptions. The weight posterior and variational objective are

$$P_N = Z_N^{-1} e^{-V(F_N(H))} U_N, \qquad G_N(Q) = D_{\mathrm{KL}}(Q \| U_N) + \mathbb{E}_Q V(F_N(H)). \tag{2}$$

Then $D_{\mathrm{KL}}(Q\|P_N) = G_N(Q) + \log Z_N$ whenever finite. Minimizing G_N is equivalent to maximizing the evidence lower bound (Blei et al., 2017). All logarithms and KL values are in nats; tolerances below are absolute. The prior's objective is bounded uniformly in width. Thus any approximation improving on it has bounded prior KL, the regime considered here.

Our family $\mathcal{S}_{N,L}$ consists of mixtures of at most L products over *each individual weight in the original parameterization*:

$$Q = \sum_{j=1}^{L} \pi_j \bigotimes_{i=1}^{N} \left(Q_{ij}^a \otimes \bigotimes_{k=1}^{d} Q_{ijk}^b \right). \tag{3}$$

We call these *scalar-factorized mixtures*. A draw chooses one component for the whole network, then draws its scalar weights independently. The shared label carries their dependence. Factors need not be Gaussian or identical; they and the mixture weights can vary with width and data. We count components, leaving factor complexity unrestricted; this is a representation cost, not a lower bound on inference time (Appendix C). At $L = 1$ this is the fully factorized, or mean-field, family. A *whole-neuron product* $Q = \bigotimes_{i=1}^{N} Q_i(a_i, b_i)$ keeps neurons independent but allows any joint distribution over each neuron's weights.

Two prior quantities distinguish the mean from its fluctuations. Shifting outgoing weights independently of incoming weights creates the mean feature ψ; the covariance kernel K describes the random output left around it:

$$\psi(h) = \mathbb{E}\varphi(b^T h), \qquad K(h, g) = \mathbb{E}[z_h z_g], \qquad \Gamma_H \sim \mathcal{N}(0, K_{HH}). \tag{4}$$

Under the prior, finite vectors of outputs converge to this Gaussian process limit (Lee et al., 2018). Statements involving a singular covariance matrix are restricted to its range; $K^\dagger$ denotes the corresponding pseudoinverse.

3 Two scales of represented dependence

The first obstruction concerns the mean. Let $A_N = N^{-1/2} \sum_i a_i$ and let J be the component label in (3). For every finite-KL mixture,

$$|\mathbb{E}_Q F_N(h) - \psi(h) \mathbb{E}_Q A_N| \le \frac{\mathrm{Lip}(\varphi)\|h\|}{\sqrt{N}} \left\{ D_{\mathrm{KL}}(Q\|U_N) + I(J; \theta_{1:N}) \right\}. \tag{5}$$

This mutual information measures how much the weights reveal about their component label and is at most $\log L$. Thus a bounded objective and $\log L = o(\sqrt{N})$ restrict the limiting mean to a multiple of ψ. Within a component, the mean output factors as $\mathbb{E}a_i \, \mathbb{E}\varphi(b_i^T h)$. Moving away from the prior feature therefore requires changes in both the outgoing mean and the incoming distribution. Their product is bounded by the component's KL cost; averaging components splits this cost into the two terms in (5).

A second restriction concerns the fluctuations around that mean. At bounded prior KL and $\log L = o(N)$, every weak limit on a fixed query set is a random mean added to an independent Gaussian with the prior covariance K (Appendix A). Mixing can add shifts but cannot remove this Gaussian fluctuation. These covariance statements concern weak limits; arbitrary factors need not have convergent second moments. In the regime $\log L = o(\sqrt{N})$, combining this description of the weak limits with the mean restriction and convexity of V gives the best limiting value

$$J_S = \min_{m \in \mathbb{R}} \left\{ \tfrac{1}{2} m^2 + \mathbb{E}V(\Gamma_H + m\psi_H) \right\}. \tag{6}$$

The first term pays for shifting the outgoing weights; the second evaluates that shift with the prior fluctuation still present. The minimizer m_* is unique.

Theorem 1 (Limit under scalar factorization)**.** *Under the assumptions of Section 2, if* $\log L_N = o(\sqrt{N})$*, then*

$$\inf_{Q \in \mathcal{S}_{N,L_N}} G_N(Q) \longrightarrow J_S.$$

Every sequence $Q_N \in \mathcal{S}_{N,L_N}$ *with* $G_N(Q_N) \le J_S + o(1)$ *has, on each fixed finite query set A, weak output limit* $\mathcal{N}(m_*\psi_A, K_{AA})$ *and convergent mean* $m_*\psi_A$*.*

For odd φ, $\psi = 0$ and every near-optimal weight distribution converges to the prior in KL. For ReLU, $\psi(h) = \|h\|/\sqrt{2\pi}$, so the family can learn a single radial feature, which can be useful on some datasets.

Whole-neuron dependence permits any mean direction supported by the kernel. Its limiting value is

$$J_B = \min_{v \in \operatorname{range} K_{HH}} \left\{ \tfrac{1}{2} v^T K_{HH}^{\dagger} v + \mathbb{E}V(\Gamma_H + v) \right\}. \tag{7}$$

Its minimizer v_* is unique. Whole-neuron products contain scalar-factorized products, so $J_B \le J_S$, but both objectives retain the prior Gaussian fluctuation.

How mixing adds a mean direction. Our construction redraws about $\sqrt{N}$ randomly selected neurons from the prior restricted to rectangular cells, so every component still factorizes over scalar weights. Under the $N^{-1/2}$ normalization, these changes can produce an order-one mean shift while the unchanged majority supplies the limiting prior fluctuations. Channel resolvability approximates this weight distribution by a finite mixture with $\log L = O(\sqrt{N})$ without enumerating neuron subsets (Yu & Tan, 2019, Lemma 1), following the sparse-signaling idea of Bloch (2016, Section V) (Appendix B).

Theorem 2 (Attaining the whole-neuron value)**.** *For every fixed* $\epsilon > 0$*, there are a finite* C_ϵ *and* $Q_N \in \mathcal{S}_{N,L_N}$ *with* $\log L_N \le C_\epsilon \sqrt{N} + O(1)$ *and* $\limsup_N G_N(Q_N) \le J_B + \epsilon$*. If* $\sqrt{N} \ll \log L_N \ll N$*, then*

$$\inf_{Q \in \mathcal{S}_{N,L_N}} G_N(Q) \longrightarrow J_B.$$

If $J_S > J_B$ *and* $0 < \epsilon < J_S - J_B$*, the smallest* L *for which some* $Q \in \mathcal{S}_{N,L}$ *satisfies* $G_N(Q) \le J_B + \epsilon$ *has* $\log L_{\min} = \Theta(\sqrt{N})$*.*

In Gaussian regression, a whole-neuron approximation attaining J_B has limiting mean

$$m_B(h) = K_{hH}(K_{HH} + \Sigma)^{-1} y, \tag{8}$$

the Gaussian-process posterior mean (Rasmussen & Williams, 2006). Its limiting output distribution retains covariance K, whereas the Gaussian-process posterior subtracts $K_{hH}(K_{HH} + \Sigma)^{-1} K_{Hg}$ from $K(h, g)$. Thus attaining J_B corrects the mean without reproducing posterior contraction.

Approximating the full weight posterior. The limiting KL from this approximation to the exact posterior is $\Delta = J_B + \log \mathbb{E} e^{-V(\Gamma_H)}$. Every sequence of whole-neuron mixtures with $\log L = o(N)$ has $\liminf_N D_{\mathrm{KL}}(Q_N \| P_N) \ge \Delta$. The residual Gaussian fluctuation prevents posterior contraction. A zero gap would make the limiting posterior a mean-shifted prior Gaussian. Then V would be affine on the Gaussian support; because $V \ge 0$, it would have to be constant. Thus a nonconstant likelihood leaves a positive gap (Appendix C).

Theorem 3 (Component count for posterior approximation)**.** *If* V *is nonconstant on the support of* Γ_H*, then* $\Delta > 0$*. For every fixed* $0 < \epsilon < \Delta$*, the smallest* L *for which some* $Q \in \mathcal{S}_{N,L}$ *satisfies* $D_{\mathrm{KL}}(Q \| P_N) \le \epsilon$ *obeys* $\log L_{\min}(N, \epsilon) = \Theta(N)$*. For every fixed* $\epsilon > 0$*,* $\log L_{\min}(N, \epsilon) = O(N)$*. The same component-count statements hold when* $\mathcal{S}_{N,L}$ *is replaced by mixtures of at most* L *whole-neuron products.*

A fixed partition of each neuron's weight space into M rectangular cells gives at most M^N scalar-factorized components. Quadratic growth of V controls the expected negative-log-likelihood error uniformly in width (Appendices A–C).

For Gaussian regression, the posterior-KL floor depends on the inputs and noise, but not on the labels (Appendix C). The opening two-input example has mean-learning gap $J_S - J_B = 1$ and posterior floor $\Delta = 1 - \log 2$. Zero labels remove only the mean gap, so a benign mean-learning problem can still require the larger posterior component count. The posterior component-count result extends to nonconvex likelihoods under the same growth bound (Appendix C.3); the mean results use convexity.

4 Gaussian correlations and nonlinear neuron dependence

A *full Gaussian* permits correlations across all weights while imposing a jointly Gaussian distribution. Neither it nor the whole-neuron family contains the other. Their intersection is the *Gaussian-per-neuron* family, with one Gaussian factor per neuron.

For ReLU at bounded prior KL, limiting Gaussian-per-neuron means span only the radial feature $\psi(h)$ and linear features $h/2$. Outgoing mean shifts produce the first; outgoing–incoming covariances produce the second. In Gaussian regression, the limiting optimum is ridge regression on these features with residual covariance K, including for nonidentical neurons (Appendix D). Arbitrary whole-neuron factors can supply additional mean directions.

The full-Gaussian output distribution. Factorization can restrict predictive uncertainty (Foong et al., 2020), and depth can change expressiveness (Farquhar et al., 2020). Here replacing a full Gaussian by the product of its neuron marginals preserves every output mean and cannot increase prior KL. Thus correlations across neurons do not enlarge the limiting mean space $\mathrm{span}\{\psi(h), h\}$. Asymptotically, they can change variance only in the collective outgoing direction $\psi(h)N^{-1/2}\sum_i a_i$; the remaining nonlinear residual approaches its prior distribution.

For any full Gaussian Q_N over the ReLU weights with uniformly bounded $D_{\mathrm{KL}}(Q_N \| U_N)$, set $s_N = \mathrm{Var}_{Q_N}(N^{-1/2}\sum_i a_i)$. The induced output distribution $Q_N^{F_N(A)}$ on a fixed query set A satisfies

$$Q_N^{F_N(A)} \approx_{W_2} \mathcal{N}\big(\mathbb{E}_{Q_N} F_N(A),\, K_{AA} + (s_N - 1)\psi_A \psi_A^T\big)\,. \tag{9}$$

Here $\approx_{W_2}$ means the 2-Wasserstein distance tends to zero, uniformly on each fixed prior-KL ball (Proposition 8, Appendix D.3).

In the introduction's two-input example, the odd output contrast $\{F_N(e_1) - F_N(-e_1)\}/\sqrt{2}$ is orthogonal to ψ. It retains prior variance while the posterior contracts it (Appendix D.3). Even a Gaussian predictive distribution can therefore require non-Gaussian dependence between weights.

Mean flexibility competes with covariance contraction. The exact regression comparison in Appendix D.4 decomposes the full-Gaussian optimum minus J_B into a nonnegative *mean-restriction cost* minus a nonnegative *covariance-contraction gain*. Only the first term depends on the labels; the second depends only on inputs and noise.

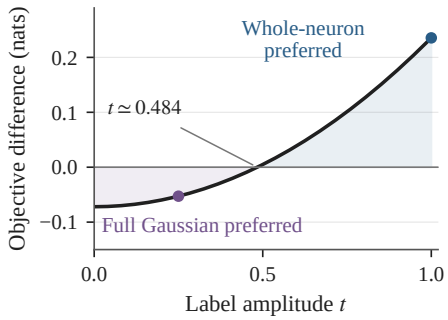
Consider inputs $(e_1, -e_1, e_2, -e_2)$ with noise covariance I_4 and labels $y = t(1, 1, -1, -1)$ of varying amplitude t. These labels are orthogonal to the Gaussian mean features: they are even on opposite inputs and sum to zero. Figure 2 shows the resulting reversal. Both family optima have attaining constructions (Appendix D.4); the fitted comparison below uses the smaller Gaussian-per-neuron family.

5 Predictive consequences of missing dependence

Missing covariance contraction already affects marginal prediction. In Gaussian regression with fixed positive noise and a fixed query distribution, whole-neuron mixtures with bounded prior KL and $\log L = o(N)$ incur positive asymptotic excess marginal log loss averaged under the exact posterior predictive whenever the data reduce variance on a set of positive query probability (Appendix E).

5.1 Correct marginals, incorrect updates

We now condition a fixed joint predictive distribution without refitting. Consider a bias-free ReLU network after observing label zero at $u = (e_1 + e_2)/\sqrt{2}$ with noise variance $\tau = .1$. We predict



Weight family	Output limit
Scalar product	Mean along ψ; prior fluctuations
Whole-neuron product	Posterior mean; prior fluctuations
Full Gaussian	Mean in $\mathrm{span}\{\psi, h\}$; only ψ can contract

Figure 2: **The better variational family changes with the labels.** Left: analytic full-Gaussian minimum minus the whole-neuron minimum, with inputs and noise fixed. The covariance benefit is constant; the mean-restriction cost grows as t^2, reversing the ordering near .484. Markers show $t = .25, 1$. Right: the distinct mean and fluctuation restrictions at the limiting optima for ReLU regression with Gaussian noise (Appendices D.4–D.5).

$Y_j = F_N(e_j) + \varepsilon_j$, $j = 1, 2$, using shared weights and independent noise of variance $\gamma = .1$. The exact limiting predictive distribution P is centered Gaussian, with marginal variance v and covariance $p < 0$: the observation reduces variance along $(1, 1)$ more than along $(1, -1)$. Its conditional mean is $\mathbb{E}_P[Y_2 \mid Y_1] = (p/v)Y_1$: a positive first response lowers the forecast of the second. This predictor has squared risk $R_{\text{exact}} = v - p^2/v$.

The ELBO-optimal full Gaussian weight approximation induces a different limiting predictive distribution. Scaling each already-noisy response using the known posterior variance v makes both predictive marginals exactly correct, yet the calibrated distribution still moves the conditional mean in the wrong direction and has larger squared error than ignoring the first response (Figure 3). Appendix G gives the conditional log-loss comparison and two controls: the calibrated prior and a weight distribution with reduced outgoing-weight variance.

For scalar mixtures, let $\mathcal{I}_N^S(Q)$ be the infimum of $I(J; \theta_{1:N})$ over representations with all scalar weights independent given the component label J. It is at most $\log L$ for an L-component mixture.

At bounded prior KL, scalar mixtures with $\mathcal{I}_N^S = o(\sqrt{N})$ have limiting outputs $\Gamma_A + \psi_A M$, where M is independent of Γ_A and has finite second moment (Appendix F). At the equal-norm inputs e_1 and e_2, ψ has the same value, so the random shift moves both outputs together. Full Gaussians can contract that shared part, but neither family yields a decreasing conditional forecast after marginal calibration.

Proposition 4 (Marginal calibration and conditional risk)**.** *Let Q_N have bounded $D_{\mathrm{KL}}(Q_N \| U_N)$. Suppose either $\mathcal{I}_N^S(Q_N) = o(\sqrt{N})$ or each Q_N is jointly Gaussian over all network weights. Take any joint limit of the two responses and apply nondecreasing maps to its two coordinates so that both marginals are exactly $\mathcal{N}(0, v)$. Denote the transformed law by $\widetilde{Q}$ and set $g(z) = \mathbb{E}_{\widetilde{Q}}[Y_2 \mid Y_1 = z]$. Then g is nondecreasing, and its risk under the exact limiting predictive law P satisfies*

$$\mathbb{E}_P\{Y_2 - g(Y_1)\}^2 \geq v = R_{\text{exact}} + p^2/v.$$

The independent calibrated marginals, which ignore Y_1, attain risk v. In the Gaussian case, $g(z) = \kappa z$ for $\kappa > 0$, so the inequality is strict.

The proposition concerns limiting noisy predictions. It does not assert finite-width conditional convergence or a weight distribution realizing the calibrated outputs (Appendix G).

5.2 A CONDITIONAL READOUT PRESERVES THE UPDATE

The approximation of Hron et al. (2022) keeps the exact conditional posterior of the outgoing (readout) weights a given the incoming (hidden-layer) weights b. In Gaussian regression, the latent outputs at the observed inputs are Ψa, where $\Psi_{li} = \varphi(b_i^T h_l)/\sqrt{N}$. This conditional posterior is

$$a \mid b, y \sim \mathcal{N}\big(\Psi^T(\Sigma + \Psi\Psi^T)^{-1} y,\, I - \Psi^T(\Sigma + \Psi\Psi^T)^{-1}\Psi\big)\,. \tag{10}$$

The covariance correction has rank at most the number of observations, but couples readout weights across neurons and depends on incoming weights. Drawing the hidden-layer weights from their prior and using this conditional posterior approaches the full weight posterior under the fixed-data empirical-kernel assumptions of Hron et al. (2022). The hidden-layer marginal can remain close to its prior while the conditional readout represents the dependence needed for the update.

Figure 3 compares covariance directions at equal prior KL. The feature-dependent construction applies the optimal full Gaussian's contraction strength along the hidden features of the observed input. The exact conditional readout (10) uses more prior KL (Appendix G.5).

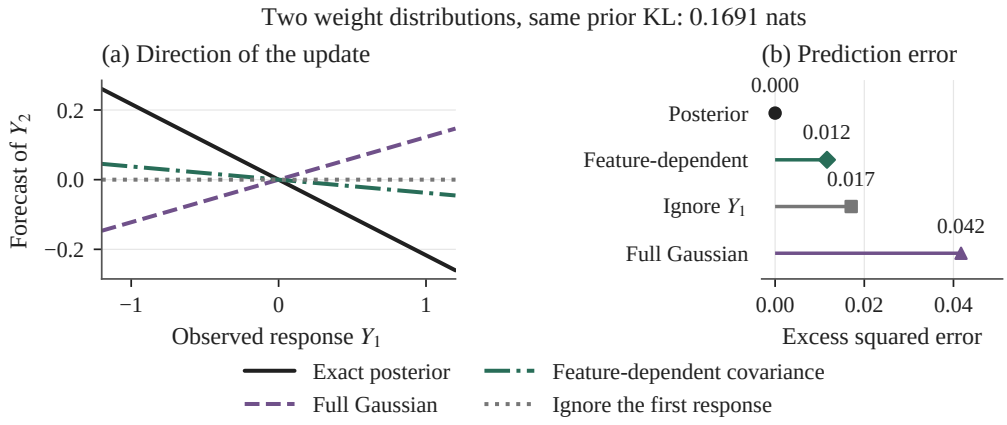


Figure 3: **The covariance direction matters at the same prior KL.** Width limits after calibrating already-noisy responses to the same marginals. The full Gaussian and feature-dependent covariance have prior KL .1691 before calibration and zero mean functions at every finite width. (a) Conditional forecasts. (b) Excess squared error under P relative to its exact predictor. The full Gaussian shown has the lowest calibrated limiting risk among full Gaussians under that prior-KL bound, yet does worse than ignoring Y_1. A feasible feature-dependent covariance improves on ignoring Y_1.

The conditional readout (10) recovers the exact limiting risk.

6 Mean learning and conditional prediction at finite width

We fit ReLU networks with $h = (x, 1)$, a Gaussian hidden bias and no separate output bias, on two two-point regressions adapted from Coker et al. (2022). The symmetric data have opposite labels at $x = \pm 1$ and noise variance .025; the radial-label data have positive labels proportional to ψ, which scalar-factorized means can fit (Appendix J).

At widths $32, 256, 2048, 16384$, a fully factorized Gaussian and a Gaussian-per-neuron product each start at the prior or a shared random diagonal Gaussian, giving 32 primary fits. Analytic Gaussian moments remove Monte Carlo objective noise, but optimization remains nonconvex (Appendix J).

A correct mean can coexist with predictive error. On the symmetric dataset at width 16,384, both starts give five-query average predictive KLs near 1.186 nats for the fully factorized Gaussian and $.764$ for Gaussian-per-neuron, with respective limiting losses 1.187 and $.769$ (Figure 4a; Appendix K.2). At $x = 0$, symmetry makes the posterior mean zero at every width. The widest prior-start fits match it but retain near-prior variance, with predictive KL about $.419$ nats. A finite-width variance bound for whole-neuron products requires at least 93.7% of prior latent variance at the prior-KL costs of the four symmetric fits at width 16,384 (Appendix J). At width 32, random starts nearly tie; scalar prior starts at widths 32 and 256 have verified negative curvature.

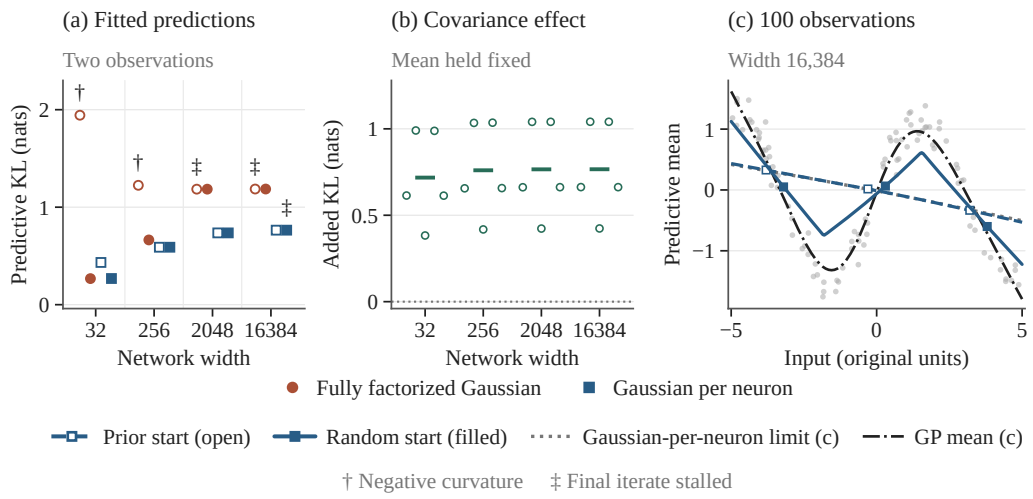


Figure 4: **Finite fits separate mean learning, covariance effects and initialization.** (a) Forward KL estimates from the finite posterior predictive to all sixteen fitted predictives on the symmetric dataset, averaged over five queries. Points use the lowest-objective accepted iterates; † marks verified negative curvature and ‡ final-iterate stalling. (b) Added predictive KL from restoring prior covariance in the conditional readout, with its mean fixed. Points show five queries $(-2, -1, 0, 1, 2)$ at four widths, with integration jackknife SEs; bars show query averages (SE at most .0011 nats). (c) Gaussian-per-neuron fits to 100 observations (gray) at width 16,384. Dotted: this family's limiting mean; dash-dotted: Gaussian-process (GP) posterior mean. Responses and predictions are standardized.

Replacing the conditional covariance in (10) by the identity preserves the entire finite-width mean function. This restoration of prior covariance (Figure 4b) increases marginal predictive loss in all 20 symmetric query/width comparisons, by .72–.77 nats averaged over queries. At width 16,384, the .766-nat increase approaches its .7663-nat limit. Holding the mean exactly fixed isolates this covariance effect; it need not equal the finite posterior mean.

On the radial-label data, both widest scalar fits follow the predicted radial mean; Gaussian-per-neuron fits leave that direction and attain lower objectives. Of the 32 fits, 15 have stalled final iterates (Appendix J and its accompanying records). These $L = 1$ comparisons do not test mixture rates.

Finite width and initialization. On a fixed 100-observation sine regression adapted from Coker et al. (2022), the widest prior-start Gaussian-per-neuron fit approaches its limiting mean, with neurons tied by deterministic fitting. The random-start fit has a lower objective, and its mean lies outside the limiting span (Figure 4c). Both Gaussian-per-neuron fits have higher objective than a fully factorized member of their family. Only the prior start meets the prespecified limiting-mean agreement criteria; the width needed for both initializations remains unresolved. Appendix K.5 bounds the prior-start error and attributes most of the random fit's departure to the two highest-KL neurons' contributions beyond the first-order mean expansion.

Binary decisions without marginal calibration. In the two-query example of Section 5.1, replace the Gaussian observation by one positive and one negative label at u, with success probability $\mathrm{sigmoid}(\beta F_N(u))$. We reuse the two distributions of Figure 3 without refitting; they and the posterior assign probability $1/2$ to each future label. At $\beta = 4$, the feature-dependent approximation gives the Bayes decision at the same prior KL: classification error falls by 2.81 and 3.49 percentage points at widths 32 and 256 relative to ignoring the first label. The full Gaussian gives the opposite decision at $\beta = 4$ and the Bayes decision at $\beta = 1$. Training and future labels use the same β (Appendix G.7).

A separate width-limit result concerns variational optima for this binary likelihood: for sufficiently large fixed β, every near-optimal sequence of mixtures with a fixed number of full-Gaussian components retains the wrong conditional update (Appendix G.8).

7 EVIDENCE-BASED PRIOR SELECTION

The ELBO subtracts posterior KL from log evidence. This approximation penalty can favor priors whose posteriors the family represents more closely (Turner & Sahani, 2011). We study this effect in separate logistic preference heads on fixed language-model features (Melo et al., 2024), using three 1,024-comparison subsets, A, B and C, of Reddit TL;DR summary comparisons (Stiennon et al., 2020).

We compare isotropic Gaussian priors with standard deviations 1 and .3. Bounds on the model evidence favor scale one on each subset, yet every scale-one product mixture with finite ELBO and up to a million components has lower ELBO than a feasible scale-.3 product (Figure 5). Factors may be arbitrary in the original coordinates. Mixing L products improves the best product ELBO by at most $\eta_1 \log L$, where likelihood curvature determines $\eta_1 < 1$ at scale one (Appendix L.1). We also bound the fitted Gaussian product's gap to the best arbitrary product (Appendix L.2).

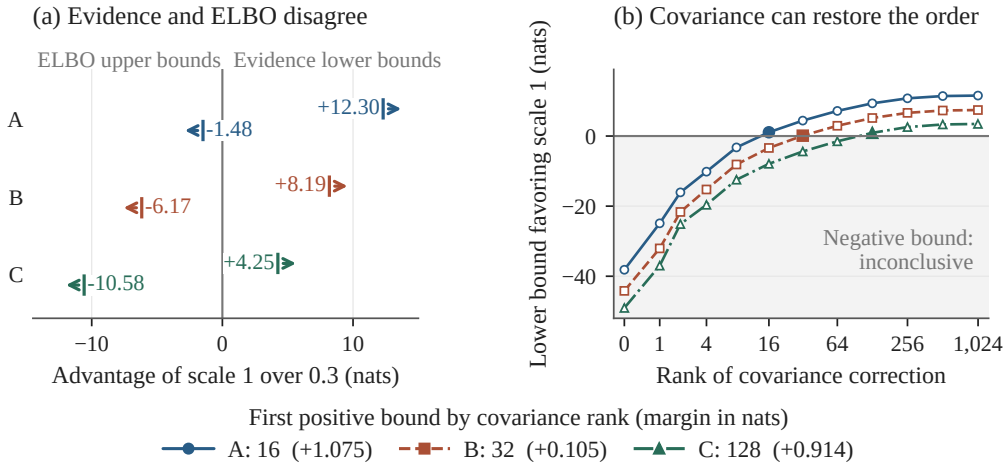


Figure 5: **Dependence restrictions can reverse the preferred prior.** (a) One-sided bounds on the advantage of scale 1 over .3. Right-pointing arrows give lower bounds on the log-evidence difference; left-pointing arrows give upper bounds on the ELBO difference between any eligible scale-1 product mixture with $L \leq 10^6$ and a feasible scale-.3 product. (b) Covariance corrections from the full Laplace fit restore the evidence ordering; filled markers give the smallest tested ranks with a positive bound. Other tested ranks are hollow; negative bounds are inconclusive. The rank axis uses $\log_2(\text{rank} + 1)$ spacing. Numerical evaluations have no rigorous error enclosures.

These ranks need not be minimal; we truncate full Laplace covariances. A Gaussian diagonal in the observed feature row-space basis also restores all three orderings (Appendix L.3).

On 1,000 previously explored development prompts, scale-.3 diagonal fits also predict worse: mean loss .6629 versus .6366 nats across the subsets. At scale one, diagonal, row-space-diagonal and full-Gaussian mean losses differ by at most .00011 nats; MAP also favors scale one. These fitted comparisons do not bound the product family's predictive risk (Appendix L.5).

8 CONCLUSION

Mean learning, uncertainty contraction and conditional prediction can require different forms of dependence. The mixture hierarchy quantifies the first two; the family comparison shows why Gaussian correlations alone need not suffice. Correct means or marginal forecasts can coexist with incorrect joint predictions. Dependence restrictions can also reverse evidence-based prior selection. Finite-width onset, growing datasets and greater depth remain open.

AI USE STATEMENT

We developed this work with assistance from generative AI tools, including OpenAI's GPT models. In iterative discussions, we raised conceptual questions, explored connections to the literature, suggested experiments and critically examined the emerging ideas. The tools helped formulate theoretical models, hypotheses and mathematical claims, and supplied derivations and proof arguments, including complete arguments for some results. This included developing the connection to channel resolvability used to establish the sharp mixture-size bound in Theorem 2. We also used AI tools to find and understand related theoretical work and to design and implement experiments, including code for data preprocessing and synthetic-data generation. They assisted with analysis and interpretation of results, figure preparation, and refining the manuscript.

We reviewed the proofs, worked through additional derivations and checked relevant claims against primary sources. We also recomputed analytic examples and checked the consistency of the numerical results. AI tools assisted with these checks. We retain responsibility for the final content and its verification.

REPRODUCIBILITY STATEMENT

The appendices provide the proofs and experiment protocols. The anonymous computational supplement includes setup instructions, tested environments, a result-to-command map and scripts with fixed inputs for all five main figures. Its CPU commands have been tested from a relocated archive extraction. The supplementary code recomputes the preference study's product-approximation bounds, evidence comparisons and covariance-rank results from saved states and feature rows (`supplement/preference_replay/`). Laplace covariance factors are recomputed from the saved modes and designs. This is numerical replay, without refitting or interval certification. The neural saved-state package (`supplement/neural_predictive_replay/`) recomputes objectives and analytic moments for all 32 primary fits and 32 refinements, and reaggregates the saved predictive integrals and sample-block statistics. It does not repeat fitting, sampling or predictive-density integration. A third package (`supplement/many_observation_replay/`) replays all 24 named endpoints from the 12-fit, 100-observation study, including objectives, analytic gradients, moments and KL concentration. Reference coefficients and GP predictions are supplied inputs. Each package gives its tested environment and calculation scope. The additional `supplement/conditional_updates_replay/` reconstructs finite conditional-prediction summaries from saved integration statistics and preference scores from saved probabilities. These replays do not repeat fitting, sampling or integration.

REFERENCES

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-Oriented Bayesian Active Learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7331–7348. PMLR, 2023. URL https://proceedings.mlr.press/v206/bickfordsmith23a.html.

Freddie Bickford Smith, Adam Foster, and Tom Rainforth. Making Better Use of Unlabelled Data in Bayesian Active Learning. In *International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 847–855. PMLR, 2024. URL https://proceedings.mlr.press/v238/bickford-smith24a.html.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://arxiv.org/abs/1601.00670.

Matthieu R. Bloch. Covert communication over noisy channels: A resolvability perspective. *IEEE Transactions on Information Theory*, 62(5):2334–2354, 2016. doi: 10.1109/TIT.2016.2530089. URL https://arxiv.org/abs/1503.08778v3.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*,

volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622. PMLR, 2015. URL https://proceedings.mlr.press/v37/blundell15.html.

Alberto Bordino, Stefano Favaro, and Sandra Fortini. Non-asymptotic approximations of Gaussian neural networks via second-order Poincaré inequalities. In *Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference*, volume 253 of *Proceedings of Machine Learning Research*, pp. 45–78. PMLR, 2024. URL https://proceedings.mlr.press/v253/bordino24a.html.

Sourav Chatterjee. Fluctuations of eigenvalues and second order Poincaré inequalities. *Probability Theory and Related Fields*, 143:1–40, 2009. doi: 10.1007/s00440-007-0118-6. URL https://arxiv.org/abs/0705.1224v3.

Youngmin Cho and Lawrence K. Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009. URL https://papers.nips.cc/paper/2009/file/5751ec3e9a4feab575962e78e006250d-Paper.pdf.

Beau Coker, Wessel P. Bruinsma, David R. Burt, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field Bayesian neural networks ignore the data. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5276–5333. PMLR, 2022. URL https://proceedings.mlr.press/v151/coker22a.html.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981. doi: 10.1214/aos/1176345462. URL https://doi.org/10.1214/aos/1176345462.

Sebastian Farquhar, Lewis Smith, and Yarin Gal. Liberty or depth: Deep Bayesian neural nets do not need complex weight posterior approximations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4346–4357, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2dfe1946b3003933b7f8ddd71f24dbb1-Abstract.html.

Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. On the expressiveness of approximate inference in Bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15897–15908. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b6dfd41875bc090bd31d0b1740eb5b1b-Paper.pdf.

Yoav Gelberg, Tycho F. A. van der Ouderaa, Mark van der Wilk, and Yarin Gal. Variational inference failures under model symmetries: Permutation invariant posteriors for Bayesian neural networks. In *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pp. 233–248. PMLR, 2024. URL https://proceedings.mlr.press/v251/gelberg24a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Jiri Hron, Roman Novak, Jeffrey Pennington, and Jascha Sohl-Dickstein. Wide Bayesian neural networks have a simple weight posterior: theory and accelerated sampling. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8926–8945. PMLR, 2022. URL https://proceedings.mlr.press/v162/hron22a.html.

Tom Huix, Szymon Majewski, Alain Durmus, Eric Moulines, and Anna Korba. Variational inference of overparameterized Bayesian neural networks: A theoretical and empirical study, 2022. URL https://arxiv.org/abs/2207.03859.

Tommi S. Jaakkola and Michael I. Jordan. Improving the mean field approximation via the use of mixture distributions. In Michael I. Jordan (ed.), *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, pp. 163–173. Springer, Dordrecht, 1998. doi: 10.1007/978-94-011-5014-9_6. URL https://link.springer.com/chapter/10.1007/978-94-011-5014-9_6.

Richard Kurle, Ralf Herbrich, Tim Januschowski, Yuyang Wang, and Jan Gasthaus. On the detrimental effect of invariances in the likelihood for variational inference. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4531–4542, 2022. doi: 10.52202/068431-0327. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1d18bc97cd1a3c5f0d9d1d382cd1ce91-Abstract-Conference.html.

Daniel Lacker. Independent projections of diffusions: Gradient flows for variational inference and optimal mean field approximations. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 62(1):638–666, 2026. doi: 10.1214/24-AIHP1525. URL https://arxiv.org/abs/2309.13332v2.

Daniel Lacker, Sumit Mukherjee, and Lane Chun Yeung. Mean field approximations via log-concavity. *International Mathematics Research Notices*, 2024(7):6008–6042, 2024. doi: 10.1093/imrn/rnad302. URL https://doi.org/10.1093/imrn/rnad302.

Hugo Lavenant and Giacomo Zanella. Convergence rate of random scan coordinate ascent variational inference under log-concavity. *SIAM Journal on Optimization*, 34(4):3750–3761, 2024. doi: 10.1137/24M1670627. URL https://arxiv.org/abs/2406.07292v2.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.00165.

Rui Li, S. T. John, and Arno Solin. Improving hyperparameter learning under approximate inference in Gaussian process models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19595–19615. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23m.html.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989. doi: 10.1007/bf01589116. URL https://doi.org/10.1007/BF01589116.

Wei Liu, Ge Xu, and Biao Chen. The common information of N dependent random variables. In *48th Annual Allerton Conference on Communication, Control, and Computing*, pp. 836–843, 2010. doi: 10.1109/ALLERTON.2010.5706995. URL https://arxiv.org/abs/1010.3613.

luckeciano. reddit-features-hermes. Hugging Face dataset, 2024. URL https://huggingface.co/datasets/luckeciano/reddit-features-hermes.

David J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448. URL https://doi.org/10.1162/neco.1992.4.3.448.

Luckeciano C. Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. Deep Bayesian active learning for preference modeling in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 118052–118085, 2024. doi: 10.52202/079017-3749. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d5e256c988bdee59a0f4d7a9bc1dd6d9-Paper-Conference.pdf.

Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2593–2602. PMLR, 2019. URL https://proceedings.mlr.press/v89/molchanov19a.html.

Sebastian W. Ober and Laurence Aitchison. Global inducing point variational posteriors for Bayesian neural networks and deep Gaussian processes. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8248–8259. PMLR, 2021. URL https://proceedings.mlr.press/v139/ober21a.html.

James Odgers, Ben Riegler, Siddharth Swaroop, and Vincent Fortuin. Gaussian mean field variational inference can overestimate predictive variance. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.25745v1.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Brendan O'Donoghue, Botao Hao, Dieterich Lawson, Morteza Ibrahimi, Xiuyuan Lu, Vikranth Dwaracherla, and Benjamin Van Roy. The Neural Testbed: Evaluating Joint Predictions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12554–12565, 2022. doi: 10.52202/068431-0912. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5141f6bc105d30edbae48f1d2e0b1e66-Abstract.html.

Gábor Pituk, Vik Shirvaikar, and Tom Rainforth. Do Bayesian neural networks actually behave like Bayesian models? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49420–49458. PMLR, 2025. URL https://proceedings.mlr.press/v267/pituk25a.html.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 324–333. PMLR, 2016. URL https://proceedings.mlr.press/v48/ranganath16.html.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. URL https://gaussianprocess.org/gpml/.

Tommy Rochussen and Vincent Fortuin. Amortising inference and meta-learning priors in neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cf8c2436c68ab2511ce9103ceb07a492-Abstract-Conference.html.

Artem Sobolev and Dmitry P. Vetrov. Importance weighted hierarchical variational inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/5737c6ec2e0716f3d8a7a5c4e0de0d9a-Abstract.html.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Teknium. OpenHermes 2.5 Mistral 7B, 2023. URL https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B. Model card.

Richard Eric Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time series models. In David Barber, A. Taylan Cemgil, and Silvia Chiappa (eds.), *Bayesian Time Series Models*, chapter 5, pp. 104–124. Cambridge University Press, 2011. doi: 10.1017/CBO9780511984679.006. URL https://doi.org/10.1017/CBO9780511984679.006.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu (eds.), *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://aclanthology.org/W17-4508/.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5660–5669. PMLR, 2018. URL https://proceedings.mlr.press/v80/yin18b.html.

Lei Yu and Vincent Y. F. Tan. Rényi resolvability and its applications to the wiretap channel. *IEEE Transactions on Information Theory*, 65(3):1862–1897, 2019. doi: 10.1109/TIT.2018.2885318. URL https://arxiv.org/abs/1707.00810v3.

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4):550–560, 1997. doi: 10.1145/279232.279236. URL https://doi.org/10.1145/279232.279236.

SUPPORTING MATERIAL

The appendices follow the main argument. The table locates the proofs and experiments that illustrate their consequences.

Table 1: Guide to the theoretical and empirical support.

Main result or comparison	Supporting material
Two mixture scales and the opening example	Appendices A–C
Gaussian families and their optimal ordering	Appendix D
Marginal and conditional predictive restrictions	Appendices E–G
Conditional readout and finite-width evidence	Appendices H–K, including K.5
Preference-head evidence and predictions	Appendix L

The posterior component-count extension beyond convex likelihoods is in Appendix C.3. For conditional prediction, Appendix G.5 compares covariance directions at the same prior KL; Appendices G.6 and G.7 give the finite calibrated forecasts and binary decisions. Appendix G.8 treats variational optima of fixed-size Gaussian mixtures. The finite variance bound used in the experiments is in Appendix I.1, followed by its semi-implicit specialization.

A GAUSSIAN-CONVOLUTION CLOSURE AND WHOLE-NEURON STABILITY

We first prove the whole-neuron result used by both component-count bounds. Let $Z = z(\theta) \in \mathbb{R}^r$ be centered with covariance I_r and $\mathbb{E}e^{\delta\|Z\|} < \infty$ for some $\delta > 0$. Set $T_N = N^{-1/2}\sum_i Z_i$ under the iid group prior U_N. For the network, restrict z_H to the actual covariance range and whiten. The bound $\|z_H\| \le C(1 + a^2 + \|b\|^2)$ gives the required exponential moment. It also holds jointly with finitely many query outputs and a. Rank zero is trivial and is excluded from strict-gap statements.

Lemma 5 (Closure with bounded KL to the prior)**.** *Suppose Q_N is a mixture of at most L_N products across whole neurons, $D_{\mathrm{KL}}(Q_N\|U_N) \le C_0$, and $\log L_N = o(N)$. Its aggregate laws are tight, and every weak subsequential limit of T_N is $\mathcal{N}(0, I_r) * H$ for some probability law H.*

Proof. For component j, let $d_j = D_{\mathrm{KL}}(Q_j\|U_N)$. The finite mixture chain rule gives

$$\sum_j \pi_j d_j = D_{\mathrm{KL}}(Q_N\|U_N) + I(J;\theta_{1:N}) \le C_0 + \log L_N = o(N). \tag{11}$$

Choose $\eta_N \downarrow 0$ slowly enough that $(C_0 + \log L_N)/(N\eta_N) \to 0$. Components with $d_j > N\eta_N$ have vanishing total mass. For a fixed clipping radius B, put $Z^B = Z\mathbf{1}_{\{\|Z\|\le B\}} - \mathbb{E}[Z\mathbf{1}_{\{\|Z\|\le B\}}]$, $T_N^B = N^{-1/2}\sum_i Z_i^B$ and $\Sigma_B = \mathrm{Cov}_U(Z^B)$. Pinsker, additivity of component KL across groups, and Cauchy–Schwarz give

$$\|\,\mathrm{Cov}_{Q_j}(T_N^B) - \Sigma_B\| \le C_{B,r}\sqrt{d_j/N} = o(1)$$

uniformly over retained components. In this estimate the squares of component means are bounded by the same per-group total-variation terms.

For a retained component, write $Y_{ij} = Z_i^B - \mathbb{E}_{Q_j} Z_i^B$ and $m_j = N^{-1/2}\sum_i \mathbb{E}_{Q_j} Z_i^B$. Let W_{ij} be independent centered Gaussians with the respective covariances of Y_{ij}, independent also of the Y_{ij}, and put $G_j = N^{-1/2}\sum_i W_{ij}$ and $A_j = \mathrm{Cov}(G_j)$. Since $\|Z^B\| \le 2B$ and $\|Y_{ij}\| \le 4B$, $\mathbb{E}\|Y_{ij}\|^3 + \mathbb{E}\|W_{ij}\|^3 \le C_{B,r}$. In the ith replacement, the common partial sum $S_{ij} = m_j + N^{-1/2}(\sum_{k<i} W_{kj} + \sum_{k>i} Y_{kj})$ is independent of both summands being compared. For a bounded C^3 test f with bounded derivatives, Taylor expansion about S_{ij} cancels their first two moments and leaves a remainder at most $\|D^3 f\|_\infty(\mathbb{E}\|Y_{ij}\|^3 + \mathbb{E}\|W_{ij}\|^3)/(6N^{3/2})$. Summing these bounds gives

$$\left|\mathbb{E}f\left(m_j + N^{-1/2}\sum_i Y_{ij}\right) - \mathbb{E}f(m_j + G_j)\right| \le C_{B,r}\|D^3 f\|_\infty N^{-1/2}.$$

Here and below derivative norms are uniform operator norms, so the bounds are independent of the translation m_j.

Let $G_B \sim \mathcal{N}(0, \Sigma_B)$ and $G_{j,t} \sim \mathcal{N}(0, (1-t)\Sigma_B + tA_j)$. Differentiating the Gaussian density and integrating twice by parts gives

$$\begin{aligned}&\mathbb{E}f(m_j + G_j) - \mathbb{E}f(m_j + G_B)\\ &\qquad = \frac{1}{2}\int_0^1 \mathbb{E}\,\mathrm{tr}\big[(A_j - \Sigma_B)D^2 f(m_j + G_{j,t})\big]\ \mathrm{d}t.\end{aligned}$$

The formula also holds for singular covariances by first adding κI_r to both endpoints and then taking $\kappa \downarrow 0$. The preceding covariance bound therefore makes its absolute value at most $C_{B,r}\|D^2 f\|_\infty \sqrt{\eta_N}$.

To pass uniformly to bounded Lipschitz tests, let $f_\delta = f * \rho_\delta$, where $\rho_\delta(x) = \delta^{-r}\rho(x/\delta)$ for a fixed smooth probability density ρ supported in the unit ball. For $\|f\|_\infty \le 1$ and $\mathrm{Lip}(f) \le 1$, $\|f_\delta - f\|_\infty \le \delta$ and $\|D^k f_\delta\|_\infty \le C_r \delta^{1-k}$ for $k = 2, 3$. Let w_N be the discarded component mass and let $H_{N,B}$ put mass π_j at each retained m_j and the discarded mass at zero. The two estimates and averaging now give

$$\begin{aligned}&d_{\mathrm{BL}}\big(\mathcal{L}_{Q_N}(T_N^B), \mathcal{N}(0,\Sigma_B) * H_{N,B}\big)\\ &\qquad \le 2w_N + 2\delta + C_{B,r}\left(\delta^{-2}N^{-1/2} + \delta^{-1}\sqrt{\eta_N}\right),\end{aligned}$$

where d_{BL} takes the supremum over the stated test class. For each fixed B, first take $N \to \infty$ and then $\delta \downarrow 0$. The clipping radius is sent to infinity only afterward, using the uniform estimate below.

The KL bound on the full mixture controls the error from removing clipping. Under the prior,

$$\sup_N \mathbb{E}_U\|T_N - T_N^B\|^2 \le \mathbb{E}\|Z\mathbf{1}_{\{\|Z\|>B\}}\|^2 \longrightarrow 0.$$

For every event A, binary entropy data processing gives

$$Q_N(A)\log\frac{1}{U_N(A)} \le C_0 + \log 2.$$

Indeed, set $q_B = \mathbb{E}\|Z\mathbf{1}_{\{\|Z\|>B\}}\|^2$. For fixed $\epsilon > 0$ and $0 < q_B < \epsilon^2$, Chebyshev and the entropy inequality give

$$\sup_N Q_N(\|T_N - T_N^B\| > \epsilon) \le \frac{C_0 + \log 2}{\log(\epsilon^2/q_B)} \longrightarrow 0.$$

The case $q_B = 0$ is immediate by absolute continuity. Thus clipping changes the aggregate law by a bounded-Lipschitz distance tending to zero uniformly in N. The same entropy inequality and the prior second-moment bounds make the laws of T_N and T_N^B tight: their prior second moments are bounded by r.

Fix a subsequence on which $T_N \Rightarrow \nu$. At a countable sequence $B_k \to \infty$, extract a common further subsequence with clipped limits $\nu_k = \mathcal{N}(0, \Sigma_{B_k}) * H_k$. The shift laws are tight whenever the corresponding convolutions are tight, since for independent G, M,

$$\Pr(\|M\| > 2R) \le \Pr(\|G + M\| > R) + \Pr(\|G\| > R).$$

The uniform clipping estimate gives $\nu_k \Rightarrow \nu$. The Gaussian covariances $\Sigma_{B_k} \to I_r$ are bounded, so the same inequality makes (H_k) tight. Along a further subsequence $H_k \Rightarrow H$, continuity of convolution yields $\nu = \mathcal{N}(0, I_r) * H$. This retains all mixing mass and proves the claimed closure. □

Information budgets and continuous labels. The same closure holds for a standard-Borel label with a measurable whole-neuron product kernel whenever $D_{\mathrm{KL}}(Q_N\|U_N) + I(J;\theta_{1:N}) = o(N)$ and global prior KL is bounded. Replace the finite sum in (11) by its component integral. Markov's inequality discards components with KL above $N\eta_N$ at vanishing mass; the clipped replacement estimates are uniform on the remainder, and the mixture's KL bound removes clipping as above. The stability argument below uses only this closure. The scalar mean bound likewise integrates when the information is $o(\sqrt{N})$.

Stability of the whole-neuron reference. We now use the closure result to identify the best mean shift and control objective excess relative to its realizing product distribution. Let $W \geq 0$ be convex, C^1, and at most quadratic on the whitened space. Write

$$J(\mu) = \tfrac{1}{2}\|\mu\|^2 + \mathbb{E}W(G+\mu), \qquad \mu_* = \arg\min J, \quad G \sim \mathcal{N}(0, I_r).$$

The objective is strongly convex and coercive. Convex quadratic growth implies $\|\nabla W(t)\| \leq C'(1+\|t\|)$: apply the supporting-line inequality in each signed direction at distance $1+\|t\|$. Thus $\mathbb{E}\nabla W(G+\mu_*) = -\mu_*$. The tilted energy $\widetilde{W}(t) = W(t) + \mu_*^T t$ has a finite global lower bound, because averaging the supporting tangent at $Y = G + \mu_*$ gives

$$\widetilde{W}(t) \geq \mathbb{E}[W(Y) - \nabla W(Y)^T Y] > -\infty.$$

Its Gaussian convolution is minimized at μ_*, so $\mathbb{E}_{\mathcal{N}(0,I_r)*H}\widetilde{W} \geq \mathbb{E}\widetilde{W}(G+\mu_*)$.

Define $M(t) = \mathbb{E}_U e^{t^T Z}$ and $R_N = e^{\mu_*^T T_N - N \log M(\mu_*/\sqrt{N})} U_N$. This is one whole-neuron product for large N. The local cumulant expansion $\log M(t) = \|t\|^2/2 + O(\|t\|^3)$ yields its mean shift, covariance, CLT and uniformly bounded fourth aggregate moments. Consequently $G_N(R_N) \to J(\mu_*)$. For any fixed vector s,

$$\log \mathbb{E}_{R_N} e^{s^T T_N} \longrightarrow s^T \mu_* + \tfrac{1}{2}\|s\|^2. \tag{12}$$

For a bounded-objective competitor all terms in the exact identity

$$G_N(Q_N) - G_N(R_N) = D_{\mathrm{KL}}(Q_N \| R_N) + \mathbb{E}_Q \widetilde{W}(T_N) - \mathbb{E}_R \widetilde{W}(T_N) \tag{13}$$

are finite. Indeed $W \geq 0$ bounds prior KL directly, the prior exponential moment bounds first moments by entropy duality, and finite objective bounds the energy. Lemma 5, Portmanteau for the bounded-below $\widetilde{W}$, and reference moment convergence show that the last difference has non-negative liminf. Therefore every such sequence satisfying $G_N(Q_N) \leq G_N(R_N) + o(1)$ has $D_{\mathrm{KL}}(Q_N \| R_N) \to 0$, and a fixed objective excess ϵ gives the corresponding limsup-KL bound. Approximate infima and admissibility of R_N prove the sharp value J_B for whole-neuron mixtures with $\log L_N = o(N)$.

Query means and bounded predictions. The joint cumulant expansion for observed and query outputs identifies the reference mean as $K_{AH}K_{HH}^{\dagger}v_*$ and residual kernel K_{AA}. Data processing and Pinsker control differences in expectations of bounded test functions of the query outputs. For a scalar query $F_N(h)$, entropy duality at each fixed $t > 0$ gives

$$\limsup(\mathbb{E}_Q F_N(h) - m_B(h)) \leq \epsilon/t + tK(h,h)/2.$$

Apply the same argument to the negative query and optimize t to obtain $\sqrt{2\epsilon K(h,h)}$. At $\epsilon = 0$, let $t \downarrow 0$.

These arguments do not transfer arbitrary second moments. In the normalized one-query ReLU network $T_N = \sqrt{2/N}\sum_i a_i(w_i)_+$ and a fixed logistic energy, define S_N by replacing the first neuron of R_N with independent $a_1, w_1 \sim \mathcal{N}(0, N^2)$, each with variance N^2. Every other neuron retains its original joint law, and neurons remain independent. Mix $Q_N = (1 - N^{-3})R_N + N^{-3}S_N$. Then $D_{\mathrm{KL}}(Q_N \| R_N) = O(N^{-1})$ and the Lipschitz energy difference is $O(N^{-3/2})$, but the additional variance is $N^{-3} \cdot N^4/N = 1$. The output variance tends to 2, while the reference variance tends to 1. The example lies in the larger whole-neuron class and shows that weight-space KL stability does not by itself imply convergence of second moments. The scalar-factorized mixture results do not use such an implication.

B Scalar confinement and square-root attainment

Weight-space stability in Theorems 1 and 2. If $\log L_N = o(\sqrt{N})$, let $S_{m_*,N}$ keep the incoming prior and take independent $a_i \sim \mathcal{N}(m_*/\sqrt{N}, 1)$. Every $Q_N \in \mathcal{S}_{N,L_N}$ with $G_N(Q_N) \leq J_S + o(1)$ satisfies $D_{\mathrm{KL}}(Q_N \| S_{m_*,N}) \to 0$. If $\sqrt{N} \ll \log L_N \ll N$, every sequence within $o(1)$ of the variational infimum converges in weight-space KL to

$$R_N \propto \exp\Big(v_*^T K_{HH}^{\dagger} F_N(H)\Big)\, U_N, \tag{14}$$

which exists for sufficiently large N. On each fixed finite query set A, the respective weak output limits are $\mathcal{N}(m_*\psi_A, K_{AA})$ and $\mathcal{N}(K_{AH}K_{HH}^{\dagger}v_*, K_{AA})$; the output means also converge. These statements concern arbitrary factors and do not assert convergence of their second moments.

B.1 Proof of Theorem 1

For one component and one neuron, write d_a, d_b for outgoing and incoming prior KL. Gaussian entropy duality gives $|\mathbb{E}a| \le \sqrt{2d_a}$. Gaussian concentration of the Lipschitz function $\varphi(b^T h)$ gives

$$|\mathbb{E}\varphi(b^T h) - \psi(h)| \le \mathrm{Lip}(\varphi)\|h\|\sqrt{2d_b}.$$

Independence of outgoing and incoming variables and $2\sqrt{d_a d_b} \le d_a + d_b$ bound the component's deviation from $\psi(h)\mathbb{E}A_N$ by its total KL times $\mathrm{Lip}(\varphi)\|h\|/\sqrt{N}$. Averaging and using (11) proves (5). Products over the original scalar weights satisfy this independence condition; dependence among incoming weights does not change it.

Apply Lemma 5 to the augmented observable (a, z_H). Bounded global KL gives first-moment uniform integrability. One direct justification uses entropy duality for truncated absolute linear functionals: prior aggregate exponential moments are eventually bounded for every fixed coefficient, so the tail bound can be made small by taking that coefficient large after the width limit. Along a subsequence with $\mathbb{E}_Q A_N \to m$, (5) gives $\mathbb{E}_Q F_N(H) \to m\psi_H$. Since A_N is exactly standard normal under the prior, $D_{\mathrm{KL}}(Q\|U_N) \ge (\mathbb{E}_Q A_N)^2/2$. The convolution closure, Portmanteau and Jensen then give

$$\liminf G_N(Q_N) \ge \tfrac{1}{2}m^2 + \mathbb{E}V(\Gamma_H + m\psi_H) \ge J_S.$$

The outgoing tilt $S_{m,N}$ has exactly $D_{\mathrm{KL}}(S_{m,N}\|U_N) = m^2/2$. Writing $a_i = a_i^0 + m/\sqrt{N}$ yields a prior Gaussian CLT plus $mN^{-1}\sum_i \varphi(b_i^T h) \to m\psi(h)$. Uniform fourth moments give convergence of the expected negative log likelihood, establishing attainment of J_S. For stability use the exact identity

$$G_N(Q) - G_N(S_{m_*,N}) = D_{\mathrm{KL}}(Q\|S_{m_*,N}) + \mathbb{E}_Q[V(F_N(H)) + m_* A_N] - \mathbb{E}_S[V(F_N(H)) + m_* A_N].$$

With $\ell(m) = \mathbb{E}V(\Gamma_H + m\psi_H)$, scalar optimality gives $-m_* \in \partial\ell(m_*)$ and $\ell(m) + m_* m \ge \ell(m_*) + m_*^2$. The augmented closure and first-moment transfer therefore make the last difference nonnegative in liminf. This proves weight-space KL convergence and the fixed-excess version. The query cumulant argument from the preceding appendix gives weak convergence of the output distribution and convergence of its means. It does not require $V(F) + m_* A$ to be globally bounded below on the unrestricted augmented space.

For Gaussian regression, the two minimizers specialize to $m_* = (\psi_H^T \Sigma^{-1} y)/(1 + \psi_H^T \Sigma^{-1}\psi_H)$ and $v_* = K_{HH}(K_{HH} + \Sigma)^{-1} y$.

B.2 A bounded construction using rectangular weight cells

Let $g_*(\theta) = v_*^T K_{HH}^{\dagger} z_H(\theta)$. Then $\mathbb{E}g_* = 0$, $\mathbb{E}g_*^2 = v_*^T K_{HH}^{\dagger} v_*$ and $\mathbb{E}[z_H g_*] = v_*$. Finite rectangular partitions of the original weight coordinates generate their Borel sigma field. Their conditional expectations $g = \mathbb{E}[g_* \mid \mathcal{B}]$ approximate g_* in L^2. Each such score is bounded and centered. Hence

$$\tfrac{1}{2}\mathbb{E}g^2 + \mathbb{E}V(\Gamma_H + \mathbb{E}[z_H g]) \longrightarrow J_B$$

as the partition is refined. The energy limit follows from deterministic mean convergence and the Gaussian quadratic-growth bound.

Let the positive-probability cells be B_k, with probabilities u_k and score values g_k. Choose $t > \max(0, -\min_k g_k)$ and set $w_k = u_k(1 + g_k/t)$. These are probability weights. At sufficiently large width,

$$q_N = (1 - t/\sqrt{N})U_1 + \frac{t}{\sqrt{N}}\sum_k w_k U_1(\cdot \mid B_k) = (1 + g/\sqrt{N})U_1, \qquad S_N = q_N^{\otimes N}.$$

Conditioning on a rectangular cell preserves independence of the scalar weights. Boundedness of g gives

$$D_{\mathrm{KL}}(S_N\|U_N) \to \tfrac{1}{2}\mathbb{E}g^2, \quad \mathbb{E}_{S_N}F_N(H) = \mathbb{E}[z_H g], \quad G_N(S_N) \to \tfrac{1}{2}\mathbb{E}g^2 + \mathbb{E}V(\Gamma_H + \mathbb{E}[z_H g]).$$

The last step uses the constructed triangular-array CLT and uniform fourth moments, not a moment assertion for arbitrary competitors.

B.3 A FINITE-MIXTURE CONSTRUCTION FROM CHANNEL RESOLVABILITY

We view each neuron as one use of a channel whose input is a label and whose output is the neuron's weights. Label 0, with probability $1-t/\sqrt{N}$, outputs a prior draw; label k, with probability $tw_k/\sqrt{N}$, outputs a draw from its cell conditional $U_1(\cdot \mid B_k)$. Given the output cell, the conditional distribution of the neuron weights is independent of the label. The channel therefore reduces without loss to a finite cell-output channel. For $0 < \rho \le 1$, its information-density moment is exactly

$$A_N(\rho) = \sum_k u_k(1+g_k/\sqrt{N})^{-\rho}\left[1 - t/\sqrt{N} + (t/\sqrt{N})w_k u_k^{-1-\rho}\right] = 1 + \frac{B_\rho}{\sqrt{N}} + O(N^{-1}),$$

where $B_\rho = t(\sum_k w_k u_k^{-\rho} - 1)$. The finite-length covering bound of Yu & Tan (2019, Lemma 1), applied to N independent channel uses and L iid codewords, gives

$$\mathbb{E}_{\mathcal{C}} D_{\mathrm{KL}}(Q_{\mathcal{C}} \| S_N) \le \frac{A_N(\rho)^N}{\rho L^\rho}. \tag{15}$$

A codeword specifies one product distribution over all original scalar weights. Thus $Q_{\mathcal{C}}$ is an equally weighted mixture with L components. For $\log L = C\sqrt{N} + O(1)$, the right side vanishes whenever $\rho C > B_\rho$. Since $B_\rho/\rho \to t\sum_k w_k \log(1/u_k)$ as $\rho \downarrow 0$, any strict coefficient above this finite quantity suffices. This agrees with the exact label-information identity

$$I(J_{1:N}; \theta_{1:N}) = t\sqrt{N}\sum_k w_k \log(1/u_k) - D_{\mathrm{KL}}(S_N \| U_N).$$

The entropy of the label string itself has the larger $\sqrt{N}\log N$ order. Covering the output rather than reproducing that string is the established sparse-signaling distinction (Bloch, 2016).

The objective transfer is exact:

$$\mathbb{E}_{\mathcal{C}} G_N(Q_{\mathcal{C}}) = G_N(S_N) + \mathbb{E}_{\mathcal{C}} D_{\mathrm{KL}}(Q_{\mathcal{C}} \| S_N). \tag{16}$$

Indeed $\mathbb{E}_{\mathcal{C}} Q_{\mathcal{C}} = S_N$, so expected energy and the change-of-reference term average exactly. All terms are finite for the finite cells and Gaussian prior. Some fixed codebook therefore gives a mixture whose objective is at most this average, or within an arbitrarily small slack of it. Choosing a fixed partition for the desired tolerance proves the first part of Theorem 2.

For $\log L_N/\sqrt{N} \to \infty$, the component count required by any fixed partition is eventually available. Let its tolerance tend to zero after the width limit. The whole-neuron lower bound in Appendix A applies to this scalar-factorized subclass when $\log L_N = o(N)$, proving value convergence to J_B. Its stability identity also applies although R_N need not be a scalar-factorized mixture with the permitted component count. Finally, if $J_S > J_B$ and $0 < \epsilon < J_S - J_B$, the construction with tolerance $\epsilon/2$ gives a feasible mixture with $\log L = O(\sqrt{N})$. A subsequence on which the minimum $\log L$ is $o(\sqrt{N})$ would contradict the scalar lower bound of Theorem 1. This proves the two-sided order with the stated strict gap and sufficiently large widths.

B.4 THE EXACT COMPONENT CURVE IN THE OPENING EXAMPLE

Use the bias-free ReLU network, observations $(e_1, 1), (-e_1, -1)$ and noise covariance $I_2/2$. Write $T_N = (F_N(e_1), F_N(-e_1))$, $y = (1, -1)$, and $S_N = \sum_i a_i b_{i1}$; then $V(T) = \|T - y\|^2$ and $G_N(U_N) = 3$ exactly.

Corollary 6 (A worked component curve)**.** *If* $\log L_N/\sqrt{N} \to C \in [0, \infty)$ *and* $c = \min(C, 1)$*, then*

$$\inf_{Q \in \mathcal{S}_{N, L_N}} G_N(Q) \to j(c) := 2 + (1-c)^2.$$

Every eligible sequence $Q_N \in \mathcal{S}_{N,L_N}$ *with* $G_N(Q_N) \le j(c) + o(1)$ *satisfies* $D_{\mathrm{KL}}(Q_N \| R_{c,N}) \to 0$*, where*

$$\frac{dR_{c,N}}{dU_N} = e^{cS_N/\sqrt{N} - \Lambda_N(c)}, \qquad \Lambda_N(c) = -\frac{N}{2}\log(1 - c^2/N), \quad N > c^2. \tag{17}$$

On the observed inputs, the mean under Q_N tends to $(c/2, -c/2)$, and the output distribution converges weakly to $\mathcal{N}((c/2, -c/2), I_2/2)$. These statements also hold with $c = 1$ when $\log L_N/\sqrt{N} \to \infty$ and $\log L_N = o(N)$. For fixed $0 < \epsilon < 1$, the least count admitting $G_N(Q) \le 2 + \epsilon$ satisfies $\log L_{\min}(N, \epsilon)/\sqrt{N} \to 1 - \sqrt{\epsilon}$.

Proof. Lower bound under scalar factorization. Put $d = D_{\mathrm{KL}}(Q\|U_N)$ and $h = I(J;\theta)$. The mixture identity (11), scalar independence, and the Gaussian mean-entropy bound give

$$|\mathbb{E}_Q S_N| \le \frac{1}{2}\sum_j \pi_j \sum_i \{(\mathbb{E}_j a_i)^2 + (\mathbb{E}_j b_{i1})^2\} \le d + h \le d + \log L.$$

Finite d and L imply finite KL and second moments for every positive-weight component. ReLU gives the exact identity $T_{N,+} - T_{N,-} = S_N/\sqrt{N}$. Competitive objectives bound d; thus a limiting mean μ has $|\mu_+ - \mu_-| \le C$. The prior neuron moment-generating function is $M(u,v) = \frac{1}{2}(1-u^2)^{-1/2} + \frac{1}{2}(1-v^2)^{-1/2}$. Hence $N \log M(t/\sqrt{N}) \to \|t\|^2/4$ for each fixed t. Summing over the four signed coordinate directions gives eventual exponential-norm bounds at every fixed coefficient. Entropy duality applied to truncated absolute coordinates, first taking the tail cutoff and then the coefficient large, transfers first moments along weak subsequences. The same duality gives $\liminf d \ge \sup_t\{t^T\mu - \|t\|^2/4\} = \|\mu\|^2$. Lemma 5, Portmanteau and Jensen therefore yield

$$\liminf G_N(Q_N) \ge \|\mu\|^2 + 1 + \|\mu - y\|^2 = 3 + \sigma^2 + u^2 - 2u, \qquad \sigma = \mu_+ + \mu_-, \quad u = \mu_+ - \mu_-.$$

Minimizing over $|u| \le C$ gives $\sigma = 0, u = c$ and $j(c)$. This uses a quadratic-energy lower bound, not covariance convergence.

Upper bound using Gaussian product mixtures. For a fixed $k \ge 0$, put $\rho = k/\sqrt{N}$. Under $R_{k,N}$ the pairs (a_i, b_{i1}) are independent centered Gaussians with covariance $(1-\rho^2)^{-1}\left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$; all unused coordinates remain at prior. Direct Gaussian moments give

$$D_{\mathrm{KL}}(R_{k,N}\|U_N) = \frac{k^2}{1 - k^2/N} + \frac{N}{2}\log(1 - k^2/N) \to k^2/2, \qquad \mathbb{E}_R T_N = \frac{(k/2, -k/2)}{1 - k^2/N}.$$

The covariance tends to $I_2/2$, and uniform fourth moments give $G_N(R_{k,N}) \to j(k)$.

For $k > 0$ represent each pair by independent standard normals Z_i, A_i, B_i:

$$a_i = \sqrt{\frac{\rho}{1-\rho^2}}Z_i + \frac{A_i}{\sqrt{1+\rho}}, \qquad b_{i1} = \sqrt{\frac{\rho}{1-\rho^2}}Z_i + \frac{B_i}{\sqrt{1+\rho}}.$$

Conditioning on $Z_{1:N}$ makes all original scalar weights independent.

A codebook of L iid latent vectors therefore defines a scalar-factorized mixture. For one pair the informative Gaussian correlation has $r^2 = 2\rho/(1+\rho)$. Diagonalizing its joint Gaussian information density gives $\imath = -\frac{1}{2}\log(1-r^2) + \frac{r}{2}(G_1^2 - G_2^2)$ with independent standard G_1, G_2. Consequently

$$\log \mathbb{E} e^{t\imath_N} = -\frac{N}{2}\{t\log(1-r^2) + \log(1 - t^2 r^2)\} = k(t + t^2)\sqrt{N} + O(1).$$

The version for continuously distributed codewords of (15) follows directly by conditioning on the generating codeword and applying Jensen to the other $L - 1$ likelihood ratios:

$$\mathbb{E}_{\mathcal{C}} D_{\mathrm{KL}}(Q_{\mathcal{C}}\|R_{k,N}) \le \mathbb{E}\log(1 + e^{\imath_N}/L) \le \frac{\mathbb{E} e^{t\imath_N}}{tL^t}.$$

For $\log L = C'\sqrt{N} + O(1)$ with $C' > k$, take $0 < t < \min(1, C'/k - 1)$; the bound vanishes. Moreover (16) holds exactly with reference distribution $R_{k,N}$: the average mixture distribution is $R_{k,N}$, and both $V(T)$ and $\log(dR_{k,N}/dU_N)$ are integrable under that Gaussian weight distribution. Fubini and the finite expected KL justify the identity for this unbounded quadratic energy. Some fixed codebook therefore gives a mixture with objective at most the average, with arbitrarily small slack if needed. For $0 < C \le 1$, choose $k < C' < C$, take the width limit, then let $k \uparrow C$. For $C > 1$ take $k = 1 < C' < C$; for $C = 0$ use the prior. The construction uses a strict inequality $C' > k$.

Stability, including the rate boundary. Put $W_c(T) = V(T) + c(T_+ - T_-)$, a bounded-below quadratic. Exactly,

$$G_N(Q) - G_N(R_{c,N}) = D_{\mathrm{KL}}(Q \| R_{c,N}) + (\mathbb{E}_Q - \mathbb{E}_R) W_c(T_N).$$

For a closure limit with mean coordinates (σ, u), Jensen gives $\mathbb{E}W_c \geq F_c(\sigma, u) := 1 + \sigma^2/2 + (u-2)^2/2 + cu$. At $(0, c)$ its gradient is $(0, 2c-2)$. If $c = C < 1$, feasibility $u \leq c$ makes the supporting-plane remainder nonnegative; if $c = 1$ the gradient is zero; if $C = 0$, feasibility fixes $u = 0$. Portmanteau and reference moment convergence make the expectation difference nonnegative in liminf. Absolute $o(1)$ excess forces $D_{\mathrm{KL}}(Q_N \| R_{c,N}) \to 0$ by subsequence compactness. Pinsker and first-moment transfer give the claimed weak convergence and limiting mean. When $C = \infty$, the same argument at $c = 1$ uses closure under $\log L_N = o(N)$, without a mean constraint.

Finally, $j(c)$ decreases strictly on $[0, 1]$. A subsequence of feasible minimum counts below $1 - \sqrt{\epsilon}$ by a fixed margin contradicts the lower bound; any fixed coefficient above it has limiting value strictly below $2 + \epsilon$ and supplies a feasible mixture at all sufficiently large widths. This proves the fixed-tolerance count without assuming a finite-width optimizer. It gives neither a shrinking-tolerance rate nor a width threshold. At the plateau, $\log Z_N \to -1 - \log 2$, so the posterior-KL floor remains $1 - \log 2$. No covariance convergence for arbitrary near-optimal factors at other queries is asserted. □

C Component count for full posterior approximation

C.1 A positive posterior-KL gap

The exact posterior-output law converges to the prior Gaussian tilted by e^{-V}: the likelihood is bounded and continuous and its normalizer converges to $Z = \mathbb{E}e^{-V(\Gamma_H)} > 0$. For bounded-posterior-KL competitors, entropy duality under P_N with truncations of V first proves finite expected energy, since $\mathbb{E}_{P_N} e^V = 1/Z_N$. The Gibbs identity then bounds global prior KL. The whole-neuron lower bound therefore applies to every mixture of whole-neuron products with $\log L = o(N)$, including scalar-factorized mixtures, and gives liminf posterior KL at least Δ.

After whitening, the floor is

$$\Delta = D_{\mathrm{KL}}\big(\mathcal{N}(\mu_*, I_r) \,\big\|\, Z^{-1} e^{-W} \phi_r\big)\,.$$

Here ϕ_r denotes the standard Gaussian density on $\mathbb{R}^r$. If $\Delta = 0$, equality of densities would make W affine on the whole whitened support. A nonnegative affine function there is constant. Thus nonconstancy on the actual support implies $\Delta > 0$. For regression, write $K_{HH} = BB^T$ on its actual range and $A = B^T \Sigma^{-1} B$. Completing the square gives covariance $(I + A)^{-1}$ and the same mean as the reference. The Gaussian KL formula yields

$$\Delta = \tfrac{1}{2}\{\operatorname{tr} A - \log\det(I + A)\}. \tag{18}$$

This calculation also proves the exact values and reference curves in Figure 1.

For sampling, $V \geq 0$ allows prior proposals accepted with probability $e^{-V(F_N(H))}$. The overall acceptance probability is $Z_N \to Z > 0$, so the expected number $1/Z_N$ of prior proposals remains bounded as width grows. This proposal count differs from the mixture component count.

C.2 A mixture construction under quadratic growth

Use the whitened prior observables Z_i and T_N from Appendix A. Given $\eta > 0$, choose a finite rectangular partition of the original weight coordinates $\mathcal{B}$ with M positive cells such that $\mathbb{E}\|Z - \mathbb{E}[Z \mid \mathcal{B}]\|^4 \leq \eta^4$. Such partitions are dense in L^4. Put $S_N = \mathbb{E}[T_N \mid \mathcal{B}^N]$ and $R_i = Z_i - \mathbb{E}[Z_i \mid \mathcal{B}_i]$. The centered independent-sum fourth-moment expansion gives

$$\mathbb{E}\|T_N - S_N\|^4 \leq c_r \left\{ N^{-1} \mathbb{E}\|R_1\|^4 + (\mathbb{E}\|R_1\|^2)^2 \right\} \leq 2c_r \eta^4.$$

Prior fourth moments of T_N are uniformly bounded, and conditional Jensen gives the same bound for S_N. The quadratic gradient bound implies $|W(x) - W(y)| \leq C(1 + \|x\| + \|y\|)\|x - y\|$. Conditional least-squares projection and Cauchy–Schwarz yield

$$\mathbb{E}\operatorname{Var}(W(T_N) \mid \mathcal{B}^N) \leq \mathbb{E}[W(T_N) - W(S_N)]^2 \leq C' \eta^2,$$

uniformly in width and in the chosen partition.

For each full cell label b, write $u_b = U_N(\mathcal{B}^N = b)$ and $v_b = \mathbb{E}_U[W(T_N) \mid b]$. Define

$$\bar{Z}_N = \sum_b u_b e^{-v_b}, \qquad Q_N = \sum_b \frac{u_b e^{-v_b}}{\bar{Z}_N} U_N(\cdot \mid b).$$

There are at most M^N cells. The prior and cells both factor, so every conditional is an admissible product distribution over the original scalar weights. Within a cell the log density ratio to the posterior is $W(T_N) - v_b + \log(Z_N/\bar{Z}_N)$; averaging proves the weight-space identity

$$D_{\mathrm{KL}}(Q_N \| P_N) = \log(Z_N/\bar{Z}_N). \tag{19}$$

Conditional Jensen and $|(e^{-v})''| \le 1$ for $v \ge 0$ imply $0 \le Z_N - \bar{Z}_N \le C'\eta^2/2$. Also $Z_N \ge e^{-\mathbb{E}_U W(T_N)} \ge e^{-C(1+r)} =: z_0 > 0$. Taking η small gives

$$D_{\mathrm{KL}}(Q_N \| P_N) \le -\log\left(1 - \frac{C'\eta^2}{2z_0}\right), \qquad \log L_N \le N \log M.$$

This proves the $O(N)$ upper bound on the log component count for every fixed positive tolerance. A subsequence with a fixed tolerance below Δ and minimum log count $o(N)$ would contradict the preceding lower bound. Together these bounds prove Theorem 3. The construction is scalar-factorized and hence also a mixture of whole-neuron products; the same upper and lower orders hold for that larger family. The constants depend on the fixed source and tolerance and do not provide a practical finite-width mixture construction or a sharp positive-rate coefficient.

C.3 Posterior approximation for nonconvex likelihoods

The $\Theta(N)$ order persists for continuous nonnegative energies of at most quadratic growth, although the scalar and whole-neuron mean theorems do not thereby extend. For the upper bound, choose an L^2 cell approximation. The aggregate residual is uniformly small in probability. Prior fourth moments, quadratic growth and uniform continuity on compact sets imply $\sup_N \mathbb{E}|W(T_N) - W(S_N)| \to 0$. Let $v_b = \mathbb{E}[W(T_N) \mid b]$. Conditional Jensen and the fact that e^{-v} is 1-Lipschitz on $[0, \infty)$ give $Z_N - \bar{Z}_N \le 2\mathbb{E}|W(T_N) - W(S_N)|$; (19) then proves the same $O(N)$ upper bound.

For the lower bound, let $p_\infty = Z^{-1}e^{-W}\phi_r$ and $\mathcal{C}_r = \{\phi_r * H\}$. This convolution class is weakly closed: tightness of $G + M_n$ forces tightness of M_n through $\Pr(\|M_n\| > 2R) \le \Pr(\|G + M_n\| > R) + \Pr(\|G\| > R)$. Every nontrivial Gaussian location mixture has unbounded density ratio to ϕ_r. Indeed some bounded set of positive mixing mass satisfies $u^T m \ge c > 0$; along su, its contribution to the ratio is at least a positive constant times $e^{sc - R^2/2}$. In contrast, $p_\infty/\phi_r \le 1/Z$. Hence if p_∞ belonged to $\mathcal{C}_r$, only $H = \delta_0$ would be possible, and that would force W to be constant. Thus, for a nonconstant energy on the actual support,

$$0 < \delta_* := \inf_{\nu \in \mathcal{C}_r} D_{\mathrm{KL}}(\nu \| p_\infty) < \infty.$$

Positivity follows from Pinsker and weak closedness; finiteness follows by testing the prior Gaussian. Closure, data processing and lower semicontinuity give the floor δ_* in the regime $\log L = o(N)$. Every fixed tolerance below it therefore costs $\Theta(N)$ in log count. For convex W, Jensen and entropy duality show that the best convolution is the translated reference, so $\delta_* = \Delta$. For nonconvex likelihoods the floor need not have an explicit formula, and this argument identifies no optimal mean or variational reference.

D Gaussian families: mean features, fluctuations and variational optima

D.1 The Gaussian-per-neuron optimum

The network, likelihood assumptions and notation are those of Section 2, specialized to ReLU. Here H also denotes the matrix with rows h_l^T.

Proposition 7 (Gaussian factors over ReLU neurons)**.** *For the ReLU network, let each neuron have an arbitrary nondegenerate joint Gaussian distribution, independently across neurons. The limiting optimum is*

$$\min_{m\in\mathbb{R},\, w\in\mathbb{R}^d} \left\{ \tfrac{1}{2}(m^2 + \|w\|^2) + \mathbb{E}V\big(\Gamma_H + m\psi_H + Hw/2\big) \right\}.$$

Every sequence with vanishing absolute objective excess has its mean network output converging to the minimizing mean above. Its output distribution converges weakly to the Gaussian with that mean and covariance K, and its output covariance also converges to K.

The two mean features come from the first-order Gaussian score directions a and ab_k; other perturbations have vanishing aggregate mean at bounded prior KL. The result allows nonidentical neurons. This family is a restricted subclass of whole-neuron products and need not attain J_B. For regression it is ridge regression on $\Phi(h) = [\psi(h), h^T/2]^T$, with mean kernel $J(h,g) = \psi(h)\psi(g) + h^T g/4$. Here J describes the available mean features and K the residual output fluctuations. The difference $K - J$ is positive semidefinite: the mean features are orthogonal projections of the full neuron observable.

Huix et al. (2022) study N^{-1}-normalized networks and scale the KL term with the sample-size-to-width ratio. The fixed $N^{-1/2}$ normalization here retains the prior output fluctuations described above.

D.2 Proof of Proposition 7

Under a fixed global Gaussian prior-KL bound C, every neuron's mean is bounded and every covariance eigenvalue lies in one compact subinterval of $(0, \infty)$. The Gaussian KL formula further gives

$$\sum_i \left(\|\mu_i\|^2 + \|\Sigma_i - I\|_F^2 \right) \le C'(C).$$

Write Φ_{G} and φ_{G} for the standard normal CDF and density. For $Y = b^T h$, let $\mu_Y = \mathbb{E}Y$ and $s_Y^2 = \mathrm{Var}(Y)$. The Gaussian Stein identity is

$$\mathbb{E}[aY_+] = \mu_a \mathbb{E}[Y_+] + \mathrm{Cov}(a, Y)\Phi_{\mathrm{G}}(\mu_Y / s_Y).$$

On the compact parameter set its first derivative at the prior is $\mu_a\psi(h) + \Sigma_{a,b}h/2$, and the second-order remainder is bounded by the squared parameter displacement. Summing and dividing by $\sqrt{N}$ leaves only $m_N = N^{-1/2}\sum_i \mu_{i,a}$ and $w_N = N^{-1/2}\sum_i \Sigma_{i,a,b}$ in the mean.

The analogous second-moment map is Lipschitz on that set. Cauchy–Schwarz bounds its averaged change by $O_C(N^{-1/2})$; aggregate subtraction of squared neuron means vanishes. Output covariance therefore tends to K. Uniform neuron fourth moments give the multivariate Lyapunov CLT and uniform aggregate fourth moments, giving convergence of the expected negative log likelihood. Convexity of $\mathrm{tr}\,\Sigma - \log\det\Sigma$ and of squared means gives a lower bound on KL to the prior in terms of the averaged Gaussian parameters. Expanding that bound at I gives

$$D_{\mathrm{KL}}(Q_N \| U_N) \ge \tfrac{1}{2}m_N^2 + \tfrac{1}{2}\|w_N\|^2 + o_C(1).$$

The two symmetric outgoing/incoming covariance blocks account for the coefficient $1/2$ of $\|w_N\|^2$.

The matching construction gives every neuron mean $(m/\sqrt{N}, 0)$ and covariance

$$I + N^{-1/2}\begin{pmatrix} 0 & w^T \\ w & 0 \end{pmatrix}.$$

It is positive definite for $N > \|w\|^2$, has exactly the desired aggregate mean, and has prior KL $m^2/2 - (N/2)\log(1 - \|w\|^2/N) \to (m^2 + \|w\|^2)/2$. The same moment bound gives convergence of the expected negative log likelihood. Strong convexity identifies the unique coefficient optimum and forces all near-optimal coefficient subsequences to it. This proves the proposition, including its actual covariance conclusion within this Gaussian family.

D.3 Arbitrary Gaussian correlations: means and fluctuations

The shallow ReLU architecture and Gaussian prior remain fixed; all correlations between network weights are allowed. The distinction between independence error and Gaussian approximation error

is already explicit in Farquhar et al. (2020); their depth-based existence results change the architecture and do not impose the present bound on KL to the prior. Normal approximation of smooth Gaussian functionals follows from classical second-order Poincaré inequalities (Chatterjee, 2009), already applied to shallow Gaussian networks by Bordino et al. (2024). We characterize the covariance reachable at bounded total prior KL. The self-contained proof below allows ReLU and arbitrary weight correlations.

The mean reduction is exact. For any law Q with neuron marginals Q_i, put $\bar{Q} = \bigotimes_i Q_i$. Additivity of the network output and the product-prior entropy identity give

$$\mathbb{E}_Q F_N(h) = \mathbb{E}_{\bar{Q}} F_N(h), \qquad D_{\mathrm{KL}}(\bar{Q}\|U_N) = \sum_i D_{\mathrm{KL}}(Q_i\|U_1) \le D_{\mathrm{KL}}(Q\|U_N).$$

If Q is jointly Gaussian, so is $\bar{Q}$. The set of predictive means induced by Gaussian weight distributions at a fixed prior-KL budget is therefore unchanged by allowing cross-neuron correlations. The marginal expansion in the preceding proof gives, uniformly for $D_{\mathrm{KL}}(Q_N\|U_N) \le C$,

$$\mathbb{E}_Q F_N(h) = m_N \psi(h) + w_N^T h/2 + O_C(\|h\| N^{-1/2}),$$

with bounded coefficients. This does not identify the objective or spread of a joint Gaussian: cross-neuron covariance can change both.

Proposition 8 (Full Gaussian fluctuation law)**.** *For the ReLU network, let $Q_N = \mathcal{N}(\eta_N, \Sigma_N)$ be any joint Gaussian distribution over the network weights with $D_{\mathrm{KL}}(Q_N\|U_N) \le C$, for fixed C. For a fixed finite query set A, let $\mu_N = \mathbb{E}_Q F_N(A)$ and $s_N = \mathrm{Var}_Q(N^{-1/2}\sum_i a_i)$. Then*

$$W_2\Big(Q_N^{F_N(A)}, \mathcal{N}(\mu_N, K_{AA} + (s_N - 1)\psi_A\psi_A^T)\Big) \to 0$$

uniformly on this KL ball. Its covariance differs from the displayed matrix by $O_{C,d,A}(N^{-1/2})$. Convergence of (μ_N, s_N) gives a fixed Gaussian limit; the width-dependent approximation does not require these parameters to converge.

Proof. The Gaussian KL formula bounds $\|\eta_N\|^2$ and every covariance eigenvalue above and away from zero. Write these spectral bounds as $\ell I \preceq \Sigma_N \preceq uI$. Couple $X = \eta_N + TG$ and $G \sim \mathcal{N}(0, I)$, with $T = \Sigma_N^{1/2}$. Since $(\sqrt{\lambda} - 1)^2 \le \lambda - 1 - \log\lambda$,

$$\mathbb{E}\|X - G\|^2 = \|\eta_N\|^2 + \|T - I\|_F^2 \le 2C, \qquad \|\Sigma_N - I\|_F = O_C(1). \tag{G1}$$

We first establish a block covariance bound. Suppose the full covariance of a Gaussian vector has spectrum in $[\ell, u]$, and partition the vector into blocks of size $d+1$. A stack of block functions $f_i(X_i)$ with uniformly bounded second moments then has covariance operator norm bounded independently of the number of blocks.

Whiten each block and let R be the resulting correlation matrix. Its diagonal blocks are identities and $\|R\|_{\mathrm{op}} \le u/\ell$. The degree-one Hermite coefficients form a bounded block-diagonal map, so that part of the covariance has bounded norm. For the remaining Hermite degrees, a block pair has cross-covariance norm at most $M\rho_{ij}^2$, where $\rho_{ij} = \|R_{ij}\|_{\mathrm{op}} \le 1$ and M bounds second moments. To see this, for whitened blocks W_i, W_j expand

$$\mathbb{E} e^{s^T W_i - \|s\|^2/2} e^{t^T W_j - \|t\|^2/2} = e^{s^T R_{ij} t}.$$

Singular-value rotations of R_{ij} give zero cross expectation between different Hermite degrees and contraction at most ρ_{ij}^k in degree k. If $f_i^{(k)}$ is the degree-k projection, Parseval and Cauchy–Schwarz therefore give the remainder bound

$$\left\| \mathrm{Cov}\left(\sum_{k\ge 2} f_i^{(k)}, \sum_{k\ge 2} f_j^{(k)} \right) \right\|_{\mathrm{op}} \le \sum_{k\ge 2} \rho_{ij}^k \|f_i^{(k)}\|_2 \|f_j^{(k)}\|_2 \le M\rho_{ij}^2.$$

The row bound

$$\sum_j \rho_{ij}^2 \le \sum_j \|R_{ij}\|_F^2 = \mathrm{tr}(R^2)_{ii} \le (d+1)(u/\ell)^2$$

then proves the claimed operator bound, also for vector-valued functions.

For one nonzero query h, remove the outgoing first-degree term:

$$R_N(\theta;h) = F_N(h) - \psi(h)A_N, \qquad A_N = N^{-1/2}\sum_i a_i.$$

Its weak gradient block is

$$\nabla_i R_N = N^{-1/2}\big((b_i^T h)_+ - \psi(h),\ a_i \mathbf{1}_{\{b_i^T h>0\}} h\big).$$

The block bound gives $\|\operatorname{Cov}(\nabla R_N(X))\|_{\mathrm{op}} = O(N^{-1})$. The expected gradient has squared norm $O(N^{-1})$ as well. For its outgoing entries this follows from ReLU Lipschitzness and (G1). For the hidden entries use, with $Y_i = b_i^T h$,

$$\mathbb{E}[a_i \mathbf{1}_{\{Y_i>0\}}] = \eta_{a_i}\Phi_{\mathrm{G}}(\mu_{Y_i}/\sigma_{Y_i}) + \frac{\operatorname{Cov}(a_i, Y_i)}{\sigma_{Y_i}}\varphi_{\mathrm{G}}(\mu_{Y_i}/\sigma_{Y_i}).$$

Here $\sigma_{Y_i} \geq \sqrt{\ell}\|h\|$, and the squared weight means and same-neuron covariance changes have bounded sum. To obtain the required gradient estimate, put $B = T - I$ and $g = \nabla R_N(X;h)$. Then

$$\begin{aligned}\mathbb{E}\|B^T g\|^2 &= \operatorname{tr}\big(B^T \operatorname{Cov}(g) B\big) + \|B^T \mathbb{E} g\|^2 \\ &\leq \|\operatorname{Cov}(g)\|_{\mathrm{op}}\|B\|_F^2 + \|B\|_{\mathrm{op}}^2\|\mathbb{E}g\|^2 = O(N^{-1}).\end{aligned} \tag{G2}$$

For the gradient change, outgoing entries are controlled by (G1). A hidden entry changes by a displacement term and a sign-change term. For $\delta = X - G$, a sign change requires either $|G_{b_i}^T h| \leq \epsilon\|h\|$ or $|\delta_{b_i}^T h| > \epsilon\|h\|$. The prior G_{a_i} is independent of G_{b_i}, and for any affine Gaussian D jointly Gaussian with a standard Gaussian Z, $\mathbb{E}[Z^2 D^2] \leq 3\mathbb{E}[D^2]$. Thus

$$\frac{1}{N}\sum_i \mathbb{E}\big[G_{a_i}^2 \mathbf{1}_{\{\text{sign change}\}}\big] \leq c\epsilon + \frac{c}{N\epsilon^2},$$

and consequently

$$\mathbb{E}\|\nabla R_N(X;h) - \nabla R_N(G;h)\|^2 \leq c\left(N^{-1} + \epsilon + (N\epsilon^2)^{-1}\right).$$

Define $D_N(G;h) = R_N(\eta_N + TG;h) - R_N(G;h)$. Its weak gradient with respect to the independent standard Gaussian coordinates is

$$\nabla_G D_N = (T-I)^T \nabla R_N(X;h) + \nabla R_N(X;h) - \nabla R_N(G;h).$$

The first-order Gaussian Poincaré inequality $\operatorname{Var}(D_N) \leq \mathbb{E}\|\nabla_G D_N\|^2$, with $\epsilon = N^{-1/3}$, therefore gives $\operatorname{Var}(D_N) = O(N^{-1/3})$. Since $\mathbb{E}R_N(G;h) = 0$, subtracting $\mathbb{E}D_N$ and summing over the fixed query set gives

$$R_N(X;A) - \mathbb{E}R_N(X;A) = R_N(G;A) + \xi_N, \qquad \mathbb{E}\|\xi_N\|^2 = O(N^{-1/3}). \tag{G3}$$

The ReLU products and their affine compositions have square-integrable weak first derivatives. Gaussian Poincaré extends to these functions by Sobolev approximation; no second derivative of ReLU is used. Zero query inputs contribute zero coordinates and can be omitted.

The prior residual is an iid normalized sum of vectors $r(G_i) = G_{a_i}[(G_{b_i}^T A)_+ - \psi_A]$, with covariance $K_0 = K_{AA} - \psi_A\psi_A^T$ and finite fourth moments. It is orthogonal to every linear weight coordinate. Let $L_N = e_N^T TG = \sum_i v_{i,N}^T G_i$, where e_N averages the outgoing coordinates. Its variance is $s_N \leq u$. Let $I_N = \{i : \|v_{i,N}\| > N^{-1/8}\}$ and $\tau_N = \sum_{i\in I_N}\|v_{i,N}\|^2$. Since $\sum_i \|v_{i,N}\|^2 = s_N \leq u$, we have $|I_N| \leq uN^{1/4}$ and

$$\mathbb{E}\left\|N^{-1/2}\sum_{i\in I_N} r(G_i)\right\|^2 = \frac{|I_N|}{N}\operatorname{tr} K_0 = O(N^{-3/4}).$$

Thus these blocks can be removed from the residual. Their linear contribution is exactly $\mathcal{N}(0,\tau_N)$ and is independent of all remaining blocks; it is retained in the final sum. For the remaining blocks, each Cramér–Wold combination of $(r(G_i)/\sqrt{N}, v_{i,N}^T G_i)$ has fourth-moment sum at most

$$c\left(N^{-1} + \sum_{i\notin I_N}\|v_{i,N}\|^4\right) \leq c(N^{-1} + uN^{-1/4}) \longrightarrow 0.$$

Along any subsequence with $s_N \to s$, extract a further subsequence with $\tau_N \to \tau$. The triangular-array CLT and exact zero cross covariance give independent Gaussian limits with covariance K_0 and variance $s - \tau$ for the remaining pair, allowing zero variance. Adding the independent removed linear contribution gives variance s; restoring the negligible removed residual proves

$$(R_N(G; A), L_N) \Rightarrow (Z_0, Z_1), \qquad Z_0 \sim \mathcal{N}(0, K_0), \quad Z_1 \sim \mathcal{N}(0, s), \quad Z_0 \text{ independent of } Z_1.$$

The final limit is the same for every such further subsequence. The pair's second moment is exactly $\operatorname{tr} K_0 + s_N$, so weak convergence also gives W_2 convergence, including singular query directions.

Equation (G3) transfers this convergence through

$$F_N(X; A) - \mu_N = R_N(G; A) + \psi_A L_N + \xi_N.$$

If uniformity failed on a fixed KL ball, a violating sequence would have a subsequence with $s_N \to s$. The preceding argument applies to that sequence, while the centered comparison Gaussians also converge in W_2. The triangle inequality contradicts the violation. Translation invariance of W_2 and exact centering remove any need for μ_N to converge.

The estimate (G3) is sufficient for this qualitative limit. We derive the sharper $O(N^{-1/2})$ covariance rate separately, using Hermite projections. Let $z_i = a_i(b_i^T A)_+$ and $J_i = \mathbb{E}\nabla_i z_i$. Its exact first Hermite projection is $J_i^T(X_i - \eta_i)$. The earlier bounds give $J_i = J_0 + \Delta_i$, where $J_0 = (\psi_A^T, 0)^T$ and $\sum_i \|\Delta_i\|_F^2 = O(1)$. The first-degree covariance $N^{-1} J^T \Sigma_N J$ has baseline $s_N \psi_A \psi_A^T$; its cross terms are $O(N^{-1/2})$ and its quadratic perturbation is $O(N^{-1})$. Subtract the actual first projections and write $q_i = z_i - \mathbb{E}z_i - J_i^T(X_i - \eta_i)$. These remainders have Hermite rank at least two. With $R_{ij} = \Sigma_{ii}^{-1/2} \Sigma_{ij} \Sigma_{jj}^{-1/2}$, their normalized off-diagonal covariance sum is bounded by

$$\frac{c}{N} \sum_{i \neq j} \|R_{ij}\|_{\mathrm{op}}^2 \leq \frac{c}{N\ell^2} \|\Sigma_N - I\|_F^2 = O(N^{-1}).$$

For diagonal terms put $d_i^2 = \mathbb{E}\|X_i - G_i\|^2$, so $\sum_i d_i^2 = O(1)$. Lipschitzness and Gaussian fourth moments give $\|z_i(X_i) - z_i(G_i)\|_2 \leq c d_i$; the average covariance change is therefore $O(N^{-1} \sum_i d_i) = O(N^{-1/2})$. The subtracted first-degree covariances have the same bound. Different Hermite degrees have zero cross covariance. This proves the rate. □

The Gaussian cost and two exact consequences. Data processing to A_N gives

$$D_{\mathrm{KL}}(Q_N \| U_N) \geq \tfrac{1}{2}(\mathbb{E}_Q A_N)^2 + h(s_N), \qquad h(s) = \tfrac{1}{2}(s - 1 - \log s).$$

Keeping hidden weights at prior and setting outgoing covariance $I + (s-1)ee^T$, where $e = \mathbf{1}/\sqrt{N} \in \mathbb{R}^N$, attains the spread-only cost $h(s)$ with mean zero. It has finite covariance $K + (s-1)[K/N + (1 - 1/N)\psi\psi^T]$.

For the introduction's two-input regression example, the unit odd contrast is orthogonal to ψ, and therefore retains Gaussian variance $1/2$ in every limiting output distribution induced by Gaussian weights with bounded prior KL. The exact posterior has variance $1/4$ in that contrast. Data processing and joint weak lower semicontinuity of KL give a floor $h(2) = (1 - \log 2)/2$; mean mismatch only increases it. Bounded posterior KL implies bounded prior KL through $V \geq 0$ and the uniform normalizer lower bound, so this also covers approximate infima over all joint Gaussian weight distributions.

D.4 Arbitrary labels and general observation noise

For Gaussian regression, we can now write the limiting optimum for arbitrary labels and observation noise. Let the observation noise have fixed covariance $\Sigma \succ 0$, let $\Phi = [\psi_H, H/2]$, $J = \Phi\Phi^T$, $K = K_{HH}$, and $\beta = \psi_H^T \Sigma^{-1} \psi_H$. The exact limiting infimum $\lim_N \inf_{Q \text{ Gaussian}} D_{\mathrm{KL}}(Q \| P_N)$ is

$$\begin{aligned} &\frac{1}{2} y^T \left[(J + \Sigma)^{-1} - (K + \Sigma)^{-1}\right] y + \frac{1}{2} \operatorname{tr}\left(\Sigma^{-1}(K - \psi_H \psi_H^T)\right) \\ &\qquad + \frac{1}{2} \log(1 + \beta) - \frac{1}{2} \log\det(I + \Sigma^{-1} K). \end{aligned} \tag{20}$$

The quadratic form in y is the cost of the restricted Gaussian mean features. The remaining terms give the spread cost. The gain from Gaussian covariance contraction $[\beta - \log(1+\beta)]/2$ is independent of the labels; the total error and its ordering against the whole-neuron family need not be.

To prove the lower bound, let $\bar{Q}$ be the product of a Gaussian Q's neuron marginals. Their predictive means are exactly equal, and

$$D_{\mathrm{KL}}(Q\|U_N) = D_{\mathrm{KL}}(Q\|\bar{Q}) + D_{\mathrm{KL}}(\bar{Q}\|U_N).$$

Write $s_N = \mathrm{Var}_Q\, A_N$ and $\bar{s}_N = \mathrm{Var}_{\bar{Q}}\, A_N$. On a bounded prior-KL ball, $\sum_i (\mathrm{Var}_Q\, a_i - 1)^2 = O(1)$, so $\bar{s}_N = 1 + O(N^{-1/2})$. Data processing to A_N, whose mean is the same under both laws, gives

$$D_{\mathrm{KL}}(Q\|\bar{Q}) \geq h(s_N/\bar{s}_N) = h(s_N) + o(1).$$

The last error is uniform because both variances lie in a fixed positive compact interval. The fluctuation law and the equality of means give

$$G_N(Q) = G_N(\bar{Q}) + D_{\mathrm{KL}}(Q\|\bar{Q}) + (s_N - 1)\beta/2 + o(1).$$

The Gaussian-per-neuron optimum is $J_{\mathrm{G,prod}} = \frac{1}{2}\,\mathrm{tr}(\Sigma^{-1}K) + \frac{1}{2}y^T(J+\Sigma)^{-1}y$. Minimizing $h(s) + (s-1)\beta/2$ over $s > 0$ yields $s_* = (1+\beta)^{-1}$ and value $-[\beta - \log(1+\beta)]/2$. Competitive objectives bound prior KL through $V \geq 0$, so this controls the infimum over all joint Gaussian weight distributions.

For an upper construction attaining mean and spread together, let (m, w) minimize the Gaussian-feature ridge objective. Put $u = \mathbf{1}/\sqrt{N}$ and choose a Gaussian with

$$\mathbb{E}a = mu, \quad \mathbb{E}b = 0, \quad \mathrm{Cov}(b) = I,$$
$$\mathrm{Cov}(a) = I + (s-1)uu^T, \quad \mathrm{Cov}(a_i, b_j) = \mathbf{1}_{\{i=j\}} w^T/\sqrt{N}.$$

Writing $\delta_N = \|w\|^2/N$, its Schur complement is positive whenever $\delta_N < \min(1, s)$, which holds for all sufficiently large N. Stein's identity gives the exact mean $m\psi_H + Hw/2$ and $\mathrm{Var}\, A_N = s$. Its exact prior KL is

$$\frac{m^2}{2} + \frac{1}{2}\left[s - 1 - \log(s - \delta_N) - (N-1)\log(1-\delta_N)\right] \longrightarrow \frac{m^2 + \|w\|^2}{2} + h(s).$$

The fluctuation law therefore attains the lower bound at $s = s_*$. Subtracting the fixed-data Gaussian posterior normalizer proves (20). Strict convexity in s also implies $s_N \to s_*$ for every absolute-$o(1)$ Gaussian near optimum.

D.5 The two- and four-input comparisons

In the introduction's two-input example, Gaussian mean features attain the Bayesian mean. Covariance contraction improves the objective from $J_B = 2$ to 1.928006614, leaving posterior-KL floor $.234859433$.

For the four-input example, take $(e_1, -e_1, e_2, -e_2)$ in $\mathbb{R}^2$, $y = t(1, 1, -1, -1)^T$, and noise covariance I_4. The bias-free ReLU kernel has diagonal $1/2$, opposite-axis entries zero and perpendicular entries $1/(2\pi)$. Thus $Ky = \lambda y$ for $\lambda = 1/2 - 1/\pi$, $\mathrm{tr}\, K = 2$, and $\|\psi_H\|^2 = 2/\pi$. The labels are orthogonal to every Gaussian mean feature. The Gaussian-per-neuron limiting optimum is consequently $1 + 2t^2$; full Gaussian covariance contracts the collective outgoing direction at $s_* = (1 + 2/\pi)^{-1}$, gaining $g = [2/\pi - \log(1 + 2/\pi)]/2$. Hence the family optima are

$$J_{\mathrm{G}}(t) = 1 + 2t^2 - g, \qquad J_B(t) = 1 + \frac{2t^2}{1+\lambda}.$$

At $t = 1$ these are $2.928006614\ldots$ and $2.692491099\ldots$, respectively, versus 3 for Gaussian-per-neuron products. The ordering reverses at $|t| = \sqrt{g(1+\lambda)/(2\lambda)} = .483857475\ldots$. The Gaussian construction above and the whole-neuron reference in Appendix A attain these values. Nonlinear dependence within neurons and collective output-variance contraction are complementary; neither family ranks above the other in general.

E PREDICTIVE LOSS UNDER A FIXED QUERY DISTRIBUTION

This appendix derives the predictive consequence of the asymptotic hierarchy for queries drawn from a fixed distribution. The training data (H, y) and noise covariance Σ are fixed; the activation is globally Lipschitz, and the regression noise satisfies $\Sigma \succ 0$. At a query x, let $P_{N,x}$ and $Q_{N,x}$ denote the laws of $F_N(x) + \eta$ under the posterior and an approximation, respectively, where $\eta \sim \mathcal{N}(0, \gamma)$ is independent and $\gamma > 0$ is fixed. The same approximation is chosen before an independent query $X \sim \rho$ is drawn. The Borel probability distribution ρ is fixed across widths and may be chosen using the observed training data. The lower bound and the explicit construction below need no moment assumption on ρ. The equalities for all near-optimal sequences additionally assume $\mathbb{E}_\rho \|X\|^2 < \infty$. Define the expected predictive KL

$$\mathcal{R}_N(Q_N) = \mathbb{E}_{X\sim\rho} D_{\mathrm{KL}}(P_{N,X} \| Q_{N,X}). \tag{21}$$

This is the posterior-expected excess marginal log loss for the observed query and answer. It is not a frequentist risk, a score after discarding the query, or a joint score for a future batch sharing one weight draw.

Let $\mathcal{I}_N^B(Q)$ be the infimum of $I(J; \theta_{1:N})$ over standard-Borel latent representations with measurable kernels and $Q(\cdot \mid J = j) = \bigotimes_i q_{ij}(a_i, b_i)$. Define $\mathcal{I}_N^S(Q)$ by requiring each q_{ij} to factorize over all of its original scalar coordinates, as in (3). These are grouped, continuous-variable versions of the multivariate common-information variational form (Liu et al., 2010). Thus $\mathcal{I}_N^B(Q) \le \mathcal{I}_N^S(Q)$; a small whole-neuron information cost does not imply a small scalar-factorized cost. Neither infimum needs to be attained: choosing representations within one nat preserves the asymptotic information conditions. An infimum is $+\infty$ if no finite-information representation exists.

In Gaussian regression, put

$$\begin{aligned} v(x) &= K_{xx} + \gamma, \\ d(x) &= K_{xH}(K_{HH} + \Sigma)^{-1} K_{Hx}, \\ b(x) &= \frac{1}{2}\left[-\frac{d(x)}{v(x)} - \log\left(1 - \frac{d(x)}{v(x)}\right)\right]. \end{aligned} \tag{22}$$

Here $v(x)$ is the prior predictive variance, $d(x)$ is the variance removed by conditioning on the data, and $b(x)$ is the excess Gaussian log loss from retaining variance $v(x)$ at the correct mean. The Schur complement and $\gamma > 0$ give $0 \le d(x) < v(x)$.

Lemma 9 (Projection onto a Gaussian location-mixture family)**.** *Let $P = \mathcal{N}(m, c)$ and q_0 be the density of $\mathcal{N}(m, v)$, with $0 < c \le v$. For any probability law ν on $\mathbb{R}$, let $q = \int q_t \, \nu(dt)$, where q_t is the density of $\mathcal{N}(m + t, v)$. Then*

$$D_{\mathrm{KL}}(P\|q) \ge D_{\mathrm{KL}}(P\|q_0) = \tfrac{1}{2}\{c/v - 1 - \log(c/v)\}.$$

Proof. The Gaussian likelihood ratio satisfies

$$\mathbb{E}_P \frac{q_t(X)}{q_0(X)} = \mathbb{E}_P \exp\left\{ \frac{t(X-m)}{v} - \frac{t^2}{2v} \right\} = \exp\left\{ -\frac{(v-c)t^2}{2v^2} \right\} \le 1.$$

Tonelli's theorem therefore gives $\mathbb{E}_P(q/q_0) \le 1$. Concavity of the logarithm yields

$$D_{\mathrm{KL}}(P\|q) - D_{\mathrm{KL}}(P\|q_0) = -\mathbb{E}_P \log(q/q_0) \ge -\log \mathbb{E}_P(q/q_0) \ge 0.$$

No moment condition on the mixing law is needed. The expectations are well defined: some bounded interval has positive ν mass, giving a Gaussian quadratic lower bound on $\log q$, while $q \le (2\pi v)^{-1/2}$. □

Proposition 10 (Predictive lower bound and attainment)**.** *Assume $\sup_N D_{\mathrm{KL}}(Q_N\|U_N) < \infty$. If $\mathcal{I}_N^B(Q_N) = o(N)$, then*

$$\liminf_{N\to\infty} \mathcal{R}_N(Q_N) \ge \mathbb{E}_\rho b(X). \tag{23}$$

An explicit whole-neuron mean-only product has zero component-label information, bounded KL to the prior, and attains this lower bound in the limit. Moreover, if $G_N(Q_N) \to J_B$ and $\mathbb{E}_\rho\|X\|^2 < \infty$, every whole-neuron sequence satisfying the same information condition obeys

$$\mathcal{R}_N(Q_N) \longrightarrow \mathbb{E}_\rho b(X). \tag{24}$$

Under that input-moment assumption, for representations that are conditionally independent over the original scalar weights, with $\mathcal{I}_N^S(Q_N) = o(\sqrt{N})$ and $G_N(Q_N) \to J_S$, the limit is instead

$$\mathbb{E}_\rho\left[b(X) + \frac{\{m_B(X) - m_*\psi(X)\}^2}{2v(X)}\right]. \tag{25}$$

Here m_* is selected by the training objective; it is not the coefficient that would minimize the separately chosen predictive score $\mathcal{R}_N$.

Proof and resource consequence. The fixed-query closure theorem says that a whole-neuron mixture sequence with bounded KL to the prior with component-label information $o(N)$ has, along every weakly convergent subsequence, a prior Gaussian residual plus an independent mean shift. The limiting posterior query law has mean $m_B(x)$ and variance $K_{xx} - d(x)$ before the future noise is added. At each fixed query, the prior output laws are uniformly tight, and the bounded KL to the prior makes the approximate output laws tight as well: for an event E with $0 < U_N(E) < 1$, the binary entropy bound gives $Q_N(E) \le [D_{\mathrm{KL}}(Q_N\|U_N) + \log 2]/\log(1/U_N(E))$. After adding the same future noise, apply Lemma 9 with $c = v(x) - d(x) > 0$ and $v = v(x)$ to every subsequential limiting law. Joint lower semicontinuity of relative entropy gives the pointwise bound $\liminf_N D_{\mathrm{KL}}(P_{N,x}\|Q_{N,x}) \ge b(x)$. Nonnegativity and Fatou's lemma give (23).

For attainment, write $\alpha = (K_{HH} + \Sigma)^{-1}y$ and $u(b) = \alpha^T z_H(1, b)$. The existing mean-only product keeps the incoming prior and shifts each outgoing weight by $u(b)/\sqrt{N}$. It has zero component-label information, its KL to the prior is $\frac{1}{2}\mathbb{E}u(b)^2$, and its mean function is exactly $m_B(x)$ at every width. Positive Gaussian convolution and the fixed-data posterior density bound give pointwise predictive-KL convergence. There is also an envelope uniform over *all* queries. Put

$$T = \mathbb{E}u(b)^2, \qquad A_0 = \tfrac{1}{2}\{y^T\Sigma^{-1}y + \mathrm{tr}(\Sigma^{-1}K_{HH})\},$$
$$W_N = N^{-1/2}\sum_i a_i u(b_i), \qquad J_N = N^{-1}\sum_i u(b_i)^2.$$

This weight distribution, denoted Q_N^{mean}, has exact log density ratio $W_N - J_N/2$ to the prior, while $Z_N \ge e^{-A_0}$ and $dP_N/dU_N \le e^{A_0}$. Since $\mathbb{E}_U W_N^2 = \mathbb{E}_U J_N = T$, Cauchy–Schwarz gives

$$D_{\mathrm{KL}}(P_N\|Q_N^{\mathrm{mean}}) \le A_0 + e^{A_0/2}\sqrt{T} + \tfrac{1}{2}e^{A_0}T.$$

Data processing bounds every noisy query KL by this same finite constant. Bounded convergence proves attainment for any fixed ρ. This approximation has KL $T/2$ to the prior; attainment is not asserted for every smaller prescribed KL ceiling. Its exact mean is the limiting GP posterior mean, which need not equal the finite neural posterior mean.

For a near-optimal sequence, weight-space KL stability gives total-variation convergence to the exact finite exponential-tilt reference R_N (or the scalar reference $S_{m_*,N}$), hence the corresponding latent query limit. These references need not equal the finite mean-only distribution. The posterior density is bounded by a fixed multiple of the prior density, whose fixed-query fourth moments are uniformly bounded. Positive Gaussian smoothing gives locally uniform positive density convergence; tightness supplies a Gaussian lower density bound. The resulting quadratic log-ratio bound and posterior second-moment uniform integrability prove pointwise forward-KL convergence.

To integrate over queries, the exact references satisfy $\mathbb{E}_{R_N}F_N(x)^2 \le C_R(1 + \|x\|^2)$ uniformly for sufficiently large widths, with the same bound for $S_{m_*,N}$. Indeed, with $u(b) = \alpha^T z_H(1, b)$, the whole-neuron tilt has incoming density proportional to $e^{u(b)^2/(2N)}$ relative to its Gaussian prior and conditional outgoing law $\mathcal{N}(u(b)/\sqrt{N}, 1)$; Gaussian exponential moments and Lipschitz growth give the bound. Put $r(x) = 2\sqrt{C_R(1 + \|x\|^2)}$. Markov's inequality and total-variation convergence give $Q_N\{|F_N(x)| \le r(x)\} \ge 1/2$ uniformly in x for sufficiently large N. The density after adding future noise therefore obeys

$$q_{N,x}(z) \ge \tfrac{1}{2}(2\pi\gamma)^{-1/2}e^{-(|z|+r(x))^2/(2\gamma)}.$$

The posterior predictive density is at most $(2\pi\gamma)^{-1/2}$, and its second moment is at most $C_P(1+\|x\|^2)$. Consequently

$$D_{\mathrm{KL}}(P_{N,x}\|Q_{N,x}) \le \log 2 + \frac{\mathbb{E}_{P_{N,x}}(|Y| + r(x))^2}{2\gamma} \le C(1 + \|x\|^2).$$

Dominated convergence proves the whole-neuron and scalar-factorized limits; the latter Gaussian KL adds the squared mean mismatch in (25).

Finally, for every fixed bound C on KL to the prior and tolerance $0 \le \epsilon < \mathbb{E}_\rho b(X)$, there are $c > 0$ and N_0 such that any whole-neuron representation with $\mathcal{R}_N(Q_N) \le \epsilon$ for $N \ge N_0$ has $\mathcal{I}_N^B(Q_N) \ge cN$. A finite mixture with L_N components has $\mathcal{I}_N^B(Q_N) \le \log L_N$, so its component count must obey $\log L_N \ge cN$. This is a representation-information statement, not a runtime lower bound or a guarantee about a trained optimizer. □

The conditional-readout approximation keeps incoming weights at their prior and samples outgoing weights from their exact conditional posterior. Its forward predictive KL vanishes in the same fixed-data, positive-noise regime. Whenever the predictive floor is positive, this vanishing loss is incompatible with simultaneously bounded KL to the prior and $o(N)$ whole-neuron label information. This established comparison identifies the dependence restriction responsible for the floor; it is not a new inference algorithm.

F Output-distribution limits under factorization and nonconvex losses

The output-distribution characterization retains Section 2's network and prior but uses no likelihood or optimality assumption. A final Student-t example shows why convexity matters for the single-shift optimum.

F.1 The joint outgoing-coordinate and residual law

Fix a finite query set $A = (x_1, \dots, x_r)$ and put

$$A_N = N^{-1/2} \sum_i a_i, \qquad D_N(A) = F_N(A) - \psi_A A_N, \qquad K_A^\perp = K_{AA} - \psi_A \psi_A^T.$$

Here $K_A^\perp$ is positive semidefinite: it is the prior covariance of the single-neuron vector $[a\{\varphi(b^T x) - \psi(x)\}]_{x \in A}$. Write $\gamma_1 = \mathcal{N}(0,1)$.

Proposition 11 (Output limits under conditional factorization)**.** *Suppose Q_N has a representation with component label J_N satisfying*

$$Q_N(d\theta_{1:N} \mid J_N = j) = \bigotimes_{i=1}^N (q_{ij}^a \otimes q_{ij}^b)(d\theta_i), \qquad \sup_N d_N < \infty, \quad i_N = o(\sqrt{N}), \tag{26}$$

where $d_N = D_{\mathrm{KL}}(Q_N \| U_N)$ and $i_N = I(J_N; \theta_{1:N})$. The label may be finite or standard Borel with a measurable conditional kernel. In particular, every mixture in (3) *satisfies the stated factorization. The augmented laws are tight, and every subsequential weak limit has the form*

$$(A_N, D_N(A)) \;\Rightarrow\; (Z + M, G_A^\perp), \qquad Z \sim \gamma_1, \quad G_A^\perp \sim \mathcal{N}(0, K_A^\perp), \tag{27}$$

where $Z, M, G_A^\perp$ are mutually independent and $\mathbb{E}M^2 < \infty$. Thus the corresponding latent-output limit is $\psi_A(Z + M) + G_A^\perp$. The mixing law may depend on the subsequence. For $q = \mathcal{L}(Z + M)$, along that subsequence,

$$D_{\mathrm{KL}}(q \| \gamma_1) \le \liminf_N d_N. \tag{28}$$

Conversely, every square-integrable M admits a representation obeying (26) *and attaining this joint limit, with prior KL exactly $D_{\mathrm{KL}}(q\|\gamma_1)$ at every width. If $L_N \to \infty$ and $\log L_N = o(\sqrt{N})$, finite mixtures with at most L_N components also attain the limit, with prior KL converging to $D_{\mathrm{KL}}(q\|\gamma_1)$.*

Proof of the limiting-law assertion. Let $g_x(b) = \varphi(b^T x) - \psi(x)$, and let c_B clip a scalar to $[-B, B]$. The separably centered bounded single-neuron observable is

$$Y^B(a,b) = \left(c_B(a),\, [c_B(a)\{c_B(g_x(b)) - \mathbb{E}_U c_B(g_x(b))\}]_{x \in A}\right).$$

It is prior-centered. Its prior covariance is $\mathrm{diag}(v_B, \Sigma_B^\perp)$, with $v_B = \mathbb{E}c_B(a)^2 \to 1$ and $\Sigma_B^\perp \to K_A^\perp$. The cross-covariance is exactly zero because each incoming factor is prior-centered. Independence of a, b and their Gaussian moments imply $Y^B \to (a, [a g_x(b)]_{x \in A})$ in prior L^2.

For component j, let $D_j = D_{\mathrm{KL}}(Q_N(\cdot \mid j)\|U_N)$. The entropy chain rule and conditional factorization give

$$\int D_j \, \mathbb{P}_{J_N}(dj) = d_N + i_N, \qquad D_j = \sum_i (d_{ij}^a + d_{ij}^b),$$

where the latter terms are the outgoing and incoming prior KL values. These identities also hold for the stated standard-Borel labels; finite $d_N + i_N$ makes the component KL finite almost everywhere. Both clipped factors are centered under their own priors. Gaussian concentration and entropy duality bound their component means by $\sqrt{2d_{ij}^a}$ and $\mathrm{Lip}(\varphi)\|x\|\sqrt{2d_{ij}^b}$, respectively. They remain independent within a component. Hence, writing $D_N^B(A)$ for the residual coordinates of $N^{-1/2}\sum_i Y_i^B$,

$$\int \left\|\mathbb{E}[D_N^B(A) \mid J_N = j]\right\| \mathbb{P}_{J_N}(dj) \le C_A \frac{d_N + i_N}{\sqrt{N}} \longrightarrow 0, \tag{29}$$

uniformly in B, where one may take $C_A = \mathrm{Lip}(\varphi)\sum_{x\in A}\|x\|$.

Apply the bounded-observable replacement in the proof of Lemma 5 to the joint vector Y^B. Its hypotheses follow from the same chain rule: choose $\eta_N \downarrow 0$ with $(d_N + i_N)/(N\eta_N) \to 0$ and discard components with $D_j > N\eta_N$, whose mass vanishes. At fixed B, the proof's Pinsker and Lindeberg estimates compare the retained mixture with a Gaussian of covariance $\mathrm{diag}(v_B, \Sigma_B^{\perp})$ plus its component-mean law, uniformly over aggregate means and allowing singular covariance. Equation (29) removes the residual coordinates of that mean law. The comparison therefore has the form

$$(\sqrt{v_B}Z + M_{N,B}, G_B^{\perp}), \qquad G_B^{\perp} \sim \mathcal{N}(0, \Sigma_B^{\perp}),$$

with all three variables independent.

Clipping is removed by the same joint argument in the lemma. Prior centering, independence across neurons and $Y^B \to (a, [ag_x]_{x\in A})$ in prior L^2 give

$$\sup_N \mathbb{E}_U \left\| (A_N, D_N(A)) - N^{-1/2}\sum_i Y_i^B \right\|^2 \longrightarrow 0.$$

The bounded global d_N transfers this error to Q_N-probability and ensures tightness. The independent shift laws are tight because $\mathbb{P}(|M| > 2t)\mathbb{P}(|G| \le t) \le \mathbb{P}(|M + G| > t)$. Taking width subsequences at each fixed clipping radius and then sending $B \to \infty$ gives (27), including independence of the residual. Applying replacement to the joint vector is essential; separate marginal limits would not establish that independence.

Finally, A_N is exactly standard normal under U_N at every width. Entropy duality at $0 < t < 1/2$ yields

$$t\mathbb{E}_Q A_N^2 \le d_N - \tfrac{1}{2}\log(1 - 2t).$$

Data processing to A_N and lower semicontinuity give (28). The second-moment bound passes as an inequality to the weak limit. Conditioning on M gives $\mathbb{E}(Z + M)^2 = 1 + \mathbb{E}M^2$, so M is square-integrable. These steps do not assert convergence of competitor output covariances. □

All Gaussian residuals are understood on their actual covariance supports. A prior-null linear combination of the original observables vanishes under every finite-prior-KL law by absolute continuity. The argument thus includes zero features and singular $K_A^{\perp}$. To discuss observations and future queries together, apply it to their fixed finite union. It does not assert process-level convergence on an unbounded or width-dependent query set.

F.2 Exact construction and finite-mixture approximation

Draw $M \sim \pi$ with $\mathbb{E}M^2 < \infty$ and set

$$a_i = \xi_i + M/\sqrt{N}, \qquad \xi_i \overset{\mathrm{iid}}{\sim} \gamma_1,$$

leaving every incoming scalar independently at prior, also independently of M and the ξ_i. Conditional on M, this is a product over all original scalar weights. Denote its weight distribution by $S_{\pi,N}$ and put $q = \gamma_1 * \pi$. Directly dividing each shifted Gaussian density by its prior density gives

$$\frac{dS_{\pi,N}}{dU_N} = r_\pi(A_N), \qquad r_\pi(a) = \frac{dq}{d\gamma_1}(a) = \int e^{am - m^2/2}\pi(dm). \tag{30}$$

This is an exactly normalized finite-width density, because the prior law of A_N is exactly γ_1. Its orthogonal outgoing coordinates and its incoming coordinates retain their prior conditional law. The conditional Gaussian KL and the entropy chain rule give

$$\begin{aligned} D_{\mathrm{KL}}(S_{\pi,N}\|U_N) &= D_{\mathrm{KL}}(q\|\gamma_1), \\ I(M;\theta_{1:N}) &= \tfrac{1}{2}\mathbb{E}M^2 - D_{\mathrm{KL}}(q\|\gamma_1) = I(M;Z+M) < \infty. \end{aligned} \tag{31}$$

These are representation-information identities, not an assertion that this representation minimizes information at every finite width.

For each fixed finite A, write

$$F_N(A) = F_N^0(A) + MB_N(A), \qquad B_N(A) = N^{-1}\sum_i [\varphi(b_i^T x)]_{x\in A}.$$

The prior joint CLT for $(N^{-1/2}\sum_i \xi_i, F_N^0(A))$, together with its fixed second moments, gives convergence in W_2 to $(Z, \psi_A Z + G_A^{\perp})$. Also $B_N(A) \to \psi_A$ in L^2. Since M is independent of all prior variables,

$$\mathbb{E}\|M(B_N(A)-\psi_A)\|^2 = \mathbb{E}M^2\,\mathbb{E}\|B_N(A)-\psi_A\|^2 \longrightarrow 0.$$

Couple M unchanged through the prior convergence. This proves the converse in Proposition 11, and gives W_2 convergence for this construction. Expectations of continuous losses of at most quadratic growth therefore converge under it. Finite-support π gives a literal finite mixture; general π uses the stated continuous label.

Finite mixtures with the same limiting cost. Couple finite-support quantizations M_k to M with $\mathbb{E}|M_k - M|^2 \to 0$, and put $q_k = \mathcal{L}(Z + M_k)$, $q = \mathcal{L}(Z + M)$. Data processing for the coupled Gaussian channels gives

$$D_{\mathrm{KL}}(q_k\|q) \le \tfrac{1}{2}\mathbb{E}|M_k - M|^2 \to 0, \qquad q_k \longrightarrow q \quad \text{in } W_2.$$

Choose $R < \infty$ with $p = \mathbb{P}(|M| \le R) > 0$. Equation (30) gives $\log p - R^2/2 - R|a| \le \log r_\pi(a) \le a^2/2$. Thus $\log r_\pi$ is continuous with at most quadratic absolute growth, and

$$D_{\mathrm{KL}}(q_k\|\gamma_1) = D_{\mathrm{KL}}(q_k\|q) + \mathbb{E}_{q_k}\log r_\pi \longrightarrow D_{\mathrm{KL}}(q\|\gamma_1).$$

A sufficiently slow diagonal choice of these finite-support constructors fits any prescribed $L_N \to \infty$. Its information is at most $\log L_N$, its prior KL converges as stated, and the joint law converges. This proves the finite-mixture converse. For the explicit constructors, W_2 convergence also gives convergence of expected continuous losses of at most quadratic growth. It makes no second-moment assertion for arbitrary competitors.

F.3 A ROBUST-LIKELIHOOD BOUNDARY

The distinction between a mean shift and a spread is visible in an ordinary nonconvex likelihood. Consider the one-input ReLU network with prior output variance $K = 1/2$ and mean feature $\psi = 1/\sqrt{2\pi}$. Two responses are $+4$ and -4, with a Student-t observation model having three degrees of freedom and scale $1/2$. Up to its common normalizer, the loss is

$$V(f) = 2\log\left(1 + \frac{(f-4)^2}{3/4}\right) + 2\log\left(1 + \frac{(f+4)^2}{3/4}\right).$$

For a deterministic outgoing-mean shift m, let

$$f(m) = \tfrac{1}{2}m^2 + \mathbb{E}V(\Gamma + \psi m), \qquad \Gamma \sim \mathcal{N}(0, 1/2).$$

The objective is even, and direct differentiation gives $f''(m) \ge 1 - 2/(3\pi) > 0$. Hence $m = 0$ is the unique global optimum among all single shifts.

If instead $M \sim \mathcal{N}(0, s-1)$ with $s > 1$, the output variance becomes $1/2 + (s-1)/(2\pi)$ and the KL to the prior is $(s - 1 - \log s)/2$. Gaussian differentiation gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=1}\left[\tfrac{1}{2}(s - 1 - \log s) + \mathbb{E}V\left(\sqrt{\tfrac{1}{2} + \tfrac{s-1}{2\pi}}\,Z\right)\right] = \frac{\psi^2}{2}\mathbb{E}V''(\Gamma) < 0.$$

For the sign, put $c = 3/4$ and $A = 8(36 - c)/(36 + c)^2$. On $|\Gamma| \le 2$, both absolute residuals lie in $[2, 6]$, and $V''(\Gamma) \le -A$. Everywhere $V'' \le 32/3$. The Gaussian Mills bound gives $\mathbb{P}(|\Gamma| > 2) \le \epsilon = e^{-4}/(2\sqrt{\pi})$, so $\mathbb{E}V''(\Gamma) \le -A(1 - \epsilon) + (32/3)\epsilon < 0$. This random outgoing-mean shift has label information $\frac{1}{2} \log s$, independent of width. It therefore improves this nonconvex objective using constant information, without contracting the prior residual, even though no deterministic mean shift does. This example does not extend the convex square-root theorem to nonconvex losses.

G Conditional prediction and marginal calibration

This appendix studies what remains unresolved after each response has the correct predictive distribution. The two responses share one neural weight draw, so their joint law determines how observing one should change the forecast of the other. We condition a fixed joint distribution; sequential posterior refitting tests a different operation (Pituk et al., 2025).

Task-dependent effects of lowering posterior temperature are already established (Odgers et al., 2026); here restrictions on attainable neural output laws constrain conditional prediction after exact marginal calibration.

G.1 A training observation that changes the sign of dependence

Retain the Gaussian weight prior and network normalization in Section 2, with $d = 2$, $\varphi(t) = t_+$, and no bias. Observe $y = 0$ at $u = (e_1 + e_2)/\sqrt{2}$ with noise variance $\tau > 0$. For the future pair, use

$$Y_j = F_N(e_j) + \varepsilon_j, \qquad \varepsilon_1, \varepsilon_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \gamma), \qquad \gamma > 0,$$

where the noises are independent of the shared weights. For two nonzero inputs with angle θ, the ReLU prior kernel is the order-one arc-cosine kernel with our weight normalization (Cho & Saul, 2009):

$$K(h, g) = \frac{\|h\|\|g\|}{2\pi}\{\sin\theta + (\pi - \theta)\cos\theta\}.$$

Thus, writing

$$a = \tfrac{1}{2}, \qquad c = \frac{1}{2\pi}, \qquad k = \frac{3\pi + 4}{8\pi\sqrt{2}}, \qquad \delta = \frac{k^2}{a + \tau}, \qquad v = a - \delta + \gamma, \qquad p = c - \delta, \tag{32}$$

the limiting posterior law of the noisy pair is

$$P = \mathcal{N}(0, C_P), \qquad C_P = \begin{pmatrix} v & p \\ p & v \end{pmatrix}. \tag{33}$$

It follows from the joint prior CLT and bounded Gaussian-likelihood reweighting. Positive future noise makes C_P positive definite. The posterior dependence is negative precisely when

$$0 < \tau < \tau_* := k^2/c - a = .396363938\ldots. \tag{34}$$

At $\tau = 1/10$, the sign $\delta > c$ is equivalent to $45\pi^2 - 72\pi + 80 > 0$; the quadratic has negative discriminant and positive leading coefficient. The sign does not depend on numerical rounding.

The geometry is simple in the common and difference directions $(1, 1)/\sqrt{2}$ and $(1, -1)/\sqrt{2}$. Conditioning subtracts $\delta \mathbf{1}\mathbf{1}^T$ from the latent prior covariance: it reduces common-direction variance by 2δ and leaves difference-direction variance unchanged. When the former falls below the latter, the off-diagonal becomes negative. Each marginal variance averages these two directional variances and does not reveal how the contraction was allocated.

G.2 Gaussian comparisons and marginal calibration

On the zero-label source, the full-Gaussian optimum in Appendix D.4 has zero limiting mean. Its outgoing-sum variance minimizes $h(s) + (s - 1)\beta/2$, where $h(s) = (s - 1 - \log s)/2$ and $\beta = \psi(u)^2/\tau = c/\tau$. At $\tau = \gamma = .1$, the minimizer is $s_* = (1 + 5/\pi)^{-1}$. An attaining Gaussian keeps incoming weights at prior and uses outgoing covariance $I + (s_* - 1)ee^T$, $e = \mathbf{1}/\sqrt{N}$, at exact prior KL $h(s_*) = .1690627389\ldots$ for every width. Its noisy limit has variance $w_* = a + \gamma + (s_* - 1)c$

and covariance s_*c. Scaling each already-noisy response by $\sqrt{v/w_*}$ matches the target marginals, with conditional slope

$$\kappa_* = s_*c/w_* = .1222738771\ldots > 0, \qquad \kappa_P = p/v = -.2170237825\ldots.$$

Optimality concerns the original variational objective, before calibration. The latter need not correspond to a weight distribution in the network.

The unchanged prior. For this zero response, the arbitrary whole-neuron reference is the prior. It has zero prior KL and zero scalar label information. Its noisy pair has variance $a + \gamma$ and covariance c. Scaling each response by $\sqrt{v/(a+\gamma)}$ again matches both target marginals, giving

$$C_{\mathrm{cal}} = \begin{pmatrix} v & q_{\mathrm{cal}} \\ q_{\mathrm{cal}} & v \end{pmatrix}, \qquad q_{\mathrm{cal}} = vc/(a+\gamma) > 0. \tag{35}$$

G.3 Reducing outgoing-weight variance

A different operation keeps incoming weights at prior and draws independent $a_i \sim \mathcal{N}(0, s)$. The choice

$$s = 1 - \delta/a \in (0, 1) \tag{36}$$

matches both limiting latent variances, hence the noisy marginals as well. Its noisy covariance is

$$C_Q = \begin{pmatrix} v & q \\ q & v \end{pmatrix}, \qquad q = sc > 0. \tag{37}$$

This scalar product pays prior KL $Nh(s) = .08492117203\ldots N$. Its negative-ELBO difference from the prior is $Nh(s) + a(s-1)/(2\tau)$, minimized at $s_N = N/(N + a/\tau) \to 1$. Fixed scalar variance reduction therefore leaves the bounded-KL regime. Scaling the latent output before adding noise gives this limit, whereas scaling the already-noisy prior gives (35). The two output operations should not be identified.

A common risk comparison. For any centered Gaussian forecast with marginal variance v and covariance q, the conditional slope is q/v. Its squared risk under P and excess over the exact conditional mean are

$$R(q) = v + (q^2 - 2pq)/v, \qquad R(q) - R_P = (q-p)^2/v, \qquad R_P = v - p^2/v. \tag{38}$$

Ignoring Y_1 sets $q = 0$ and gives risk v. Every $q > 0$ does worse, since $p < 0$. Equality of the first marginals makes joint forward KL equal to the target-averaged conditional forward KL:

$$D_{\mathrm{KL}}(P\|Q) = \frac{1}{2}\left\{\frac{v+p}{v+q} + \frac{v-p}{v-q} - 2 + \log\frac{v^2-q^2}{v^2-p^2}\right\}. \tag{39}$$

Its derivative in q is $(q-p)(v^2+q^2)/(v^2-q^2)^2 > 0$ for $0 \le q < v$. Independence thus also minimizes conditional log loss among the Gaussian laws with these marginals and nonnegative covariance, giving

$$D_{\mathrm{KL}}(P\|\mathcal{N}(0, vI_2)) = -\tfrac{1}{2}\log(1 - p^2/v^2) = .02412230064\ldots. \tag{40}$$

This last minimization is within the Gaussian family, not every calibrated law in the next proposition.

Table 2 evaluates these formulas at $\tau = \gamma = .1$ and includes the matched-cost feature-dependent construction developed in Appendix G.5. All rows are analytic limiting predictions, evaluated under P; they are not finite fitted endpoints.

G.4 Increasing marginal calibration of restricted limiting laws

The next result covers two distinct weight families. Its premises apply before output calibration; the transformed predictive distribution need not correspond to a weight distribution in the stated network model.

Table 2: **Conditional forecasts with the same individual marginals.** Analytic limits at $\tau = \gamma = .1$, with squared risk and conditional forward KL evaluated under P. The feature-dependent construction is defined in Appendix G.5.

Forecast	Slope	Squared risk	Conditional KL (nats)
Exact posterior	$-.217024$	.345171	0
Independent calibrated marginals	0	.362232	.024122
Feature-dependent covariance + calibration	$-.037831$	.356802	.016617
Full Gaussian + calibration	$+.122274$	.386872	.058707
Calibrated noisy prior	$+.265258$	.429425	.125254
Reduced outgoing-weight variance	$+.230435$	.417697	.105734

Proposition 12 (Proposition 4 with its equality case)**.** *Assume* (34)*. Let Q_N be a sequence with $\sup_N D_{\mathrm{KL}}(Q_N \| U_N) < \infty$ and either (i) $\mathcal{I}_N^S(Q_N) = o(\sqrt{N})$, with the infimum of component-label information $\mathcal{I}_N^S$ defined in Appendix E, or (ii) a jointly Gaussian distribution over all network weights. Choose any subsequential weak limit of its noisy pair (Y_1, Y_2). Let T_1, T_2 be deterministic nondecreasing coordinate maps taking the two marginals of this limit to $\mathcal{N}(0, v)$, and denote the transformed joint law by $\widetilde{Q}$. Its conditional mean $g(z) = \mathbb{E}_{\widetilde{Q}}[Y_2 \mid Y_1 = z]$ has a nondecreasing version and*

$$\mathbb{E}_P\{Y_2 - g(Y_1)\}^2 \geq v, \qquad \mathbb{E}_P\{Y_2 - g(Y_1)\}^2 - R_P \geq p^2/v. \tag{41}$$

Equality in the first inequality requires $g = 0$ almost surely. The independent law with the calibrated marginals, which ignores Y_1, attains this risk. In case (ii), $g(z) = \kappa z$ for some $\kappa > 0$, and both risk inequalities are strict.

Proof. For case (ii), data processing to the outgoing sum gives $D_{\mathrm{KL}}(Q_N \| U_N) \geq h(s_N)$, so s_N stays in a compact subset of $(0, \infty)$. The means are bounded. From any weakly convergent output subsequence, extract one with $s_N \to s > 0$ and mean $\mu_N \to \mu$. Proposition 8, with independent future noise, identifies its limit as

$$\mathcal{N}\big(\mu, (a + \gamma - c)I_2 + sc\mathbf{1}\mathbf{1}^T\big)\,.$$

Its common variance is $w_s = a + \gamma + (s-1)c > 0$ and covariance $sc > 0$. The unique nondecreasing map from either continuous Gaussian marginal to $\mathcal{N}(0, v)$ is $T_j(z) = \sqrt{v/w_s}(z - \mu_j)$ almost surely. The calibrated conditional mean is therefore $g(z) = \kappa_s z$, with $\kappa_s = sc/w_s > 0$. Directly,

$$\mathbb{E}_P(Y_2 - \kappa_s Y_1)^2 = v(1 + \kappa_s^2) - 2p\kappa_s > v,$$

which proves this case. No finite-width conditional limit is needed.

For case (i), Proposition 11 applies after choosing representations within one nat of $\mathcal{I}_N^S$. At the equal-norm inputs e_1, e_2, the feature means agree, so its common shift absorbs their common factor $\psi(e_j)$. The limiting noisy pair therefore has the form

$$Z = (G_1 + M, G_2 + M), \qquad G \sim \mathcal{N}\left(0, \begin{pmatrix} a+\gamma & c \\ c & a+\gamma \end{pmatrix}\right), \qquad M \perp G, \quad \mathbb{E}M^2 < \infty. \tag{42}$$

Put $A_0 = a + \gamma$ and $\rho = c/A_0 \in (0, 1)$. A Gaussian regression decomposition gives

$$Z_2 = \rho Z_1 + (1 - \rho)M + \eta,$$

where η is Gaussian and independent of (M, Z_1). Writing π for the law of M, its conditional law given $Z_1 = x$ is proportional to $\exp\{-(x - m)^2/(2A_0)\}\pi(dm)$. For $x' > x$, the likelihood ratio of these conditional laws is a positive constant times $\exp\{(x' - x)m/A_0\}$, an increasing function of m. Thus $M \mid Z_1 = x$ increases in stochastic order with x. The coefficients ρ and $1 - \rho$ are nonnegative, and independent addition of η preserves that order. Consequently $Z_2 \mid Z_1 = x$ is stochastically nondecreasing in x.

The source marginals have strictly positive continuous densities by Gaussian convolution. A deterministic nondecreasing map from either one to a nondegenerate Gaussian agrees almost surely with

its strictly increasing quantile map. In particular, T_1 is invertible almost surely. Conditioning through T_1 and applying increasing T_2 therefore preserve the stochastic order. The conditional mean g is nondecreasing. Moreover Jensen's inequality gives $\mathbb{E}_{\widetilde{Q}_1} g(Y_1)^2 \le v$. The equality $\widetilde{Q}_1 = P_1 = \mathcal{N}(0, v)$ gives the same integrability under the target first marginal.

Under P, $\mathbb{E}[Y_2 \mid Y_1] = (p/v)Y_1$, hence

$$\mathbb{E}_P\{Y_2 - g(Y_1)\}^2 = v + \mathbb{E}_P g(Y_1)^2 - 2(p/v)\mathbb{E}_P[Y_1 g(Y_1)].$$

For an independent copy Y_1',

$$2\,\mathrm{Cov}_P(Y_1, g(Y_1)) = \mathbb{E}_P[(Y_1 - Y_1')\{g(Y_1) - g(Y_1')\}] \ge 0.$$

Since $\mathbb{E}_P Y_1 = 0$ and $p < 0$, both terms added to v are nonnegative. This proves the bound and its equality condition. □

Case (i) allows arbitrary, non-Gaussian, nonidentical factors in the original scalar parameterization; case (ii) imposes no label-information bound. Neither requires near-optimality or a likelihood assumption on Q_N. The conclusion concerns calibrated limiting laws. It does not cover arbitrary whole-neuron shifts, decreasing maps, joint dependence repairs, or all finite-width conditional Bayes predictors. Independence is a further joint operation, and a repair that constructs the required negative dependence can improve on it.

G.5 Changing the covariance direction at matched prior KL

The conditional readout in (10) contracts along the observed features. To separate that direction from the amount of contraction, put $r_i = (b_i^T u)_+/\sqrt{N}$, $t = \|r\|^2$, and

$$d_N(b) = \begin{cases} r/\sqrt{t}, & t > 0, \\ e, & t = 0, \end{cases} \qquad e = \mathbf{1}/\sqrt{N}.$$

Keep incoming weights at their prior. For $0 < s \le 1$, compare the fixed Gaussian outgoing covariance $I + (s-1)ee^T$ with the conditional covariance $I + (s-1)d_N(b)d_N(b)^T$. Denote the two weight distributions by $Q^G_{N,s}$ and $Q^F_{N,s}$. Their entire mean functions are zero and, by the conditional KL identity,

$$D_{\mathrm{KL}}(Q^G_{N,s}\|U_N) = D_{\mathrm{KL}}(Q^F_{N,s}\|U_N) = h(s) = \tfrac{1}{2}(s - 1 - \log s). \tag{43}$$

The identity holds at every width: both conditional covariances have eigenvalues $s, 1, \dots, 1$. The fallback on $t = 0$, an event of probability 2^{-N}, is needed for exact cost equality.

Let $B = k^2/a$, with a, c, k from (32), and let $\Phi_N \in \mathbb{R}^{N\times 2}$ have columns $((b_i^T e_j)_+/\sqrt{N})_{i=1}^N$, $j = 1, 2$. The feature Gram law of large numbers gives the noisy limit of $Q^F_{N,s}$:

$$C_F(s) = \begin{pmatrix} w_F & q_F \\ q_F & w_F \end{pmatrix}, \qquad w_F = a + \gamma + (s-1)B, \quad q_F = c + (s-1)B.$$

Indeed, the conditional covariance is the empirical query Gram plus $(s-1)\Phi_N^T d_N d_N^T \Phi_N + \gamma I$, which converges almost surely to $C_F(s)$. Bounded conditional characteristic functions give distributional convergence. The fourth moments are bounded by $3/4 + 15/(4N) + 3\gamma + 3\gamma^2$, so second moments also converge. Calibrating the already-noisy limit gives slope $\kappa_F = q_F/w_F$.

The Gaussian comparison covers the whole family. At a fixed prior-KL bound C on the weight distribution before calibration, data processing to $N^{-1/2}\sum_i a_i$ gives $h(s_N) \le C$. If $s_-(C) \le 1$ is the smaller root of $h(s) = C$, every calibrated full-Gaussian limit in Proposition 8 has slope

$$\kappa_G(s) = \frac{cs}{a + \gamma - c + cs}, \qquad s \ge s_-(C).$$

Both this slope and its target risk $R(\kappa) = R_P + v(\kappa - p/v)^2$ increase with s in this range. The fixed-direction construction at $s_-(C)$ therefore attains the smallest calibrated limiting risk among all full Gaussians under that ceiling.

At $s = s_* = \pi/(\pi + 5)$, both costs are .169063 nats. The feature-dependent slope is $-.037831$, with squared risk .356802 and conditional KL .016617 nats. These improve on independence (.362232 and

.024122) and on the best Gaussian under the same ceiling (.386872 and .058707); see Figure 3. The prior-KL charge concerns the original weight distributions, before calibration. No weight realization of the calibrated outputs at that same cost is asserted.

The established conditional readout has covariance $I - rr^T/(\tau + t)$ and prior KL $\frac{1}{2}\mathbb{E}\{\log(1 + t/\tau) - t/(t + \tau)\}$. For this model, the prior KL tends to .479213 nats. The predictive limit is P, as given by the conditional-readout construction of Hron et al. (2022). The constant-eigenvalue construction instead reaches the same limit at $s = \tau/(a + \tau) = 1/6$, without being the same finite weight distribution. Its slope is zero at $s = 1 - c/B$; between that value and $1/6$, it improves on independence. This interval is specific to the construction, not a universal cost threshold.

G.6 Full marginal calibration at finite width

At widths 32 and 256, we integrate over all 8,192 existing incoming-weight draws per width. Let $A = \Phi_N^T \Phi_N$ be the query Gram and $z = \Phi_N^T r$ the query–training cross-Gram column. The fixed and feature constructions have noisy component covariances $A + (s_* - 1)mm^T + \gamma I$ and $A + (s_* - 1)zz^T/t + \gamma I$, respectively, where $m = \Phi_N^T e$ and the second uses mm^T when $t = 0$. Their incoming weights are integrated under the prior. The exact posterior uses component covariance $H = A - zz^T/(\tau + t) + \gamma I$, weighted by $(1 + t/\tau)^{-1/2}$. The ordinary conditional readout uses the same H with prior incoming weights.

For each resulting candidate mixture Q_M and target mixture P_M, apply $T_j = F^{-1}_{P_M,j} \circ F_{Q_M,j}$ to the already-noisy response. These maps match the full numerical-mixture marginals, not only their variances. We compute the actual nonlinear conditional mean after this transformation. Calibration uses the target marginals; it is an oracle comparison. Averaging each incoming state with its coordinate swap enforces the model's exchange symmetry without adding independent draws.

Table 3: **Finite conditional forecasts after full marginal calibration.** Differences in squared risk, with one block-jackknife numerical standard error. Negative values favor the first forecast.

N	Fixed − ignore	Feature − ignore	Feature − fixed
32	$.021676 \pm .000238$	$-.002471 \pm .000076$	$-.024147 \pm .000204$
256	$.024288 \pm .000066$	$-.004999 \pm .000033$	$-.029286 \pm .000066$

The approximate simultaneous 95% intervals are centered at estimates using all 8,192 draws, without bias correction. Their half-width is the block-jackknife standard error times the $1 - .05/12$ quantile of Student's t distribution with 15 degrees of freedom: a two-sided Bonferroni adjustment for three contrasts at two widths. All six exclude zero; both unsymmetrized query directions retain these signs. At widths 32 and 256, the ordinary readout has risks .335016 and .343968, versus .334993 and .343967 for the posterior conditional mean. Its original KL differs from the two matched constructions.

Each block-jackknife deletion recomputes the posterior weights, marginal maps and nonlinear forecasts. We divide all 8,192 draws into sixteen blocks of 512. Additional checks use eight-block halves and paired deletions from disjoint halves. Independent integration directly over response values reproduces all 35 reported statistics, including squared risks and risk contrasts, for each of the 110 retained configurations within 7.9×10^{-14}. These checks reuse the same incoming draws and give no rigorous quadrature error enclosures.

We bound truncation error separately. Let v_P be the finite posterior's noisy marginal variance. Its fourth moment is bounded by $M_4 = \sqrt{6}(1.08 + 3.75/N)$, and $v_P \le .6\sqrt{6}$. Conditional Jensen and Cauchy–Schwarz bound the outer risk tail at marginal probability ϵ by $3\sqrt{M_4\epsilon}$. The inner omission has L^2 bound $d = (M_4\epsilon)^{1/4}$ and risk allowance $4\sqrt{v_P}d + d^2$. Taking the larger of the population bounds and empirical mixture moments, the selected normal-quantile cutoffs 7.5 and 12 give outer allowances below 1.30×10^{-6} and inner allowances below 4.90×10^{-8} per law. For a risk difference, we add the allowances for both laws. These bounds do not cover error from the sampled incoming-weight integral or floating-point arithmetic.

G.7 BINARY CONDITIONAL DECISIONS

Keep the network, Gaussian prior and queries e_1, e_2, u, but observe one positive and one negative label at u with logistic link $f_\beta(t) = \mathrm{sigmoid}(\beta t)$. The same fixed scale β applies to training and future labels; no Gaussian output noise or marginal calibration is used. Future labels are independent conditional on the weights. The likelihood is $\ell_\beta(F_N(u)) = f_\beta(F_N(u))f_\beta(-F_N(u))$. Outgoing sign reversal preserves this likelihood and each centered candidate. Every single-label marginal is therefore exactly $1/2$.

Let $r_P = \mathbb{E}_P[(2Y_1 - 1)(2Y_2 - 1)]$ and define r_Q similarly for a candidate. The complete pair table and its conditional log loss are

$$P(11) = P(00) = (1 + r_P)/4, \qquad P(10) = P(01) = (1 - r_P)/4, \tag{44}$$

$$\mathcal{L}(r_P, r_Q) = -\tfrac{1+r_P}{2} \log \tfrac{1+r_Q}{2} - \tfrac{1-r_P}{2} \log \tfrac{1-r_Q}{2}. \tag{45}$$

The candidate repeats the first label if $r_Q > 0$ and predicts its complement if $r_Q < 0$; its classification error is $(1 - \mathrm{sign}(r_Q)r_P)/2$. A fixed tie rule has error $1/2$. Correct sign gives the Bayes decision, not necessarily the Bayes probability.

For the finite posterior, with latent query outputs X_1, X_2 and training output U, compute

$$e_\beta = \mathbb{E}[\ell_\beta(U) \mid b], \qquad n_\beta = \mathbb{E}[\ell_\beta(U) \tanh(\beta X_1/2) \tanh(\beta X_2/2) \mid b], \qquad r_P = \frac{\mathbb{E}_b n_\beta}{\mathbb{E}_b e_\beta}.$$

The ratio uses separately averaged numerator and denominator, retaining the incoming posterior reweighting. The two candidates are Q^G_{N,s_*} and Q^F_{N,s_*} from (43); s_* was retained from the regression comparison, not optimized for logistic likelihood. Their exact prior KL remains .169063 nats and their mean functions remain zero before observing the first future label.

Table 4: **The useful conditional direction changes with the likelihood scale.** Both specified approximations have exact half-probability single-label marginals and prior KL .169063. Ignoring the first label gives error 50% and log loss $\log 2$. Log-loss differences are in 10^{-3} nats, followed by approximate simultaneous 95% block-jackknife intervals, rounded outward. All eight intervals exclude zero. These are numerical finite-posterior comparisons, not fitted or finite all-family optima.

			Error (%)		Log loss − ignoring (10^{-3} nats)	
N	β	r_P	Gaussian	Feature	Gaussian	Feature
32	1	$+0.02121$	48.94	51.06	-0.190 $[-0.197, -0.183]$	$+0.038$ $[+0.034, +0.042]$
32	4	-0.05624	52.81	47.19	$+6.063$ $[+5.827, +6.299]$	-0.450 $[-0.507, -0.394]$
256	1	$+0.02174$	48.91	51.09	-0.197 $[-0.199, -0.195]$	$+0.076$ $[+0.074, +0.077]$
256	4	-0.06989	53.49	46.51	$+6.870$ $[+6.802, +6.938]$	-1.123 $[-1.148, -1.098]$

All four width–scale cells use sixteen blocks of 512 saved incoming draws. For their eight forecast-minus-independence contrasts, the same interval rule uses the $1 - .05/16$ quantile. Gaussian integration conditional on the training output includes the $t = 0$ case, where its likelihood is $1/4$ and the query covariance remains A. Both query integration orders and all prescribed refinements were retained. An independent Fourier calculation uses every incoming state; the largest aggregate correlation and log-loss discrepancies are below 3.7×10^{-9} and 2.4×10^{-10}, respectively. One incoming state has a larger Hermite discrepancy (3.24×10^{-7} in its signed numerator); no states were dropped. Block deletions rebuild the entire evidence ratio and decision scores. The approximate family intervals describe integration uncertainty on these previously exposed draws, not rigorous population coverage. At scale one the fixed Gaussian is better; at scale four the feature direction is better. At width one the two specified constructions coincide, so these results do not assert a useful update at every width or a finite all-Gaussian optimum.

G.8 FIXED-SIZE GAUSSIAN MIXTURES AT THE VARIATIONAL OPTIMUM

The binary example also allows a comparison with a richer family. Let $\mathcal{M}_{N,L}$ contain all mixtures of at most L nondegenerate joint Gaussian distributions over the weights, with arbitrary means,

covariances and mixture weights. The following result uses the full-Gaussian fluctuation law in Proposition 8. For a candidate Q_N, let $\mathcal{L}(Q_N)$ be the expected negative log probability it assigns to Y_2 after observing Y_1, with expectation under the exact finite posterior predictive distribution.

Proposition 13 (Fixed Gaussian mixtures and conditional prediction)**.** *Fix $L \geq 1$, incoming dimension and $\beta > 0$ in the preceding binary model. Let $Q_N \in \mathcal{M}_{N,L}$ satisfy $G_N(Q_N) \leq \inf_{Q \in \mathcal{M}_{N,L}} G_N(Q) + \epsilon_N$, where the absolute tolerance $\epsilon_N \to 0$. There is $\Delta_{L,\beta} > 0$, independent of the sequence, such that*

$$\liminf_N \mathrm{Cov}_{Q_N}(Y_1, Y_2) \geq \Delta_{L,\beta} > 0.$$

If the exact posterior's label correlation tends to $r_\beta < 0$, then its target-averaged conditional log loss satisfies

$$\liminf_N \{\mathcal{L}(Q_N) - \log 2\} \geq 4|r_\beta|\Delta_{L,\beta} > 0.$$

Here $\log 2$ is the loss of the correct marginal forecast that ignores Y_1. The approximate marginals need not be correct.

Proof. Rotate the incoming coordinates into the training direction u and its orthogonal complement. Write $X = (a_i, b_i^T u)_i$ and call the remaining coordinates W. The prior factorizes as $U_X U_W$ and the likelihood depends only on X. Projection $\Pi Q = Q_X U_W$ preserves the Gaussian mixture family and its component count, while the KL chain rule gives

$$G_N(Q) - G_N(\Pi Q) = D_{\mathrm{KL}}(Q \| \Pi Q).$$

Thus $\mathrm{TV}(Q_N, \Pi Q_N) \leq \sqrt{\epsilon_N/2}$. The prior is feasible and has bounded objective, so both distributions have prior KL at most a fixed C eventually.

Write $\Pi Q_N = \sum_j p_{j,N} G_{j,N}$. The mixture identity yields $\sum_j p_j D_{\mathrm{KL}}(G_j \| U_N) \leq C + \log L$. For fixed $0 < \eta < 1/L$, components with $p_j \geq \eta$ therefore lie on the fixed Gaussian KL ball $C_\eta = (C + \log L)/\eta$. The other components have total mass at most $L\eta$ and binary covariance at least $-1/4$ each.

Every projected component is invariant under reflection in the $e_1 - e_2$ direction. Its two query marginals agree exactly at finite width. Its between-component covariance contribution is consequently a variance. Proposition 8 approximates each retained component by

$$Z_\ell = m + \sqrt{cs}\, Z_0 + \sqrt{a - c}\, Z'_\ell, \quad \ell = 1, 2,$$

with independent standard normals, exact common mean m, and $s = \mathrm{Var}(N^{-1/2} \sum_i a_i) > 0$. The covariance of $f_\beta(Z_1), f_\beta(Z_2)$ is strictly positive: it is the variance of the strictly increasing convolved sigmoid evaluated at $m + \sqrt{cs} Z_0$.

At least one component has weight at least $1/L$ and KL at most $L(C + \log L)$. The Gaussian mean bound in Appendix D.3 and the KL spectral bounds put (m, s) in a fixed compact set with s bounded away from zero. The continuous positive covariance has a minimum $\kappa > 0$ there. The sigmoid and its coordinate product are bounded Lipschitz functions. If $\tau_N(C_\eta) \to 0$ denotes the uniform Wasserstein error, total covariance and the projection bound give

$$\mathrm{Cov}_{Q_N}(Y_1, Y_2) \geq \kappa/L - \frac{\beta(2 + \sqrt{2})}{4} \tau_N(C_\eta) - L\eta/4 - 3\sqrt{\epsilon_N/2}.$$

Take the width limit at fixed η, then let $\eta \downarrow 0$.

For the actual conditional forecasts $q_+ = Q_N(Y_2 = 1 \mid Y_1 = 1)$ and $q_- = Q_N(Y_2 = 1 \mid Y_1 = 0)$, $q_+ - q_- = \mathrm{Cov}_Q(Y_1, Y_2)/\mathrm{Var}_Q(Y_1)$. The strict covariance bound makes this difference eventually at least $4\,\mathrm{Cov}_Q(Y_1, Y_2)$. The convex tangent of Bernoulli cross-entropy at forecast $1/2$ gives $\mathcal{L}(Q_N) - \log 2 \geq -r_N(\beta)(q_+ - q_-)$. This proves the second assertion without calibration or a continuity argument for an unbounded log loss. □

Increasing the shared logistic scale concentrates the training logit near zero and sharpens the future sigmoid links, exposing a negatively correlated Gaussian residual. In the width-limit posterior, the

training output U has density proportional to $\varphi_a(U)\ell_\beta(U)$, where φ_a is the centered normal density with variance a. Conditional on U, the query pair has common mean $(k/a)U$ and residual covariance

$$R = \begin{pmatrix} a-B & c-B \\ c-B & a-B \end{pmatrix}, \qquad B = k^2/a.$$

Its off-diagonal is negative because $B-c = (9\pi^2-8\pi+16)/(64\pi^2) > 0$. Rescaling $T = \beta U$ leaves an integrable density dominated by $\varphi_a(0)\,\mathrm{sigmoid}(T)\,\mathrm{sigmoid}(-T)$. The signed-label product converges to the product of the residual Gaussian signs. Dominated convergence therefore gives

$$\lim_{\beta\to\infty} r_\beta = \frac{2}{\pi}\arcsin\frac{c-B}{a-B} < 0.$$

Hence every sufficiently large fixed scale has a negative target correlation. The same change of variables makes the balanced training-pair probability of order $1/\beta$; this observation pattern becomes rarer as the scale grows. Independent quadrature at the displayed scale four gives $r_4 = -.0719342628835384$, but does not rigorously enclose its sign or locate the analytic threshold. Width is taken to infinity with scale fixed. The result gives neither a growing-L rate nor a useful finite onset or numerical value of $\Delta_{L,\beta}$.

H THE CONDITIONAL-READOUT COMPARISON

Let $b = (b_1, \ldots, b_N)$ collect the incoming weights, with prior U_b, and let P_N^b be their marginal under the posterior. Retain the feature matrix Ψ from (10) and set $\widehat{K} = \Psi\Psi^T$. Drawing b from U_b and the outgoing weights from the exact conditional in (10) defines Q_N^{cond}. The conditional evidence is

$$Z_b = \frac{|\Sigma|^{1/2}}{|\Sigma+\widehat{K}|^{1/2}}\exp\Big[-\tfrac{1}{2}y^T(\Sigma+\widehat{K})^{-1}y\Big].$$

Since its readout conditional is exact, the KL chain rule gives

$$D_{\mathrm{KL}}(Q_N^{\mathrm{cond}}\|P_N) = D_{\mathrm{KL}}(U_b\|P_N^b) = \log\mathbb{E}_{U_b}Z_b - \mathbb{E}_{U_b}\log Z_b \longrightarrow 0.$$

Indeed $\widehat{K} \to K_{HH}$ in probability, $0 < Z_b \le 1$, and

$$-\log Z_b \le \tfrac{1}{2}\,\mathrm{tr}(\Sigma^{-1}\widehat{K}) + \tfrac{1}{2}y^T\Sigma^{-1}y.$$

Finite feature fourth moments give uniform integrability of the right side, proving both expectation limits. This is the conditional-readout result of Hron et al. (2022, Theorem 2.1). The construction does not require a Gaussian joint distribution or an explicit mixture of scalar-factorized components. The exactly mean-preserving finite intervention evaluated in this paper is defined separately in Appendix K.

I FINITE VARIANCE BOUNDS AND SEMI-IMPLICIT MIXTURES

The finite bound used in Section 6 depends on the KL to the prior and the information carried by the mixture label. We first give this bound for whole-neuron mixtures, then relate its cost to a sampled semi-implicit objective. The network follows Section 2.

I.1 A FINITE LOWER BOUND ON LATENT-OUTPUT VARIANCE

We bound latent variance at finite width without assuming optimization. At a fixed query x, put $K_x = K(x,x) > 0$ and define the prior transform and its variance bound by

$$\Lambda_x(t) = \mathbb{E}_b\left(1 + 2t\frac{\varphi(b^Tx)^2}{K_x}\right)^{-1/2}, \qquad \mathcal{V}_x(a) = \sup_{t>0}\frac{-a-\log\Lambda_x(t)}{t}. \tag{46}$$

The same result holds for any measurable incoming feature $v_x(b)$ with $0 < \mathbb{E}v_x(b)^2 < \infty$ and a local absolute exponential moment of $av_x(b)$ under the independent standard-Gaussian readout prior. The present Lipschitz/Gaussian model has this moment automatically.

Proposition 14 (Finite variance from the total component cost)**.** *Let $\Xi \sim \pi$ be a label on a standard Borel space, with a measurable conditional distribution $\bigotimes_i q_{i,\xi}$ over network weights Θ_N. Its marginal is $Q = \int \bigotimes_i q_{i,\xi}\, \pi(d\xi)$. For this whole-neuron product representation, put $d = D_{\mathrm{KL}}(Q\|U_N)$, $i = I(\Xi;\Theta_N)$. If $d + i < \infty$, then at every width*

$$\mathrm{Var}_Q\, F_N(x) \geq K_x \mathcal{V}_x((d+i)/N). \tag{47}$$

Proof. For the normalized prior neuron output $X = a\varphi(b^T x)/\sqrt{K_x}$, integrate out the Gaussian readout to get, for every real center m,

$$\mathbb{E}_U e^{-t(X-m)^2} = \mathbb{E}_b\left[(1+2tR)^{-1/2} e^{-tm^2/(1+2tR)}\right] \leq \Lambda_x(t), \qquad R = \varphi(b^T x)^2/K_x.$$

Entropy duality, centered at $\mathbb{E}_q X$, gives $D_{\mathrm{KL}}(q\|U_1) \geq -t\,\mathrm{Var}_q\, X - \log\Lambda_x(t)$. The local exponential moment supplies finite first moments for finite-KL laws; an infinite variance satisfies the lower bound trivially. Use the same t for every neuron and label. Conditional independence, total variance, and the component-KL chain rule yield

$$\frac{d+i}{N} \geq -t\frac{\mathrm{Var}_Q\, F_N(x)}{K_x} - \log\Lambda_x(t).$$

Taking the supremum proves (47). □

The underlying variance frontier is sharp in the *total* cost $d + i$ at one query. The centered whole-neuron tilt $dq_t/dU_1 = e^{-tX^2}/\Lambda_x(t)$ attains its dual value, using KL to the prior with $i = 0$. It generally couples scalar weights and depends on x. This does not supply a matching sufficient component count at bounded KL to the prior, sharpness for scalar-factorized or Gaussian components, or simultaneous attainment at several queries. Variance alone also gives no finite predictive-KL bound: rare large tails can retain variance while forward KL from a light-tailed target vanishes.

I.2 From a semi-implicit objective to the marginal distribution

Hierarchical variational models represent dependence by placing a distribution over the parameters of a simpler family (Ranganath et al., 2016). The number of distributions used to evaluate a semi-implicit objective may differ from the number of components in the fitted marginal. The standard mixture-entropy identity connects these quantities to the neural limits above. The expected finite-mixture ELBO interpretation is established by Molchanov et al. (2019, Theorem 1); related mutual-information bounds are given by Sobolev & Vetrov (2019, Supplementary Appendix F).

The following mixture identity needs only a nonnegative likelihood energy. Let $\Xi \sim \pi_N$ be a label on a standard Borel space, and let $q_{N,\xi}$ be a measurable family of distributions that factorize over the individual weights in the original parameterization. The marginal weight distribution obtained by integrating out the label is $h_N = \int q_{N,\xi}\, \pi_N(d\xi)$. For L independent labels, define

$$\bar{q}_{N,L} = \frac{1}{L}\sum_{j=1}^{L} q_{N,\Xi_j}, \qquad S_{N,L} = \mathbb{E}_{\Xi_{1:L}} G_N(\bar{q}_{N,L}). \tag{48}$$

This is the expected negative semi-implicit objective whose density estimate includes the conditional distribution that generated the weight sample, up to the fixed likelihood normalization. Our L is $K + 1$ in Yin & Zhou (2018, Eq. (9)). Indeed, by exchangeability its value also equals

$$\mathbb{E}_{\Xi_{1:L}}\mathbb{E}_{\theta\sim q_{N,\Xi_1}}\left[\log\frac{d\bar{q}_{N,L}}{dU_N}(\theta) + V(F_N(H;\theta))\right].$$

Equation (48) gives an unambiguous nonnegative extended-value definition when the objective is infinite. The marginal h_N can have a continuous mixing distribution; it is not generally an L-component mixture.

Finite expected score controls the actual marginal. For each realized collection, write

$$j(\Xi_{1:L}) = \frac{1}{L}\sum_{k=1}^{L} D_{\mathrm{KL}}(q_{N,\Xi_k}\|\bar{q}_{N,L}), \qquad 0 \le j \le \log L.$$

The finite-mixture chain rule and the nonnegative energy give

$$\frac{1}{L}\sum_{k} G_N(q_{N,\Xi_k}) = G_N(\bar{q}_{N,L}) + j(\Xi_{1:L}). \tag{49}$$

This identity holds in the extended nonnegative sense. If $S_{N,L} < \infty$, Tonelli's theorem therefore gives $\mathbb{E}_\pi G_N(q_{N,\Xi}) = S_{N,L} + \mathbb{E}j < \infty$. The continuous-mixture chain rule is then justified without subtracting infinities. In the joint distribution $\Xi \sim \pi_N$, $\Theta_N \mid \Xi \sim q_{N,\Xi}$, it yields

$$\begin{aligned} D_{\mathrm{KL}}(h_N\|U_N) + I(\Xi;\Theta_N) &= S_{N,L} + \mathbb{E}j - \mathbb{E}_{h_N} V(F_N(H)) \\ &\le S_{N,L} + \log L. \end{aligned} \tag{50}$$

Convexity of KL and linearity of expected energy also give $G_N(h_N) \le S_{N,L}$. The information in the actual hierarchical label is thus bounded by $S_{N,L} + \log L$, not by $\log L$ alone.

Proposition 14 and the nonincreasing function $\mathcal{V}_x$ therefore imply that the actual marginal in (48) obeys

$$\boxed{\mathrm{Var}_{h_N} F_N(x) \ge K_x \mathcal{V}_x\left(\frac{S_{N,L} + \log L}{N}\right).} \tag{51}$$

The conclusion also holds when the hierarchy's conditionals factorize only across whole neurons.

Equivalently, reducing the variance to at most $(1-\eta)K_x$, with $0 < \eta < 1$, requires

$$S_{N,L} + \log L \ge N\mathcal{C}_x(\eta), \qquad \mathcal{C}_x(\eta) = \sup_{t\ge 0}\{-t(1-\eta) - \log \Lambda_x(t)\}. \tag{52}$$

These bounds depend on the training data through its nonnegative energy and expected score; they require no globally optimal fit. They also apply to restricted parameterizations satisfying the stated factorization. They do not infer an expected score from one sampled mixture.

Satisfying (52) is only a necessary condition under the score ceiling.

An optimizer-free predictive consequence. For the fixed Gaussian-regression data and positive future noise of Proposition 10, if $\sup_N S_{N,L_N} < \infty$ and $\log L_N = o(N)$, then

$$\liminf_N \mathbb{E}_{X\sim\rho} D_{\mathrm{KL}}(P_{N,X}\|h_{N,X}) \ge \mathbb{E}_\rho b(X). \tag{53}$$

Indeed $D_{\mathrm{KL}}(h_N\|U_N) \le S_{N,L_N}$, while its displayed whole-neuron label information is at most $S_{N,L_N} + \log L_N = o(N)$. This consequence follows from the earlier closure theorem for output distributions, independently of the finite variance argument. It identifies neither the mean nor equality at the floor.

J GAUSSIAN FITTING WITH ANALYTIC OBJECTIVES: FULL SPECIFICATION

The network is (1) with ReLU and incoming input $h = (x, 1)$. In this appendix and Appendix K, each neuron's incoming vector is written (w_i, b_i), an input weight and a hidden bias, rather than Section 2's $b_i \in \mathbb{R}^2$. The symmetric source has $(x, y) = (-1, -1), (1, 1)$ and noise variance 0.025. For the radial-label source, labeled "Counterexample" in the numerical records and displays, let $\lambda = (1, \sqrt{2})/\sqrt{2\pi}$, $\beta = 0.9\sqrt{2}|\lambda_1 - \lambda_2|$, and $C = 1.1 \cdot 4(1 + 1.5 + 1.75 + \beta^2)/\beta^2$. The two inputs are $x = 0$ and $x = 1$, with labels $\sqrt{C}\lambda$ and observation-noise variance $1/C$. We retain the constants 1.5 and 1.75 from the counterexample construction in the public code of Coker et al. (2022), although our fitted network has no separate output bias. We do not standardize or round these labels.

Both approximate posteriors have free Gaussian means for all weights. The fully factorized Gaussian has diagonal covariance; the Gaussian-per-neuron approximation has a block-diagonal covariance with one full positive-definite 3×3 block for each (a_i, w_i, b_i). Widths are $32, 256, 2048, 16384$. The

prior start is exact. Paired random starts draw independent means from $\mathcal{N}(0, 1)$ and variances $100/G$, where $G \sim \mathrm{Gamma}(101, 1)$ has unit scale, following Coker et al. (2022, Appendix L) with $\nu = 100$. Both families begin at the same diagonal Gaussian. The draws index optimization starts, not tasks. Float64 full-batch L-BFGS (Liu & Nocedal, 1989) has a cap of 4096 outer steps. Stationarity requires four consecutive fresh global gradient norms $\leq 10^{-6}$ and objective changes $\leq 10^{-10}(1 + |G_N|)$; 64 bit-identical steps above tolerance mean stalling. No primary fit exhausts the cap. We select the lowest-objective accepted iterate and report stopping status for the final iterate.

We refine each final endpoint by restarting the optimizer history, evaluating the same KL with cancellation-resistant arithmetic, and subtracting a constant from the line-search objective. Applying these changes together means that the refinement does not isolate the effect of arithmetic; it also does not replace the primary endpoints. Independent saved-parameter reconstruction checks all 64 primary and refined endpoints. Exact seeds, optimizer settings, endpoint and stopping records, and numerical discrepancies are in `supplement/neural_fit_records/`. These floating-point checks and analytic tail bounds do not give rigorous numerical enclosures.

The scalar-family mean reference $m_*\psi$ was specified before fitting. The feature reference for the Gaussian-per-neuron fits was derived after fitting and evaluated from the model and dataset specification, without fitting to the curves. Both are compared on 201 equally spaced points on $[-2, 2]$. The Gaussian-process posterior describes the width limit; this comparison does not establish a finite-width global optimum.

Mean comparisons at width 16,384. Both scalar initializations give essentially the same mean. It is nearly zero on the symmetric dataset; on the radial-label source its grid RMSE to $m_*\psi(x)$ is about 0.01327. The Gaussian-per-neuron fits escape this direction. On the symmetric dataset, their observed-input means are close to their own limiting slope $0.952381x$. On the radial-label source, both Gaussian-per-neuron fits remain far from the scalar mean limit and attain lower objectives. For both starts, the selected mean has grid RMSE 0.11992 to the Gaussian-per-neuron reference and 0.54918 to the limiting GP posterior mean, to the displayed precision. Both wide fitted families have output variances near the prior's, while the limiting GP posterior contracts near observations. Both scalar fits on this source have weight-space KL of about 6.28 nats to the product Gaussian with outgoing weights $a_i \sim \mathcal{N}(m_*/\sqrt{N}, 1)$ and independent prior incoming weights and biases. These endpoint comparisons do not establish convergence or global optimality.

A finite variance bound at the fitted costs. Proposition 14 applies to every mixture of whole-neuron products at a cost $d + i$, where d is KL to the prior and i is component-label information. For ReLU with augmented input $h = (x, 1)$, the prior variance is $K_x = (1 + x^2)/2$, and the normalized transform in that proposition is independent of x:

$$\Lambda(t) = \tfrac{1}{2} + \tfrac{1}{2}\mathbb{E}_{G\sim\mathcal{N}(0,1)}(1 + 4tG^2)^{-1/2}.$$

Every saved fit is one whole-neuron product, so $i = 0$. Table 5 evaluates $\mathcal{V}(d/N)$ at each fit's actual KL and compares it with its smallest observed variance ratio on the 201 queries. The widest symmetric fits cannot contract much at these costs; the bound is vacuous for the two smallest radial-label widths. All 64 primary/refined states satisfy their respective bounds, but the refinements are numerical checks of the same 32 fits.

The bound overlays are in `supplement/neural_fit_records/variance_bounds.json`; full variance curves and saved-state checks accompany the separately supplied `supplement/neural_predictive_replay/`. These are latent-variance restrictions, not a finite predictive-KL bound or evidence of optimizer convergence. They were evaluated on the two-point sources, not the 100-observation experiment.

K Finite predictive evaluation and mean fitting

We evaluate all 16 symmetric-source primary fits from Appendix J. This predictive study followed inspection of the fitted moment curves; the mean-preserving covariance intervention followed the first predictive results. Both are exploratory analyses of one fixed source and add no training replication.

Scores, reference comparisons, stopping flags and numerical records are in `supplement/finite_predictive_records/`.

Table 5: **Finite-width restrictions on variance at the attained KL cost.** Fractions of prior latent variance. Each range spans the four primary fits (two families and two starts), not a confidence interval. The observed value for each fit is its minimum across 201 queries. The lower bounds are floating-point evaluations of Proposition 14, without interval enclosures.

Source	Width	Lower-bound range	Observed range
Symmetric	32	0.0002–0.0258	0.1553–0.2246
Symmetric	256	0.1380–0.3891	0.5864–0.6208
Symmetric	2,048	0.7913–0.8551	0.9152–0.9184
Symmetric	16,384	0.9373–0.9796	0.9884–0.9889
Radial-label	32	0.0000–0.0000	0.0128–0.0748
Radial-label	256	0.0000–0.0000	0.1382–0.5305
Radial-label	2,048	0.1024–0.1858	0.8623–0.8647
Radial-label	16,384	0.5008–0.6226	0.9502–0.9760

K.1 Source, score, and finite predictive laws

Use the symmetric regression source and Gaussian weight prior specified in Appendix J, with $H = (-1, 1)$, $y = (-1, 1)^T$ and training-noise variance $\tau = .025$. A future response is $Y_x = F_N(x) + \epsilon$, with independent $\epsilon \sim \mathcal{N}(0, \gamma)$ and $\gamma = .025$, including at observed inputs. Queries are uniform on $\mathcal{X} = \{-2, -1, 0, 1, 2\}$. Writing $p_N(\cdot \mid x)$ for the actual finite posterior predictive density and $q_N(\cdot \mid x)$ for a candidate's density, the score is

$$\mathcal{L}_N(Q) = \frac{1}{5} \sum_{x \in \mathcal{X}} \int p_N(z \mid x) \log \frac{p_N(z \mid x)}{q_N(z \mid x)} \, dz. \tag{54}$$

This is the excess posterior-expected logarithmic loss (Gneiting & Raftery, 2007) in nats, with smaller values better. It is a forward predictive KL, distinct from the reverse weight-space KL used in fitting. The average concerns five marginal predictions, not their joint law or new-data generalization.

Let $B = (w_i, b_i)_{i=1}^N$ denote the incoming weights, with prior p_0, and set

$$r_i(x) = (w_i x + b_i)_+, \qquad S_{ji} = \frac{r_i(H_j)}{\sqrt{N}}, \qquad s_x = \frac{(r_1(x), \ldots, r_N(x))^T}{\sqrt{N}}.$$

Thus $S \in \mathbb{R}^{2 \times N}$ and $F_N(H) = Sa$; S is the feature matrix denoted by Ψ in (10). Define $T = SS^T$, $A = T + \tau I_2$, $u_x = s_x^T S^T$, and $v_x = s_x^T s_x$. Integrating the outgoing weights gives

$$a \mid B, y \sim \mathcal{N}\big(S^T A^{-1} y, I_N - S^T A^{-1} S\big), \tag{55}$$

$$\mu_B(x) = u_x A^{-1} y, \qquad c_B(x) = v_x - u_x A^{-1} u_x^T,$$

$$w(B) = \det(I_2 + T/\tau)^{-1/2} \exp\{-y^T A^{-1} y/2\}. \tag{56}$$

If ϕ_ν is the centered Gaussian density of variance ν, then

$$p_N(z \mid x) = \frac{\mathbb{E}_{B \sim p_0}[w(B) \phi_{c_B(x)+\gamma}(z - \mu_B(x))]}{\mathbb{E}_{B \sim p_0} w(B)}. \tag{57}$$

The training cross-kernel, hidden bias, and incoming evidence tilt are all retained. The finite posterior is therefore the target throughout; the Gaussian-process (GP) posterior is only a limiting reference.

Each fitted neuron has Gaussian mean $(\mu_{a,i}, \mu_{B,i})$ and covariance $\Sigma_i = \begin{pmatrix} d_i & c_i \\ c_i^T & C_i \end{pmatrix}$. Its conditional readout mean and variance are

$$\mu_{a|B,i} = \mu_{a,i} + c_i C_i^{-1} (B_i - \mu_{B,i}), \qquad \nu_i = d_i - c_i C_i^{-1} c_i^T = \det \Sigma_i / \det C_i > 0.$$

Given B, the noisy output is Gaussian with mean and variance

$$m_Q(B, x) = N^{-1/2} \sum_i r_i(x) \mu_{a|B,i}, \qquad v_Q(B, x) = \gamma + N^{-1} \sum_i r_i(x)^2 \nu_i.$$

Integration over the fitted incoming marginals gives positive Gaussian mixtures for the finite laws. The scalar family is the diagonal special case. No Gaussian replacement based only on output moments is used.

Three additional references help interpret the fits. The unchanged neural prior measures the cost of making no update. The whole-neuron mean-only approximation keeps the incoming prior and independently sets

$$a_i \mid B_i \sim \mathcal{N}\left(\frac{r_i(H)^T\alpha}{\sqrt{N}}, 1\right), \qquad \alpha = (K_{HH} + \tau I_2)^{-1} y, \tag{58}$$

where $r_i(H) = (r_i(-1), r_i(1))^T$. Its mean equals $m_{\mathrm{GP}}(x) = K_{xH}\alpha$ at every width. Its conditional latent variance is v_x, and it is generally not a Gaussian-per-neuron approximation. Finally, the limiting GP predictor is $\mathcal{N}(m_{\mathrm{GP}}(x), C_D(x,x) + \gamma)$, where $C_D(x,x) = K_{xx} - K_{xH}(K_{HH} + \tau I_2)^{-1} K_{Hx}$.

K.2 Fitted predictions and the correct-mean example

The primary-iterate selection and stopping rules are those in Appendix J, fixed before predictive evaluation; no refinement or predictive best-of-starts replaces an endpoint. Thirteen final iterates met the stationarity rule and three stalled; neither establishes optimality. The scalar prior starts at widths 32 and 256 are approximately stationary states with verified negative curvature. The Hessian restricted to the orthonormal two-neuron raw mean-contrast subspace has minimum eigenvalues -3.659821 and -1.177828, respectively; saved-state perturbations decrease the objective. These checks retain the learned, nonzero hidden-bias means. Positive curvature in the same block at the two larger widths is not a test of positive semidefiniteness of the full Hessian. No other endpoint is marked as a saddle.

At width 16,384, the scalar fits' five-query average predictive KL exceeds that of the Gaussian-per-neuron fits by about .42189 nats for both starts (paired numerical jackknife SE .000073). At width 32 the random starts nearly tie: $-.000896$ nats (SE .000457), while the scalar prior start is worse than the unchanged prior. The fitted comparison establishes neither universal family dominance nor monotonic harm with width.

The central query gives an exact mean comparison. The transformation $(a_i, w_i, b_i) \mapsto (-a_i, -w_i, b_i)$ preserves the prior and sends $F_N(x)$ to $-F_N(-x)$. It exchanges the two squared training errors and changes the sign of $F_N(0)$. Thus the exact finite posterior predictive mean at $x = 0$ is zero at every width. The widest prior-start fitted means are also zero, yet their predictive KLs there are approximately 0.41906 (scalar) and 0.41922 (Gaussian-per-neuron). This demonstrates correct mean with an inaccurate predictive law. At finite width, distributional shape as well as variance can contribute to this discrepancy.

On the same five queries, limiting Gaussian losses are 1.18656 (scalar), .76855 (Gaussian-per-neuron) and .76627 (whole-neuron mean-only). The Gaussian-per-neuron mean $x/(1 + 2\tau)$ follows from kernel $J(h, g) = \psi(h)\psi(g) + h^T g/4$ and differs from the GP mean. Its limiting mean restriction adds $.0022835$ nats to the .7662702 contraction cost; this does not establish optimality of a finite-width fit.

K.3 An exactly mean-preserving covariance intervention

The conditional-readout approximation uses (55) while retaining $B \sim p_0$. Its predictive density is the unweighted mixture $q_{\mathrm{ord}}(z \mid x) = \mathbb{E}_0 \phi_{c_B(x)+\gamma}(z - \mu_B(x))$. It differs from the exact posterior predictive because the incoming-weight distribution is not reweighted by $w(B)$. The covariance-ablated approximation retains the same incoming-weight distribution and conditional readout mean, but restores the prior readout covariance:

$$\begin{aligned} Q_{\mathrm{ord}}: &\quad a \mid B \sim \mathcal{N}(S^T A^{-1} y, I_N - S^T A^{-1} S), \\ Q_{\mathrm{abl}}: &\quad a \mid B \sim \mathcal{N}(S^T A^{-1} y, I_N). \end{aligned} \tag{59}$$

Both have mean $\mathbb{E}_0 \mu_B(x)$ for *every* input, exactly at each finite width. Neither mean need equal the finite posterior mean or the GP mean. Comparing (58) with the conditional readout changes finite means as well as covariance.

For query columns collected in $G = (s_{x_1}, \ldots, s_{x_k})$, the conditional output covariance increase from the ablation is $G^T S^T A^{-1} S G \succeq 0$. Since the conditional mean map is unchanged, its expectation

is also the unconditional covariance increase. The intervention generally changes cross-query covariances and the mixture's distributional shape; the measured scores remain marginal scores. The ablated law is not a product-neuron variational endpoint: its conditional mean depends on the entire incoming configuration through $A(B)^{-1}$.

We evaluate the loss added by removing the update directly as

$$\Delta_N = \mathcal{L}_N(Q_{\mathrm{abl}}) - \mathcal{L}_N(Q_{\mathrm{ord}}) = \frac{1}{5}\sum_{x\in\mathcal{X}} \mathbb{E}_{p_N(\cdot|x)} \log \frac{q_{\mathrm{ord}}(Y_x \mid x)}{q_{\mathrm{abl}}(Y_x \mid x)}. \tag{60}$$

The target-entropy term cancels. All 20 query/width contrasts are positive, ranging from .383 to 1.041 nats; their query averages range from .718 to .766 nats. The complete averages and paired numerical standard errors are in `covariance_intervention.csv`. This comparison isolates the cost of removing the conditional covariance update on this fixed source; it does not assign a covariance-only fraction to every fitted law's error.

K.4 Numerical construction and uncertainty

Each width uses 8192 incoming draws per law in eight independent blocks. The fitted approximations share standard-normal draws transformed through their incoming marginals; all reference laws share prior incoming draws. Widths use distinct streams, independent of fit selection. The later covariance intervention reuses the same draws without refitting or resampling.

We sum the target's Gaussian component numerators and evidence weights before normalization, and renormalize after every half-sample or deleted-block calculation. Candidate mixtures are unweighted. Densities use log-sum-exp and composite Gauss–Legendre quadrature. The integration interval is enlarged to meet the population-tail calculation below; neither truncated-density renormalization nor clipping of KL estimates is used. The accompanying `numerical_settings.json` and `reported_checks.json` retain seed keys, quadrature settings and checks.

For a score or paired contrast, average over queries first and let $d_{(-b)}$ denote its value after deleting block b from target and candidate streams together. The block-jackknife numerical standard error (Efron & Stein, 1981) is

$$\widehat{\mathrm{SE}}_{\mathrm{JK}} = \left\{ \frac{7}{8}\sum_{b=0}^{7} (d_{(-b)} - \bar{d}_{(-\cdot)})^2 \right\}^{1/2}.$$

This preserves dependence across queries and approximations and measures integration variability, not uncertainty over datasets or trained models. The largest average-score SE is .00331 nats at width 32 and .000080 at width 16384. Declared sampling and quadrature checks pass; the 32768-draw fallback is unused. Independent saved-state reconstruction checks that all 16 predictors use the selected primary fit for their stated width, family and start, and reproduces their latent-output means and variances at the five queries. Independent integration checks the conditional-readout and GP scores and the covariance contrasts.

Density estimation limits the smallest scores. At width 16384, the conditional-readout KL estimate changes from 1.65×10^{-10} with paired draws to 9.52×10^{-8} with disjoint halves, close to the posterior half-to-half discrepancy. The GP split diagnostic crosses zero. These values establish neither a small-score ordering nor a width rate. Split-mixture Jensen gaps diagnose density estimation but do not bound plug-in bias; subtracting a split-target distance is not a justified correction. Quadrature agreement concerns the sampled mixtures and does not enclose their error relative to the finite laws.

Separate control of omitted population tails. For this source, the nonnegative ReLU training features imply the evidence ceiling $w(B) \le W = \tau/e$. One way to see it is to rotate A into symmetric and antisymmetric coordinates, writing $A = \left(\begin{smallmatrix} s & c \\ c & r \end{smallmatrix}\right)$ and $t = r - c^2/s$. Then $s \ge r \ge t > 0$ and $w = \tau e^{-1/t}/\sqrt{st} \le \tau/e$. For M independent incoming draws, Hoeffding's inequality (Hoeffding, 1963) gives

$$Z_N = \mathbb{E}_0 w(B) \ge Z_L := \bar{w} - W\sqrt{\frac{\log(1/\delta)}{2M}}, \qquad \delta = 0.01/8,$$

with failure probability at most δ. A union bound gives at least 0.99 simultaneous coverage across the four widths and the two fixed possible budgets $M \in \{8192, 32768\}$ under the stipulated independent sampling law. This is not an arbitrary-stopping guarantee.

On this event, the noisy posterior density satisfies $p_N \le p_{0,N}/Z_L$ and its latent second moment is at most K_{xx}/Z_L, where $K_{xx} = (1+x^2)/2$. For a candidate latent second moment bound M_Q, set $R^2 = 2\max\{M_Q, K_{xx}/Z_L\}$, $C_0 = \log 2 + R^2/\gamma$, and $C_1 = 1/\gamma$. At least half of either latent law lies within its corresponding second-moment radius. Gaussian convolution then gives $|\log(p_N(z \mid x)/q_N(z \mid x))| \le C_0 + C_1 z^2$. The bounds used are the fitted exact second moments, K_{xx} for the prior, $81K_{xx}$ for the conditional readout and its covariance ablation, $K_{xx}(1+12\|\alpha\|^2)$ for the mean-only approximation, and the explicit GP second moment.

The exact finite-prior noisy moment-generating function is

$$M_{0,N,x}(t) = e^{\gamma t^2/2}\left[\frac{1}{2} + \frac{1}{2}\left(1 - \frac{(1+x^2)t^2}{N}\right)^{-1/2}\right]^N, \qquad |t| < \sqrt{\frac{N}{1+x^2}}.$$

Prior symmetry and exponential tilting yield, for every admissible $t > 0$,

$$\int_{|z|>L} p_N(z \mid x)\left|\log\frac{p_N(z \mid x)}{q_N(z \mid x)}\right| dz \le \frac{2e^{-tL}}{Z_L}\{C_0 M_{0,N,x}(t) + C_1 M''_{0,N,x}(t)\}. \tag{61}$$

The event supports every interval radius, so choosing an interval from the saved samples is allowed. The largest floating evaluation of this bound on the chosen intervals is about 5.58×10^{-13}. Independent high-precision derivative checks support its numerical evaluation, but do not provide rigorous numerical error bounds. Equation (61) controls the omitted absolute *population* log-ratio integral on the normalizer event. It supplies no bound on Monte Carlo density error inside the interval or on total numerical score error.

K.5 Finite width and initialization

A limiting mean characterization leaves open the width needed for agreement in fitted networks. We fixed the scalar and Gaussian-per-neuron mean predictions before fitting this 100-observation dataset.

Source, families, and fitting. We regenerate the seed-0 sine-regression dataset of Coker et al. (2022) in float64, standardizing inputs and responses with training moments. The likelihood variance remains $.025$ after standardization; the transformed generation variance is $.04256$. The network, Gaussian prior and two variational families are those of Appendix J, with a hidden bias and no separate output bias. At each width $N \in \{256, 2048, 16384\}$, both families start at the prior or at the paired random diagonal law specified there, using seed $230923001 + N$. Prior starts keep all neurons identical throughout deterministic fitting, restricting that search to the tied-neuron subfamily.

Float64 L-BFGS uses analytic Gaussian moments and minimizes

$$G_N(Q) = D_{\mathrm{KL}}(Q\|U_N) + \frac{\sum_{j=1}^{100}(\mathbb{E}_Q F_N(x_j) - y_j)^2}{.05} + \frac{\sum_{j=1}^{100}\mathrm{Var}_Q\, F_N(x_j)}{.05}. \tag{62}$$

The cap, stationarity, stalling and iterate-selection rules follow Appendix J. Five final states meet stationarity and seven stall; no fit hits the cap. Three selected states have gradients above tolerance despite stationary final states; their mean curves differ by at most 1.21×10^{-8}.

A prospective mean comparison. The 201 queries span $[-5, 5]$ in original input units. Using the standardized input x, the scalar and neuron references are ridge regressions on $\psi(x) = \sqrt{1+x^2}/\sqrt{2\pi}$ and $[\psi(x), x/2, 1/2]$, respectively, with penalty $.025$. The scalar reference is a near-zero control (RMS $.004807$). Before fitting, we required both widest neuron fits to have RMS error below $.171623$, one fifth of the neuron–GP gap, and to be closer to their own mean reference than to the GP mean. The GP reference is a limiting-model mean, not the exact finite posterior. The prior-start fit meets both criteria, with RMS errors $.009555$ to its limiting mean and $.855816$ to the GP mean. The random-start fit meets neither: its errors are $.470892$ and $.395703$, respectively. Agreement under both initializations was therefore not observed.

All six random starts are closer to GP than to their own limiting mean. The widest random neuron mean has RMS residual $.466797$ outside the entire limiting feature span, so its departure is not just a difference in the optimal coefficients. These are fixed-grid function comparisons, with no test-label evaluation. Figure 4c shows the widest neuron means; Table 6 retains all 12 fits.

Table 6: **All twelve prescribed fits.** Scalar: products of scalar Gaussians; neuron: products of joint Gaussian distributions over individual neurons. G_N is the total objective in nats; RMS mean errors average 201 fixed queries in standardized response units. These metrics use the selected state; stopping status belongs to the final state. Neither status certifies a minimum, and the starts are not dataset replications. Selected/final gradients, KL and objective components are preserved for all 24 states in the accompanying `ENDPOINTS.csv`.

Width	Family	Start	Final status	G_N	RMS mean error	
					Limiting mean	GP mean
256	Scalar	Prior	Stationary	2178.432	0.30342	0.84923
256	Scalar	Random	Stalled	607.726	0.92117	0.08204
256	Neuron	Prior	Stalled	1801.780	0.28536	0.71711
256	Neuron	Random	Stalled	596.161	0.88378	0.07998
2,048	Scalar	Prior	Stationary	2869.127	0.00260	0.89554
2,048	Scalar	Random	Stalled	1546.259	0.82146	0.11451
2,048	Neuron	Prior	Stationary	2682.649	0.05208	0.84580
2,048	Neuron	Random	Stalled	1525.863	0.77949	0.10870
16,384	Scalar	Prior	Stationary	3630.484	0.00030	0.89586
16,384	Scalar	Random	Stalled	3019.039	0.55567	0.38066
16,384	Neuron	Prior	Stationary	3480.085	0.00956	0.85582
16,384	Neuron	Random	Stalled	3034.952	0.47089	0.39570

What the lower objective buys. A descriptive decomposition, added after fitting began, separates mean fit from variance contraction. At the widest neuron fit, random minus prior gives

$$\Delta G_N = -445.13238 = \underbrace{832.32853}_{\text{prior KL}} - \underbrace{1304.03679}_{\text{mean-error term}} + \underbrace{26.57588}_{\text{variance term}}.$$

The random fit improves the training mean, paying more KL and slightly more variance. Every matched width/family pair has this sign pattern.

By comparison, the finite Gaussian-per-neuron construction in Appendix D, evaluated at the limiting ridge coefficients, is feasible but not optimal: the widest prior-start fit improves on it mainly by contracting variance. The widest random-start fully factorized Gaussian also has a lower objective than the Gaussian-per-neuron fit (3019.039 versus 3034.952), although the Gaussian-per-neuron family contains that fully factorized fit; the richer fit is therefore not at its family optimum.

At this width, the limiting mean therefore does not predict the fitted mean independently of initialization. At fixed data, bounded prior KL restricts the limiting mean span; near-global objective values further identify its optimal mean. Neither statement supplies a useful finite-width threshold here. The prior objective is 4000 independently of width; the widest random neuron fit spends 1111.901 in KL and remains within that bound.

A finite mean diagnostic. An exact Gaussian-moment calculation separates the expansion remainder from disagreement with the optimal limiting coefficients. For a nondegenerate Gaussian neuron at a nonzero query h, let $\mathbf{b}_i$ denote its incoming-weight vector, equal to $(w_i, b_i)^T$ in this experiment. Write $A_i = \mathbb{E}a_i$, $u_i = (\mathbb{E}\mathbf{b}_i)^T h$, $s_i^2 = \mathrm{Var}(\mathbf{b}_i^T h)$, $c_i = \mathrm{Cov}(a_i, \mathbf{b}_i^T h)$ and $n = \|h\|$. Let φ_{G} and Φ_{G} be the standard normal density and CDF, with $\phi_0 = \varphi_{\mathrm{G}}(0)$ and $t_i = u_i/s_i$. The exact mean contribution f_i and its first-order part g_i are

$$f_i(h) = A_i\{s_i\varphi_{\mathrm{G}}(t_i) + u_i\Phi_{\mathrm{G}}(t_i)\} + c_i\Phi_{\mathrm{G}}(t_i),$$
$$g_i(h) = A_i n\phi_0 + c_i/2, \qquad r_i(h) = f_i(h) - g_i(h).$$

The difference $\mathbb{E}(\mathbf{b}_i^T h)_+ - s_i\phi_0 - u_i/2$ is the integral of $\Phi_{\mathrm{G}}(v/s_i) - 1/2$ from zero to u_i, hence lies between zero and $\min\{|u_i|/2, \phi_0 u_i^2/(2s_i)\}$. Together with $|\Phi_{\mathrm{G}}(t_i) - 1/2| \le \min\{1/2, \phi_0|t_i|\}$,

this gives $|r_i(h)| \le B_i(h)$, where

$$B_i(h) = |A_i| \left[\phi_0 |s_i - n| + \min\left\{ |u_i|, \frac{|u_i|}{2} + \frac{\phi_0 u_i^2}{2 s_i} \right\} \right] + |c_i| \min\left\{ \frac{1}{2}, \frac{\phi_0 |u_i|}{s_i} \right\}. \tag{63}$$

At $h = 0$, $f_i = g_i = r_i = B_i = 0$ by continuity. Put $g_N = N^{-1/2} \sum_i g_i$, which lies in the limiting mean span. For the fixed ridge reference m_{ref} above, the RMS over the 201 queries therefore satisfies

$$\|\mathbb{E}_Q F_N - m_{\mathrm{ref}}\|_{\mathrm{RMS}} \le \|g_N - m_{\mathrm{ref}}\|_{\mathrm{RMS}} + \left\| N^{-1/2} \sum_i B_i \right\|_{\mathrm{RMS}}. \tag{64}$$

The first term measures coefficient mismatch; the second bounds the intrinsic expansion remainder. The following values use the selected states. For the widest prior-start neuron fit, their sum is $.0253433 + .0232262 = .0485695$, below the prespecified $.171623$ threshold; the actual error is $.0095551$. The random-start bound 7.7300766 is inconclusive, and its actual error $.4708919$ still fails the criterion.

To locate the remaining mean shape, rank neurons by their KL to the prior and retain the exact remainders of a set S in $g_N + N^{-1/2} \sum_{i \in S} r_i$. Its error is $N^{-1/2} \sum_{i \notin S} r_i$, bounded pointwise by $N^{-1/2} \sum_{i \notin S} B_i$. In the widest random neuron fit, the two highest-KL neurons carry 76.5250% of KL. The remaining residual's component outside the limiting span has RMS $.0093819$ (originally $.4667972$); its full RMS is $.0731752$, bounded by $.1252917$. Those two neurons contribute at most $.252\%$ of total query variance. Cancellation matters: for the widest random scalar fit, the remaining RMS outside its scalar mean span rises from $.548458$ to $.570113$ after one retained term and $.926878$ after two, before falling to $.0167015$ after four.

This is a post-fit algebraic localization, not a deletion intervention or a prospective width rule. The mean identity and bound require only Gaussian neuron marginals, even with cross-neuron correlations; the variance shares use the independence of neurons in these fits. Mean agreement also does not establish covariance convergence: the widest prior-start neuron fit retains only $.63935$–$.69988$ of prior query variance. The floating-point evaluations have no rigorous error enclosures. All 24 endpoint diagnostics and exceptional-set curves are recorded in `many_observation_support/finite_mean_diagnostic.json`, alongside the earlier KL summaries in `many_observation_support/kl_concentration.csv`.

L From product-mixture capacity to preference-head model selection

The preference-head comparison needs three ingredients: a bound for arbitrary product mixtures, a bound on the fitted Gaussian's gap to the best product, and an independent ordering of the true evidence. We give them in that order, followed by the structured-covariance controls.

L.1 A convex extension for arbitrary product components

Let $\pi = \bigotimes_{i=1}^d \pi_i$, where $\pi_i(\mathrm{d}x_i) \propto e^{-V_i(x_i)}\,\mathrm{d}x_i$ and $V_i'' \ge \kappa_i > 0$. For a convex C^2 likelihood energy f, define

$$P(\mathrm{d}x) = Z^{-1} e^{-f(x)} \pi(\mathrm{d}x), \qquad \mathcal{G}(Q) = D_{\mathrm{KL}}(Q\|\pi) + \mathbb{E}_Q f = D_{\mathrm{KL}}(Q\|P) - \log Z. \tag{65}$$

Thus $-\mathcal{G}$ is the evidence lower bound (ELBO). All energies below use natural logarithms and include the normalized prior. Suppose the coordinate second derivatives of f are globally bounded. Let $Q_* = \bigotimes_i q_i^*$ be its finite-energy product optimum, with

$$q_i^*(\mathrm{d}x_i) \propto e^{-\phi_i(x_i)} \pi_i(\mathrm{d}x_i), \qquad \phi_i(x_i) = \mathbb{E}_{Q_{*,-i}} f(x_i, X_{-i}), \qquad 0 \le \phi_i'' \le b_i. \tag{66}$$

The existence, uniqueness and fixed-point characterization of the product optimum under strong log-concavity are established ingredients (Lacker et al., 2024, Corollary 1.4); they do not require Gaussian variational marginals. The coordinate-convexity and projected-score arguments used here also build on existing results (Lacker, 2026; Lavenant & Zanella, 2024).

Proposition 15 (Convex product-mixture bound)**.** *Under these assumptions, put* $\eta = \max_i b_i/(\kappa_i + b_i) < 1$*. For a finite mixture* $Q = \sum_{j=1}^{L} w_j Q_j$ *of arbitrary coordinate-product laws with finite energies, and its component label* J*,*

$$\mathcal{G}(Q) - \mathcal{G}(Q_*) \geq (1-\eta) D_{\mathrm{KL}}(Q\|Q_*) - \eta I(J;X) \geq -\eta \log L. \tag{67}$$

In particular, allowing a mixture with L *product components improves the best single-product ELBO by at most* $\eta \log L$*.*

Proof. First let Q be a product. Independently for each coordinate, optimally couple $X_i \sim Q_i$ to $Y_i \sim q_i^*$ for quadratic transport. Write $d_i = D_{\mathrm{KL}}(Q_i\|q_i^*)$, $W_i^2 = \mathbb{E}(X_i - Y_i)^2$, and

$$B_i = \mathbb{E}\{\phi_i(X_i) - \phi_i(Y_i) - \phi_i'(Y_i)(X_i - Y_i)\}.$$

The entropy tangent inequality at q_i^*, together with the curvature of V_i, gives $d_i \geq \kappa_i W_i^2/2 + B_i$. The curvature bound on ϕ_i gives $0 \leq B_i \leq b_i W_i^2/2$. Consequently,

$$d_i - B_i \geq \frac{\kappa_i}{\kappa_i + b_i} d_i. \tag{68}$$

For completeness, the first inequality follows by writing relative entropy as entropy plus the expected potential $V_i + \phi_i$: displacement convexity bounds the entropy difference by its tangent, the linear terms cancel against $\partial_i \log q_i^* = -V_i' - \phi_i'$, and the two potential remainders are at least $\kappa_i W_i^2/2$ and B_i.

The fixed-point density gives

$$\mathcal{G}(Q) - \mathcal{G}(Q_*) = D_{\mathrm{KL}}(Q\|Q_*) + (\mathbb{E}_Q - \mathbb{E}_{Q_*})\Big(f - \sum_i \phi_i\Big).$$

Expand the second term along the independent couplings. Its first-order terms cancel because $\mathbb{E}[\partial_i f(Y) \mid X_i, Y_i] = \phi_i'(Y_i)$. The full f-Bregman remainder is nonnegative by convexity. Hence the product objective gap is at least $\sum_i (d_i - B_i)$, which is at least $(1-\eta) D_{\mathrm{KL}}(Q\|Q_*)$ by (68). For the mixture, the two exact identities

$$\mathcal{G}(Q) = \sum_j w_j \mathcal{G}(Q_j) - I(J;X), \qquad \sum_j w_j D_{\mathrm{KL}}(Q_j\|Q_*) = D_{\mathrm{KL}}(Q\|Q_*) + I(J;X)$$

give (67). Approximation by smooth finite-energy products extends the coupling argument to its stated domain. □

The classical entropy-of-mixing bound has coefficient one (Jaakkola & Jordan, 1998). The curvature coefficient sharpens it when η is small, but can remain positive for a separable energy.

For the logistic head, $z \sim N(0, I_d)$ and

$$f_\alpha(z) = \sum_{\ell=1}^{n} \log\{1 + \exp(-y_\ell B_{\alpha,\ell} z)\}, \qquad B_\alpha = \alpha X, \quad y_\ell \in \{-1, 1\}. \tag{69}$$

This is equivalent to weights $w \sim N(0, \alpha^2 I_d)$ acting on X. The standardization includes the prior-scale change exactly. Since the logistic second derivative is at most $1/4$, we may take

$$b_i = \frac{1}{4}\|B_{\alpha,:,i}\|^2, \qquad \eta_\alpha = \frac{\max_i b_i}{1 + \max_i b_i}. \tag{70}$$

The component family remains the products in the original feature coordinates. Rotating those coordinates changes that family. This extension covers convex heads with fixed features, not joint inference over the nonlinear language-model weights.

L.2 Relating a fitted Gaussian to the best arbitrary product

A fitted diagonal Gaussian is only one product law. To use Proposition 15 for *every* product mixture, we must bound the gap between that attained Gaussian and Q_*. The following computable specialization of projected-score control supplies that step. It uses the projected log-Sobolev/convexity

framework of Lacker (2026, Theorem 2.5), rather than treating the Gaussian fit as the arbitrary-product optimum.

For $U(x) = \|x\|^2/2 + f(x)$ and a smooth product q, define

$$e_i(x_i) = \partial_i \log q_i(x_i) + \mathbb{E}_q[\partial_i U(X) \mid X_i = x_i], \qquad \mathcal{I}_{\mathrm{proj}}(q) = \sum_i \mathbb{E}_q e_i(X_i)^2.$$

Along the independent coordinate transport from q to Q_*, entropy is displacement convex and U is 1-strongly convex. Completing the square in its first-order lower bound gives

$$0 \le \mathcal{G}(q) - \mathcal{G}(Q_*) \le \frac{1}{2}\mathcal{I}_{\mathrm{proj}}(q). \tag{71}$$

Indeed the lower bound for $\mathcal{G}(Q_*)$ is $\mathcal{G}(q) + \mathbb{E}\sum_i e_i(X_i)(Y_i - X_i) + \frac{1}{2}\mathbb{E}\|Y - X\|^2 \ge \mathcal{G}(q) - \frac{1}{2}\mathbb{E}\|e(X)\|^2$.

For $q = \bigotimes_i N(\mu_i, \sigma_i^2)$, let g_μ, g_σ be the gradients of its exact Gaussian objective with respect to the mean and *standard deviation*. Gaussian integration by parts identifies these with the constant and standardized linear Hermite coefficients of e_i. The remaining coefficients have degree at least two. The Gaussian Hermite expansion therefore gives

$$\mathbb{E}e_i^2 \le g_{\mu,i}^2 + g_{\sigma,i}^2 + \frac{\sigma_i^4}{2}\mathbb{E}\psi_i''(X_i)^2, \qquad \psi_i(x_i) = \mathbb{E}_q[\partial_i f(X) \mid X_i = x_i].$$

For (69), write $s(t) = (1 + e^{-t})^{-1}$, $T_i = \sum_\ell |B_{\alpha,\ell i}|^3$, and $v_\ell = \mathbb{E}_q[s''(-y_\ell B_{\alpha,\ell} X)^2]$. Conditional Jensen and weighted Cauchy–Schwarz give $\mathbb{E}\psi_i''(X_i)^2 \le T_i \sum_\ell |B_{\alpha,\ell i}|^3 v_\ell$. Therefore the observable quantity

$$\rho(q) = \frac{1}{2}\left\{\|g_\mu\|^2 + \|g_\sigma\|^2 + \frac{1}{2}\sum_i \sigma_i^4 T_i \sum_\ell |B_{\alpha,\ell i}|^3 v_\ell\right\} \tag{72}$$

upper-bounds $\mathcal{G}(q) - \mathcal{G}(Q_*)$. Each v_ℓ is a one-dimensional Gaussian expectation; the global bound $s''(t)^2 \le 1/108$ is a looser alternative. The gradient terms are retained, so exact numerical stationarity is unnecessary.

Combining (72) with Proposition 15 gives the computable lower bound used below:

$$\boxed{\inf_{Q \in \mathcal{M}_L^{\mathrm{prod}}} \mathcal{G}_\alpha(Q) \ge \mathcal{G}_\alpha(q) - \rho(q) - \eta_\alpha \log L.} \tag{73}$$

Here $\mathcal{M}_L^{\mathrm{prod}}$ allows arbitrary component marginals and weights, not just fitted Gaussian components.

L.3 Data, training-set sizes, and the evidence reversal

The source posts come from the Reddit TL;DR corpus (Völske et al., 2017), with human summary comparisons from Stiennon et al. (2020).

We use the public `luckeciano/reddit-features-hermes` release (Melo et al., 2024; luckeciano, 2024), containing pairs of 4,096-dimensional OpenHermes-2.5-Mistral-7B response embeddings (Teknium, 2023). Chosen-minus-rejected differences are normalized by the median norm over the original training source, including subsequently designated query features. Randomly orienting each pair and changing its sign consistently leaves the likelihood and evidence unchanged. We fit only the head. The released-feature file and preparation are pinned in `supplement/preference_data_and_settings/`; backbone and tokenizer extraction revisions were not independently established.

Exact prompt groups define data roles. After excluding groups exposed in earlier work, the original training source supplies a training pool, 1,000 development prompts (5,122 comparisons), and 2,000 separate query prompts (10,321 comparisons). The query outcomes had already appeared in an acquisition study and are reused descriptively. Neither the original test nor out-of-distribution split is used; exclusion of touched prompts also changes the source population.

A preliminary study crosses three training orders, five counts $\{64, 256, 1024, 4096, 16384\}$, and scales $\{.3, 1, 3\}$, giving 45 MAP fits. The accompanying `reported_map_development.csv`

Table 7: **Evidence and product mixtures rank the scales oppositely.** Total nats for each fixed 1,024-comparison prefix, using the evidence bounds and $L = 10^6$ margin defined in this subsection. These are numerical evaluations of proved inequalities, without rigorous numerical error bounds.

Prefix	$-\log Z_{0.3}$ lower	$-\log Z_1$ upper	$\mathcal{G}_{0.3}(q_{0.3,\mathrm{diag}})$	$L = 10^6$ margin
A	676.568714	664.264987	690.444292	1.482997
B	677.841167	669.651610	691.976652	6.171303
C	682.334084	678.086526	696.932294	10.580292

records scale-one development performance, averaging within each prompt and then equally across the 1,000 prompts and three orders. These fits share data; varying training order or sample size does not independently replicate the evidence comparison below.

The evidence experiment uses the 1,024-label prefix of each order, denoted A, B and C. Exact split and prefix indices and seed rules are supplied in `data_selection.json` and its data files.

At scale one, we fit a diagonal Gaussian in the original coordinates, a diagonal Gaussian in the observed row-space basis, an unrestricted Gaussian in that row space, and the ordinary Laplace approximation (MacKay, 1992). The complete observed span has rank 1,021–1,022. Each structured family retains the original independent Gaussian prior on the orthogonal complement. Thus "full Gaussian" means full covariance on every data-dependent direction, not a truncated-rank approximation. The accompanying `reported_family_objectives.csv` gives all four objectives, the original-coordinate gap allowance ρ and the curvature coefficient η_1 at their reported precision.

To establish which scale the *true evidence* prefers, we also need a bound independent of the variational family. Put $\Phi_\alpha(z) = \|z\|^2/2 + f_\alpha(z)$. Strong convexity implies

$$\Phi_\alpha(z) \geq \Phi_\alpha(z_*) + \tfrac{1}{2}\|z - z_*\|^2, \qquad -\log Z_\alpha \geq \inf_z \Phi_\alpha(z) \geq \Phi_\alpha(z_0) - \tfrac{1}{2}\|\nabla\Phi_\alpha(z_0)\|^2.$$

The first evidence inequality follows by integrating the Gaussian upper envelope of $e^{-\Phi_\alpha}$, including its $(2\pi)^{-d/2}$ prior normalization. On the other side, $-\log Z_1 \leq \mathcal{G}_1(q_{\mathrm{full}})$ for any attained full Gaussian. Table 7 separates these two bounds for every prefix: the true evidence prefers scale one to scale 0.3. The product-mixture margin $\mathcal{G}_1(q_{\mathrm{diag}}) - \rho - \eta_1 \log L - \mathcal{G}_{0.3}(q_{0.3,\mathrm{diag}})$ is positive at $L = 10^6$ for each prefix. By (73), every scale-one product mixture of at most 10^6 components in the original coordinates therefore has worse ELBO than that feasible scale-0.3 product.

The coefficient-one $\log L$ bound already excludes $L = 1024$ on all prefixes and $L = 10^6$ on B and C. On A at $L = 10^6$, the curvature coefficient changes the margin from $-.399$ to $+1.483$ nats. Row-space-diagonal Gaussians and full Laplace also recover the evidence ordering. Inference-dependent model-selection bias is established (Turner & Sahani, 2011; Li et al., 2023); this example adds a whole-family exclusion and bounds on the true evidence ordering, not a predictive-risk theorem.

Structured covariance restores the model choice. Retain the scale-one Laplace mean and its strongest covariance contractions, restoring prior covariance in the remaining directions. In the predeclared grid, the first ranks whose negative ELBOs fall below the lower bound on $-\log Z_{.3}$ are 16, 32 and 128 for A, B and C, with margins 1.075258, .105306 and .914342 nats. Every distribution at scale .3 has negative ELBO at least $-\log Z_{.3}$, so each displayed approximation beats every distribution at that scale and restores the correct two-scale choice. Rank two fails this comparison on all three prefixes. These sufficient ranks are not proved minimal and do not establish cheaper fitting: the full Laplace fits and spectral decompositions are computed first.

L.4 A PRODUCT-MIXTURE BOUND FOR THE SEMI-IMPLICIT OBJECTIVE

Apply the expected sampled-mixture definition (48) to the finite-dimensional energy (65). Each conditional is a product in the original coordinates, so every realized iid L-component mixture obeys (73). Taking expectations gives the same lower bound on S_L, the expected negative semi-implicit objective. Its density estimate includes the component that generated the sample; the continuous marginal itself need not have L components. For prefix C at scale one and $L = 128$, this yields

$S_{128} \geq 715.287160$ nats, above the feasible scale-.3 score 696.932294. The bound applies to every eligible product hierarchy at this count; it establishes neither sampled-mixture attainment nor a predictive-risk ordering.

L.5 Predictive consequences of the prior choice

For each of three 1,024-comparison training subsets from one data source, we evaluate six fitted predictors on the same 5,122 comparisons in 1,000 development prompts. This exploratory evaluation reuses prompts examined earlier. Training and development do not overlap by comparison or prompt; the original test and OOD sources are unused. Metrics average comparisons within prompts, then weight prompts equally; the table averages the three subset values.

Table 8: **Predictive scores for the fitted preference models.** Means over three training subsets from one data source, evaluated on the same development prompts used in earlier analyses. Lower log loss and Brier scores are better; higher accuracy is better. This comparison concerns these fitted models, not every member of their families.

Predictor	Prior scale	Log loss (nats)	Brier	Accuracy (%)
Diagonal Gaussian	.3	0.662923	0.235141	61.786
Diagonal Gaussian	1	0.636566	0.223112	63.968
Row-space diagonal	1	0.636473	0.223060	64.074
Full Gaussian	1	0.636463	0.223057	64.067
MAP	.3	0.662638	0.235007	61.886
MAP	1	0.637010	0.223233	64.098

For all three training subsets, the scale-one diagonal Gaussian has lower development log loss than the scale-.3 fit, by .02754, .02888 and .02265 nats. The mean difference is .02636 nats. At scale one, the mean log losses of the diagonal, row-space-diagonal and full-Gaussian fits differ by at most .000104 nats. MAP also favors scale one. The row-space-diagonal Gaussian uses the full observed row-space basis; it is not one of the truncated-rank repairs in Figure 5. All eighteen rows, including Brier scores and accuracy, accompany the numerical record. Exact $1/2$ ties receive half accuracy credit, including the comparison whose feature difference is identically zero.

The prior-scale comparison and this development population were already explored, so these scores are descriptive. They concern fitted predictors; the family-wide evidence bounds do not imply a predictive-risk ordering.

L.6 Numerical and computational scope

Gaussian fits use L-BFGS-B (Zhu et al., 1997) with order-256 Gauss–Hermite expectations, checked at order 512; independent scalar reconstruction uses Gauss–Legendre quadrature. At the smallest passing rank margin (.105306 nats, B at rank 32), both discrepancies are below 3×10^{-11} nats. These floating-point evaluations retain optimization residuals and provide no rigorous interval bounds. The data/settings package contains settings, checks and four-thread fit and workflow times excluding feature extraction; it is not a complete executable fitting archive. The separately supplied `supplement/preference_replay/` recomputes product bounds, evidence comparisons and full rank-grid objectives from saved feature rows and fitted parameters, without refitting or feature extraction. It does not recompute the development predictions in Appendix L.5.