Routing Changes the Prior, Not the Function Class: Exact Bayesian Complexity of a Two-Branch Regressor

Ali Mehrabian
University of British Columbia
alimehrabian619@ece.ubc.ca

Mahdi Mazloum
University of Pittsburgh
sem322@pitt.edu

Abstract

A two-branch routed regressor with an affine gate represents exactly the same functions as ordinary quadratic regression. Since routing does not enlarge the function class, the two models differ only in the prior that the routed parameterization induces on the quadratic coefficients. For a Gaussian prior on the routed coordinates, we derive an explicit one-dimensional integral for the induced coefficient density: integrating along the fibers of the routing map produces a density with a logarithmic singularity on the linear submodel. The posterior over quadratic coefficients under the routed model is then exactly the quadratic-regression posterior under this pushforward prior, so in this model the effect of routing is entirely a matter of prior geometry. When the true regression function is linear ($q_2 = 0$), we compute the real log canonical threshold (RLCT) and its multiplicity. For smooth compactly supported priors reaching the set of true parameters, we consider two support regimes. The pair is $(3/2, 2)$ if the support reaches the collapsed stratum, where the gate is constant and the two branches have equal slopes, and $(3/2, 1)$ if the support is disjoint from that stratum. The Gaussian prior has full support and gives $(3/2, 2)$. Compared with quadratic regression under a smooth coefficient prior and the same known noise variance, the routed model therefore gains a $\log\log n$ term in the log Bayes factor on linear truth. The same multiplicity-2 product singularity appears locally for sigmoid gates and for d-dimensional inputs.

1 Introduction

Mixture-of-experts (MoE) models and other routed architectures, from classical mixtures of experts to modern conditional-computation networks (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022), are singular statistical models. Latent expert assignments, redundant experts, and inactive routes can make their parameters non-identifiable and invalidate the classical asymptotics of regular models (Jiang and Tanner, 1999; Rousseau and Mengersen, 2011; Ho and Nguyen, 2019). Singular learning theory gives a precise language for such degeneracies and shows that they can change the Bayesian complexity of a model in ways that parameter counting does not capture (Watanabe, 2009, 2018; Drton and Plummer, 2017).

Computing the exact Bayesian complexity of routed architectures is hard, partly because routing changes both the function class and the way it is parameterized. To separate the second effect from the first, we study a minimal routed regressor: two scalar linear branches combined by a single affine gate. This model is tractable precisely because routing does *not* enlarge its function class. After a change of coordinates, the regression functions it can represent are exactly those of ordinary quadratic regression.

We ask how Bayesian inference changes when a Gaussian prior is placed on the routed coordinates rather than a smooth prior on the quadratic coefficients. In this exactly solvable setting, the routed model and the quadratic baseline differ only in their prior on the coefficients. For the routed model,

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

this prior is the pushforward of a product Gaussian prior on the routed coordinates through the routing map; for the baseline, it is a smooth prior placed directly on the coefficients.

We derive an explicit fiber-integral formula for this induced prior. A product Gaussian prior on the routed coordinates pushes forward to a coefficient density with a logarithmic singularity on the linear submodel (Theorem 5.3), and the posterior over quadratic coefficients under the routed model is exactly the quadratic-regression posterior under this pushforward prior (Proposition 6.1). Compared with quadratic regression under a smooth coefficient prior and the same known noise variance, the singularity gives the routed model a $\log\log n$ advantage in log marginal likelihood when the truth lies on the linear submodel, and only an $O_p(1)$ difference otherwise, where $O_p(1)$ denotes a term that is bounded in probability. This gain comes from comparing a singular prior with a smooth one; it is not a prior-invariant property of quadratic regression.

The broader point is that the coordinates in which a model is parameterized can matter for Bayesian inference even when they do not change what the model can express. In our model, the routing coordinates induce a singular prior on the observable coefficients, and this alone changes the asymptotics of the marginal likelihood and hence Bayesian model selection.

Contributions.

- **Geometry and local normal forms.** We show that the routed model represents exactly the quadratic regression functions, describe the fibers of its parameter-to-coefficient map and the points where this map degenerates, and derive exact local normal forms of the population loss, including the form $u^2 + v^2 + 8w^2z^2$ at the collapsed stratum (Theorems 4.1 and 5.1).
- **Induced prior and posterior equivalence.** We derive an explicit fiber-integral formula for the pushforward of a product Gaussian prior on the routed coordinates. The resulting density ρ_G has a logarithmic singularity whose leading coefficient A_G is an explicit function of the prior scales (Proposition E.1); at unit scales, $A_G(0,0) = 1/(4\pi^2)$. For every dataset, the coefficient posterior of the routed model equals the quadratic-regression posterior under ρ_G (Proposition 6.1). This links the prior calculation to the evidence and contraction results.
- **Bayesian complexity and model selection.** For compactly supported priors we compute the RLCT in two support regimes; we also compute it for the Gaussian prior on the routed coordinates. On linear truth, a compactly supported prior reaching W_0 gives global RLCT $(3/2, 2)$ if its support reaches the collapsed stratum S_C and $(3/2, 1)$ if its support is disjoint from S_C; the Gaussian prior has full support and gives $(3/2, 2)$ (Table 1). With the Gaussian prior, the routed model therefore gains a $\log\log n$ advantage in log Bayes factor on linear truth over quadratic regression with a smooth coefficient prior and the same known noise variance (Theorem 6.4).
- **Local robustness.** The multiplicity-2 product singularity at the collapsed stratum persists locally when the affine gate is replaced by a sigmoid gate (Propositions 5.4 and F.4) and when the input is d-dimensional (Proposition 5.5). For sigmoid gates this is a local statement about the leading pole; the exact pushforward identity in coefficient space does not carry over.

Scope. Two caveats apply throughout. First, the RLCT classification for compactly supported priors assumes that the support reaches W_0 and, on linear truth, either reaches S_C or is disjoint from it (Theorem 5.2). Second, the $\log\log n$ gain compares a Gaussian prior on the routed coordinates with smooth priors on the coefficients. It is neither a prior-invariant statement about quadratic regression nor a claim that the coefficient posterior contracts more slowly; in fact, the coefficient posterior contracts at the usual $n^{-1/2}$ rate (Proposition 6.2).

2 Related Work

Mixture-of-experts models have a long history (Jacobs et al., 1991; Jordan and Jacobs, 1994; Waterhouse et al., 1995; Yuksel et al., 2012), and conditional-computation architectures built on them now run at very large scale (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022). On the theoretical side, the identifiability problems caused by expert routing have been known for a long time (Jiang and Tanner, 1999). Convergence rates have been established for finite mixtures (Chen, 1995) and for mixtures of experts (Ho et al., 2022; Nguyen et al., 2024). More recent work shows how MoE models trained with stochastic gradient descent (SGD) detect latent cluster structure (Kawata

et al., 2025) and measures MoE expressivity using tropical geometry (Su et al., 2026). We do not try to match the generality of these settings. Instead, we study a single product-type routing singularity in a model that can be solved exactly.

Singular learning theory (SLT) describes non-identifiable models through birational invariants such as the real log canonical threshold (RLCT) and its multiplicity, rather than through parameter counts (Watanabe, 2001, 2009, 2018). These invariants determine the asymptotics of the Bayes free energy and underlie information criteria such as WAIC, WBIC (Watanabe, 2010, 2013), and sBIC (Drton and Plummer, 2017).

Exact stochastic complexities are known for several classical models, including reduced-rank regression, restricted Boltzmann machines, and various mixture models (Aoyagi and Watanabe, 2005; Aoyagi, 2010; Watanabe and Watanabe, 2006, 2022; Kurumadani, 2024). The product degeneracy in our model is closely related to the one in reduced-rank regression (Aoyagi and Watanabe, 2005), so this routed model belongs to the familiar SLT family of product-type singularities. The mechanism is shared, but we do not claim that the RLCT pairs coincide. Other recent work measures the complexity of neural networks through local learning coefficients (Lau et al., 2025) and studies low-rank Bayesian neural networks (Toure and Stephens, 2026). Outside SLT, asymptotic analyses of overfitted and finite mixtures also show how such degeneracies change posterior and estimation behavior (Rousseau and Mengersen, 2011; Ho and Nguyen, 2019).

Conceptually, our results concern the prior geometry that a parameterization induces. A smooth one-to-one reparameterization, with the prior transformed accordingly, leaves a Bayesian model unchanged (Jeffreys, 1946; Gelman, 2004), but placing independent default priors on the coordinates of two different parameterizations leads to genuinely different procedures. Our setting is related in spirit to work on priors near embedded submodels (Johnson and Rossell, 2010), but here the mechanism is explicit: a simple Gaussian prior on the routed coordinates pushes forward to a logarithmically singular prior on the coefficients. The resulting $\log\log n$ evidence gap lies between the $O(1)$ differences produced by smooth positive priors and the $O(\log n)$ differences produced by a change in model dimension.

Lippl et al. (2022) characterize the implicit bias of gradient descent in gated linear networks through an architecture-dependent norm. We instead study the prior and the marginal likelihood induced on a fixed function class, without reference to a training algorithm.

A large literature on stochastic optimization relates gradient noise to diffusion limits and Gibbs-type stationary measures (Kushner and Yin, 2003; Li et al., 2019; Mandt et al., 2017). We use this connection only as motivation. Our one result in this direction (Corollary G.3 in the appendix) concerns the low-temperature Gibbs posterior and makes no claim about SGD dynamics.

3 Setup

Definition 3.1 (Minimal two-branch routed model)**.** Let $x \in \mathbb{R}$. The parameter is $\theta = (a_0, a_1, b_0, b_1, r) \in \mathbb{R}^5$, with branch functions

$$e_1(x) = a_0 + a_1 x, \qquad e_2(x) = b_0 + b_1 x,$$

and affine gate

$$g(x; r) = \tfrac{1}{2} + rx.$$

The model output is

$$f(x; \theta) = g(x; r)e_1(x) + \big(1 - g(x; r)\big)e_2(x). \tag{1}$$

We use an affine gate because it keeps the model algebraically solvable. The gate acts here as a weighting function in a regression model, not as a mixture probability, so it may take values outside $[0, 1]$. The central singularity does not depend on this choice. Since $\sigma(rx) = \frac{1}{2} + \frac{r}{4}x + O(r^3x^3)$, a sigmoid gate produces the same leading product $r\delta_1$ in the quadratic Taylor coefficient of the output, and Proposition 5.4 shows that the same local multiplicity-2 product singularity appears at the collapsed stratum of the sigmoid model. What is lost with a sigmoid gate is the exact pushforward identity in coefficient space, because the output is no longer a polynomial in x. The explicit pushforward density and the dataset-by-dataset equivalence in Proposition 6.1 rely on the affine form (1).

We work in the canonical coordinates

$$s_0 = \frac{a_0 + b_0}{2}, \quad \delta_0 = \frac{a_0 - b_0}{2}, \quad s_1 = \frac{a_1 + b_1}{2}, \quad \delta_1 = \frac{a_1 - b_1}{2}.$$

From now on we identify θ with $(s_0, \delta_0, s_1, \delta_1, r)$, unless we refer explicitly to the branch coordinates (a_0, a_1, b_0, b_1, r). In these coordinates,

$$f(x;\theta) = s_0 + \big(s_1 + 2r\delta_0\big)x + 2r\delta_1 x^2. \tag{2}$$

The parameter-to-function map is therefore

$$\Phi(s_0, \delta_0, s_1, \delta_1, r) = \big(s_0,\ s_1 + 2r\delta_0,\ 2r\delta_1\big) = (p_0, p_1, p_2),$$

and the model output is the quadratic polynomial $p_0 + p_1 x + p_2 x^2$. The rest of the paper studies the fibers of Φ and the priors that Φ induces on the coefficients when a prior on θ is pushed through it.

The data are generated by the realizable Gaussian regression model

$$y = q(x) + \varepsilon, \qquad q(x) = q_0 + q_1 x + q_2 x^2, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \qquad x \sim \mathcal{N}(0, 1).$$

The population Kullback–Leibler (KL) loss is

$$K(\theta) = \frac{1}{2\sigma^2}\mathbb{E}\big[(f(x;\theta) - q(x))^2\big].$$

With $\Delta_k = p_k - q_k$, this becomes

$$K(\theta) = \frac{1}{2\sigma^2}\begin{pmatrix}\Delta_0\\ \Delta_1\\ \Delta_2\end{pmatrix}^{\top}\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 3\end{pmatrix}\begin{pmatrix}\Delta_0\\ \Delta_1\\ \Delta_2\end{pmatrix}. \tag{3}$$

The matrix in (3) is the Gram matrix $G = \mathbb{E}_x[\phi(x)\phi(x)^\top]$ with $\phi(x) = (1, x, x^2)^\top$. It is positive definite, so $K(\theta) = 0$ if and only if $\Phi(\theta) = (q_0, q_1, q_2)$.

RLCT relative to a compactly supported prior. Let $\varphi \geq 0$ be a smooth compactly supported prior density and define

$$\zeta_\varphi(s) = \int_{\mathbb{R}^5} K(\theta)^s \varphi(\theta)\, d\theta, \qquad \Re(s) > 0.$$

When the support reaches W_0, the meromorphic continuation of ζ_φ has its leading (largest) pole at $s = -\lambda$, with order m; the pair (λ, m) is the RLCT and its multiplicity relative to φ (Watanabe, 2009, 2018). In this model the set of true parameters $W_0 = \{\theta : K(\theta) = 0\}$ is unbounded, so the support of φ matters: it determines which parts of W_0 contribute to the asymptotics of the evidence. We say that the support of φ *reaches* a subset $S \subseteq W_0$ if φ is strictly positive on a neighborhood of some point of S.

Gaussian prior on the routed coordinates. For positive scale parameters $\tau_{s_0}, \tau_{\delta_0}, \tau_{s_1}, \tau_{\delta_1}, \tau_r$, define the product Gaussian prior in canonical coordinates by

$$\varphi_G(\theta) = \prod_{u \in \{s_0, \delta_0, s_1, \delta_1, r\}} \frac{1}{\sqrt{2\pi\tau_u^2}} \exp\left(-\frac{u^2}{2\tau_u^2}\right). \tag{4}$$

This prior is smooth and strictly positive on $\mathbb{R}^5$, and its tails decay exponentially. Unlike a compactly supported prior, it puts positive mass on a neighborhood of every point of W_0, including the collapsed stratum S_C of Proposition 4.2 when $q_2 = 0$. For this prior we consider the zeta function

$$\zeta_G(s) = \int_{\mathbb{R}^5} K(\theta)^s \varphi_G(\theta)\, d\theta, \qquad \Re(s) > 0.$$

The integral converges absolutely for $\Re(s) > 0$ because $K \geq 0$ is a polynomial and φ_G has exponentially decaying tails. Under φ_G, the *global RLCT* refers to the leading pole of ζ_G and its order.

Proposition 3.2 (Expressivity)**.** *The map $\Phi : \mathbb{R}^5 \to \mathbb{R}^3$ is surjective. That is, every quadratic polynomial $p_0 + p_1 x + p_2 x^2$ is the output of the routed model for some θ.*

The proof is a direct construction. It is given in Appendix B, together with the proofs of the results in Section 4.

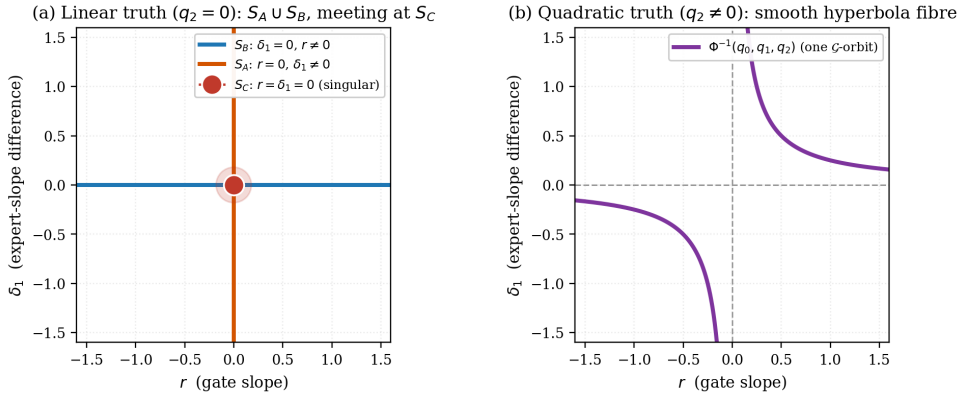

Figure 1: Projection of the true-parameter set $W_0 = \Phi^{-1}(q_0, q_1, q_2)$ onto the (r, δ_1) plane for (a) linear truth, $q_2 = 0$, and (b) quadratic truth, $q_2 \neq 0$. In panel (a), the red point is the projection of the collapsed stratum S_C.

4 Geometry of the Parameter Fibers

Theorem 4.1 (Fiber geometry and quotient identifiability)**.** *The map Φ is invariant under the two global transformations*

$$(s_0, \delta_0, s_1, \delta_1, r) \mapsto \big(s_0, \delta_0/\lambda, s_1, \delta_1/\lambda, \lambda r\big), \qquad \lambda \in \mathbb{R}^*, \tag{5}$$

$$(s_0, \delta_0, s_1, \delta_1, r) \mapsto \big(s_0, \delta_0 - t, s_1 + 2rt, \delta_1, r\big), \qquad t \in \mathbb{R}. \tag{6}$$

These generate a gauge group $\mathcal{G} \cong \mathbb{R}^* \ltimes \mathbb{R}$.

Moreover:

(i) *If $p_2 \neq 0$, the fiber $\Phi^{-1}(p_0, p_1, p_2)$ is a single $\mathcal{G}$-orbit. Hence the model is identifiable modulo $\mathcal{G}$ on the open set $\{p_2 \neq 0\}$.*

(ii) *If $p_2 = 0$, the fiber is the union of two branches,*

$$\{r = 0,\ s_0 = p_0,\ s_1 = p_1\} \quad \text{and} \quad \{\delta_1 = 0,\ s_0 = p_0,\ s_1 = p_1 - 2r\delta_0\},$$

which meet along $\{r = 0, \delta_1 = 0, s_0 = p_0, s_1 = p_1\}$.

Proposition 4.2 (Singular locus of the parameter-to-function map)**.** *The Jacobian of Φ drops rank exactly on*

$$\Sigma_\Phi = \{(s_0, \delta_0, s_1, \delta_1, r) : r = 0,\ \delta_1 = 0\}.$$

When $q_2 = 0$, the true set $W_0 = \Phi^{-1}(q_0, q_1, 0)$ decomposes into the smooth pieces

$$S_A = \{s_0 = q_0,\ s_1 = q_1,\ r = 0,\ \delta_1 \neq 0\},$$
$$S_B = \{s_0 = q_0,\ s_1 = q_1 - 2r\delta_0,\ \delta_1 = 0,\ r \neq 0\},$$

and the one-dimensional collapsed piece

$$S_C = \{s_0 = q_0,\ s_1 = q_1,\ r = 0,\ \delta_1 = 0\}.$$

Splitting S_A by the sign of δ_1 and S_B by the sign of r gives connected strata.

Interpretation. The global symmetries (5)–(6) should be distinguished from the extra freedom that appears only on $\{r = 0\}$ or $\{\delta_1 = 0\}$; the distinction matters both for identifiability and for the RLCT. The computations are elementary, but they explain why linear truth ($q_2 = 0$) is geometrically different from quadratic truth ($q_2 \neq 0$), as Figure 1 shows. For $q_2 = 0$, the constraint $2r\delta_1 = 0$ gives the cross $S_A \cup S_B \cup S_C$ in panel (a), with the collapsed stratum S_C at the origin. For $q_2 \neq 0$, the constraint $2r\delta_1 = q_2$ gives the smooth hyperbola in panel (b), a single $\mathcal{G}$-orbit that does not intersect either axis. The crossing at the origin of panel (a) comes from the product $r\delta_1$, which is also the source of the normal form $u^2 + v^2 + 8w^2z^2$ and of the multiplicity-2 RLCT.

5 Exact Singular Complexity

We now compute the RLCT of the routed model in two complementary ways. For compactly supported priors, we resolve the singularities locally (Theorem 5.1) and read off the RLCT from the resulting normal forms (Theorem 5.2). For the Gaussian prior, we compute its pushforward to coefficient space, which has a logarithmic singularity (Theorem 5.3); this gives the same raised multiplicity without any compactness assumption. We begin with the local normal forms used in the compact-support calculation.

Theorem 5.1 (Local normal forms)**.** (a) *At every point $\theta^* \in W_0$ such that $d\Phi_{\theta^*}$ has rank 3 (equivalently, $r^* \neq 0$ or $\delta_1^* \neq 0$), the loss is Morse–Bott with three quadratic normal directions. Equivalently, after a smooth local change of coordinates $(\eta, t) \in \mathbb{R}^3 \times \mathbb{R}^2$ adapted to W_0,*

$$K(\theta) = \frac{1}{2\sigma^2}\big(\eta_1^2 + \eta_2^2 + \eta_3^2\big).$$

In particular, this applies at every true parameter when $q_2 \neq 0$, and at every point of $S_A \cup S_B$ when $q_2 = 0$.

(b) *If $q_2 = 0$ and $\theta^* = (q_0, \delta_0^*, q_1, 0, 0) \in S_C$, write perturbations as $\varepsilon = \theta - \theta^*$. Then the polynomial change of variables*

$$\begin{aligned} u &= \varepsilon_1 + 2\varepsilon_4\varepsilon_5, \\ t &= \varepsilon_2, \\ v &= \varepsilon_3 + 2\delta_0^*\varepsilon_5 + 2\varepsilon_2\varepsilon_5, \\ w &= \varepsilon_4, \\ z &= \varepsilon_5 \end{aligned} \tag{7}$$

has Jacobian determinant 1 and gives the exact identity

$$K(\theta) = \frac{1}{2\sigma^2}\big(u^2 + v^2 + 8w^2z^2\big). \tag{8}$$

The coordinate t is tangent to S_C and does not appear in the loss.

Intuition. Theorem 5.1 reflects the redundancy of the routed parameterization. When $q_2 \neq 0$, the target is genuinely quadratic, so the gate must depend on the input ($r \neq 0$) and the two branches must have different slopes ($\delta_1 \neq 0$). Near such points Φ is a submersion, and the loss is a non-degenerate quadratic form in the three normal directions.

When the target is linear ($q_2 = 0$), the model can represent it in two ways: the gate can ignore the input ($r = 0$), or the two branches can share a slope ($\delta_1 = 0$). The stratum S_C is where both happen at once ($r = \delta_1 = 0$). The term $8w^2z^2$ in (8) records this. Once the two regular directions u and v are removed, the loss behaves like the squared product of the distances to $\{r = 0\}$ and to $\{\delta_1 = 0\}$, and it is this product structure that raises the multiplicity of the RLCT.

Theorem 5.2 (RLCT under compactly supported priors)**.** *Let φ be a smooth compactly supported prior whose support reaches W_0. If $q_2 = 0$, assume additionally that either the support reaches S_C or $\operatorname{supp}\varphi \cap S_C = \varnothing$.*

(a) *If $q_2 \neq 0$, then $(\lambda, m) = (3/2, 1)$.*
(b) *If $q_2 = 0$, then at points where φ is positive, the local RLCT is $(3/2, 1)$ on $S_A \cup S_B$ and $(3/2, 2)$ on S_C.*
(c) *Consequently, since the global RLCT is the smallest local λ together with the largest multiplicity among the points that attain it (Watanabe, 2009, Theorem 6.4), the global RLCT relative to φ for $q_2 = 0$ is*

$$(\lambda, m) = \begin{cases} (3/2, 2), & \text{if the support reaches } S_C, \\ (3/2, 1), & \text{if } \operatorname{supp}\varphi \cap S_C = \varnothing. \end{cases}$$

These are the two support regimes for compactly supported priors considered here. Priors that vanish on S_C while their support meets S_C are not classified by this theorem.

Theorem 5.3 (Pushforward of the Gaussian prior and its RLCT)**.** *Let μ_G be the pushforward of the Gaussian prior φ_G in* (4) *under the coefficient map Φ. Then μ_G has a locally integrable density ρ_G*

with respect to Lebesgue measure, so that

$$\int_{\mathbb{R}^5} F(\Phi(\theta))\,\varphi_G(\theta)\,d\theta = \int_{\mathbb{R}^3} F(p)\,\rho_G(p)\,dp$$

for every bounded measurable F. Moreover:

(a) *On the open set $\{p_2 \neq 0\}$, the density ρ_G is smooth and has the fiber-integral representation*

$$\rho_G(p_0,p_1,p_2) = C_G e^{-p_0^2/(2\tau_{s_0}^2)} \int_0^\infty \frac{\exp\left(-\frac{r^2}{2\tau_r^2} - \frac{p_2^2}{8\tau_{\delta_1}^2 r^2} - \frac{p_1^2}{2(\tau_{s_1}^2+4\tau_{\delta_0}^2 r^2)}\right)}{r\sqrt{\tau_{s_1}^2+4\tau_{\delta_0}^2 r^2}}\,dr \qquad (9)$$

with the constant $C_G = ((2\pi)^2\tau_{s_0}\tau_{\delta_1}\tau_r)^{-1} > 0$.

(b) *If $q_2 \neq 0$, then ρ_G is smooth and strictly positive in a neighborhood of (q_0,q_1,q_2). Consequently, the global RLCT under φ_G is $(3/2,1)$.*

(c) *If $q_2 = 0$, then in a neighborhood of $(q_0,q_1,0)$ and for $p_2 \neq 0$,*

$$\rho_G(p_0,p_1,p_2) = A_G(p_0,p_1)\log\frac{1}{|p_2|} + B_G(p_0,p_1,p_2), \qquad (10)$$

where A_G is continuous with $A_G(q_0,q_1) > 0$ and B_G is locally bounded. In particular, ρ_G extends across $p_2 = 0$ as a locally integrable density, and the global RLCT under φ_G is $(3/2,2)$.

Appendix D.1 sketches the arguments for Theorems 5.2 and 5.3, including the fiber integral that produces the logarithmic singularity. Complete proofs are in Appendices C–E.

Proposition 5.4 (Sigmoid gate: local product singularity at the collapsed stratum)**.** *Let $K_\sigma(\theta) = \frac{1}{2\sigma^2}\mathbb{E}_x[(f_\sigma(x;\theta) - q(x))^2]$ be the population KL loss with sigmoid gate $g_\sigma(x;r) = \sigma(rx) = (1+e^{-rx})^{-1}$, and assume $q_2 = 0$. Let $\theta^* = (q_0,0,q_1,0,0)$ be the point of the collapsed stratum with $\delta_0 = 0$. Then there is an analytic local diffeomorphism Ψ near θ^* such that, in coordinates $\Psi(\theta) = (u,t,v,w,z)$,*

$$2\sigma^2 K_\sigma(\theta) = u^2 + v^2 + c\,w^2z^2 + R(u,t,v,w,z), \qquad c = \tfrac{1}{2},$$

where R vanishes to order at least 5 at the origin. The principal part $u^2+v^2+c\,w^2z^2$ is non-degenerate with respect to its Newton polyhedron and has leading pole pair $(3/2,2)$. Thus the sigmoid model has the same multiplicity-2 product singularity at θ^*, in the sense that its local zeta function at θ^* has leading pole pair $(3/2,2)$. For symmetric covariates whose bounded support contains an interval, K_σ is real-analytic near θ^* and this pair is the ordinary local RLCT.

At the level of the leading Taylor terms, the same multiplicity-2 product structure arises for any gate $\widetilde{g}(x;r)$ with $\widetilde{g}(x;0) = 1/2$ and $\partial_r\widetilde{g}(x;0) = \alpha x$ for some $\alpha \neq 0$; the sigmoid $\sigma(rx)$ is the case $\alpha = 1/4$. Proposition 5.4 is local at θ^*, and it does not extend the global pushforward identity of Proposition 6.1 to sigmoid gates.

Proposition 5.5 (Local RLCT at the collapsed stratum for d-dimensional inputs)**.** *With $x \sim \mathcal{N}(0,I_d)$, two linear branches $e_i(x) = a_{i,0} + a_{i,1}^\top x$ and gate $g(x;r) = \frac{1}{2} + r^\top x$ ($a_{i,0} \in \mathbb{R}$, $a_{i,1}, r \in \mathbb{R}^d$), and linear truth $q(x) = q_0 + q_1^\top x$, the model output is $f(x;\theta) = s_0 + (s_1 + 2\delta_0 r)^\top x + 2(r^\top x)(\delta_1^\top x)$. At the collapsed-stratum point $\theta^* = (q_0,0,q_1,0,0)$ a polynomial change of variables with Jacobian 1 gives, in coordinates $(u,t,v,w,z) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d$ centered at θ^*,*

$$2\sigma^2 K(\theta) = u^2 + \|v\|^2 + 4\big(\|w\|^2\|z\|^2 + (w^\top z)^2\big), \qquad (11)$$

with t tangent. The local RLCT at θ^* is $((2d+1)/2,\ 2)$, recovering Theorem 5.2(b) when $d = 1$. The multiplicity $m = 2$ does not depend on d, while λ grows linearly in d.

Consequently, a prior that is positive at θ^* and supported in a small neighborhood of it contributes a single $\log\log n$ term in every dimension. Symbolic computations in the supplementary material confirm (11) and the unit Jacobian for $d \in \{1,2,3,4\}$. For $d \geq 2$, however, the model no longer represents all quadratic functions of x. For sigmoid gates on linear truth, Proposition F.4 in Appendix F.2 describes the full zero set and the resulting leading pole pair for compactly supported priors.

Table 1: Global RLCT pairs (λ, m) for the support regimes in Theorem 5.2 and under the full-support Gaussian prior φ_G. The compact priors listed reach W_0; on linear truth their supports either reach S_C or are disjoint from it.

Regime	Prior	(λ, m)
$q_2 \neq 0$	compactly supported (reaching W_0) or Gaussian φ_G	$(3/2, 1)$
$q_2 = 0$	compactly supported, reaching W_0, disjoint from S_C	$(3/2, 1)$
$q_2 = 0$	compactly supported, reaching S_C	$(3/2, 2)$
$q_2 = 0$	Gaussian φ_G	$(3/2, 2)$

6 Bayesian Consequences

We now turn to the consequences for Bayesian inference. Since Φ maps onto the full family of quadratic regression functions (Proposition 3.2), the routed model with an affine gate is, in coefficient space, quadratic regression with a particular prior. The routed and quadratic models therefore differ only through their priors, and this difference is what changes the asymptotics of the evidence in Theorem 6.4.

Proposition 6.1 (Exact equivalence of coefficient posteriors)**.** *For every dataset $(x_i, y_i)_{i=1}^n$ and every bounded measurable $h : \mathbb{R}^3 \to \mathbb{R}$,*

$$\frac{\int_{\mathbb{R}^5} h(\Phi(\theta)) \prod_{i=1}^n p(y_i \mid x_i, \theta)\, \varphi_G(\theta)\, d\theta}{\int_{\mathbb{R}^5} \prod_{i=1}^n p(y_i \mid x_i, \theta)\, \varphi_G(\theta)\, d\theta} = \frac{\int_{\mathbb{R}^3} h(p) \prod_{i=1}^n p(y_i \mid x_i, p)\, \rho_G(p)\, dp}{\int_{\mathbb{R}^3} \prod_{i=1}^n p(y_i \mid x_i, p)\, \rho_G(p)\, dp}.$$

That is, for every dataset, the posterior over quadratic coefficients induced by the routed model is exactly the quadratic-regression posterior under the prior ρ_G.

Proof. Apply the pushforward identity in Theorem 5.3 with

$$F(p) = h(p) \prod_{i=1}^n p(y_i \mid x_i, p)$$

for the numerator and with $F(p) = \prod_{i=1}^n p(y_i \mid x_i, p)$ for the denominator. □

Proposition 6.2 (Posterior contraction at the regular rate)**.** *Assume linear truth $q = (q_0, q_1, 0)$, and let Π_n^ρ denote the posterior on quadratic coefficients induced by the routed model under φ_G, or equivalently the quadratic-regression posterior under ρ_G. Then, for every deterministic sequence $M_n \to \infty$,*

$$\Pi_n^\rho\Big(\|p - q\| > M_n n^{-1/2} \mid (x_i, y_i)_{i=1}^n\Big) \to 0$$

in P_q-probability. If P_p denotes the joint law of (x, y) under coefficient p and h denotes the Hellinger distance, then for some finite constant C depending only on σ^2 and G,

$$\Pi_n^\rho\Big(h(P_p, P_q) > C M_n n^{-1/2} \mid (x_i, y_i)_{i=1}^n\Big) \to 0$$

in P_q-probability.

The proofs of Proposition 6.2 and of the two theorems below are in Appendix G.

Theorem 6.3 (Marginal-likelihood asymptotics)**.** *Let $(x_i, y_i)_{i=1}^n$ be i.i.d. from the realizable Gaussian model above, let*

$$Z_n = \int_{\mathbb{R}^5} \prod_{i=1}^n p(y_i \mid x_i, \theta)\, \varphi(\theta)\, d\theta,$$

where $p(y \mid x, \theta)$ is the Gaussian conditional density with mean $f(x; \theta)$ and variance σ^2. Let $p^\star(y \mid x)$ denote the true conditional density.

(a) *For a smooth compactly supported prior φ satisfying the support assumptions of Theorem 5.2,*

$$\log Z_n = \sum_{i=1}^n \log p^\star(y_i \mid x_i) - \lambda \log n + (m - 1) \log\log n + O_p(1), \tag{12}$$

where (λ, m) is the RLCT pair relative to φ given by Theorem 5.2.

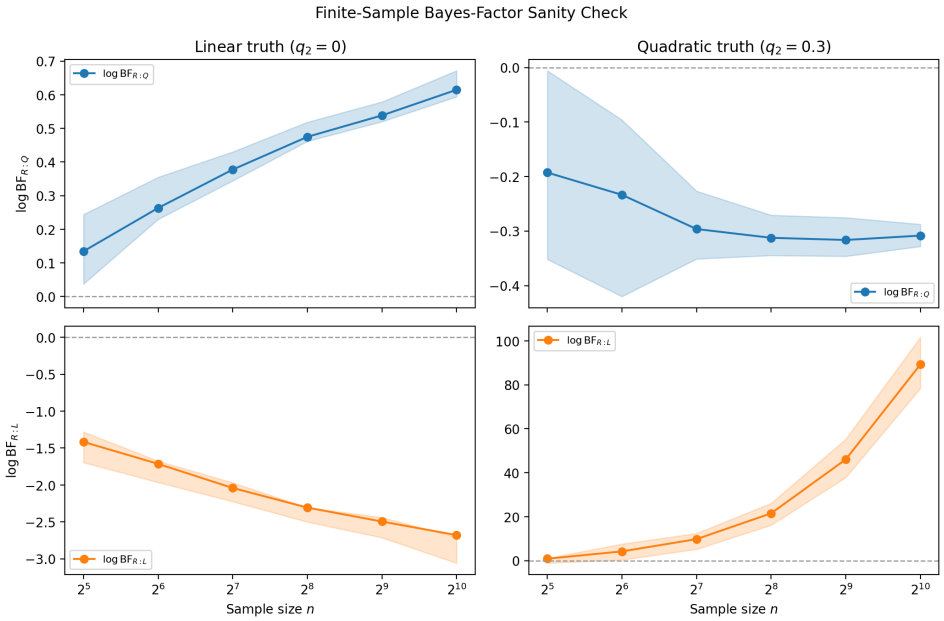

Figure 2: Mean log Bayes factors $\log \mathrm{BF}_{R:Q}$ (top) and $\log \mathrm{BF}_{R:L}$ (bottom) at unit prior scales, for linear truth $q_2 = 0$ (left) and quadratic truth $q_2 = 0.3$ (right). Each point averages 40 datasets; shaded bands show interquartile ranges.

(b) *Under the Gaussian prior φ_G,*

$$\log Z_n = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - \frac{3}{2} \log n + \mathbf{1}[q_2 = 0] \log \log n + O_p(1). \tag{13}$$

Theorem 6.4 (Bayes factors against baselines with smooth priors)**.** *Let M_R be the routed model with the Gaussian prior φ_G; by Proposition 6.1, in coefficient space this is quadratic regression with prior ρ_G. Let M_Q be ordinary quadratic regression*

$$\beta_0 + \beta_1 x + \beta_2 x^2$$

with a prior density π_Q that is smooth and strictly positive near the true coefficient vector, and let M_L be ordinary linear regression

$$\alpha_0 + \alpha_1 x$$

with a prior density π_L that is smooth and strictly positive near the relevant true or pseudo-true parameter. Both baseline models use the same known noise variance σ^2 as the routed model. Write $Z_{R,n}, Z_{Q,n}, Z_{L,n}$ for the three marginal likelihoods and $\mathrm{BF}_{A:B,n} = Z_{A,n}/Z_{B,n}$ for Bayes factors.

Then:

(a) *If $q_2 = 0$, all three models are well specified, and*

$$\log \mathrm{BF}_{R:Q,n} = \log \log n + O_p(1), \qquad \log \mathrm{BF}_{R:L,n} = -\frac{1}{2} \log n + \log \log n + O_p(1).$$

Thus the routed model beats quadratic regression with a smooth prior by a $\log \log n$ margin, but it still loses to linear regression.

(b) *If $q_2 \neq 0$, the routed and quadratic models are well specified while the linear model is misspecified. Writing*

$$\kappa_L = \inf_{\alpha_0, \alpha_1} \frac{1}{2\sigma^2} \mathbb{E}\Big[\big(q_0 + q_1 x + q_2 x^2 - \alpha_0 - \alpha_1 x\big)^2\Big] = \frac{q_2^2}{\sigma^2}, \tag{14}$$

we have

$$\log \mathrm{BF}_{R:Q,n} = O_p(1), \qquad \log \mathrm{BF}_{R:L,n} = n\kappa_L + O_p(\sqrt{n}).$$

Thus the routed and quadratic models agree up to an $O_p(1)$ term, and both beat linear regression exponentially fast.

Interpretation. Theorem 6.4 compares the singular prior ρ_G, induced by a Gaussian on the routed coordinates, with smooth priors placed directly on the coefficients. Routing does not enlarge the function class here; its only effect is to make the coefficient prior logarithmically singular. This raises the order of the leading pole and produces the $\log\log n$ advantage on linear truth. Because the effect comes from the geometry of the prior, it is not a prior-invariant property of quadratic regression, and by Proposition 6.2 it does not slow down posterior contraction in coefficient space. In finite samples the effect is modest: on linear truth, Figure 2 gives a geometric mean of $\mathrm{BF}_{R:Q}$ of about 1.86 at $n = 1024$. Further numerical checks are collected in Appendix H.

Finite-sample posterior shape. The singular prior also changes the posterior at moderate sample sizes. Consider a one-dimensional Gaussian sequence model for the curvature coefficient p_2 at unit prior scales. Under the induced prior on p_2, which is a product-normal distribution, the posterior mean shrinks observations near zero by about 19% at $n = 10^3$, compared with 0.01% under a Gaussian prior with the same variance. This calculation concerns only the marginal prior of p_2; it is not a result about the joint regression posterior or about predictive rates. Appendix G.1 gives the derivation and also quantifies the joint posterior mass near the linear submodel.

Model selection. Applied to the five routed coordinates, BIC would charge a penalty of $(5/2)\log n$. The singular analysis gives the correct leading penalty $(3/2)\log n$, and on the linear submodel the multiplicity-2 pole of the pushforward prior adds the $\log\log n$ term. This term changes the marginal likelihood; by Proposition 6.2, the coefficient posterior still contracts at the regular $n^{-1/2}$ rate, also in predictive Hellinger distance.

7 Discussion, Limitations, and Conclusion

In an exactly solvable routed model, we have shown that a Gaussian prior on the routing coordinates induces a logarithmically singular prior on the regression coefficients, even though routing leaves the function class unchanged. On linear truth, this gives the routed model a $\log\log n$ advantage in log Bayes factor over quadratic regression with a smooth coefficient prior. The gain differs in kind from the $O(1)$ effect of changing a smooth prior and the $O(\log n)$ effect of changing the model dimension, but it is numerically small: $\log\log 10^3 \approx 1.93$ and $\log\log 10^6 \approx 2.63$. It acts on the marginal likelihood through the pole multiplicity and does not slow posterior contraction in coefficient space.

Scope and limitations. The explicit singular prior (Proposition E.1) and the exact equivalence of coefficient posteriors (Proposition 6.1) are proved only for the two-branch model with scalar input, affine gate, and Gaussian prior. The product singularity persists locally for sigmoid gates on linear truth (Propositions 5.4 and F.4) and for d-dimensional inputs (Proposition 5.5), but the exact pushforward to coefficient space does not; for sigmoid gates with Gaussian covariates, we identify only the leading pole of the local zeta function. We do not prove a prior-invariant Bayes-factor theorem or extend the global pushforward to d dimensions, and we make no quantitative claims about models with more than two branches, softmax gating, or finite-time SGD dynamics. The input distribution is fixed independently of the model parameters; jointly learned routing features are outside the analysis. Our results concern Bayesian model selection, not predictive performance or the dynamics of gradient descent. The main risk in interpreting them is extrapolating these exact asymptotics for a small model to large routing architectures.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding for this work and have no competing interests to disclose.

References

M. Aoyagi. Stochastic Complexity and Generalization Error of a Restricted Boltzmann Machine in Bayesian Estimation. *Journal of Machine Learning Research*, 11:1243–1272, 2010.

M. Aoyagi and S. Watanabe. Stochastic Complexities of Reduced Rank Regression in Bayesian Estimation. *Neural Networks*, 18(7):924–933, 2005.

J. Chen. Optimal Rate of Convergence for Finite Mixture Models. *Annals of Statistics*, 23(1):221–233, 1995.

M. Drton and M. Plummer. A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society: Series B*, 79(2):323–380, 2017.

W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

A. Gelman. Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545, 2004.

N. Ho and X. L. Nguyen. Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures of Distributions. *SIAM Journal on Mathematics of Data Science*, 1(4):730–758, 2019.

N. Ho, C.-Y. Yang, and M. I. Jordan. Convergence Rates for Gaussian Mixtures of Experts. *Journal of Machine Learning Research*, 23(323):1–81, 2022.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

H. Jeffreys. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A*, 186:453–461, 1946.

W. Jiang and M. A. Tanner. On the identifiability of mixtures-of-experts. *Neural Networks*, 12(9):1253–1258, Nov. 1999.

V. E. Johnson and D. Rossell. On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests. *Journal of the Royal Statistical Society: Series B*, 72(2):143–170, 2010.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

R. Kawata, K. Matsutani, Y. Kinoshita, N. Nishikawa, and T. Suzuki. Mixture of Experts Provably Detect and Learn the Latent Cluster Structure in Gradient-Based Learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 29390–29448, 2025.

B. J. K. Kleijn and A. W. van der Vaart. The Bernstein–von Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

Y. Kurumadani. Learning Coefficients in Semi-Regular Models. *arXiv preprint arXiv:2406.02646*, 2024.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

E. Lau, Z. Furman, G. Wang, D. Murfet, and S. Wei. The Local Learning Coefficient: A Singularity-Aware Complexity Measure. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 244–252, 2025.

D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations*, 2021.

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

S. Lippl, L. F. Abbott, and S. Chung. The Implicit Bias of Gradient Descent on Generalized Gated Linear Networks. *arXiv preprint arXiv:2202.02649*, 2022.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

H. Nguyen, N. Ho, and A. Rinaldo. On Least Square Estimation in Softmax Gating Mixture of Experts. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 37707–37735, 2024.

J. Rousseau and K. Mengersen. Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models. *Journal of the Royal Statistical Society: Series B*, 73(5):689–710, 2011.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Y. Su, H. Tang, Z. Gong, and Y. Liu. Sparsity is Combinatorial Depth: Quantifying MoE Expressivity via Tropical Geometry. *arXiv preprint arXiv:2602.03204*, 2026.

M. D. Toure and D. A. Stephens. Singular Bayesian Neural Networks. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 121841–121911, 2026.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

S. Watanabe. Algebraic Analysis for Nonidentifiable Learning Machines. *Neural Computation*, 13(4):899–933, 2001.

K. Watanabe and S. Watanabe. Stochastic Complexities of Gaussian Mixtures in Variational Bayesian Approximation. *Journal of Machine Learning Research*, 7:625–644, 2006.

S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

S. Watanabe. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

S. Watanabe. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14(27):867–897, 2013.

S. Watanabe. *Mathematical Theory of Bayesian Statistics*. CRC Press, 2018.

T. Watanabe and S. Watanabe. Asymptotic Behavior of Bayesian Generalization Error in Multinomial Mixtures. *arXiv preprint arXiv:2203.06884*, 2022.

S. R. Waterhouse, D. J. C. MacKay, and T. Robinson. Bayesian Methods for Mixtures of Experts. In *Advances in Neural Information Processing Systems*, volume 8, pages 351–357, 1995.

S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

Appendix roadmap. Appendix A collects the notation. Appendix B proves the results on expressivity and fiber geometry, and Appendix C proves the local normal forms. Appendix D proves the RLCT theorem for compactly supported priors and sketches the main RLCT arguments, and Appendix E derives the pushforward density and the RLCT under the Gaussian prior. Appendix F treats the sigmoid gate at the collapsed stratum; Appendix F.1 covers d-dimensional inputs, and Appendix F.2 describes the full zero set of the sigmoid model on linear truth. Appendix G proves the evidence, contraction, and Bayes-factor results. It also contains the WBIC expansion, the robustness of the RLCT under smooth changes of the prior, the finite-sample curvature calculation, the case of unknown noise variance (Appendix G.2), and the low-temperature Gibbs corollary (Corollary G.3). Appendix H collects the numerical illustrations.

A Notation

Table 2 collects the symbols used throughout the paper, together with their first definition.

Table 2: Notation used throughout the paper.

Symbol	Meaning	Defined in
$x \in \mathbb{R}$	scalar input feature, $x \sim \mathcal{N}(0,1)$	Definition 3.1
θ	routed parameter, identified with canonical coordinates after the change of variables	Section 3
(a_0, a_1, b_0, b_1, r)	original branch coordinates	Definition 3.1
$(s_0, \delta_0, s_1, \delta_1, r)$	canonical coordinates used henceforth	Section 3
$\Phi(\theta) = (p_0, p_1, p_2)$	parameter-to-coefficient map of the model	Section 3
p_0, p_1, p_2	constant, linear, and quadratic coefficients	Section 3
$q = (q_0, q_1, q_2)$	true target coefficients	Section 3
σ^2	noise variance of the realizable Gaussian regression	Section 3
$G = \mathbb{E}[\phi(x)\phi(x)^\top]$	Gram matrix of the features $\phi(x) = (1, x, x^2)^\top$	Section 3
$K(\theta)$	population KL loss against the truth	(3)
$W_0 = \{K(\theta) = 0\}$	true-parameter set	Proposition 4.2
S_A, S_B, S_C	smooth strata within $\{r = 0\}$, $\{\delta_1 = 0\}$, and their intersection	Proposition 4.2
Σ_Φ	singular locus of Φ	Proposition 4.2
$\mathcal{G} \cong \mathbb{R}^* \ltimes \mathbb{R}$	gauge group of Φ	Theorem 4.1
φ	smooth compactly supported prior	Section 3
φ_G, ρ_G	Gaussian prior on the routed coordinates; its pushforward density	(4), Theorem 5.3
$\tau_{s_0}, \tau_{\delta_0}, \tau_{s_1}, \tau_{\delta_1}, \tau_r$	Gaussian prior scales on the routed coordinates	(4)
A_G, B_G, C_G	coefficient, remainder, and normalization constant of the density expansion	Proposition E.1
$\zeta_\varphi(s), \zeta_G(s)$	zeta functions defining the RLCT	Section 3
(λ, m)	RLCT pair (real log canonical threshold, multiplicity)	Theorem 5.2
Z_n, $Z_{R,n}, Z_{Q,n}, Z_{L,n}$	marginal likelihoods of routed, quadratic, linear models	Theorems 6.3 and 6.4
$\mathrm{BF}_{A:B,n}$	Bayes factor $Z_{A,n}/Z_{B,n}$	Theorem 6.4
$\mathrm{WBIC}(n)$	widely applicable Bayesian information criterion	Proposition G.1
κ_L	misspecification penalty of the linear model	(14)
η	temperature; inverse temperature $1/\eta$	Corollary G.3
π_η, Z_η	Gibbs posterior and its normalizer	Section G.3
U_A, U_B, U_C	open neighborhoods of compact subsets of the strata	Corollary G.3
(u, v, w, z, t)	coordinates in the local normal form near S_C	Theorem 5.1(b)
(η_1, η_2, η_3)	Morse–Bott normal coordinates near S_A or S_B	Theorem 5.1(a)

B Algebra and Fiber Geometry

Proof of Proposition 3.2. Fix $(p_0, p_1, p_2) \in \mathbb{R}^3$.

If $p_2 \neq 0$, choose $r = 1$, $\delta_1 = p_2/2$, $s_0 = p_0$, $\delta_0 = 0$, and $s_1 = p_1$. Then

$$\Phi(p_0, 0, p_1, p_2/2, 1) = (p_0, p_1, p_2).$$

If $p_2 = 0$, choose $r = 0$, $s_0 = p_0$, $s_1 = p_1$, and any δ_0, δ_1 (for instance 0). Then

$$\Phi(p_0, 0, p_1, 0, 0) = (p_0, p_1, 0).$$

Hence every quadratic polynomial has a preimage. □

Proof of Theorem 4.1. The invariance of (5) and (6) is immediate from

$$\Phi(s_0, \delta_0, s_1, \delta_1, r) = \big(s_0,\ s_1 + 2r\delta_0,\ 2r\delta_1\big).$$

For (5),

$$2(\lambda r)(\delta_0/\lambda) = 2r\delta_0, \qquad 2(\lambda r)(\delta_1/\lambda) = 2r\delta_1.$$

For (6),

$$\big(s_1 + 2rt\big) + 2r(\delta_0 - t) = s_1 + 2r\delta_0, \qquad 2r\delta_1 \text{ is unchanged.}$$

Writing S_λ and T_t for the transformations in (5) and (6), one also checks

$$S_\lambda \circ T_t \circ S_\lambda^{-1} = T_{t/\lambda},$$

so the generated symmetry group is the semidirect product $\mathcal{G} \cong \mathbb{R}^* \ltimes \mathbb{R}$.

Now fix two points θ_1, θ_2 in the same fiber with $p_2 \neq 0$. Then $2r_i\delta_{1,i} = p_2 \neq 0$, so $r_1, r_2 \neq 0$. Apply the scaling action to θ_1 with $\lambda = r_2/r_1$ so that the router coordinates agree. The equality of p_2 then forces the rescaled δ_1 to equal $\delta_{1,2}$. Writing $\delta'_{0,1}$ for the rescaled value of δ_0, choose the translation parameter

$$t = \delta'_{0,1} - \delta_{0,2}$$

to match δ_0; the equality of $p_1 = s_1 + 2r\delta_0$ then forces s_1 to match as well. Thus the whole fiber is one $\mathcal{G}$-orbit.

For $p_2 = 0$, the defining equations are

$$s_0 = p_0, \qquad s_1 + 2r\delta_0 = p_1, \qquad 2r\delta_1 = 0.$$

The last equation gives the union of the two branches $r = 0$ and $\delta_1 = 0$, which intersect along $r = \delta_1 = 0$. □

Proof of Proposition 4.2. Differentiating Φ gives

$$J_\Phi = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2r & 1 & 0 & 2\delta_0 \\ 0 & 0 & 0 & 2r & 2\delta_1 \end{pmatrix}.$$

The first row is always nonzero. The last two rows are linearly independent if either $r \neq 0$ or $\delta_1 \neq 0$, and become dependent exactly when $r = 0$ and $\delta_1 = 0$. Hence $\operatorname{rank}(J_\Phi) = 3$ off Σ_Φ and $\operatorname{rank}(J_\Phi) = 2$ on Σ_Φ.

When $q_2 = 0$, the true set is defined by

$$s_0 = q_0, \qquad s_1 + 2r\delta_0 = q_1, \qquad r\delta_1 = 0.$$

The last relation splits the set into $r = 0$ or $\delta_1 = 0$. This yields the two smooth two-dimensional pieces S_A and S_B and the one-dimensional intersection piece S_C. Splitting S_A and S_B by sign gives connected smooth strata. □

C Local Normal Forms

Proof of Theorem 5.1. Fix a base point $\theta^* \in W_0$ and write perturbations as

$$\varepsilon = \theta - \theta^* = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5)$$

in the coordinates $(s_0, \delta_0, s_1, \delta_1, r)$.

Generic rank-3 case. Assume first that $d\Phi_{\theta^*}$ has rank 3. By the constant-rank theorem, there are local coordinates $(\xi, t) \in \mathbb{R}^3 \times \mathbb{R}^2$ centered at θ^* such that

$$\Phi(\theta) - \Phi(\theta^*) = \xi.$$

Since $\theta^* \in W_0$, we have $\Phi(\theta^*) = (q_0, q_1, q_2)$, and therefore

$$K(\theta) = \frac{1}{2\sigma^2}\, \xi^\top G\, \xi.$$

Because $G \succ 0$, a linear change of variables $\eta = L\xi$ reduces this quadratic form to

$$K(\theta) = \frac{1}{2\sigma^2}\big(\eta_1^2 + \eta_2^2 + \eta_3^2\big).$$

This proves part (a) in full generality.

Explicit coordinates on the smooth strata when $q_2 = 0$. For completeness, we also write out the normal directions on S_A and S_B explicitly.

Smooth branch S_A. Let $\theta^* = (q_0, \delta_0^*, q_1, \delta_1^*, 0)$ with $\delta_1^* \neq 0$. Then

$$\Delta_0 = \varepsilon_1, \tag{15}$$
$$\Delta_1 = \varepsilon_3 + 2\delta_0^*\varepsilon_5 + 2\varepsilon_2\varepsilon_5, \tag{16}$$
$$\Delta_2 = 2\delta_1^*\varepsilon_5 + 2\varepsilon_4\varepsilon_5. \tag{17}$$

Along W_0, the tangent directions are ε_2 and ε_4. Restrict to the transverse slice $\varepsilon_2 = \varepsilon_4 = 0$. On that slice,

$$(\Delta_0, \Delta_1, \Delta_2) = L_A(\varepsilon_1, \varepsilon_3, \varepsilon_5) = \big(\varepsilon_1,\ \varepsilon_3 + 2\delta_0^*\varepsilon_5,\ 2\delta_1^*\varepsilon_5\big),$$

and L_A is invertible because $\delta_1^* \neq 0$. Since the Gram matrix in (3) is positive definite and L_A is invertible, the matrix

$$\frac{1}{2\sigma^2} L_A^\top G L_A$$

is positive definite on the transverse slice. Therefore the Hessian of K is non-degenerate in the normal directions and vanishes in the tangent directions. By the Morse–Bott lemma, there are local coordinates $(\eta, t) \in \mathbb{R}^3 \times \mathbb{R}^2$ with

$$K = \frac{1}{2\sigma^2}\big(\eta_1^2 + \eta_2^2 + \eta_3^2\big).$$

Smooth branch S_B. Let $\theta^* = (q_0, \delta_0^*, q_1 - 2r^*\delta_0^*, 0, r^*)$ with $r^* \neq 0$. Then

$$\Delta_0 = \varepsilon_1,$$
$$\Delta_1 = \varepsilon_3 + 2r^*\varepsilon_2 + 2\delta_0^*\varepsilon_5 + 2\varepsilon_2\varepsilon_5,$$
$$\Delta_2 = 2r^*\varepsilon_4 + 2\varepsilon_4\varepsilon_5.$$

Here ε_2 and ε_5 parameterize the true-set component. On the transverse slice $\varepsilon_2 = \varepsilon_5 = 0$,

$$(\Delta_0, \Delta_1, \Delta_2) = L_B(\varepsilon_1, \varepsilon_3, \varepsilon_4) = \big(\varepsilon_1,\ \varepsilon_3,\ 2r^*\varepsilon_4\big),$$

and L_B is invertible because $r^* \neq 0$. The same Morse–Bott argument applies.

Collapsed branch S_C. Let $\theta^* = (q_0, \delta_0^*, q_1, 0, 0)$. Then

$$\Delta_0 = \varepsilon_1, \tag{18}$$
$$\Delta_1 = \varepsilon_3 + 2\delta_0^*\varepsilon_5 + 2\varepsilon_2\varepsilon_5, \tag{19}$$
$$\Delta_2 = 2\varepsilon_4\varepsilon_5. \tag{20}$$

Substituting (18)–(20) into (3) yields

$$\begin{aligned} 2\sigma^2 K &= \Delta_0^2 + \Delta_1^2 + 3\Delta_2^2 + 2\Delta_0\Delta_2 \\ &= \varepsilon_1^2 + \big(\varepsilon_3 + 2\delta_0^*\varepsilon_5 + 2\varepsilon_2\varepsilon_5\big)^2 + 12\varepsilon_4^2\varepsilon_5^2 + 4\varepsilon_1\varepsilon_4\varepsilon_5 \\ &= \big(\varepsilon_1 + 2\varepsilon_4\varepsilon_5\big)^2 + \big(\varepsilon_3 + 2\delta_0^*\varepsilon_5 + 2\varepsilon_2\varepsilon_5\big)^2 + 8\varepsilon_4^2\varepsilon_5^2. \end{aligned}$$

This is exactly (8) under the change of variables (7). The inverse map is

$$\varepsilon_1 = u - 2wz, \qquad \varepsilon_2 = t, \qquad \varepsilon_3 = v - 2\delta_0^* z - 2tz, \qquad \varepsilon_4 = w, \qquad \varepsilon_5 = z,$$

so the map is a polynomial automorphism. Its Jacobian matrix is triangular with ones on the diagonal, hence has determinant 1. □

D RLCT of the Collapsed Normal Form

Proposition D.1 (RLCT of the product normal form)**.** *The function*

$$h(u, v, w, z) = u^2 + v^2 + w^2 z^2$$

has local RLCT $(3/2, 2)$ *at the origin.*

Proof. Multiplying the integrand by a smooth positive density does not change the location or order of the leading pole, so it suffices to work on $D_\varepsilon \times [-\varepsilon, \varepsilon]^2$ with constant density, where $D_\varepsilon = \{(u, v) : u^2 + v^2 < \varepsilon^2\}$.

Define

$$\zeta(s) = \int_{D_\varepsilon \times [-\varepsilon,\varepsilon]^2} h(u, v, w, z)^s \, du \, dv \, dw \, dz.$$

Use polar coordinates in (u, v):

$$u = R\cos\phi, \qquad v = R\sin\phi, \qquad du\, dv = R\, dR\, d\phi.$$

Then

$$\zeta(s) = 2\pi \int_0^\varepsilon \int_{-\varepsilon}^\varepsilon \int_{-\varepsilon}^\varepsilon \big(R^2 + w^2 z^2\big)^s R \, dw \, dz \, dR.$$

For $a = w^2 z^2$,

$$\int_0^\varepsilon (R^2 + a)^s R \, dR = \frac{(\varepsilon^2 + a)^{s+1} - a^{s+1}}{2(s + 1)}.$$

Hence

$$\zeta(s) = \frac{\pi}{s + 1}\big(A(s) - B(s)\big),$$

where

$$A(s) = \int_{-\varepsilon}^\varepsilon \int_{-\varepsilon}^\varepsilon \big(\varepsilon^2 + w^2 z^2\big)^{s+1} \, dw \, dz,$$

and

$$B(s) = \int_{-\varepsilon}^\varepsilon \int_{-\varepsilon}^\varepsilon |w|^{2s+2} |z|^{2s+2} \, dw \, dz = \frac{4\varepsilon^{4s+6}}{(2s + 3)^2}.$$

The function $A(s)$ is holomorphic near $s = -3/2$ because the integrand is bounded and holomorphic there. The apparent pole at $s = -1$ cancels because

$$A(-1) = \iint 1 \, dw \, dz = 4\varepsilon^2 = B(-1).$$

Therefore the first non-removable pole comes from $B(s)$ at $2s + 3 = 0$, namely $s = -3/2$, and its order is exactly 2.

Thus the local RLCT is $(\lambda, m) = (3/2, 2)$. □

Proof of Theorem 5.2. If $q_2 \neq 0$, then $2r\delta_1 = q_2$ on W_0, so every true parameter lies away from Σ_Φ. By Theorem 5.1, the loss is Morse–Bott with three quadratic normal directions at every point of W_0, in particular at every such point where φ is positive. These local contributions have RLCT $(3/2, 1)$, and so does the global integral.

Assume now that $q_2 = 0$. By Theorem 5.1, every point of S_A or S_B where φ is positive has local normal form $\eta_1^2 + \eta_2^2 + \eta_3^2$, hence local RLCT $(3/2, 1)$. Every point of S_C where φ is positive has the exact normal form $u^2 + v^2 + 8w^2 z^2$; the constant factor 8 does not change the RLCT, so Proposition D.1 gives local RLCT $(3/2, 2)$.

The global RLCT relative to φ is obtained by taking the smallest local exponent among strata reached by the support and, among those with minimal exponent, the largest multiplicity (Watanabe, 2009, 2018). This yields the stated cases. □

D.1 Proof Sketches for the Main RLCT Theorems

Theorems 5.2 and 5.3 are proved by complementary methods: local resolution of singularities for compactly supported priors, and explicit integration along the fibers of Φ for the Gaussian prior. We summarize the main mechanism here; full proofs are in Appendices D and E.

Compactly supported priors. The asymptotics of the evidence are governed by the local geometry of W_0. When $q_2 \neq 0$, every true parameter lies in the generic Morse–Bott stratum and has local RLCT $(3/2, 1)$. When $q_2 = 0$ and the support reaches S_C, the exact normal form $u^2 + v^2 + 8w^2z^2$ at S_C from Theorem 5.1(b) governs the singularity: integrating the local zeta function $\int (u^2 + v^2 + 8w^2z^2)^s \, du\, dv\, dw\, dz$ in polar coordinates on (u, v) isolates a radial integral $\int R^{2s+2} \log(1/R)\, dR$, which produces a leading pole at $s = -3/2$ with multiplicity $m = 2$.

Gaussian prior and the fiber integral. For Theorem 5.3, we push the prior forward along Φ to coefficient space. The change of variables $(s_0, \delta_0, s_1, \delta_1, r) \mapsto (p_0, p_1, p_2, \delta_0, r)$ has Jacobian $2r$, so $d\theta = \frac{1}{2|r|}\, dp_0\, dp_1\, dp_2\, d\delta_0\, dr$. The singular factor $1/|r|$ is the geometric source of the routing degeneracy. Integrating the Gaussian prior over δ_0 (a Gaussian convolution) and using the symmetry in r gives the fiber-integral representation

$$\rho_G(p_0, p_1, p_2) \propto \int_0^\infty \frac{1}{r} \exp\left(-\frac{r^2}{2\tau_r^2} - \frac{p_2^2}{8\tau_{\delta_1}^2 r^2} - \frac{p_1^2}{2(\tau_{s_1}^2 + 4\tau_{\delta_0}^2 r^2)} \right) \frac{dr}{\sqrt{\tau_{s_1}^2 + 4\tau_{\delta_0}^2 r^2}}. \tag{21}$$

For $p_2 \neq 0$, the factor $\exp(-p_2^2/(8\tau_{\delta_1}^2 r^2))$ damps the $1/r$ pole and yields a smooth positive density, hence RLCT $(3/2, 1)$. For $p_2 \to 0$ (linear truth), the substitution $u = |p_2|/(\sqrt{8}\tau_{\delta_1} r)$ extracts the leading singularity $\rho_G = A_G(p_0, p_1) \log(1/|p_2|) + B_G$, so the routed model is exactly ordinary quadratic regression with a logarithmically singular prior on p_2. Integrating this density against the quadratic loss raises the order of the leading pole to 2, which gives the global RLCT $(3/2, 2)$.

E Global Gaussian Prior and Fiber Integration

Proposition E.1 (Pushforward density of the Gaussian prior)**.** *Under the Gaussian prior* (4)*, the pushforward measure of $\varphi_G(\theta)\, d\theta$ under Φ is absolutely continuous with respect to Lebesgue measure. Its density ρ_G is smooth on $\{p_2 \neq 0\}$, is given there by* (9)*, and* (9) *holds with*

$$C_G = \frac{1}{(2\pi)^2 \tau_{s_0} \tau_{\delta_1} \tau_r}.$$

Moreover, if

$$A_G(p_0, p_1) = \frac{C_G}{\tau_{s_1}} \exp\left(-\frac{p_0^2}{2\tau_{s_0}^2} - \frac{p_1^2}{2\tau_{s_1}^2} \right),$$

then for $p_2 \neq 0$ in a neighborhood of $p_2 = 0$,

$$\rho_G(p_0, p_1, p_2) = A_G(p_0, p_1) \log \frac{1}{|p_2|} + B_G(p_0, p_1, p_2),$$

where B_G is locally bounded. In particular, after redefining ρ_G on the measure-zero plane $\{p_2 = 0\}$, one has $\rho_G \in L^1_{\mathrm{loc}}(\mathbb{R}^3)$.

Proof. The proof has four steps: (1) a polynomial change of variables that exposes the singular Jacobian factor $1/|r|$; (2) a closed-form Gaussian convolution in δ_0; (3) a small-r analysis that isolates the logarithmic singularity at $p_2 = 0$; and (4) evaluation of the leading integral.

Step 1: Singular change of variables. Let F be bounded and measurable. On the sets $\{r > 0\}$ and $\{r < 0\}$, use the change of variables

$$(s_0, \delta_0, s_1, \delta_1, r) \longmapsto (p_0, p_1, p_2, \delta_0, r) = (s_0,\ s_1 + 2r\delta_0,\ 2r\delta_1,\ \delta_0,\ r).$$

Its Jacobian determinant is $2r$, so $d\theta = \frac{1}{2|r|}\, dp_0\, dp_1\, dp_2\, d\delta_0\, dr$. Therefore

$$\int_{\mathbb{R}^5} F(\Phi(\theta)) \varphi_G(\theta)\, d\theta = \int_{\mathbb{R}^3} F(p) \rho_G(p)\, dp,$$

where

$$\rho_G(p) = \int_{\mathbb{R}} \int_{\mathbb{R}} \frac{\varphi_G\big(p_0, \delta_0, p_1 - 2r\delta_0, \frac{p_2}{2r}, r\big)}{2|r|}\, d\delta_0\, dr.$$

The factor $1/|r|$ is the geometric source of the singularity: it is integrable in r only because the Gaussian factor $\exp(-p_2^2/(8\tau_{\delta_1}^2 r^2))$ damps the small-r region whenever $p_2 \neq 0$.

Step 2: Closed-form Gaussian convolution in δ_0. Integrating out δ_0 is a one-dimensional Gaussian convolution:

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\left(-\frac{\delta_0^2}{2\tau_{\delta_0}^2} - \frac{(p_1 - 2r\delta_0)^2}{2\tau_{s_1}^2} \right) d\delta_0 \\
&= \sqrt{\frac{2\pi\tau_{\delta_0}^2\tau_{s_1}^2}{\tau_{s_1}^2 + 4\tau_{\delta_0}^2 r^2}} \exp\left(-\frac{p_1^2}{2(\tau_{s_1}^2 + 4\tau_{\delta_0}^2 r^2)} \right).
\end{aligned}$$

Substituting this identity and using the symmetry in r yields (9).

Step 3: Small-r analysis isolating the logarithmic singularity. To identify the singularity at $p_2 = 0$, write

$$\rho_G(p_0, p_1, p_2) = \int_0^\infty e^{-a^2/r^2}\, \frac{h(p_0, p_1, r)}{r}\, dr, \qquad a = \frac{|p_2|}{2\sqrt{2}\tau_{\delta_1}},$$

where

$$h(p_0, p_1, r) = C_G e^{-p_0^2/(2\tau_{s_0}^2)} \frac{\exp\left(-\frac{r^2}{2\tau_r^2} - \frac{p_1^2}{2(\tau_{s_1}^2 + 4\tau_{\delta_0}^2 r^2)} \right)}{\sqrt{\tau_{s_1}^2 + 4\tau_{\delta_0}^2 r^2}}.$$

The function h is smooth and even in r, so $h(p_0, p_1, r) = A_G(p_0, p_1) + O(r^2)$ as $r \to 0$. Hence

$$\rho_G - A_G \int_0^1 e^{-a^2/r^2} \frac{dr}{r} = \int_0^1 e^{-a^2/r^2} \frac{h - A_G}{r}\, dr + \int_1^\infty e^{-a^2/r^2} \frac{h}{r}\, dr.$$

Both terms on the right are locally bounded: $(h - A_G)/r = O(r)$ near 0, and $h(r)$ has Gaussian decay as $r \to \infty$. The implied bounds are uniform on compact subsets of (p_0, p_1) because h is smooth in (p_0, p_1, r).

Step 4: Asymptotic evaluation of the leading integral. It remains to analyze $J(a) = \int_0^1 e^{-a^2/r^2} \frac{dr}{r}$. With the substitution $u = a/r$,

$$J(a) = \int_a^\infty e^{-u^2} \frac{du}{u} = \log \frac{1}{a} + O(1) = \log \frac{1}{|p_2|} + O(1)$$

as $a \to 0^+$. The same splitting gives $B_G(p) = B_0(p_0, p_1) + O(p_2^2[1 + |\log |p_2||])$ with B_0 smooth: replacing the damping factor by 1 in the $O(r)$ remainder costs $O(a^2 \log(1/a))$, while the $r \geq 1$ tail costs $O(a^2)$. This proves the stated decomposition for $p_2 \neq 0$. Because $A_G(p_0, p_1) \log(1/|p_2|) + B_G(p_0, p_1, p_2)$ is locally integrable near $p_2 = 0$, redefining ρ_G on $\{p_2 = 0\}$ gives a locally integrable density on all of $\mathbb{R}^3$. □

Lemma E.2 (Logarithmic coefficient density implies multiplicity two)**.** *Let Q be a positive-definite quadratic form on $\mathbb{R}^3$, let ψ be smooth with $\psi(0) > 0$, and suppose $\chi(\Delta) = \chi_0(\Delta) + O(\|\Delta\|^2[1 + |\log |\Delta_3||])$ with χ_0 smooth. Then*

$$\int_{|\Delta| < \varepsilon} Q(\Delta)^s \left(\psi(\Delta) \log \frac{1}{|\Delta_3|} + \chi(\Delta) \right) d\Delta$$

has a leading pole at $s = -3/2$ of order 2.

Proof. Write $\Delta = r\omega$ with $r > 0$ and $\omega \in S^2$. Then

$$Q(r\omega)^s = r^{2s} Q(\omega)^s, \qquad d\Delta = r^2\, dr\, d\omega,$$

and

$$\log \frac{1}{|\Delta_3|} = \log \frac{1}{r} + \log \frac{1}{|\omega_3|}.$$

Since $Q(\omega)$ is bounded above and below on the sphere and

$$\int_{S^2} \left| \log \frac{1}{|\omega_3|} \right| d\omega < \infty,$$

the only source of a double pole is the radial integral

$$\int_0^\varepsilon r^{2s+2} \log \frac{1}{r}\, dr,$$

whose meromorphic continuation has a double pole at $s = -3/2$. Replacing ψ and χ_0 by their values at zero leaves an integrand of higher radial order, holomorphic near $s = -3/2$; the stated logarithmic remainder is also integrable there. The smooth constant term contributes at most a simple pole. Because $\psi(0) > 0$, the coefficient of the double pole is nonzero. □

Proof of Theorem 5.3. The fiber-integral formula and the logarithmic decomposition come from Proposition E.1. Write

$$Q(\Delta) = \frac{1}{2\sigma^2} \Delta^\top G \Delta,$$

so that $K(\theta) = Q(\Phi(\theta) - q)$ with $q = (q_0, q_1, q_2)$ and $G \succ 0$. Hence

$$\zeta_G(s) = \int_{\mathbb{R}^3} Q(p-q)^s \rho_G(p)\, dp.$$

Fix $\varepsilon > 0$ and split $\zeta_G(s) = I_\varepsilon(s) + R_\varepsilon(s)$ into the regions $|p - q| < \varepsilon$ and $|p - q| \geq \varepsilon$. Because $p_0 = s_0$, $p_1 = s_1 + 2r\delta_0$, and $p_2 = 2r\delta_1$ are polynomial functions of Gaussian coordinates, the pushforward measure has finite moments of every order. On $|p - q| \geq \varepsilon$, the positive-definite quadratic form Q satisfies

$$0 < c_\varepsilon \leq Q(p-q) \leq C_\varepsilon(1 + |p|^2).$$

Therefore, for every compact set $S \subset \mathbb{C}$,

$$\sup_{s \in S} \left| Q(p-q)^s \mathbf{1}_{|p-q| \geq \varepsilon} \right| \leq C_S(1 + |p|^{2\sigma_+}), \qquad \sigma_+ = \max\{0, \sup_{s \in S} \Re(s)\},$$

and the right-hand side is integrable against $\rho_G(p)\, dp$. Hence $R_\varepsilon(s)$ is entire by dominated convergence. Only the local term $I_\varepsilon(s)$ can contribute to the leading pole.

If $q_2 \neq 0$, then in a small neighborhood of (q_0, q_1, q_2) the coordinate p_2 stays bounded away from 0. In (9), every derivative with respect to (p_0, p_1, p_2) is dominated by an integrable function because the factor

$$\exp\left(-\frac{p_2^2}{8\tau_{\delta_1}^2 r^2} \right)$$

suppresses the $r \to 0$ region super-polynomially, while $\exp(-r^2/(2\tau_r^2))$ suppresses $r \to \infty$. Hence ρ_G is smooth near (q_0, q_1, q_2) and strictly positive because the integrand is everywhere positive. The zeta function is therefore the zeta function of a positive-definite quadratic form against a smooth positive density in three variables, so the RLCT is $(3/2, 1)$.

If $q_2 = 0$, then Proposition E.1 gives

$$\rho_G(q + \Delta) = A_G(q_0 + \Delta_0, q_1 + \Delta_1) \log \frac{1}{|\Delta_2|} + B_G(q + \Delta).$$

Since $A_G(q_0, q_1) > 0$, Lemma E.2 applies to the coefficient-space zeta integral

$$I_\varepsilon(s) = \int_{|\Delta| < \varepsilon} Q(\Delta)^s \rho_G(q + \Delta)\, d\Delta.$$

Its leading pole is therefore at $s = -3/2$ with multiplicity 2, proving RLCT $(3/2, 2)$. □

F Sigmoid Robustness at the Collapsed Stratum

This appendix proves Proposition 5.4. Throughout, $g_\sigma(x;r) = \sigma(rx) = (1+e^{-rx})^{-1}$, so $g_\sigma - 1/2 = \frac{1}{2}\tanh(rx/2)$, and

$$g_\sigma(x;r) - \tfrac{1}{2} = \tfrac{rx}{4} - \tfrac{r^3x^3}{48} + \tfrac{r^5x^5}{480} - \cdots .$$

We compare K_σ with the affine-gate loss K_{aff} after rescaling the router by $r \mapsto r/4$, which makes the leading-order terms agree. For Gaussian covariates, we use this finite-order expansion only to identify the leading pole of the local zeta function. For a symmetric covariate law whose bounded support contains an interval, the loss is real-analytic and the same argument is an ordinary local normal-form computation.

Setup. With $q_2 = 0$, write the affine model output (using router $r/4$ in place of r in (2)) as $f_{\mathrm{aff}}(x;\theta) = s_0 + (s_1 + \frac{r\delta_0}{2})x + \frac{r\delta_1}{2}x^2$, and the sigmoid output as $f_\sigma(x;\theta) = s_0 + s_1x + 2(g_\sigma(x;r) - \frac{1}{2})(\delta_0 + \delta_1 x)$. Set $\Delta f := f_\sigma - f_{\mathrm{aff}}$.

Lemma F.1 (Higher-order sigmoid correction)**.** *At every $\theta = (s_0, \delta_0, s_1, \delta_1, r)$,*

$$\Delta f(x;\theta) = -\tfrac{r^3x^3}{24}\,(\delta_0 + \delta_1 x) + r^5x^5\,h(rx)\,(\delta_0 + \delta_1 x),$$

where h is real-analytic in a neighborhood of 0.

Proof. Direct computation: $2(g_\sigma - \frac{1}{2}) - \frac{rx}{2} = -\frac{r^3x^3}{24} + r^5x^5h(rx)$ for an analytic remainder h. Multiplying by $\delta_0 + \delta_1 x$ gives the stated identity. □

Lemma F.2 (Order-five vanishing of the loss difference)**.** *Write $\varepsilon = \theta - \theta^*$ with $\theta^* = (q_0, 0, q_1, 0, 0)$. Then*

$$K_\sigma(\theta) - K_{\mathrm{aff}}(\theta) = R(\varepsilon),$$

where R has no term of total degree below 5 in ε.

Proof. Set $\Delta_{\mathrm{aff}} := f_{\mathrm{aff}} - q$. Expanding the squares,

$$2\sigma^2\big(K_\sigma - K_{\mathrm{aff}}\big) = \mathbb{E}\big[\Delta f\,(2\Delta_{\mathrm{aff}} + \Delta f)\big] = 2\mathbb{E}[\Delta f \cdot \Delta_{\mathrm{aff}}] + \mathbb{E}[\Delta f^2].$$

By Lemma F.1, $\Delta f = O(r^3(\delta_0 + \delta_1 x))$, so $\Delta f^2 = O(r^6(\delta_0 + \delta_1 x)^2)$, and every x-moment of Δf^2 is a sum of monomials in ε of total degree at least 8, such as $r^6\delta_0^2$. For the cross term, expand

$$\Delta_{\mathrm{aff}} = \varepsilon_1 + \big(\varepsilon_3 + \tfrac{\varepsilon_2\varepsilon_5}{2}\big)x + \tfrac{\varepsilon_4\varepsilon_5}{2}\,x^2,$$

and the leading sigmoid correction $\Delta f = -\frac{\varepsilon_5^3}{24}(\varepsilon_2 + \varepsilon_4 x)x^3 + O(\varepsilon_5^5)$. Multiplying out and using the Gaussian moments $\mathbb{E}[x^k] = 0$ for odd k, $\mathbb{E}[x^4] = 3$, and $\mathbb{E}[x^6] = 15$ gives

$$\mathbb{E}[\Delta f \cdot \Delta_{\mathrm{aff}}] = -\tfrac{1}{8}\big(\varepsilon_5^3\varepsilon_2\varepsilon_3 + \varepsilon_5^3\varepsilon_4\varepsilon_1\big) - \tfrac{5}{16}\,\varepsilon_5^4\varepsilon_4^2 + (\text{terms of order} \geq 6).$$

The two leading monomials $\varepsilon_5^3\varepsilon_2\varepsilon_3$ and $\varepsilon_5^3\varepsilon_4\varepsilon_1$ have total degree 5, the $\varepsilon_5^4\varepsilon_4^2$ term has degree 6, and all higher-order Taylor remainders have degree ≥ 6. Combined with the order-8 contribution of $\mathbb{E}[\Delta f^2]$, every monomial of R has total degree at least 5. □

Proof of Proposition 5.4. The polynomial change of variables (7), applied to K_{aff} with router rescaled by $r \mapsto r/4$, has Jacobian determinant 1 and gives, by the same calculation as in the proof of Theorem 5.1(b),

$$2\sigma^2 K_{\mathrm{aff}}(\theta) = u^2 + v^2 + c\,w^2z^2, \qquad c = \tfrac{1}{2},$$

in coordinates (u, t, v, w, z) centered at θ^*. The diffeomorphism is polynomial with linear leading order

$$u = \varepsilon_1 + O(\varepsilon^2), \quad v = \varepsilon_3 + O(\varepsilon^2), \quad w = \varepsilon_4, \quad z = \varepsilon_5, \quad t = \varepsilon_2,$$

and the same is true of its inverse. Therefore the order-five remainder $R(\varepsilon)$ from Lemma F.2 pulls back to a remainder $\widetilde{R}(u, t, v, w, z)$ in normal-form coordinates with no term of total degree below 5, giving

$$2\sigma^2 K_\sigma(\theta) = u^2 + v^2 + c\,w^2z^2 + \widetilde{R}(u, t, v, w, z).$$

Total degree alone does not determine the leading pair, so we argue directly. Put $a(r) = \mathbb{E}[x\tanh(rx/2)]$, $b(r) = \mathrm{Var}(x\tanh(rx/2))$, and $\tilde{c}(r) = \mathbb{E}[(\tanh(rx/2) - a(r)x)^2]$. Orthogonal projection onto $1, x$ gives the exact identity

$$2\sigma^2 K_\sigma = U^2 + V^2 + \delta_1^2 b(r) + \delta_0^2 \tilde{c}(r), \tag{22}$$

where $U = s_0 - q_0 + \delta_1 a(r)$ and $V = s_1 - q_1 + \delta_0 a(r)$. Gaussian moments give $b(r) = r^2/2 + O(r^4)$ and $\tilde{c}(r) = r^6/96 + O(r^8)$. Since the quotients $b(r)/r^2$ and $\tilde{c}(r)/r^6$ are smooth and positive near $r = 0$, a local rescaling brings the loss to the form $U^2 + V^2 + w^2 z^2 + t^2 z^6$. Dropping the last term bounds the Laplace integral $\int e^{-\beta(U^2+V^2+w^2z^2+t^2z^6)}$ above by $C\beta^{-3/2}\log\beta$ as $\beta \to \infty$. Restricting $|U|, |V| \le C\beta^{-1/2}$, bounded t, $\beta^{-1/2} \le |z| \le C\beta^{-1/6}$ and $|w| \le C\beta^{-1/2}/|z|$ gives a matching lower bound. Resolution of this polynomial form gives $(3/2, 2)$.

For symmetric covariates with bounded support containing an interval, the orthogonal projection uses $\mathbb{E}[x^2]$ in the linear normalization. The analogous coefficients are positive because x^2 is nonconstant and x^3 is not a multiple of x in L^2. The loss is then analytic, giving the ordinary local RLCT. □

F.1 Higher-Dimensional Input

This subsection proves Proposition 5.5. With $x \sim \mathcal{N}(0, I_d)$, the model output is

$$f(x;\theta) - q(x) = \Delta_0 + \Delta_1^\top x + x^\top M x, \qquad M = 2r\delta_1^\top \in \mathbb{R}^{d\times d},$$

where $\Delta_0 = s_0 - q_0$ and $\Delta_1 = s_1 + 2\delta_0 r - q_1 \in \mathbb{R}^d$. The bilinear form $x^\top M x$ depends on M only through its symmetric part $M_s = r\delta_1^\top + \delta_1 r^\top$.

Lemma F.3 (Loss expansion in d dimensions)**.** *For $x \sim \mathcal{N}(0, I_d)$, the population KL loss satisfies*

$$2\sigma^2 K(\theta) = \Delta_0^2 + \|\Delta_1\|^2 + 8(r^\top\delta_1)^2 + 4\|r\|^2\|\delta_1\|^2 + 4\Delta_0\,(r^\top\delta_1).$$

Proof. For $x \sim \mathcal{N}(0, I_d)$ and any matrix A with symmetric part A_s, the standard Gaussian identities

$$\mathbb{E}[(x^\top A x)^2] = 2\,\mathrm{tr}(A_s^2) + (\mathrm{tr}\,A_s)^2, \qquad \mathbb{E}[\Delta_0(x^\top A x)] = \Delta_0\,\mathrm{tr}(A_s)$$

hold. We apply them with $A = M$, whose symmetric part is $M_s = r\delta_1^\top + \delta_1 r^\top$, so that $\mathrm{tr}(M_s) = 2(r^\top\delta_1)$ and $\mathrm{tr}(M_s^2) = 2(r^\top\delta_1)^2 + 2\|r\|^2\|\delta_1\|^2$. Substituting these and using that odd-order x-moments vanish yields the stated loss expansion. □

Proof of Proposition 5.5. Write perturbations $\varepsilon_1 = s_0 - q_0 \in \mathbb{R}$, $\varepsilon_2 = \delta_0 \in \mathbb{R}$, $\varepsilon_3 = s_1 - q_1 \in \mathbb{R}^d$, $\varepsilon_4 = \delta_1 \in \mathbb{R}^d$, $\varepsilon_5 = r \in \mathbb{R}^d$ centered at $\theta^* = (q_0, 0, q_1, 0, 0)$. Then $\Delta_0 = \varepsilon_1$, $\Delta_1 = \varepsilon_3 + 2\varepsilon_2\varepsilon_5$, $r = \varepsilon_5$, $\delta_1 = \varepsilon_4$, and Lemma F.3 gives

$$2\sigma^2 K(\theta) = \varepsilon_1^2 + \|\varepsilon_3 + 2\varepsilon_2\varepsilon_5\|^2 + 8(\varepsilon_4^\top\varepsilon_5)^2 + 4\|\varepsilon_5\|^2\|\varepsilon_4\|^2 + 4\varepsilon_1(\varepsilon_4^\top\varepsilon_5).$$

Define $u = \varepsilon_1 + 2(\varepsilon_4^\top\varepsilon_5)$, $v = \varepsilon_3 + 2\varepsilon_2\varepsilon_5 \in \mathbb{R}^d$, $w = \varepsilon_4 \in \mathbb{R}^d$, $z = \varepsilon_5 \in \mathbb{R}^d$, $t = \varepsilon_2$. The Jacobian is upper-triangular (in any block ordering placing u, v after their summands) with unit diagonal, so $\det = 1$. Substituting,

$$u^2 = \varepsilon_1^2 + 4\varepsilon_1(\varepsilon_4^\top\varepsilon_5) + 4(\varepsilon_4^\top\varepsilon_5)^2,$$

so $u^2 + \|v\|^2$ accounts for $\varepsilon_1^2 + 4\varepsilon_1(\varepsilon_4^\top\varepsilon_5) + 4(\varepsilon_4^\top\varepsilon_5)^2 + \|\varepsilon_3 + 2\varepsilon_2\varepsilon_5\|^2$. The residual is

$$2\sigma^2 K - u^2 - \|v\|^2 = 8(\varepsilon_4^\top\varepsilon_5)^2 + 4\|\varepsilon_5\|^2\|\varepsilon_4\|^2 - 4(\varepsilon_4^\top\varepsilon_5)^2 = 4\big(\|w\|^2\|z\|^2 + (w^\top z)^2\big),$$

which proves (11).

Local RLCT. The principal part decomposes as $h_1(u, v) + h_2(w, z)$ with $h_1 = u^2 + \|v\|^2$ and $h_2 = 4(\|w\|^2\|z\|^2 + (w^\top z)^2)$ on disjoint variable sets. The smooth Morse part h_1 has $1 + d$ quadratic directions and local RLCT $((1 + d)/2, 1)$. For h_2, polar coordinates $w = R_w\omega_w$, $z = R_z\omega_z$ with $\omega_w, \omega_z \in S^{d-1}$ give $h_2(w, z) = 4R_w^2 R_z^2(1 + (\omega_w^\top\omega_z)^2)$, and $1 \le 1 + (\omega_w^\top\omega_z)^2 \le 2$ is bounded and analytic on $S^{d-1} \times S^{d-1}$. The local zeta function factors as

$$\zeta_{h_2}(s) = 4^s \int_{S^{d-1}\times S^{d-1}} (1 + (\omega_w^\top\omega_z)^2)^s\, d\omega_w\, d\omega_z \;\cdot\; \Big(\int_0^\varepsilon R^{2s+d-1}\, dR\Big)^2,$$

the radial double integral equals $\varepsilon^{2(2s+d)}/(2s + d)^2$ and produces a double pole at $s = -d/2$, while the angular factor is entire. Hence h_2 has local RLCT $(d/2, 2)$. The local RLCT of a sum $h_1 + h_2$ of non-negative analytic functions on disjoint variable sets adds in the standard way (Watanabe, 2009, 2018): $\lambda_{h_1+h_2} = \lambda_{h_1} + \lambda_{h_2} = (2d + 1)/2$ with multiplicity $m_{h_1+h_2} = m_{h_1} + m_{h_2} - 1 = 2$. □

F.2 Full Zero Set of the Sigmoid Model on Linear Truth

This subsection proves Proposition F.4. Throughout, K_σ is the sigmoid loss of Proposition 5.4 and the truth is linear ($q_2 = 0$).

Proposition F.4 (Zero set and local pole pairs of the sigmoid model)**.** *For the sigmoid model of Proposition 5.4 on linear truth ($q_2 = 0$), the zero set is*

$$W_0^\sigma = S_A^\sigma \cup S_B^\sigma,$$

where

$$S_A^\sigma = \{r = 0,\, s_0 = q_0,\, s_1 = q_1,\, \delta_0, \delta_1 \in \mathbb{R}\},$$
$$S_B^\sigma = \{\delta_0 = \delta_1 = 0,\, s_0 = q_0,\, s_1 = q_1,\, r \in \mathbb{R}\},$$

and the Hessian degenerates along

$$S_C^\sigma = \{r = 0,\, s_0 = q_0,\, s_1 = q_1,\, \delta_1 = 0,\, \delta_0 \in \mathbb{R}\} \subset S_A^\sigma.$$

The local leading pole pairs are $(3/2, 1)$ on $S_A^\sigma \setminus S_C^\sigma$, $(3/2, 2)$ on S_C^σ, and $(2, 1)$ on $S_B^\sigma \setminus \{(q_0, 0, q_1, 0, 0)\}$. Consequently, for any smooth compactly supported prior whose support reaches S_C^σ, the leading pole pair relative to that prior is $(3/2, 2)$. For a symmetric covariate law with bounded support containing an interval, the same statement may be read as an ordinary local RLCT statement because the corresponding population loss is real-analytic in a neighborhood of the collapsed stratum.

Lemma F.5 (Geometry of W_0^σ)**.** *The zero set $W_0^\sigma = \{\theta : K_\sigma(\theta) = 0\}$ on linear truth equals*

$$W_0^\sigma = S_A^\sigma \cup S_B^\sigma,$$

where $S_A^\sigma = \{r = 0,\, s_0 = q_0,\, s_1 = q_1,\, \delta_0, \delta_1 \in \mathbb{R}\}$ is a 2-dimensional smooth component and $S_B^\sigma = \{\delta_0 = \delta_1 = 0,\, s_0 = q_0,\, s_1 = q_1,\, r \in \mathbb{R}\}$ is a 1-dimensional smooth component, intersecting at the single point $\theta^* = (q_0, 0, q_1, 0, 0)$. The Hessian of K_σ in normal directions degenerates exactly along the 1-dimensional sublocus $S_C^\sigma = \{r = 0,\, s_0 = q_0,\, s_1 = q_1,\, \delta_1 = 0,\, \delta_0 \in \mathbb{R}\} \subset S_A^\sigma$.

Proof. $K_\sigma(\theta) = 0$ if and only if $f_\sigma(\,\cdot\,;\theta) \equiv q$ on all of $\mathbb{R}$. Match Taylor coefficients at $x = 0$: with $\sigma(rx) - \frac{1}{2} = \frac{1}{2}\tanh(rx/2) = \sum_{k\geq 0} a_k r^{2k+1} x^{2k+1}$ (odd powers only, $a_0 = 1/4$), the constraint becomes $s_0 = q_0$ and $r^{2k+1}\delta_j = 0$ for $j \in \{0, 1\}$ and $k \geq 0$ together with $s_1 = q_1 - r\delta_0/2$. Either $r = 0$ (giving S_A^σ) or $r \neq 0$ (forcing $\delta_0 = \delta_1 = 0$, giving S_B^σ). The two sets meet at θ^*. The normal Hessian of K_σ at a point of S_A^σ in the directions (s_0, s_1, r) is, by direct computation with $x \sim \mathcal{N}(0, 1)$, $1/\sigma^2$ times the symmetric Gram matrix with diagonal $(1, 1, (\delta_0^*)^2/4 + 3(\delta_1^*)^2/4)$ and off-diagonals $\delta_1^*/2$ (entry (s_0, r)) and $\delta_0^*/2$ (entry (s_1, r)); the Gram matrix determinant is $(\delta_1^*)^2/2$, which is nonzero if and only if $\delta_1^* \neq 0$. □

Lemma F.6 (Local leading pairs at the strata of W_0^σ)**.** (a) *At every point of $S_A^\sigma \setminus S_C^\sigma$ (where $\delta_1 \neq 0$), the local leading pair is $(3/2, 1)$ by the Morse–Bott form in the 3 normal directions.*

(b) *At every point of $S_B^\sigma \setminus \{\theta^*\}$ (where $r \neq 0$), the local leading pair is $(2, 1)$: there are 4 normal directions $(s_0, s_1, \delta_0, \delta_1)$. The four corresponding features $1, x, 2(\sigma(rx) - \frac{1}{2}), 2x(\sigma(rx) - \frac{1}{2})$ have a positive-definite $L^2(\mathcal{N}(0, 1))$ Gram matrix for every $r \neq 0$.*

(c) *At every point of the singular sublocus S_C^σ (parameterized by $\delta_0^* \in \mathbb{R}$), the local leading pair is $(3/2, 2)$.*

Proof of (c). The decomposition in (22) holds near every $\theta_{\delta_0}^* = (q_0, \delta_0^*, q_1, 0, 0)$, not only at $\delta_0^* = 0$. After smooth positive rescaling, the local loss is $u^2 + v^2 + w^2 z^2 + (\widetilde{\delta}_0^* + t)^2 z^6$. The upper and lower Laplace bounds in the proof of Proposition 5.4 apply with t in any fixed bounded interval. Thus the leading pair remains $(3/2, 2)$ along S_C^σ. □

Proof of Proposition F.4. Lemma F.5 identifies the zero set as $W_0^\sigma = S_A^\sigma \cup S_B^\sigma$ and the degenerate sublocus inside S_A^σ as S_C^σ. The local pole pairs listed in the proposition are exactly those from Lemma F.6: $(3/2, 1)$ on $S_A^\sigma \setminus S_C^\sigma$, $(3/2, 2)$ on S_C^σ, and $(2, 1)$ on $S_B^\sigma \setminus \{\theta^*\}$.

For any smooth compactly supported prior whose support reaches S_C^σ, the standard local-to-global rule for compactly supported zeta functions takes the smallest local exponent and, among the points attaining it, the largest multiplicity. Hence the leading pole pair relative to such a prior is $(3/2, 2)$. In this sense the product singularity of the affine model persists locally for sigmoid gates on linear truth. □

G Evidence and Bayes-Factor Asymptotics

Proof of Theorem 6.3. **Part (a): compactly supported priors.** The log-likelihood of the Gaussian model is

$$\log p(y \mid x, \theta) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y - f(x;\theta))^2}{2\sigma^2},$$

which is a polynomial, hence real-analytic, in θ. The prior φ is smooth and compactly supported by assumption, it satisfies the support assumptions of Theorem 5.2, and the model is realizable because Proposition 3.2 shows that every quadratic target belongs to the model class.

Therefore Watanabe's singular marginal-likelihood theorem applies (Watanabe, 2009, Theorem 6.7); see also Watanabe (2018). It gives

$$\log Z_n = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - \lambda \log n + (m-1)\log\log n + O_p(1),$$

where (λ, m) is the RLCT pair relative to φ. Substituting the cases from Theorem 5.2 proves part (a).

Part (b): Gaussian prior. Under φ_G, the routed model pushes forward exactly to quadratic regression in coefficient space:

$$Z_n = \int_{\mathbb{R}^3} \prod_{i=1}^{n} p(y_i \mid x_i, p)\, \rho_G(p)\, dp.$$

Write $\phi(x) = (1, x, x^2)^\top$, let

$$\hat{G}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\phi(x_i)^\top,$$

let $q = (q_0, q_1, q_2)$, and let $\hat{p}_n$ be the least-squares estimator in the quadratic model. Because the squared-error loss is quadratic in p,

$$\sum_{i=1}^{n}(y_i - p^\top\phi(x_i))^2 = \sum_{i=1}^{n}(y_i - \hat{p}_n^\top\phi(x_i))^2 + n(p-\hat{p}_n)^\top\hat{G}_n(p-\hat{p}_n).$$

Hence

$$Z_n = \exp\left(\sum_{i=1}^{n}\log p(y_i \mid x_i, \hat{p}_n)\right)\int_{\mathbb{R}^3}\exp\Big(-\frac{n}{2\sigma^2}(p-\hat{p}_n)^\top\hat{G}_n(p-\hat{p}_n)\Big)\,\rho_G(p)\,dp.$$

Because the model is well specified, $\hat{p}_n = q + O_p(n^{-1/2})$ and $\hat{G}_n \to G \succ 0$ almost surely. Fix $\varepsilon > 0$ small enough that the ball $U = B_\varepsilon(q)$ lies inside the neighborhood where the relevant local description of ρ_G is valid, and decompose the coefficient integral into U and U^c.

If $q_2 \neq 0$, then ρ_G is smooth and positive near q by Theorem 5.3. A standard Laplace argument gives

$$\int_U \exp\Big(-\frac{n}{2\sigma^2}(p-\hat{p}_n)^\top\hat{G}_n(p-\hat{p}_n)\Big)\,\rho_G(p)\,dp = n^{-3/2}(C + o_p(1)),$$

where $C = (2\pi\sigma^2)^{3/2}\rho_G(q)/\sqrt{\det G} > 0$, and the contribution from U^c is exponentially small. Indeed, on an event of probability tending to one,

$$\hat{p}_n \in B_{\varepsilon/2}(q), \qquad \lambda_{\min}(\hat{G}_n) \geq g_0 > 0,$$

so for $p \in U^c$ one has $|p - \hat{p}_n| \geq \varepsilon/2$ and therefore

$$(p-\hat{p}_n)^\top\hat{G}_n(p-\hat{p}_n) \geq g_0\varepsilon^2/4.$$

Since $\rho_G(p)\,dp$ is a probability measure, the tail contribution is $O_p(e^{-cn})$ for some $c > 0$. Thus

$$\int_{\mathbb{R}^3} \exp\Big(-\frac{n}{2\sigma^2}(p-\hat{p}_n)^\top\hat{G}_n(p-\hat{p}_n)\Big)\,\rho_G(p)\,dp = n^{-3/2}(C + o_p(1)),$$

while

$$\sum_{i=1}^{n}\log p(y_i \mid x_i, \hat{p}_n) = \sum_{i=1}^{n}\log p^\star(y_i \mid x_i) + O_p(1).$$

Therefore

$$\log Z_n = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - \frac{3}{2}\log n + O_p(1).$$

If $q_2 = 0$, then Proposition E.1 gives

$$\rho_G(p) = A_G(p_0, p_1)\log\frac{1}{|p_2|} + B_G(p),$$

in U, with $A_G(q_0, q_1) > 0$ and B_G locally bounded. The contribution from U^c is again exponentially small by the same tail bound, so it suffices to analyze the local integral over U. Substitute $p = \hat{p}_n + n^{-1/2}u$ and write $D_n = \sqrt{n}\,(U - \hat{p}_n)$:

$$\begin{aligned}\int_U \exp\Big(-\frac{n}{2\sigma^2}(p-\hat{p}_n)^\top \hat{G}_n (p-\hat{p}_n)\Big)\,\rho_G(p)\,dp \\ = n^{-3/2}\int_{D_n} \exp\left(-\frac{1}{2\sigma^2}u^\top \hat{G}_n u\right)\rho_G(\hat{p}_n + n^{-1/2}u)\,du.\end{aligned}$$

Set $c_n = \sqrt{n}\,\hat{p}_{n,2} = O_p(1)$. Then

$$\log\frac{1}{|\hat{p}_{n,2} + n^{-1/2}u_2|} = \frac{1}{2}\log n + \log\frac{1}{|u_2 + c_n|}.$$

Write

$$A_n(u) = A_G(\hat{p}_{n,0} + n^{-1/2}u_0, \hat{p}_{n,1} + n^{-1/2}u_1), \qquad B_n(u) = B_G(\hat{p}_n + n^{-1/2}u).$$

Therefore the integral equals

$$n^{-3/2}\int_{D_n} \exp\left(-\frac{1}{2\sigma^2}u^\top \hat{G}_n u\right)\left[A_n(u)\left(\frac{1}{2}\log n + \log\frac{1}{|u_2 + c_n|}\right) + B_n(u)\right]du.$$

Because $\hat{G}_n \to G \succ 0$, there is $a_0 > 0$ such that, with probability tending to one,

$$\exp\left(-\frac{1}{2\sigma^2}u^\top \hat{G}_n u\right) \le e^{-a_0|u|^2}.$$

Also, for every fixed $M < \infty$,

$$\sup_{|c|\le M}\int_{\mathbb{R}^3} e^{-a_0|u|^2}\left|\log\frac{1}{|u_2 + c|}\right|du < \infty,$$

because the logarithmic singularity is integrable in one dimension and the Gaussian factor controls the tails. Since $c_n = O_p(1)$, A_G is continuous with $A_G(q_0, q_1) > 0$, and B_G is locally bounded, the term multiplying $\frac{1}{2}\log n$ converges in probability to a positive constant, while the remaining terms are $O_p(1)$. Hence the integral is

$$n^{-3/2}\big(C_n \log n + O_p(1)\big), \qquad C_n \to C > 0 \text{ in probability},$$

which yields

$$\log Z_n = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - \frac{3}{2}\log n + \log\log n + O_p(1).$$

This proves part (b). □

Proof of Proposition 6.2. By Proposition 6.1, the routed posterior projected to coefficient space is exactly the quadratic-regression posterior with prior density ρ_G:

$$\Pi_n^\rho(dp) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - p^\top\phi(x_i))^2\right\}\rho_G(p)\,dp, \qquad \phi(x) = (1, x, x^2)^\top.$$

On linear truth $q = (q_0, q_1, 0)$, Proposition E.1 gives

$$\rho_G(p) = A_G(p_0, p_1)\log\frac{1}{|p_2|} + B_G(p)$$

in a neighborhood of q, with $A_G(q_0, q_1) > 0$ and B_G locally bounded. Shrinking the neighborhood if needed, there are constants $0 < c_\rho < C_\rho < \infty$ and a ball $U = B_\varepsilon(q)$ such that

$$c_\rho \log \frac{e}{|p_2|} \le \rho_G(p) \le C_\rho \log \frac{e}{|p_2|}, \qquad p \in U,\ p_2 \neq 0.$$

Let

$$\hat{G}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)\phi(x_i)^\top,$$

and let $\hat{p}_n$ be the least-squares estimator in quadratic regression. Since $x_i \sim \mathcal{N}(0,1)$, $\hat{G}_n \to G := \mathbb{E}[\phi(x)\phi(x)^\top] \succ 0$ almost surely, and standard regular regression theory gives $\hat{p}_n - q = O_{P_q}(n^{-1/2})$. Conditional on the design,

$$\sum_{i=1}^{n} (y_i - p^\top \phi(x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{p}_n^\top \phi(x_i))^2 + n(p - \hat{p}_n)^\top \hat{G}_n (p - \hat{p}_n).$$

Cancelling the data-only likelihood at $\hat{p}_n$ from posterior ratios, define the centered normalizer

$$Z_n^\rho = \int_{\mathbb{R}^3} \exp\Big\{ -\frac{n}{2\sigma^2} (p - \hat{p}_n)^\top \hat{G}_n (p - \hat{p}_n) \Big\} \rho_G(p)\, dp.$$

Choose a deterministic sequence $L_n \to \infty$ such that $L_n = o(M_n)$ and $L_n/\sqrt{n} \to 0$, for instance $L_n = \min\{\sqrt{M_n}, n^{1/4}\}$. Let E_n be the event that $\|\hat{p}_n - q\| \le L_n/\sqrt{n}$ and the eigenvalues of $\hat{G}_n$ lie in $[g_-, g_+]$ for fixed constants $0 < g_- < g_+ < \infty$. Then $P_q(E_n) \to 1$.

On E_n, this centered normalizer is bounded below by the contribution from

$$B_n = \left\{ |p_j - \hat{p}_{n,j}| \le \frac{1}{4\sqrt{n}},\ j = 0, 1, 2 \right\} \subset U$$

for all sufficiently large n, because $L_n/\sqrt{n} \to 0$. For $p \in B_n$, the quadratic form $(p - \hat{p}_n)^\top \hat{G}_n (p - \hat{p}_n)$ is $O(n^{-1})$, so the likelihood factor is bounded below by a positive constant independent of n. Also $|p_2| \le (L_n + 1/4)/\sqrt{n}$, and $L_n \le n^{1/4}$, so $\rho_G(p) \ge c'_\rho \log n$ on B_n for all sufficiently large n. Since $\mathrm{Vol}(B_n) \asymp n^{-3/2}$,

$$Z_n^\rho \ge c_1 n^{-3/2} \log n$$

in P_q-probability for some $c_1 > 0$.

Now consider the posterior mass outside $\|p - q\| > M_n n^{-1/2}$. On E_n, the contribution from U is bounded above by

$$c_2 n^{-3/2} \int_{\|u\| > c_3 M_n} \exp(-c_4 \|u\|^2) \Big(\log n + \log(e + |u_2 + c_n|^{-1}) \Big)\, du,$$

after the change of variables $u = \sqrt{n}\,(p - \hat{p}_n)$, where $c_n = \sqrt{n}\,\hat{p}_{n,2}$ and $|c_n| \le L_n$ on E_n. Here $\|p - q\| > M_n/\sqrt{n}$ implies $\|u\| \ge M_n - L_n \ge c_3 M_n$ for large n. The shifted logarithmic singularity is uniformly integrable: splitting at $|u_2 + c| = 1$ gives

$$\sup_{c \in \mathbb{R}} \int_{\mathbb{R}^3} e^{-c_4 \|u\|^2/2} \log(e + |u_2 + c|^{-1})\, du < \infty.$$

Factoring $e^{-c_4 c_3^2 M_n^2/2}$ out of the tail bounds the displayed integral by $C(1 + \log n) e^{-c_4 c_3^2 M_n^2/2} = o(\log n)$. Thus the local numerator is $o_{P_q}(n^{-3/2} \log n)$.

The contribution from U^c is exponentially small on E_n: since $q \in U$ and $\hat{p}_n \to q$, there exists $\delta > 0$ such that

$$(p - \hat{p}_n)^\top \hat{G}_n (p - \hat{p}_n) \ge \delta \qquad \text{for all } p \in U^c,$$

while $\rho_G(p)\, dp$ is a probability measure because it is the pushforward of the Gaussian prior. Hence the U^c contribution is $O(e^{-cn})$ for some $c > 0$. Dividing the numerator bound by the denominator lower bound yields

$$\Pi_n^\rho \Big(\|p - q\| > M_n n^{-1/2} \mid (x_i, y_i)_{i=1}^n \Big) \to 0$$

in P_q-probability.

Finally, let P_p denote the joint law of (x, y) under coefficient p. For Gaussian conditionals with the same known variance σ^2,

$$h^2(P_p, P_q) \le \frac{1}{8\sigma^2}\mathbb{E}_x[(\phi(x)^\top(p-q))^2] = \frac{1}{8\sigma^2}(p-q)^\top G(p-q).$$

Hence coefficient contraction at rate $n^{-1/2}$ implies predictive Hellinger contraction at the same rate. □

Proposition G.1 (WBIC on the linear submodel)**.** *Let*

$$L_n(\vartheta) := -\frac{1}{n}\sum_{i=1}^{n}\log p(y_i \mid x_i, \vartheta),$$

where ϑ denotes either routed parameters θ or quadratic coefficients p as appropriate. Let $\mathrm{WBIC}(n) := \mathbb{E}_{\beta=1/\log n}[nL_n(\theta)]$ *denote the WBIC of Watanabe (2013) for the routed model under φ_G, and let $\hat{p}_{\mathrm{LSE}}$ be the least-squares (equivalently, maximum likelihood) estimator in quadratic regression. Then on linear truth ($q_2 = 0$),*

$$\mathrm{WBIC}(n) = nL_n(\hat{p}_{\mathrm{LSE}}) + \tfrac{3}{2}\log n + O_p(\sqrt{\log n}).$$

Proof. Write

$$Z_{\beta,n} := \int \exp\bigl(-\beta n L_n(\theta)\bigr)\varphi_G(\theta)\,d\theta = \int \Bigl(\prod_{i=1}^{n} p(y_i \mid x_i, \theta)\Bigr)^{\beta}\varphi_G(\theta)\,d\theta.$$

Differentiating under the integral gives the thermodynamic identity

$$-\partial_\beta \log Z_{\beta,n} = \mathbb{E}_\beta[nL_n(\theta)],$$

so $\mathrm{WBIC}(n) = -\partial_\beta \log Z_{\beta,n}|_{\beta=1/\log n}$.

We use the coefficient-space likelihood directly. Set $a = n\beta$, $p = \hat{p}_n + a^{-1/2}u$, and $Q_n(u) = u^\top \hat{G}_n u/(2\sigma^2)$. Exact Gaussian regression gives

$$\mathbb{E}_\beta[nL_n] - nL_n(\hat{p}_n) = \beta^{-1}\frac{\int Q_n(u)e^{-Q_n(u)}\rho_G(\hat{p}_n + a^{-1/2}u)\,du}{\int e^{-Q_n(u)}\rho_G(\hat{p}_n + a^{-1/2}u)\,du}.$$

At $\beta = 1/\log n$, the logarithmic density expansion has leading weight $\frac{1}{2}A_G(q_0, q_1)\log a$ in both integrals. The remaining logarithmic terms have bounded Gaussian-weighted integrals, uniformly for $\sqrt{a}\,\hat{p}_{n,2} = O_p(1/\sqrt{\log n})$; contributions outside the local chart are exponentially small. The ratio is therefore $3/2 + O_p(1/\log a)$, yielding the stated remainder. This calculation does not produce a deterministic $\log\log n$ term of the marginal-likelihood form. □

Corollary G.2 (Robustness of the RLCT under smooth prior multiplication)**.** *For any smooth, strictly positive function $h : \mathbb{R}^5 \to \mathbb{R}_{>0}$ such that $h = 1$ outside a compact set, the zeta function $\zeta(s) = \int_{\mathbb{R}^5} K(\theta)^s\, h(\theta)\varphi_G(\theta)\,d\theta$ has the same leading pole pair (λ, m) as $\zeta_G(s) = \int_{\mathbb{R}^5} K(\theta)^s\varphi_G(\theta)\,d\theta$, namely $(3/2, 2)$ on linear truth and $(3/2, 1)$ otherwise.*

Proof. Watanabe's birational invariance theorem (Watanabe, 2009, Theorem 6.4) shows that, for non-negative analytic K, multiplying the integrand by a smooth, strictly positive density does not move the leading pole of the local zeta function or change its multiplicity, since the resolution of singularities used to extract that pole depends only on $K^{-1}(0)$ and on the order of vanishing of the prior on it. The hypotheses are satisfied here: K is real-analytic and $h\varphi_G$ is smooth and strictly positive on a neighborhood of W_0. Outside the compact perturbation the prior is unchanged. A partition of unity combines this Gaussian contribution with the compact local integrals, preserving the global pair. □

Corollary G.2 shows that the $\log\log n$ advantage in log Bayes factor in Theorem 6.4(a) survives multiplying the prior by any smooth positive factor that equals 1 outside a compact set. Arbitrary changes of the prior along the unbounded true set would require additional control of the tails. Like the theorem, the corollary concerns the leading pole of the zeta function and its multiplicity, that is, the marginal likelihood. It does not address posterior contraction; for φ_G itself, the coefficient posterior contracts at the regular $n^{-1/2}$ rate by Proposition 6.2.

Proof of Theorem 6.4. For the regular quadratic model M_Q, the coefficient vector is a regular 3-dimensional parameter with prior density smooth and positive near the true value. Standard Laplace asymptotics therefore give

$$\log Z_{Q,n} = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - \frac{3}{2}\log n + O_p(1)$$

in both regimes (van der Vaart, 1998, Chapter 10).

If $q_2 = 0$, the linear model M_L is well specified and regular of dimension 2, so

$$\log Z_{L,n} = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - \log n + O_p(1).$$

Combining these expansions with Theorem 6.3(b) gives

$$\log \mathrm{BF}_{R:Q,n} = \log Z_{R,n} - \log Z_{Q,n} = \log\log n + O_p(1),$$

and

$$\log \mathrm{BF}_{R:L,n} = \log Z_{R,n} - \log Z_{L,n} = -\frac{1}{2}\log n + \log\log n + O_p(1).$$

If $q_2 \neq 0$, then M_L is regular but misspecified. Let $\alpha^\star = (\alpha_0^\star, \alpha_1^\star)$ be the KL minimizer. Standard misspecified regular-model asymptotics give a deterministic gap $n\kappa_L$ plus lower-order random fluctuations:

$$\log Z_{L,n} = \sum_{i=1}^{n} \log p^\star(y_i \mid x_i) - n\kappa_L + O_p(\sqrt{n}),$$

where

$$\kappa_L = \frac{1}{2\sigma^2}\mathbb{E}\Big[\big(q_0 + q_1 x + q_2 x^2 - \alpha_0^\star - \alpha_1^\star x\big)^2\Big]$$

(Kleijn and van der Vaart, 2012; van der Vaart, 1998). Under $x \sim \mathcal{N}(0,1)$, the best linear projection of $q_0 + q_1 x + q_2 x^2$ onto $\mathrm{span}\{1, x\}$ is $(q_0 + q_2) + q_1 x$, so the residual is $q_2(x^2 - 1)$ and

$$\kappa_L = \frac{1}{2\sigma^2}\mathbb{E}[q_2^2 (x^2-1)^2] = \frac{1}{2\sigma^2}\cdot 2q_2^2 = \frac{q_2^2}{\sigma^2}.$$

Using Theorem 6.3(b) again gives

$$\log \mathrm{BF}_{R:Q,n} = O_p(1), \qquad \log \mathrm{BF}_{R:L,n} = n\kappa_L + O_p(\sqrt{n}).$$

□

G.1 Finite-Sample Posterior Shape

Conditionally on r, the curvature coefficient is $p_2 \mid r \sim \mathcal{N}(0, 4\tau_{\delta_1}^2 r^2)$. Put $\tilde{s} = 2\tau_r \tau_{\delta_1}$. Integrating the conditional Gaussian density over r gives the marginal density

$$\pi_2(c) = \frac{1}{\pi \tilde{s}} K_0(|c|/\tilde{s}),$$

using $K_0(u) = \int_0^\infty e^{-u\cosh t}\, dt$. Its variance is $\tilde{s}^2$.

To isolate this marginal, consider the Gaussian sequence model $\hat{c} \mid c \sim \mathcal{N}(c, v)$ with prior π_2. This calculation ignores the dependence between p_2 and p_1 that arises because both involve r. Let $\delta_v(\hat{c}) = \mathbb{E}[c \mid \hat{c}]$. Evenness and differentiation of the Gaussian likelihood give

$$\delta_v'(0) = \frac{\mathbb{E}[U^2 \pi_2(\sqrt{v}\,U)]}{\mathbb{E}[\pi_2(\sqrt{v}\,U)]} = 1 - \frac{2}{\log(1/v) + 2\log\tilde{s} + 3\log 2 - \gamma} + O(v\log(1/v)),$$

where $U \sim \mathcal{N}(0,1)$ and γ is Euler's constant. The last step uses $K_0(z) = \log(2/z) - \gamma + O(z^2|\log z|)$, $\mathbb{E}\log|U| = -(\gamma + \log 2)/2$, and $\mathbb{E}[U^2 \log|U|] - \mathbb{E}\log|U| = 1$. With unit prior scales and $v = 1/(2n)$, direct quadrature gives slopes about 0.809 and 0.885 at $n = 10^3$ and 10^6. The variance-matched $\mathcal{N}(0,4)$ prior gives $4/(4+v)$, or 0.99988 at $n = 10^3$.

For fixed data D, let L_D be the likelihood and $Z_R(D) = \int L_D(p)\rho_G(p)\,dp$. For a compact set $B \subset \mathbb{R}^2$ of positive area inside the neighborhood where (10) holds,

$$\Pi_R(B \times [-\varepsilon,\varepsilon] \mid D) = \frac{2\varepsilon\log(1/\varepsilon)}{Z_R(D)} \int_B L_D(p_0,p_1,0)A_G(p_0,p_1)\,dp_0\,dp_1 + O(\varepsilon).$$

This follows by integrating (10), since L_D is smooth positive and $\int_{-\varepsilon}^{\varepsilon} \log(1/|t|)\,dt = 2\varepsilon[\log(1/\varepsilon) + 1]$. Under a smooth positive coefficient prior, the same mass is only of order ε.

G.2 Unknown Noise Variance: Local Consequence

Let the true variance be $v_0 > 0$ and the fitted variance be $v_0 e^t$, with K evaluated at $\sigma^2 = v_0$. For a smooth positive local prior, the joint KL loss is

$$K_{\mathrm{unk}}(\theta,t) = \tfrac{1}{2}(t + e^{-t} - 1) + e^{-t}K(\theta).$$

Its variance minimizer is $t_\star = \log(1 + 2K(\theta))$, with minimum $\frac{1}{2}\log(1+2K(\theta)) = K(\theta)(1 + O(K(\theta)))$ and $\partial_t^2 K_{\mathrm{unk}}(\theta,t_\star) = 1/2$. The Morse lemma with parameters therefore splits off one regular quadratic coordinate, which gives the local pair $(\lambda + 1/2, m)$. The contrast in multiplicity thus persists locally when the noise variance is unknown; a global statement would also require conditions on the support and tails of the prior.

G.3 Low-Temperature Local Gibbs Mass

The low-temperature Gibbs posterior at inverse temperature $1/\eta$ is $\pi_\eta(d\theta) = Z_\eta^{-1} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta$ with normalizer $Z_\eta = \int_{\mathbb{R}^5} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta$. We record an auxiliary low-temperature consequence of the same normal forms used in the RLCT analysis.

Corollary G.3 (Collapsed-stratum local Gibbs mass asymptotics)**.** *Assume $q_2 = 0$ and let φ be a smooth compactly supported prior. Let U_A, U_B, U_C be sufficiently small product-chart neighborhoods of compact subsets of S_A, S_B, S_C, respectively, with the smooth subsets bounded away from S_C, chosen so that φ is positive on each neighborhood under consideration. Then, as $\eta \to 0^+$,*

$$\begin{aligned}
\int_{U_A} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta &= \Theta(\eta^{3/2}),\\
\int_{U_B} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta &= \Theta(\eta^{3/2}),\\
\int_{U_C} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta &= \Theta\big(\eta^{3/2}\log(1/\eta)\big).
\end{aligned}$$

Consequently,

$$\frac{\int_{U_C} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta}{\int_{U_A} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta + \int_{U_B} e^{-K(\theta)/\eta}\varphi(\theta)\,d\theta} = \Theta\big(\log(1/\eta)\big).$$

Remark G.4 (Scope). Corollary G.3 is a measure-theoretic Laplace statement about local Gibbs mass ratios under a compactly supported prior. An overdamped diffusion has $e^{-K/\eta}\varphi$ as its invariant density only after adding the prior drift $\eta\nabla\log\varphi$, so we do not interpret it as a finite-time SGD theorem.

Lemma G.5 (Sharp asymptotics of the product Laplace integral)**.** *For fixed $\varepsilon > 0$,*

$$I_\beta = \int_{[-\varepsilon,\varepsilon]^2} e^{-\beta w^2 z^2}\,dw\,dz = \sqrt{\pi}\,\beta^{-1/2}\log\beta\,(1 + o(1)) \qquad \text{as } \beta \to \infty.$$

Proof. By symmetry, $I_\beta = 4\int_0^\varepsilon \int_0^\varepsilon e^{-\beta w^2 z^2}\, dw\, dz$. For fixed $z > 0$, the substitution $\tau = wz\sqrt{\beta}$ gives

$$\int_0^\varepsilon e^{-\beta w^2 z^2}\, dw = \frac{1}{z\sqrt{\beta}} \int_0^{\varepsilon z\sqrt{\beta}} e^{-\tau^2}\, d\tau.$$

Split the z-integral at $z_0 = \beta^{-1/2}$. On $(0, z_0)$ the inner integral is at most $\varepsilon z\sqrt{\beta}$, so this region contributes $O(\beta^{-1/2})$. On (z_0, ε), replace the inner integral by $\sqrt{\pi}/2$. The integrated error is $O(\beta^{-1/2})$: after $t = z\sqrt{\beta}$, the Gaussian tail divided by t is integrable on $[1, \infty)$. Hence, up to this error,

$$4\int_{z_0}^\varepsilon \frac{1}{z\sqrt{\beta}} \left(\frac{\sqrt{\pi}}{2} + o(1) \right) dz = \frac{2\sqrt{\pi}}{\sqrt{\beta}}\, \log \frac{\varepsilon}{z_0}\, (1 + o(1)) = \sqrt{\pi}\, \beta^{-1/2} \log \beta\, (1 + o(1)),$$

since $\log(\varepsilon/z_0) = \frac{1}{2}\log\beta + \log\varepsilon$ and the boundary term is $O(\beta^{-1/2})$. □

Proof of Corollary G.3. We track the leading constants throughout, which strengthens each $\Theta(\cdot)$ to a sharp asymptotic equivalence; the scalings and the mass ratio stated in the corollary follow.

Smooth branches. On S_A and S_B, Theorem 5.1 provides local coordinates $(\eta_1, \eta_2, \eta_3, t_1, t_2)$ in which $K(\theta) = (\eta_1^2 + \eta_2^2 + \eta_3^2)/(2\sigma^2)$ and (t_1, t_2) parameterize the tangent stratum. The Gaussian integral in the normal directions is

$$\int_{\mathbb{R}^3} e^{-(\eta_1^2+\eta_2^2+\eta_3^2)/(2\sigma^2\eta)}\, d\eta_1\, d\eta_2\, d\eta_3 = (2\pi\sigma^2\eta)^{3/2},$$

and integration over the smooth tangent against the smooth positive density φ contributes a finite positive constant obtained by integrating the pulled-back density, including its Jacobian, along $S_A \cap U_A$ (resp. $S_B \cap U_B$). Hence

$$\int_{U_A} e^{-K(\theta)/\eta} \varphi(\theta)\, d\theta = (2\pi\sigma^2)^{3/2}\, M_A\, \eta^{3/2}\, (1 + o(1)), \tag{23}$$

where $M_A = \int \varphi(\Theta_A(0,t))|\det D\Theta_A(0,t)|\, dt$ in a product chart $\theta = \Theta_A(\eta, t)$; the analogous identity defines M_B. These are prior masses along the strata weighted by the chart Jacobian, not Euclidean areas.

Collapsed branch. On S_C, Theorem 5.1 gives the unit-Jacobian factorization $K(\theta) = (u^2 + v^2 + 8w^2z^2)/(2\sigma^2)$ with t tangent. The Gaussian integral in (u, v) on a small box about 0 converges to $2\pi\sigma^2\eta$ to leading order, and Lemma G.5 applied to $\beta = 4/(\sigma^2\eta)$ gives

$$\begin{aligned}\int_{[-\varepsilon,\varepsilon]^2} e^{-4w^2z^2/(\sigma^2\eta)}\, dw\, dz &= \sqrt{\pi}\, \big(4/(\sigma^2\eta)\big)^{-1/2} \log\big(4/(\sigma^2\eta)\big)\, (1 + o(1)) \\ &= \tfrac{\sqrt{\pi}\,\sigma}{2}\, \eta^{1/2} \log(1/\eta)\, (1 + o(1)).\end{aligned}$$

Multiplying and integrating over the tangent coordinate t against φ gives

$$\int_{U_C} e^{-K(\theta)/\eta} \varphi(\theta)\, d\theta = \pi^{3/2}\, \sigma^3\, M_C\, \eta^{3/2} \log(1/\eta)\, (1 + o(1)), \tag{24}$$

where $M_C = \int_{T_C} \varphi(q_0, \delta_0, q_1, 0, 0)\, d\delta_0$ over the canonical tangent interval T_C.

Mass-ratio constant. Dividing (24) by (23) (and its S_B analogue), and using $(2\pi\sigma^2)^{3/2} = 2\sqrt{2}\,\pi^{3/2}\sigma^3$,

$$\frac{\int_{U_C} e^{-K/\eta}\varphi\, d\theta}{\int_{U_A} e^{-K/\eta}\varphi\, d\theta + \int_{U_B} e^{-K/\eta}\varphi\, d\theta} = \frac{1}{2\sqrt{2}} \cdot \frac{M_C}{M_A + M_B} \cdot \log(1/\eta)\, (1 + o(1)).$$

This proves both the scalings and the $\Theta(\log(1/\eta))$ asymptotics of the mass ratio, with the explicit constant $1/(2\sqrt{2})$ times the ratio of the weighted masses. □

H Numerical Illustrations

Figure 3 evaluates the one-dimensional fiber integral (9) on a log-spaced grid of p_2 values at $p_0 = p_1 = 0$. The left panel plots $\rho_G(0, 0, p_2)$ against $\log(1/|p_2|)$ together with the fitted and theoretical slopes; the right panel subtracts the leading term $A_G(0, 0) \log(1/|p_2|)$ and shows that the remainder stays bounded. None of the numerical results in this appendix is used in a proof. This one simply visualizes the logarithmic singularity of Proposition E.1.

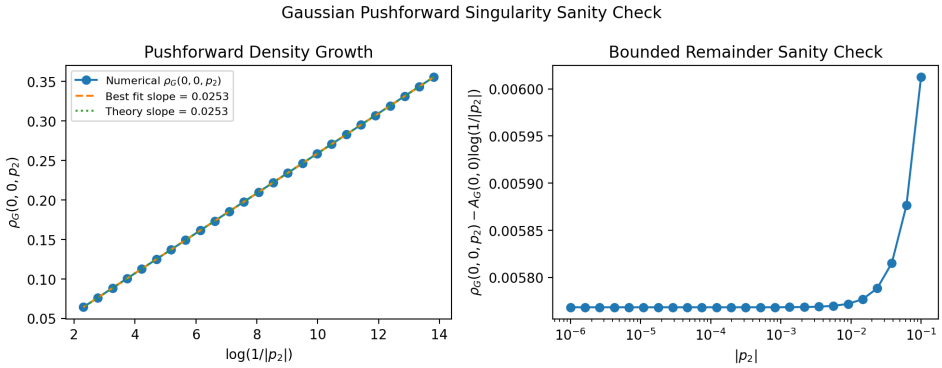

Figure 3: Numerical evaluation of $\rho_G(0, 0, p_2)$: raw growth (left) and centered remainder (right).

3D structure of the singular stratification. Figure 4 visualizes the singular stratification of W_0 in three of the routed coordinates, complementing the 2D projection of Figure 1. Here $s_0 = q_0$ and $s_1 = q_1 - 2r\delta_0$ are fixed by the true-set constraint, so the displayed surface is the true set itself in (δ_0, δ_1, r). The closures of the smooth strata, $\overline{S}_A = \{r = 0\}$ and $\overline{S}_B = \{\delta_1 = 0\}$, are 2-dimensional planes meeting on the 1-dimensional collapsed stratum S_C, which carries the multiplicity-2 singularity of Theorem 5.3. The two planes cross along S_C in the (δ_1, r) directions, which is the product geometry described by the normal form $u^2 + v^2 + 8w^2z^2$.

Numerical comparison with the constant $A_G(0, 0)$. Figure 5 numerically evaluates the fiber integral (9) at $p_0 = p_1 = 0$ on a fine log-spaced grid of $|p_2|$ and fits the leading slope by least squares, comparing the finite-grid estimate with the closed-form prediction $A_G(0, 0) = 1/(4\pi^2) \approx 0.02533$ from Proposition E.1. Panel (a) recovers the slope 0.02481 with bootstrap 95% CI $[0.02437, 0.02526]$, which is 2.0% below the theoretical value. The theoretical value lies just outside the interval, which points to a small bias from the finite grid or the quadrature. Panel (b) shows that after subtracting $A_G(0, 0) \log(1/|p_2|)$ the residual stays bounded on the well-behaved interval; the drop at the right edge is a numerical artifact from the narrowing exponential factor.

Fiber-integrand visualization. Figure 6 visualizes the integrand of (9) (equivalently (21)) as a function of the router scale r for several values of p_2. The left panel includes the reference asymptote $1/(\tau_{s_1} r)$, which matches the $p_2 = 0$ curve at small r. The $1/r$ factor responsible for the singularity at $p_2 = 0$ is visible, and the additional Gaussian factor $\exp(-p_2^2/(8\tau_{\delta_1}^2 r^2))$ damps the small-r region whenever $p_2 \neq 0$, removing the singularity. The right panel accumulates the integral up to r: it converges for every $p_2 > 0$ and diverges only at $p_2 = 0$, matching Proposition E.1.

Finite-sample Bayes factors. For Figure 2 in the main text, the marginal likelihoods of the linear and quadratic baselines are computed exactly as Gaussian integrals. For the routed model, the coordinates that enter linearly are integrated out in closed form and the router variable is integrated numerically. The figure uses unit prior scales and 40 synthetic datasets per sample size, with solid curves for empirical means and shaded bands for interquartile ranges. The top row reports $\log \mathrm{BF}_{R:Q}$ and the bottom row $\log \mathrm{BF}_{R:L}$, each under linear and quadratic truth. On linear truth, the routed model gradually gains a positive margin over quadratic regression with a smooth prior while still trailing linear regression. On quadratic truth, the routed and quadratic models stay close and both pull away from linear regression. The figure is meant only as a finite-sample illustration of Theorem 6.4.

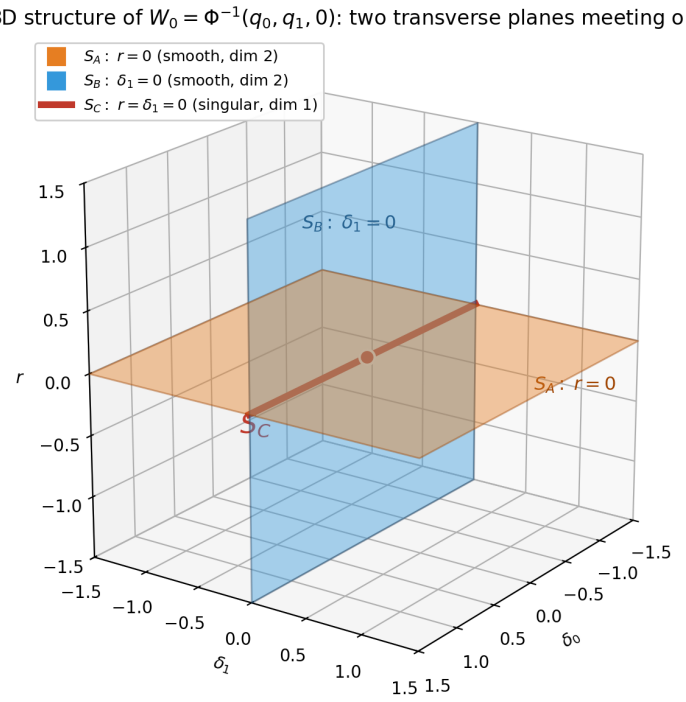

Figure 4: 3D rendering of $W_0 = \Phi^{-1}(q_0, q_1, 0)$ in (δ_0, δ_1, r) coordinates. The collapsed stratum S_C is shown in red.

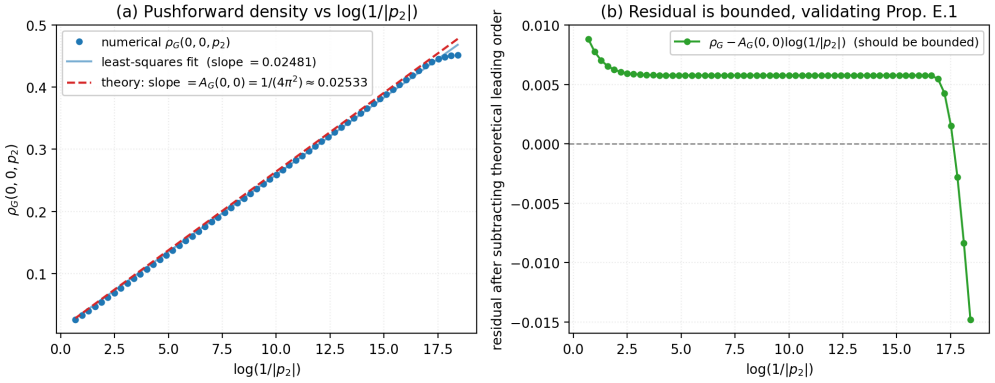

Figure 5: Numerical comparison with the constant $A_G(0, 0) = 1/(4\pi^2)$: fitted slope (a) and centered residual (b).

Bayes-factor heatmap. Figure 7 sweeps the truth signal strength q_2 and the sample size n on a grid and reports the mean of $\log \mathrm{BF}_{R:Q,n}$ over 24 replicates per cell. The red region at the bottom right shows a positive evidence margin for the routed model near linear truth, consistent with Theorem 6.4(a). In the upper rows the difference is small and negative, consistent with the bounded term in Theorem 6.4(b) for moderate q_2. The dashed curve $q_2 \propto n^{-1/4}$ is a visual guide, not a theoretical boundary established by the fixed-truth asymptotics.

Empirical verification of the asymptotic constants. Figure 8 gives a direct numerical check of Theorem 6.4(a). On linear truth ($q_2 = 0$), we compute $\log \mathrm{BF}_{R:Q,n}$ and $\log \mathrm{BF}_{R:L,n}$ for $n \in \{32, 64, 128, 256, 512, 1024, 2048, 4096\}$ over 60 replicates each, and subtract the theoretical leading order. The theorem predicts that the centered values stay bounded, and the plotted means show no clear growth over this range of n. The insets report ordinary least-squares fits of $\log \mathrm{BF} =$

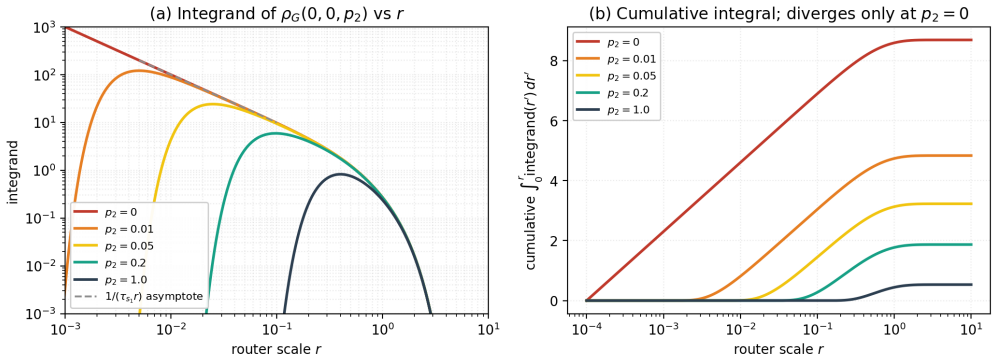

Figure 6: Integrand of (9) at $p_0 = p_1 = 0$: pointwise curves (left) and cumulative integral (right).

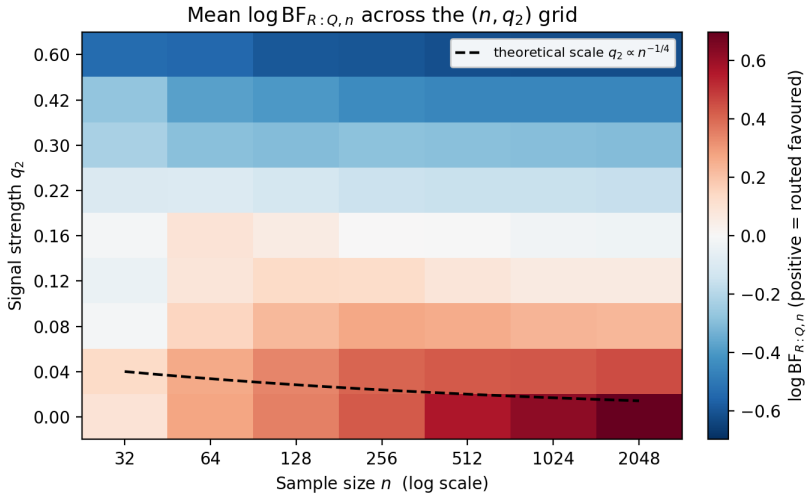

Figure 7: Mean $\log \mathrm{BF}_{R:Q,n}$ over the (n, q_2) grid. Dashed curve: $q_2 \propto n^{-1/4}$.

$a \log n + b \log\log n + c$ to the per-n means. The fitted $\log n$ coefficients are close to 0 and $-1/2$, while the fitted $\log\log n$ coefficients are sensitive to the limited range of n.

Direct empirical RLCT estimation via WBIC. Figure 9 checks numerically the RLCT λ of the routed model from Theorems 5.2 and 5.3, using the WBIC identity in the proof of Proposition G.1. We evaluate $\log Z_{\beta,n}$ with the same combination of closed-form integration and quadrature used elsewhere in this appendix and differentiate it in β by central finite differences at the WBIC inverse temperature $\beta = 1/\log n$. Proposition G.1 predicts the expansion $\mathrm{WBIC}(n) - nL_n(\hat{p}_{\mathrm{LSE}}) = \frac{3}{2}\log n + O_p(\sqrt{\log n})$ on linear truth. Panel (a) plots the pointwise estimator $\widehat{\lambda}(n)$; panel (b) compares a single-regressor fit $a \log n + c$ with the joint fit $a \log n + b \log\log n + c$. Each point is the mean over 32 replicates with an interquartile-range band. The single-regressor fit gives $\hat{a} = 1.478$ with bootstrap 95% CI $[1.464, 1.492]$, which excludes 1.5 and suggests a finite-sample bias, while the joint fit gives $(\hat{a}, \hat{b}) = (1.474, 0.024)$ with bootstrap CIs $[1.356, 1.602]$ and $[-0.778, 0.759]$, consistent with the theoretical $(3/2, 0)$. The estimate $\hat{b} \approx 0$ agrees with the absence of a deterministic $\log\log n$ term in WBIC, and the pointwise estimator reaches 1.40 at $n = 4096$, far from the value $5/2$ suggested by parameter counting. These numerical results are illustrations and do not replace the proofs.

Robustness across prior scales. Figure 10 repeats the WBIC empirical RLCT estimation across four values of the prior scale $\tau \in \{0.5, 1, 2, 4\}$. Each point is the mean over 16 replicates. The curves approach the reference value $\lambda = 3/2$, consistent with the fact that the RLCT in Theorem 5.3

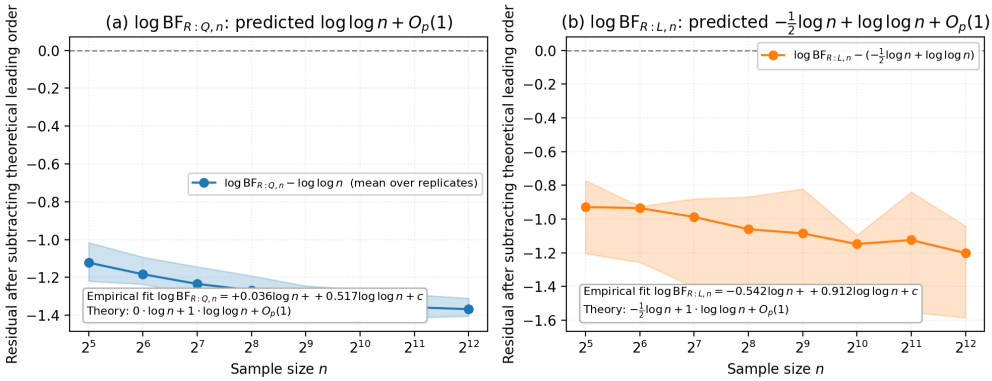

Figure 8: Centered Bayes factors on linear truth after subtracting the leading term from Theorem 6.4(a). (a) $\log \mathrm{BF}_{R:Q,n}$. (b) $\log \mathrm{BF}_{R:L,n}$.

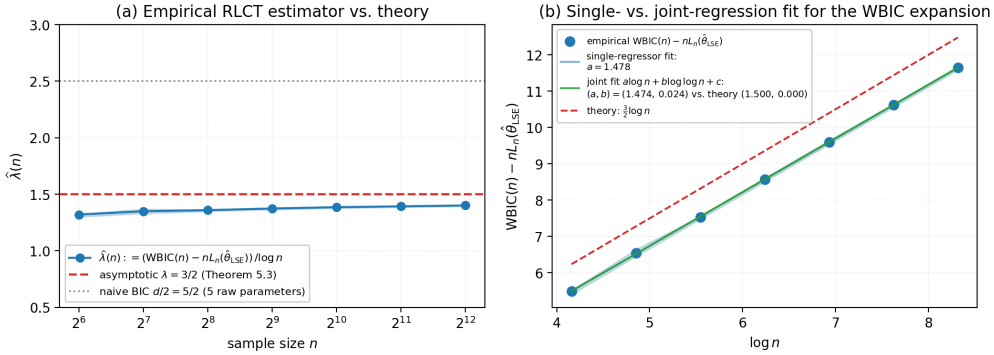

Figure 9: WBIC-based empirical RLCT check on linear truth. (a) Pointwise estimator $\widehat{\lambda}(n)$. (b) Fits to $\mathrm{WBIC}(n) - nL_n(\hat{p}_{\mathrm{LSE}})$.

does not depend on the prior scales. The corresponding single-regressor slopes of $\mathrm{WBIC} - nL_n$ are $\{1.370, 1.497, 1.463, 1.457\}$, all within 9% of 1.500, which is consistent with the $O_p(\sqrt{\log n}/\log n)$ residual scale.

Empirical coefficient-space contraction. Figure 11 illustrates posterior contraction in the observable coefficient space. By Proposition 6.1, the routed posterior projected to (p_0, p_1, p_2) is exactly the quadratic-regression posterior under ρ_G. The logarithmic singularity of the prior changes the multiplicity in the marginal likelihood but not the contraction rate (Proposition 6.2), so we compare the empirical curves with the regular $n^{-1/2}$ rate. Each point is the mean over 24 replicates with an interquartile-range band.

Companion heatmap for $\log \mathrm{BF}_{R:L}$. Figure 12 is the companion to the $\log \mathrm{BF}_{R:Q}$ heatmap of Figure 7. Each cell averages 24 replicates. Balancing the leading terms suggests the scale $|q_2| \asymp \sigma\sqrt{\log n/(2n)}$, where the linear baseline's quadratic-misspecification penalty $n\kappa_L = nq_2^2/\sigma^2$ becomes comparable to the $(1/2)\log n$ complexity gap. The dashed $n^{-1/2}$ guide in the plot suppresses this logarithmic factor. The red region at the top right illustrates the behavior for fixed nonzero curvature in Theorem 6.4(b), where $\log \mathrm{BF}_{R:L,n} \sim n\kappa_L$. The light-blue region in the bottom rows and at small n favors the linear model; at zero curvature, Theorem 6.4(a) predicts the asymptotic margin $(1/2)\log n - \log\log n$. Together, the two heatmaps illustrate both parts of Theorem 6.4 for both comparisons.

Empirical Gibbs concentration on the singular cross. Figure 13 illustrates Corollary G.3 by sampling from the Gibbs distribution $\pi_\eta \propto e^{-K(\theta)/\eta}\varphi(\theta)$ via random-walk Metropolis on $\mathbb{R}^5$ and projecting the samples onto the (r, δ_1) plane. Each panel aggregates $6 \times 30{,}000$ Metropolis samples

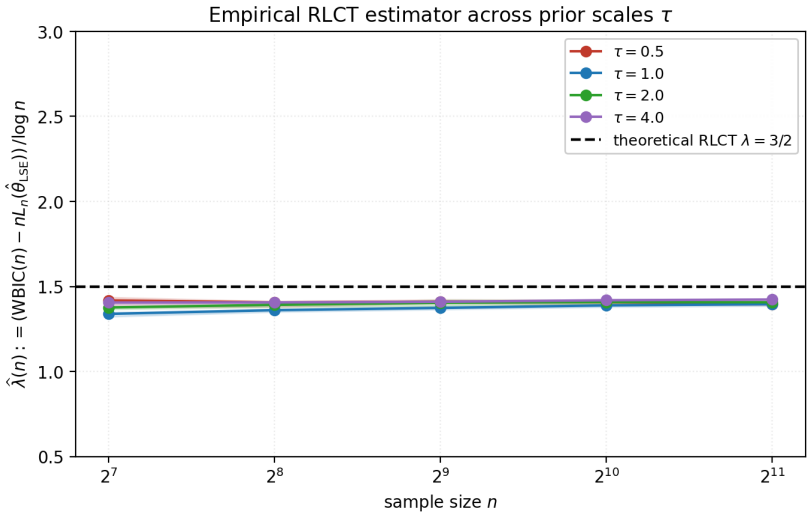

Figure 10: Empirical RLCT estimator $\widehat{\lambda}(n)$ across prior scales $\tau \in \{0.5, 1, 2, 4\}$.

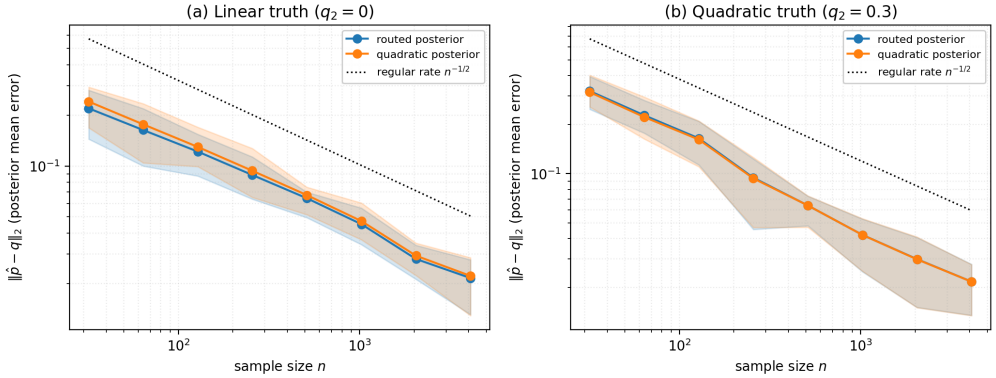

Figure 11: Empirical posterior contraction in observable coefficient space. The reference line is the regular rate $n^{-1/2}$.

and reports its acceptance rate. As η decreases from 1 to 10^{-3}, the empirical density concentrates on the singular cross $S_A \cup S_B \cup S_C = \{r = 0\} \cup \{\delta_1 = 0\}$ (cyan dashed lines), with increasing mass near the singular intersection S_C (cyan circle). This trend should not be read as a check of the constant in the mass ratio, since the corollary concerns fixed local neighborhoods and a compactly supported prior.

Symbolic verification of the central normal form. The polynomial change of variables (7), which takes the loss $K(\theta)$ near S_C to the product form $u^2 + v^2 + 8w^2z^2$ in Theorem 5.1(b), is verified symbolically in the reproducibility materials. The checks confirm that (i) substituting into the expanded loss gives exactly $u^2 + v^2 + 8w^2z^2$, (ii) the Jacobian determinant of the change of variables equals 1, and (iii) the inverse map composed with the forward map is the identity. This is an independent computer-algebra check of the central identity used in Theorem 5.2 and Corollary G.3.

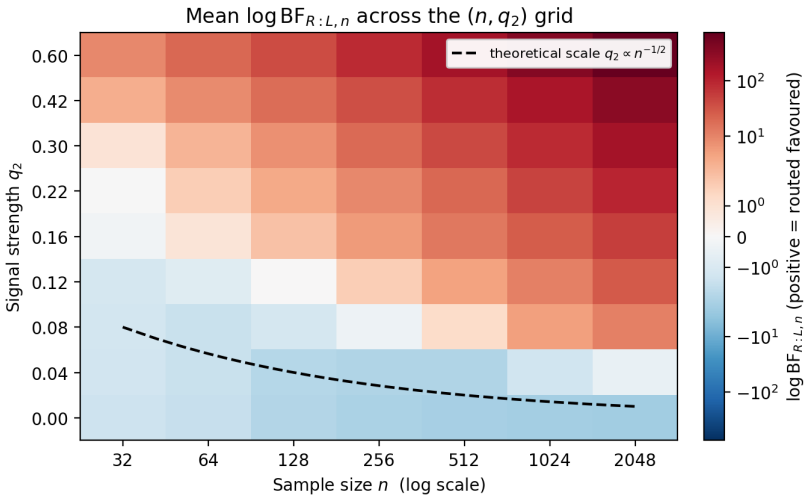

Figure 12: Mean $\log \mathrm{BF}_{R:L,n}$ over the (n, q_2) grid. Dashed curve: $q_2 \propto n^{-1/2}$, suppressing the logarithmic factor in the boundary scale.

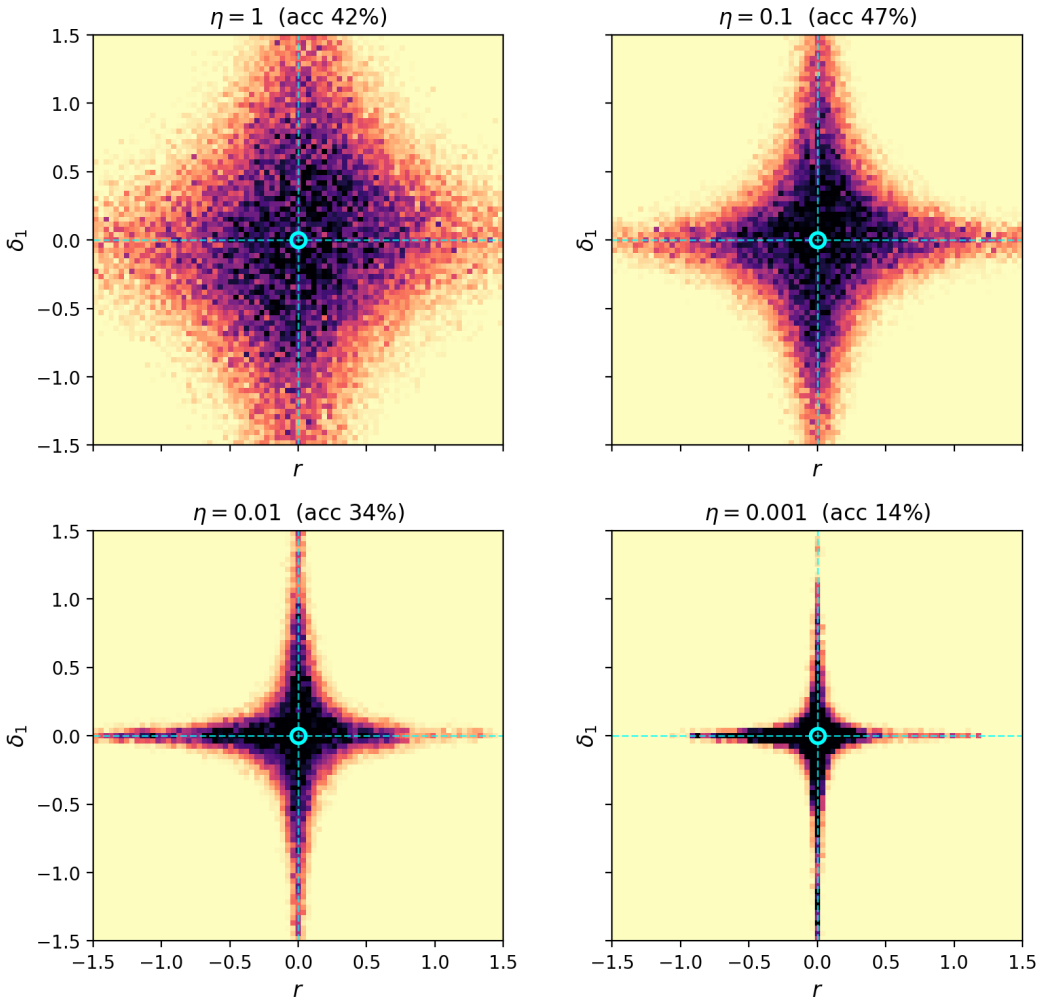

Figure 13: Empirical Gibbs density $\pi_\eta \propto e^{-K(\theta)/\eta}\varphi(\theta)$ on linear truth, projected onto the (r, δ_1) plane. Panels in reading order: $\eta = 1, 0.1, 0.01, 0.001$. Dashed cyan axes: $S_A \cup S_B$. Cyan circle: S_C.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the theoretical results for the minimal scalar routed regressor (fiber geometry, RLCTs, evidence asymptotics, the explicit pushforward prior, and the local results for sigmoid gates and d-dimensional inputs) together with their scope: the Bayes-factor comparison is not prior-invariant and applies to this realizable model. These claims match the results in Sections 1 to 7 and the proofs in the appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 states that the exact results hold only for the scalar two-branch model with an affine gate, notes that the results for sigmoid gates (Propositions 5.4 and F.4) and d-dimensional inputs (Proposition 5.5) are local, and makes no claims about softmax gating, larger architectures, or finite-time SGD dynamics. The scope paragraph in Section 1 also notes that, for compactly supported priors, the RLCT depends on whether the prior support reaches the collapsed stratum.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The model, the data-generating assumptions, and the conditions on the priors are stated in Section 3 and in each result. All theoretical results (fiber geometry, normal forms, RLCTs, the pushforward density, the sigmoid and d-dimensional results, and the Bayes-factor consequences) have complete proofs in Appendices B, C, D, E, F, and G, and Appendix D.1 gives proof sketches for the main RLCT theorems. The symbolic checks in the supplementary material only verify algebraic identities and do not replace these proofs.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The main claims are theoretical. The synthetic illustrations are described in Appendix H, and the sampling distributions, numerical grids, seeds, and integration settings are specified in the accompanying reproducibility package.

   Guidelines:

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

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The accompanying reproducibility package includes dependency documentation, code, and instructions to regenerate the figures and symbolic checks, including the scalar and d-dimensional normal forms, the sigmoid orthogonal decomposition, the curvature diagnostic, WBIC estimation, and prior-scale checks. All stochastic runs fix their seeds and require only standard scientific computing libraries; no external datasets are used.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Appendix H describes the synthetic data (scalar Gaussian features and unit-variance Gaussian noise), unit prior scales, the sensitivity check at $\tau \in \{0.5, 1, 2, 4\}$, and 16–60 independent dataset draws per sample size where applicable, as stated in the accompanying text. Deterministic geometry and quadrature illustrations do not use dataset replicates.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Stochastic Bayes-factor, WBIC, and coefficient-contraction plots report variability across independent dataset draws using interquartile bands where applicable. The WBIC and pushforward-slope fits also report bootstrap confidence intervals. Deterministic geometry and quadrature plots do not involve dataset variability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: A verification rerun of all 14 scripts took about 137 seconds on a CPU with no GPU. The supplementary notes record the CPU model, RAM, software versions, and per-script wall times for this rerun, separately from the original experiments. The work does not involve model training or large datasets.

   Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research is foundational statistical learning theory. It does not involve human subjects, crowdsourcing, personal data, or the release of high-risk dual-use artifacts, and it conforms to the Code of Ethics.

Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This foundational analysis concerns a minimal regression model and proposes no application or deployment. We do not identify a direct societal impact requiring a separate analysis.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper provides theoretical proofs and simple numerical sanity checks for a minimal proxy model; it does not release large pre-trained language models, image generators, scraped datasets, or any high-risk dual-use assets requiring safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring

that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not use or repackage any external datasets, third-party codebases, proprietary models, or third-party assets. The submitted reproducibility scripts and figures are author-created; the only external software used for the sanity checks is the standard open-source scientific computing stack listed in the reproducibility package.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The supplementary numerical and symbolic checks are new assets, documented in the accompanying reproducibility package with dependencies and instructions for use.

    Guidelines:

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

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

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

    Answer: [N/A]

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

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve human subjects research, so IRB approval is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not a component of the model or inference method. The authors independently verified the analyses, proofs, and numerical results.

Guidelines:

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