# IDENTIFIABILITY IN DEEP EVIDENTIAL REGRESSION: SCALAR BOUNDARIES AND A MULTIVARIATE GAUGE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Deep Evidential Regression (DER) gives single-pass uncertainty estimates, but different evidential parameters can induce exactly the same Student-$t$ predictive distribution while reporting different evidence and epistemic uncertainty. We characterize when architectural restrictions remove this ambiguity. If the evidence strength $\nu$ is shared across inputs, the parameters are identifiable exactly when no two distinct allowed scale functions $\beta$ are positive constant multiples of one another. If both $\nu(x)$ and $\beta(x)$ are represented by affine softplus heads on fixed features, they are jointly identifiable under an open-set density condition unless both heads are constant. For multivariate DER, we decompose the prior scale matrix into a normalized shape, which the predictive law identifies, and a scalar amplitude governed by the same scalar criterion. Experiments distinguish this structural property from finite-sample behavior. The corresponding heads yield more reproducible evidence strengths on CIFAR-rotation and UCI regression tasks, with predictive quality evaluated alongside stability. In a joint-feature QM9 experiment, the gauge-fixed parameterization reduces variation in the epistemic covariance trace relative to coupled and uncoupled Normal-Inverse-Wishart baselines at the final checkpoint. Together, these results show how architectural structure can make evidential parameters uniquely determined and more reproducible in practice.

## 1 INTRODUCTION

Deep Evidential Regression (DER; Amini et al., 2020) estimates a prediction and its uncertainty in a single forward pass. It offers a computationally attractive alternative to repeated sampling with MC-Dropout (Gal & Ghahramani, 2016) or training multiple models in a Deep Ensemble (Lakshminarayanan et al., 2017). A DER network outputs the parameters of a Normal-Inverse-Gamma (NIG) prior, whose marginal predictive distribution is Student-$t$. Quantities derived from these internal parameters are then reported as evidence strength and as aleatoric and epistemic uncertainty, and have been used as confidence signals in downstream applications (Soleimany et al., 2021).

The predictive distribution, however, need not determine those internal parameters. A change in evidence strength can be exactly compensated by a change in the prior scale, leaving every predictive probability unchanged while altering the reported uncertainty decomposition. Consequently, two models can fit the data equally well yet report different evidence and epistemic uncertainty. No amount of data can make an objective that depends only on the Student-$t$ predictive law choose a unique representation along this equivalence orbit (Theorem 3.1). This makes identifiability a prerequisite for interpreting variation in evidential parameters as data-driven rather than as an arbitrary consequence of parameterization or optimization.

Jürgens et al. (2024) establish non-uniqueness for unconstrained evidential parameters, leaving open when restrictions imposed by a shared neural-head architecture can remove the ambiguity. Related work questions the Bayesian interpretation of evidential parameters (Meinert et al., 2023) and the elicitation of faithful epistemic uncertainty by second-order scoring rules (Bengs et al., 2023). We ask when architectural structure removes the predictive equivalence itself. We call a model class *predictively identifiable* when equality of its predictive distributions implies equality of its evidential parameters, up to sets of inputs of probability zero. Unrestricted outputs can be refactored independently at every input, but neural heads share weights and functional form across inputs; a pointwise-equivalent alternative may therefore not belong to the same model class. Our analysis



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

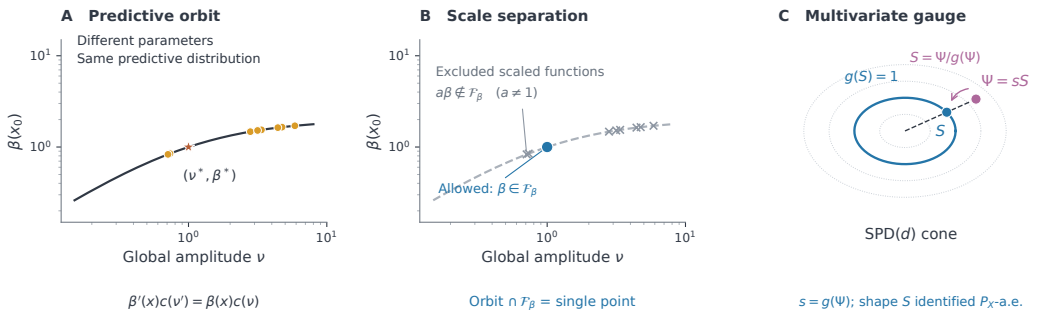


Figure 1: Predictive equivalence and model structure. (A) In the global-evidence setting, compensating changes in internal scales trace a one-dimensional orbit with the same predictive distribution. (B) A scale-separated function class intersects each orbit at most once (Theorem 3.2). (C) Normalizing the multivariate scale matrix separates its identifiable shape from the scalar ambiguity (Theorem 3.4).

isolates uniqueness of the evidential representation as a prerequisite for interpreting its internal parameters.

Our contributions are:

- **Scalar identifiability boundaries.** When evidence strength $\nu$ is shared across inputs, predictive identifiability holds if and only if no two distinct allowed scale functions $\beta$ are positive constant multiples of one another (Theorems 3.1 and 3.2). We call this property *scale-separation* and, under an open-set density condition on fixed features, give a concrete bias-free softplus-linear construction (Corollary B.2). Under the same condition, affine softplus heads for both $\nu(x)$ and $\beta(x)$ are jointly identifiable whenever at least one head is nonconstant, whereas the ambiguity remains when both are constant (Theorem 3.3). Together, these results locate the boundary between unrestricted non-identifiability and identifiability induced by a shared head architecture.
- **Multivariate identifiability by gauge fixing.** For multivariate DER with a Normal-Inverse-Wishart (NIW) prior (Meinert & Lavin, 2021), we factor the prior scale matrix into a scalar amplitude and a shape matrix with fixed normalization. The predictive law identifies the normalized shape, reducing the unresolved part to the same scalar problem. Applying the global or per-input results then yields multivariate identifiability without imposing a fixed ratio between evidence parameters (Theorems 3.4 and 3.5).
- **Empirical evaluation.** Synthetic studies distinguish recovery of known parameters from cross-seed stability. The corresponding structured heads yield more reproducible evidence strengths on CIFAR-rotation and UCI regression tasks, with predictive quality reported alongside stability. On QM9, the gauge-fixed model reduces variation in the epistemic covariance trace relative to coupled and uncoupled NIW baselines at the final checkpoint. This joint-training experiment complements the fixed-feature theory (Section 4).

Figure 1 illustrates the common geometry behind the scalar and multivariate results. Section 2 formalizes the predictive equivalence and positions the results relative to prior work; Section 3 gives the assumptions and proofs.

## 2 Problem and Prior Work

### 2.1 DER Parameterization

The full Deep Evidential Regression head outputs four functions $(\gamma(x), \nu(x), \alpha(x), \beta(x))$ parameterizing a Normal-Inverse-Gamma prior on the latent mean–variance $(\mu, \sigma^2)$. Marginalizing yields a Student-$t$ predictive

$$p(y \mid x) \;=\; \mathrm{St}_t\!\left(y;\; \mu = \gamma(x),\; \sigma^2 = \frac{\beta(x)(1+\nu(x))}{\nu(x)\,\alpha(x)},\; \mathrm{df} = 2\alpha(x)\right). \tag{1}$$

Table 1: Prior DER analyses and our structural identifiability results. The per-sample results cover affine heads with softplus activations on fixed features. Predictive injectivity and faithful second-order uncertainty elicitation are different targets.

| Work or result | Setting and conditions | Main conclusion |
|---|---|---|
| Amini et al. (2020) | Per-sample NIG head and evidence regularization | Single-pass uncertainty method; no identifiability theorem |
| Jürgens et al. (2024); Meinert et al. (2023) | Unconstrained evidential parameters | Predictive non-injectivity and limits of the uncertainty interpretation |
| Meinert & Lavin (2021) | NIW model with $\nu_W(x) = r\kappa(x)$; known fixed $r$ | The coupling determines NIW outputs from the predictive |
| Bengs et al. (2023) | Second-order uncertainty elicitation | Impossibility within the stated scoring-rule framework |
| Global scalar (Theorems 3.1 and 3.2) | Global $\nu$; positive function class $\mathcal{F}_\beta$ | Predictive injectivity iff $\mathcal{F}_\beta$ is scale-separated |
| Per-sample scalar (Theorem 3.3) | Fixed features with the density condition; softplus-affine heads; $(u, w) \neq (0, 0)$ | Unique slopes and offsets of both evidential heads |
| Global multivariate (Theorem 3.4) | Fixed features and gauge; global $\kappa$; scale-separated $\mathcal{F}_s$ | Identified normalized shape; remaining ambiguity reduces to the scalar case |
| Per-sample multivariate (Theorem 3.5) | Fixed features with the density condition; fixed gauge; softplus-affine scalar heads; $(u, w) \neq (0, 0)$ | Identified per-sample heads and normalized SPD shape |

We denote $c(t) := (1 + t)/t$ throughout, so the predictive scale-squared is $\beta(x)\, c(\nu(x))/\alpha(x)$. Following Amini et al. (2020), training minimizes $\mathbb{E}_{(x,y)\sim\mathcal{D}}[-\log p(y \mid x) + \lambda\,|y - \gamma(x)|\,(2\nu(x) + \alpha(x))]$ for some $\lambda \geq 0$. We compare $\lambda = 0$ with the same evidence regularizer at $\lambda = 0.1$ in Section 4.

### 2.2 Amplitude regimes and identifiability target

We treat two cases for the amplitude head: a global scalar $\nu \in \mathbb{R}_+$ (Theorems 3.1 and 3.2), and a per-sample softplus-affine head $\nu(x) = \mathrm{softplus}(u^\top \phi(x) + d)$ on a fixed feature map $\phi$ (Theorem 3.3). Unrestricted $\nu(x)$ admits pointwise refactoring of the predictive product: for any positive measurable $\nu_2(x)$, setting $\beta_2(x) = \beta_1(x)\, c(\nu_1(x))/c(\nu_2(x))$ leaves the Student-$t$ predictive unchanged, so the parameters cannot be recovered.

The identifiability target is the predictive family

$$\mathcal{P} \;=\; \{\mathrm{St}_t(\,\cdot\,;\ \gamma, \nu, \alpha, \beta) :\ (\gamma, \alpha, \beta) \in \mathcal{G} \times \mathcal{A} \times \mathcal{F}_\beta,\ \nu \in \mathbb{R}_+\}$$

where $\mathcal{F}_\beta$ is a class of positive measurable $\beta$-functions on $\mathcal{X}$. The predictive map is $\Theta : (\gamma, \nu, \alpha, \beta) \mapsto p(\,\cdot \mid \cdot\,; \gamma, \nu, \alpha, \beta)$; we ask whether $\Theta$ is injective up to $P_X$-null sets. A strictly proper scoring rule $S$ on $\mathbb{R}$ (e.g. the log score, Gneiting & Raftery, 2007) satisfies $Q = \arg\min_P \mathbb{E}_{y\sim Q}[S(P, y)]$. Under *realizability*, a single parameter tuple in this class represents the true conditional law $Q(\cdot \mid x)$ for $P_X$-a.e. $x$. Injectivity of $\Theta$ then makes that representation unique up to $P_X$-null sets.

**Notation.** $P_X$ denotes the marginal on $\mathcal{X}$; $\phi : \mathcal{X} \to \mathbb{R}^k$ a measurable feature map; $\mathrm{SPD}(d)$ the cone of $d \times d$ symmetric positive-definite matrices; equalities of functions are $P_X$-a.e. Table 4 (Section A) lists recurring symbols.

### 2.3 Prior work and our positioning

Table 1 distinguishes the existing analyses from the global and per-sample results established here.

**Deep Evidential Regression and its ambiguity.** Amini et al. (2020) introduced DER as a single-forward-pass method predicting four NIG hyperparameters. Training combines predictive NLL with

the evidence regularizer $\lambda|\delta|(2\nu + \alpha)$. The broader evidential-DL family is surveyed by Ulmer et al. (2023). Applications such as emotion-attribute regression (Wu et al., 2023) and loss-given-default forecasting (Nagl et al., 2022) use the evidential outputs as uncertainty signals. These uses motivate asking whether the predictive distribution determines the evidential parameters uniquely. Meinert et al. (2023) identify degeneracy in the NIG likelihood and gradient problems in the regularizer. They argue that DER's epistemic output is a relative signal rather than a posterior pseudo-count.

**Molecular and atomistic applications.** Soleimany et al. (2021) integrated evidential heads with molecular message-passing networks and SchNet for property prediction and uncertainty-guided discovery. Xu et al. (2026) combined equivariant features with evidential quantile regression to estimate directional force uncertainty for interatomic potentials. Wang et al. (2026) model force uncertainty through a full NIW covariance tensor that transforms equivariantly under rotations. Our multivariate analysis addresses the complementary question of when the predictive distribution uniquely determines its evidential parameters.

**Fixed-ratio coupling.** Meinert & Lavin (2021) extended DER to NIW priors and imposed $\nu_W(x) = r\kappa(x)$ for a fixed known $r > 0$. The predictive degrees of freedom determine $\nu_W$, and hence $\kappa$. The predictive scale matrix then determines $\Psi$. Thus this constraint gives pointwise identifiability of the NIW outputs. Our results identify independent heads through their shared function class, without prescribing this ratio.

**Training remedies and structural identifiability.** We include an auxiliary MAE loss as DER+MAE in §D.5. Other remedies address activation-induced training failures. Ye et al. (2024) identify a "fixed area" where the head activations block learning and propose a corrective regularizer. Wu et al. (2024) address "evidence contraction" with a non-saturating regularizer. These methods change the training objective. Our analysis asks when the resulting predictive distribution uniquely determines the evidential parameters. The evidence regularizer of Amini et al. favors particular points on a predictive-equivalence orbit. Scale-separation ensures that the model class contains at most one point on each orbit.

**Identifiability and second-order impossibility.** Jürgens et al. (2024) prove predictive non-injectivity for unconstrained evidential parameters across a broad class of second-order exponential-family models. Bengs et al. (2023) study second-order uncertainty elicitation. Within their scoring-rule framework, no loss incentivizes faithful epistemic uncertainty in the way proper scores support first-order prediction (Gneiting & Raftery, 2007). We characterize when the ambiguity identified by Jürgens et al. disappears within a specified function class. Under scale-separation, each predictive-equivalence orbit intersects that class at most once (Theorem 3.2).

**Uncertainty baselines and other identifiability settings.** MC-Dropout (Gal & Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017) are the canonical multi-pass methods: the former averages stochastic forward passes, while the latter aggregates independently trained models. Neither exhibits the predictive-level NIG orbit studied here. Identifiability is also studied in VAEs and ICA (Khemakhem et al., 2020), representation-disentanglement (Locatello et al., 2019), and deep latent variable models (Yacoby et al., 2022). These works concern recovery of latent generating factors, whereas our target is the internal evidential parameters. Skepticism about evidential UQ (Bengs et al., 2023; Jürgens et al., 2024; Meinert et al., 2023; Shen et al., 2024) asks whether DER's quantities are faithful; we ask the prior question of whether they are uniquely determined.

# 3 Theory

## 3.1 Scalar identifiability

The Student-$t$ predictive in Equation (1) depends on $(\nu, \beta)$ only through the product $\beta(x)\, c(\nu(x))$ with $c(t) = (1 + t)/t$. Fixing $\gamma$ and $\alpha$, any global rescaling $\nu \mapsto \nu'$ of the amplitude is then absorbed by the dual rescaling $\beta(x) \mapsto \beta(x)\, c(\nu)/c(\nu')$, leaving the predictive identical at every $x$ and the population score constant along the resulting one-parameter orbit. Proposition B.1 (Section B) records this scaling symmetry. It is one explicit form of the non-injectivity diagnosed by Jürgens et al. (2024): whenever $\mathcal{F}_\beta$ contains $\beta$ and $a \cdot \beta$ for some $a \neq 1$, the parameters lie on a one-parameter orbit of

equivalent predictives. The rescaled $\beta'$ need not itself lie in $\mathcal{F}_\beta$; the structural condition that prevents $\mathcal{F}_\beta$ from containing two distinct elements of any such orbit is the following.

**Definition 3.1** (Scale-separation with respect to $P_X$)**.** A class $\mathcal{F}_\beta$ of positive measurable functions on $\mathcal{X}$ is *scale-separated with respect to* $P_X$ if, for any $\beta_1, \beta_2 \in \mathcal{F}_\beta$ and any $a > 0$,

$$\beta_2 \;=\; a \cdot \beta_1 \;\; P_X\text{-a.e.} \quad\Longrightarrow\quad a = 1 \;\text{ and }\; \beta_1 = \beta_2 \;\; P_X\text{-a.e.}$$

**Theorem 3.1** (Non-identifiability without scale-separation)**.** *Fix nonempty parameter classes $\mathcal{G}$ for $\gamma$, $\mathcal{A}$ for $\alpha$, $\mathcal{F}_\beta \subseteq \{\beta : \mathcal{X} \to \mathbb{R}_+ \text{ measurable}\}$, and the global-$\nu$ assumption ($\nu \in \mathbb{R}_+$ a single shared scalar). If $\mathcal{F}_\beta$ is* not *scale-separated with respect to $P_X$, then there exist two distinct parameter tuples $(\gamma, \nu_1, \alpha, \beta_1), (\gamma, \nu_2, \alpha, \beta_2) \in \mathcal{G} \times \mathbb{R}_+ \times \mathcal{A} \times \mathcal{F}_\beta$ producing identical predictives $P_X$-a.e. In particular, there exists a realizable data-generating conditional $Q$ for which the population minimizer of any strictly proper* predictive *scoring rule (whose population risk depends only on the predictive law and is finite) is not unique in $\nu$ over this parameter class; hence $\nu$ is not globally identifiable from any such score. Equivalently: scale-separation of $\mathcal{F}_\beta$ is necessary for global model-class identifiability of $\nu$.*

**Theorem 3.2** (Injectivity under scale-separation)**.** *Consider the class of parameter tuples with measurable $\gamma : \mathcal{X} \to \mathbb{R}$, $\alpha : \mathcal{X} \to (1, \infty)$, $\beta \in \mathcal{F}_\beta$ (in particular $\beta(x) > 0$), and global $\nu \in (0, \infty)$. Under the global-$\nu$ assumption and with $\mathcal{F}_\beta$ scale-separated with respect to $P_X$, the map*

$$(\gamma, \nu, \alpha, \beta) \;\longmapsto\; \mathrm{St}_t(\,\cdot\,; \gamma, \nu, \alpha, \beta)$$

*is injective up to $P_X$-null sets on $\mathcal{X}$.*

Proofs in Section B. Together, the two theorems establish that scale-separation is necessary and sufficient for injectivity of the global-$\nu$ predictive map. Under realizability, any strictly proper score has its unique population minimizer at the true predictive (Gneiting & Raftery, 2007); under scale-separation, injectivity then gives $P_X$-a.e. recovery of $(\gamma, \nu, \alpha, \beta)$. Corollary B.2 (Section B) gives an implementable example using bias-free softplus linear heads: if $\phi : \mathcal{X} \to \mathbb{R}^k$ has a pushforward that admits a Lebesgue density positive almost everywhere on a non-empty open set, then $\mathcal{F}_\beta = \{x \mapsto \mathrm{softplus}(w^\top \phi(x)) : w \in \mathbb{R}^k\}$ is scale-separated by real-analytic uniqueness on $\mathbb{R}^k$. The density condition holds in our Gaussian-input synthetic setup with a full-rank projection, but can be restrictive for learned features that concentrate on low-dimensional manifolds. Diagnostics for the frozen ResNet-18 features are reported in Sections D.7 and I.

**Why per-sample $\nu(x)$ needs separate treatment.** If $\nu$ is allowed to be a trainable function, the orbit becomes the pointwise product $c(\nu(x))\beta(x)$, and any pair $(\nu_1, \beta_1), (\nu_2, \beta_2)$ matching this product is predictively equivalent. This is an infinite-dimensional ambiguity, not a scalar one, and Theorem 3.2 does not extend. The next theorem closes the gap on the softplus-affine subclass actually used by Amini et al. (2020). Here the $\nu$-head is the affine parameter pair $(u, d)$ and the $\beta$-head is $(w, b)$; the fixed feature map $\phi$ belongs to neither. The assumptions are separate: fixed features specify the function class, population thickness permits analytic continuation, and the softplus-affine form enables product factorization.

**Theorem 3.3** (Per-sample softplus-affine identifiability)**.** *Fix a measurable feature map $\phi : \mathcal{X} \to \mathbb{R}^k$ whose pushforward $P_X \circ \phi^{-1}$ admits a Lebesgue density positive almost everywhere on a non-empty open set $U \subseteq \mathbb{R}^k$. Let $s(r) := \mathrm{softplus}(r) = \log(1 + e^r)$, and for $u, w \in \mathbb{R}^k$ and $d, b \in \mathbb{R}$ define*

$$\nu_{u,d}(x) \;:=\; s(u^\top \phi(x) + d), \qquad \beta_{w,b}(x) \;:=\; s(w^\top \phi(x) + b).$$

*If $c(\nu_{u_1,d_1}(x))\,\beta_{w_1,b_1}(x) = c(\nu_{u_2,d_2}(x))\,\beta_{w_2,b_2}(x)$ $P_X$-a.e. with $(u_i, w_i) \neq (0, 0)$ for $i = 1, 2$, then $u_1 = u_2$, $d_1 = d_2$, $w_1 = w_2$, and $b_1 = b_2$. Moreover, no softplus-affine pair with $(u, w) \neq (0, 0)$ produces a constant predictive product: some direction has induced slopes $(w^\top v, u^\top v) \neq (0, 0)$, and Lemma B.3 makes the restriction along that line nonconstant. Hence the only ambiguity in the softplus-affine subclass is among constant pairs $(u, w) = (0, 0)$, where $s(b)\,c(s(d))$ admits the one-parameter level curve described in Lemma B.3.*

*Proof sketch.* Student-$t$ injectivity reduces predictive equality to equality of products $c(\nu(x))\beta(x)$. The density condition gives Lebesgue-a.e. equality on an open set in feature space, and real-analytic uniqueness extends this identity to all of $\mathbb{R}^k$. The univariate lemma (Lemma B.3, Section B) identifies the slopes and offsets of $q_{a,b,r,d}(t) := s(at + b)\,c(s(rt + d))$ from the softplus tail asymptotics

whenever $(a, r) \neq (0, 0)$. We apply this lemma along lines through the origin whose directions make neither pair of projected slopes vanish. It recovers the offsets and the projected slopes $u_i^\top v, w_i^\top v$. These projections agree on a dense set of directions, so continuity of the linear functionals gives equality of the full vectors and hence of $(u, d, w, b)$. □

The excluded case $(u, w) = (0, 0)$ makes both heads constant and recovers the scalar orbit. Allowing biases does not destroy identifiability when $(u, w) \neq (0, 0)$. When $u = 0$ and $w \neq 0$, $\nu_{0,d} \equiv s(d)$ is a global scalar, and the theorem recovers the corresponding softplus-affine special case of the global-$\nu$ result in Theorem 3.2 (Corollary B.4, Section B). The fixed-$\phi$ assumption motivates the two-stage protocol in Section K, where the trunk is frozen during evidential-head training. The factorization result depends on the specific analytic and tail properties of softplus, not positivity alone.

**Joint trunks.** Jointly training $\phi_\theta$ with the heads enlarges the effective $\beta$-class and can violate scale-separation. The fixed-feature results therefore do not provide an identifiability guarantee for this setting. Exploratory comparisons with vanilla multivariate DER are reported in Section H.

### 3.2 Multivariate gauge-fixing

The scalar obstruction reappears in multivariate DER as a scale ambiguity in the NIW scale matrix. Gauge-fixing isolates that scalar ambiguity from the identifiable SPD shape, after which the residual scalar problem is handled by Theorems 3.2 and 3.3.

For multivariate targets $y \in \mathbb{R}^d$, DER predicts a Normal-Inverse-Wishart prior $(\mu_0(x), \kappa(x), \Psi(x), \nu_W(x))$ with $\Psi \in \mathrm{SPD}(d)$, $\nu_W > d - 1$. The predictive marginalizes to a multivariate Student-$t$ with location $\mu_0$, scale matrix $\Lambda(x) = c(\kappa)\Psi(x)/\nu_t$, and df $\nu_t = \nu_W - d + 1$. Throughout this subsection $\kappa \in \mathbb{R}_+$ and $\nu_W \in (d - 1, \infty)$ are global scalars, $\phi$ is fixed, and $\Psi$ is strictly SPD; the per-sample versions appear at the end.

**Definition 3.2** (Gauge-fixed factorization)**.** A map $g : \mathrm{SPD}(d) \to \mathbb{R}_+$ is a *degree-*1 *positive-homogeneous gauge* if $g(a \cdot S) = a\, g(S)$ for all $a > 0$, $S \in \mathrm{SPD}(d)$; canonical examples are $g(S) = \mathrm{tr}(S)/d$ and $g(S) = \det(S)^{1/d}$. Any $\Psi \in \mathrm{SPD}(d)$ admits a unique factorization $\Psi = s \cdot S$ with $s = g(\Psi) > 0$ and $g(S) = 1$; we call $s$ the *scalar amplitude* and $S$ the *gauge-fixed shape*.

The multivariate Student-$t$ map $(\mu, \Lambda, \nu_t) \mapsto \text{mv-St}_t(\,\cdot\,; \mu, \Lambda, \nu_t)$ is itself injective on $\mathbb{R}^d \times \mathrm{SPD}(d) \times (0, \infty)$ via a standard density-level argument (unique mode, elliptical level sets, polynomial tail, ray-matching; Lemma C.1, Section C). Combined with gauge-fixing, this reduces multivariate identifiability to the scalar problem.

**Theorem 3.4** (Gauge-fixed multivariate identifiability)**.** *Assume fixed $\phi$, fixed degree-*1 *gauge $g$, global $\kappa \in \mathbb{R}_+$ and $\nu_W \in (d - 1, \infty)$, and strict-SPD $\Psi(x) = s(x)S(x)$ with $s \in \mathcal{F}_s$ and $g(S) = 1$. Then: (a) $\mu_0(x)$, $S(x)$, $\nu_W$ are identified $P_X$-a.e. without a scale-separation assumption; (b) $(\kappa, s)$ is identified iff $\mathcal{F}_s$ is scale-separated; (c) under scale-separation, the full predictive map $(\mu_0, \kappa, s, S, \nu_W) \mapsto \text{mv-St}_t$ is injective $P_X$-a.e.*

The implementable head uses $s(x) = \mathrm{softplus}(w^\top \phi(x))$ and $S(x) = \hat{S}(x)/g(\hat{S}(x))$. A Cholesky factor with positive diagonal gives a learnable strictly-SPD $\hat{S} = LL^\top$. Under the density condition of Corollary B.2, the scalar argument shows that $\mathcal{F}_s$ is scale-separated. We use $g(S) = \mathrm{tr}(S)/d$ by default and evaluate the determinant gauge through the Cholesky factor. This identifies the global $\kappa$ and the amplitude function without coupling $\kappa$ to $\nu_W$. If the true conditional law belongs to this model class, a strictly proper predictive score has a unique population minimizer at its parameter tuple, up to $P_X$-null sets.

**Theorem 3.5** (Per-sample softplus-affine multivariate identifiability)**.** *Under the thickness condition and a degree-*1 *gauge $g$, replace the global scalar $\kappa$ and amplitude function $s \in \mathcal{F}_s$ in Theorem 3.4 by per-sample softplus-affine heads $\kappa_{u,d}(x) = \mathrm{softplus}(u^\top \phi(x) + d)$ and $s_{w,b}(x) = \mathrm{softplus}(w^\top \phi(x) + b)$ with $(u, w) \neq (0, 0)$, and let $\nu_W$ be either a global scalar or a measurable function $\mathcal{X} \to (d - 1, \infty)$. Then $\mu_0(x), \nu_W(x), \kappa_{u,d}(x), s_{w,b}(x), S(x)$ are identified $P_X$-a.e. The normalized shape $S$ is unique; its raw head $\hat{S}$ is identified only up to positive rescaling at each $x$. The exclusion $(u, w) = (0, 0)$ is sharp.*

*Proof sketch.* By multivariate Student-$t$ injectivity, $(\mu_0(x), \Lambda(x), \nu_t(x))$ are identified $P_X$-a.e., hence $\nu_W(x)$ pointwise. Positive homogeneity of $g$ gives $g(\Lambda(x)) = c(\kappa_{u,d}(x))\, s_{w,b}(x)/\nu_t(x)$, identifying the scalar product $c(\kappa_{u,d}(x))\, s_{w,b}(x)$. Theorem 3.3 (with $\nu \leftarrow \kappa$, $\beta \leftarrow s$) recovers $(u, d, w, b)$ uniquely under $(u, w) \neq (0, 0)$. The shape $S(x) = \Lambda(x)\nu_t(x)/[c(\kappa_{u,d}(x))s_{w,b}(x)]$ follows. Sharpness inherits from Theorem 3.3; full proof in Section C. □

# 4 EXPERIMENTS

We compare the complete method with vanilla DER in cross-seed stability and predictive quality. We initialize $\nu(x)$ or $\kappa(x)$ at one to give every method the same initial evidence strength, then learn it while varying the remaining network weights and minibatch order across seeds. For a per-sample quantity $q(x)$, we compute its coefficient of variation (CV), the sample standard deviation divided by the mean, across seeds at each fixed test input and then average over inputs. A global parameter is the corresponding constant function of $x$. Synthetic experiments separately examine recovery of known parameters. Table 2 summarizes the settings; Section J.1 specifies the output heads and which parameters are learned.

Table 2: Settings used to assess stability and parameter recovery.

| Experiment | Setting | Type |
|---|---|---|
| CIFAR-rotation | frozen ResNet-18; three data realizations | frozen-feature |
| QM9 | jointly trained SchNet, three targets | joint training |
| UCI-9 | two-stage global head vs. complete vanilla methods | real-data |
| Scalar synthetic | fixed Gaussian $\phi : \mathbb{R}^{10} \to \mathbb{R}^3$, known $\beta^*$ | theorem-class |
| Per-sample synthetic | fixed Gaussian $\phi$, known softplus-affine heads | theorem-class |

## 4.1 CROSS-SEED STABILITY ON CIFAR-ROTATION FEATURES

We train each method for $1000$ epochs on frozen ResNet-18 features, using ten training seeds for each of three fixed rotation-data realizations. Our global-$\nu$ head has lower mean pointwise amplitude CV than both unregularized and regularized vanilla DER in all three settings (Figure 2). The vanilla-to-ours CV ratios range from $3.95$ to $7.37$ without the evidence regularizer and from $2.31$ to $5.11$ with it. The model's epistemic variance also has lower mean pointwise CV. Mean predictive NLL differs from unregularized vanilla by less than $0.006$ nats in each setting, while regularized vanilla has higher NLL (Section D.7). Thus the amplitude comparison is not explained by a large difference in predictive fit.

## 4.2 MULTIVARIATE STABILITY UNDER JOINT TRAINING

The QM9 comparison tests stability under jointly learned features, beyond the fixed-feature assumptions. We train SchNet on polarizability, HOMO and LUMO for $120$ epochs, with ten seeds per method and $103{,}200$ optimizer updates per run (Section H). We compare against uncoupled NIW and the coupled NIW parameterization of Meinert & Lavin (2021), which enforces $\nu_W = r\kappa$. Mean pointwise $\kappa$ CV is $3.07\%$ for ours, $27.85\%$ for uncoupled NIW and $0.82\%$ for coupled NIW. The $9.06$-fold amplitude improvement thus applies to the uncoupled baseline. For the epistemic covariance trace, ours has lower CV than both baselines: $27.58\%$, compared with $44.92\%$ and $31.78\%$, respectively. Mean NLL is $-4.689$, $-4.718$ and $-4.681$ in the same order. Coupling yields the most stable evidence strength, while our method yields the most stable epistemic component at the final checkpoint, with similar predictive fit.

**Stability and prediction on UCI-9.** Across five seeds within each standard dataset split, our two-stage global head has lower average pointwise $\nu$ CV than unregularized DER, regularized DER, and DER+MAE on all nine datasets (Table 15). Prediction and calibration vary by dataset. Relative to unregularized DER, our mean NLL is lower on naval but higher by $0.202$ nats on yacht; the full NLL and coverage comparisons are in Section F. The separate Deep Ensemble and MC-Dropout comparison in Section D.9 assesses predictive quality and semantic shift.

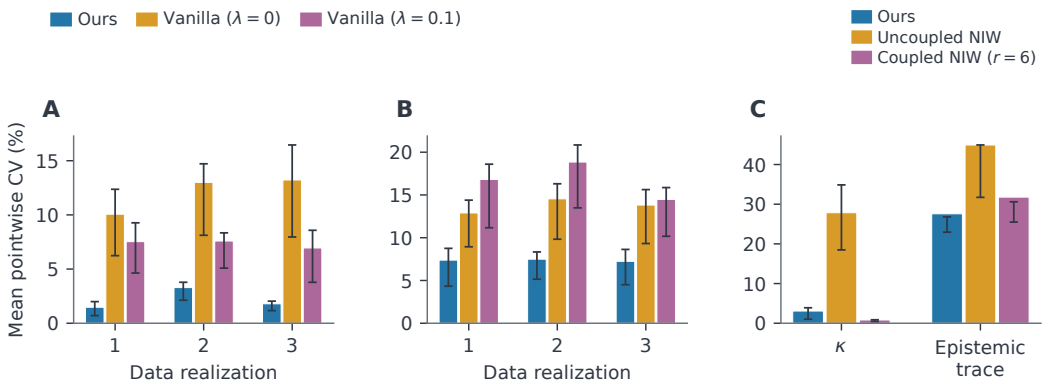


Figure 2: Cross-seed stability on CIFAR-rotation and QM9. Bars show mean pointwise CV across ten training seeds; error bars are 95% percentile bootstrap intervals from 5000 seed resamples, with test inputs held fixed. (A) Evidence strength $\nu$ and (B) epistemic variance on CIFAR-rotation, for three data realizations after 1000 epochs. (C) Evidence strength $\kappa$ and epistemic covariance trace on QM9 at 120 epochs, comparing ours with coupled and uncoupled NIW. The coupled baseline enforces $\nu_W = r\kappa$; both NIW baselines use NLL alone. Global amplitudes are evaluated as constant functions over the test inputs.

### 4.3 Scalar global-$\nu$ recovery

On fixed Gaussian features $\phi : \mathbb{R}^{10} \to \mathbb{R}^3$, we generate targets from the theorem-class $\beta^*(x) = \mathrm{softplus}(w^{*\top}\phi(x))$, with $\nu^* \in \{0.5, 1, 2, 5\}$. At $n = 8000$, across twenty data-and-training seeds, unregularized vanilla has mean pointwise $\nu$ CV 8.91–12.95 times that of our global head. Lower dispersion does not imply accurate recovery: at $\nu^* = 5$, our mean estimate remains approximately 1.18. The regularized vanilla estimates concentrate near zero, with means between 0.017 and 0.036. Both vanilla variants have slightly lower mean test NLL than ours in these settings. Per-cell results and the known-nuisance recovery diagnostic are in Section D.1.

### 4.4 Per-sample recovery (Theorem 3.3)

We generate data from softplus-affine ground-truth heads on a fixed Gaussian feature map and compare a per-sample head with a global-$\nu$ head that retains an affine softplus $\beta$. The generating heads are both nonconstant, have constant $\nu$ only, or are both constant. The fitted per-sample model learns both slope vectors in all three regimes. The global baseline fixes only the $\nu$ slope at zero. Both models learn the intercepts and start with $\nu(x) = 1$. Each regime uses twenty seeds, $n = 8000$, and 2000 Adam epochs.

Table 3: Synthetic recovery with $\nu(x) = 1$ initially: 20 seeds, $n = 8000$, 2000 Adam epochs. Mean $\pm$ sample SD; error columns are relative $L^2$. Both models use an affine softplus $\beta$ head; the global baseline fixes the $\nu$-head slope to zero and learns its intercept.

| Truth | Head | Rel. error $\nu$ | Rel. error $c(\nu)\beta$ | Rel. error $(u, d)$ | Rel. error $(w, b)$ | NLL |
|---|---|---|---|---|---|---|
| Nonconstant | Per-sample | $0.173 \pm 0.050$ | $0.415 \pm 0.199$ | $0.258 \pm 0.081$ | $0.435 \pm 0.149$ | $2.117 \pm 0.024$ |
| Nonconstant | Global $\nu$ | $0.856 \pm 0.014$ | $0.905 \pm 0.021$ | — | — | $2.128 \pm 0.026$ |
| Constant $\nu$ | Per-sample | $0.701 \pm 0.183$ | $0.315 \pm 0.146$ | $1.883 \pm 0.425$ | $0.602 \pm 0.141$ | $2.007 \pm 0.025$ |
| Constant $\nu$ | Global $\nu$ | $0.278 \pm 0.137$ | $0.184 \pm 0.088$ | — | — | $2.007 \pm 0.023$ |
| Both constant | Per-sample | $0.568 \pm 0.164$ | $0.247 \pm 0.074$ | $1.531 \pm 0.340$ | $1.331 \pm 0.235$ | $2.052 \pm 0.022$ |
| Both constant | Global $\nu$ | $0.286 \pm 0.092$ | $0.260 \pm 0.066$ | — | — | $2.055 \pm 0.029$ |

For nonconstant ground truth, the per-sample head has $\nu$ relative $L^2$ error 0.173, compared with 0.856 for the misspecified global head; the corresponding product errors are 0.415 and 0.905. Recovery

improves with sample size in this regime (Section D.2). At $n = 32000$, continuing the same five per-sample fits from 2000 to 4000 epochs reduces mean $\nu$ error from $0.147$ to $0.093$ (Section D.3).

When the true $\nu$ is constant, the correctly specified global head has lower error than the per-sample head. The latter remains identifiable under the theorem, but its finite-budget recovery is slower and is not monotone across the tested sample sizes. When both true heads are constant, the predictive identifies only their product, so a common initialization does not make the underlying representation unique.

**Interpretation.** The real-data comparisons show improved evidence-strength stability on CIFAR-rotation and UCI-9, and improved final epistemic-component stability against both NIW baselines on QM9. The synthetic results distinguish this reproducibility from parameter recovery. Identifiability alone does not guarantee finite-sample optimization or faithful epistemic uncertainty. The fitted-slope experiment in Section D.6 is retained as a separate diagnostic of the theoretical scaling direction.

## 5 Conclusion

The identifiability of Deep Evidential Regression depends on the structure of its parameter class. For a global scalar $\nu$, scale-separation of the $\beta$-class is necessary and sufficient for predictive injectivity (Theorems 3.1 and 3.2), ensuring that the scaling orbit contains a unique member of that class. For per-sample softplus-affine heads on fixed features satisfying the stated pushforward-density condition, the predictive product $c(\nu(x))\beta(x)$ uniquely determines $(u, d, w, b)$ unless $(u, w) = (0, 0)$ (Theorem 3.3). The case $u = 0, w \neq 0$ recovers the global-$\nu$ result within this subclass. A degree-1 positive-homogeneous gauge on $\mathrm{SPD}(d)$ extends both results to multivariate DER (Theorems 3.4 and 3.5). These results resolve a predictive representation ambiguity without equating identifiability with Bayesian or epistemic faithfulness.

From a common evidence initialization, the complete method has lower mean pointwise cross-seed amplitude CV than vanilla DER on CIFAR-rotation and UCI-9, and than uncoupled NIW on QM9. Coupled NIW has the lowest $\kappa$ CV on QM9, while our method has the lowest epistemic covariance-trace CV at the final checkpoint. The epistemic component is also less dispersed in the CIFAR comparisons. Predictive performance remains a separate consideration: these stability gains accompany small NLL differences on CIFAR and QM9, and larger task-dependent differences on UCI. Synthetic recovery experiments show why uniqueness and reproducibility should be distinguished from estimation accuracy: a concentrated global estimate can remain biased, and identifiable per-sample heads can require more data or optimization to recover their parameters.

**Limitations and open questions.** The fixed-feature analysis requires a density condition that may fail when learned features concentrate on a low-dimensional manifold. The joint-training QM9 result is empirical evidence beyond these assumptions, rather than an identifiability guarantee for learned representations. The stability comparisons vary network initialization and training randomness while holding the initial evidence strength fixed; they do not establish robustness to arbitrary initial evidence values. Finite-sample recovery rates, the effect of local conditioning, and identifiability under joint feature learning remain open questions.

## AI use statement

Generative AI tools assisted with manuscript drafting and editing, literature searches, research planning and mathematical reasoning, code development and debugging, experiment execution and result analysis, figure preparation, and internal review of the manuscript. The authors take full responsibility for the final text, mathematical claims, code, figures, and reported results.

## References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, 2023. URL https://proceedings.mlr.press/v202/bengs23a.html.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, 2016.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1861–1869, 2015.

Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v235/juergens24a.html.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019.

Nis Meinert and Alexander Lavin. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*, 2021. URL https://arxiv.org/abs/2104.06135.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *AAAI*, pp. 9134–9142. AAAI Press, 2023.

Boris S. Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107(3):529–530, 2020. doi: 10.1134/S0001434620030189. arXiv:1512.07276 (2015).

Matthias Nagl, Maximilian Nagl, and Daniel Rösch. Quantifying uncertainty of machine learning methods for loss given default. *Frontiers in Applied Mathematics and Statistics*, 8:1076083, 2022. doi: 10.3389/fams.2022.1076083.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014. doi: 10.1038/sdata.2014.22.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Maohao Shen, J. Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory W. Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Ava P. Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N. Bhatia, and Connor W. Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367, 2021. doi: 10.1021/acscentsci.1c00546.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=xqS8k9E75c. arXiv:2110.03051.

Zhongyao Wang, Taoyong Cui, Jiawen Zou, Shufei Zhang, Bo Yan, Wanli Ouyang, Weimin Tan, and Mao Su. Equivariant evidential deep learning for interatomic potentials. *arXiv preprint arXiv:2602.10419*, 2026. URL https://arxiv.org/abs/2602.10419.

Wen Wu, Chao Zhang, and Philip C. Woodland. Estimating the uncertainty in emotion attributes using deep evidential regression. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15681–15695, 2023. URL https://aclanthology.org/2023.acl-long.873/.

Yuefei Wu, Bin Shi, Bo Dong, Qinghua Zheng, and Hua Wei. The evidence contraction issue in deep evidential regression: Discussion and solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Han Xu, Taoyong Cui, Chenyu Tang, Jinzhe Ma, Dongzhan Zhou, Yuqiang Li, Xiang Gao, Xingao Gong, Wanli Ouyang, Shufei Zhang, and Mao Su. Evidential deep learning for interatomic potentials. *Nature Communications*, 17:937, 2026. doi: 10.1038/s41467-025-67663-y.

Yaniv Yacoby, Weiwei Pan, and Finale Doshi-Velez. Mitigating the effects of non-identifiability on inference for Bayesian neural networks with latent variables. *Journal of Machine Learning Research*, 23(244), 2022.

Kai Ye, Tiejin Chen, Hua Wei, and Liang Zhan. Uncertainty regularized evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

# A SYMBOL MAP

| Symbol | | | Meaning |
|---|---|---|---|
| $\gamma(x), \nu, \alpha(x), \beta(x), (1+t)/t$ | $c(t)$ | $=$ | NIG head; predictive scale $\sigma^2(x) = \beta(x)c(\nu)/\alpha(x)$, df $2\alpha(x)$ |
| $\phi : \mathcal{X} \to \mathbb{R}^k$, $\mathcal{F}_\beta$ | | | fixed (Stage-2 frozen) feature map; admissible $\beta$-class |
| $(u, d, w, b)$ | | | per-sample softplus-affine head |
| $\kappa, \Psi(x), \nu_W$ | | | multivariate NIW head; $\Psi \in \mathrm{SPD}(d)$, $\nu_W > d-1$ |
| $g : \mathrm{SPD}(d) \to \mathbb{R}_+$ | | | degree-1 gauge; $\Psi = s \cdot S$ with $g(S) = 1$ |
| $P_X$-a.e. | | | equality up to $P_X$-null sets in $\mathcal{X}$ |
| $s(\cdot)$ vs $s$ | | | $s(r) := \mathrm{softplus}(r)$ is used in Theorem 3.3 and Lemma B.3; $s(x) = g(\Psi(x))$ in Section 3.2 denotes the scalar amplitude. |
| $\mathbb{R}_+$ | | | always denotes $(0, \infty)$ |

Table 4: Recurring symbols across scalar (Section 3.1) and multivariate (Section 3.2) sections.

# B FULL PROOFS: SCALAR IDENTIFIABILITY

**Proposition B.1** (Scaling symmetry, global $\nu$)**.** *Fix $\gamma$ and $\alpha$. For any $\nu, \nu' > 0$ and any positive measurable $\beta : \mathcal{X} \to \mathbb{R}_+$, define $\beta'(x) := \beta(x) \cdot c(\nu)/c(\nu')$ with $c(t) = (1+t)/t$. Then the Student-t predictive in Equation* (1) *is identical under $(\gamma, \nu, \alpha, \beta)$ and $(\gamma, \nu', \alpha, \beta')$ for every $x \in \mathcal{X}$, and the population score $\mathbb{E}_{(x,y)\sim Q}[S(p_\theta(\cdot \mid x), y)]$ is constant along the one-parameter orbit $\{(\nu', \beta') : \nu' > 0\}$ within the unconstrained class of positive measurable $\beta$-functions. Whether the orbit intersects a restricted class $\mathcal{F}_\beta$ at more than one point is governed by Definition 3.1.*

*Proof.* The Student-$t$ scale-squared in Equation (1) is $\sigma^2(x) = \beta(x)c(\nu)/\alpha(x)$. Under the substitution $\beta'(x) = \beta(x)c(\nu)/c(\nu')$, we have $\sigma'^2(x) = \beta'(x)c(\nu')/\alpha(x) = \beta(x)c(\nu)/\alpha(x) = \sigma^2(x)$. The location $\gamma$ and df $2\alpha$ are unchanged. Hence the predictive density is identical at every $x$; constancy of the population score follows from the score being a functional of the predictive alone. □

**Proof of Theorem 3.1.** Suppose $\mathcal{F}_\beta$ is not scale-separated with respect to $P_X$. Then there exist $\beta_1, \beta_2 \in \mathcal{F}_\beta$ and $a > 0$, $a \neq 1$, with $\beta_2 = a \cdot \beta_1$ $P_X$-a.e. Note that the roles of $(\beta_1, \beta_2, a)$ and $(\beta_2, \beta_1, 1/a)$ are symmetric, so we may assume without loss of generality $a > 1$. The map $c : \mathbb{R}_+ \to (1, \infty)$, $c(\nu) = (1+\nu)/\nu$, is a strictly decreasing continuous bijection with $\lim_{\nu \to 0^+} c(\nu) = \infty$ and $\lim_{\nu \to \infty} c(\nu) = 1$. Hence we can choose $\nu_1 > 0$ with $c(\nu_1) > a$, and then $c(\nu_1)/a \in (1, \infty)$ lies in the range of $c$, so there is a unique $\nu_2 > 0$ with $c(\nu_2) = c(\nu_1)/a$. By Proposition B.1, the parameter tuples $(\gamma, \nu_1, \alpha, \beta_1)$ and $(\gamma, \nu_2, \alpha, \beta_2)$ produce the same predictive $P_X$-a.e. A strictly proper scoring rule minimizes its expectation over distributions $P$ uniquely at the true conditional (Gneiting & Raftery, 2007); so if two distinct parameter tuples produce the same predictive, a strictly proper score cannot distinguish them. Hence $\nu$ is not identifiable. □

**Proof of Theorem 3.2.** The Student-$t$ density at $x$ with parameters $(\mu, \sigma^2, \nu_t)$ is

$$\mathrm{St}_t(y; \mu, \sigma^2, \nu_t) \;=\; \frac{\Gamma((\nu_t+1)/2)}{\sqrt{\nu_t \pi \sigma^2}\,\Gamma(\nu_t/2)} \Big(1 + \tfrac{(y-\mu)^2}{\nu_t \sigma^2}\Big)^{-(\nu_t+1)/2}.$$

By Lemma C.1 specialized to $d = 1$ (with $\Lambda = \sigma^2$), the map $(\mu, \sigma^2, \nu_t) \mapsto \mathrm{St}_t(\cdot; \mu, \sigma^2, \nu_t)$ is injective on $\mathbb{R} \times \mathbb{R}_+ \times (0, \infty)$. Since the predictive at $x$ is $\mathrm{St}_t(\cdot; \gamma(x), \beta(x)c(\nu)/\alpha(x), 2\alpha(x))$, the triple $(\gamma(x), \beta(x)c(\nu)/\alpha(x), 2\alpha(x))$ is determined $P_X$-a.e. by the predictive. From the third component we read off $\alpha(x)$ $P_X$-a.e.; from the first $\gamma(x)$; and from the second the product $\beta(x)c(\nu)$.

We now claim the product $\beta c(\nu)$ determines $\beta$ and $\nu$ *separately*. Suppose two tuples $(\beta_1, \nu_1)$ and $(\beta_2, \nu_2)$ satisfy $\beta_1 c(\nu_1) = \beta_2 c(\nu_2)$ $P_X$-a.e. Then $\beta_2 = \beta_1 \cdot c(\nu_1)/c(\nu_2)$ $P_X$-a.e.; let $a = c(\nu_1)/c(\nu_2) > 0$. Then $\beta_2 = a\beta_1$ $P_X$-a.e. Scale-separation of $\mathcal{F}_\beta$ forces $a = 1$ and $\beta_1 = \beta_2$ $P_X$-a.e. $a = 1$ gives $c(\nu_1) = c(\nu_2)$, which by strict monotonicity and injectivity of $c$ on $\mathbb{R}_+$ forces $\nu_1 = \nu_2$. □

**Corollary B.2** (Implementable softplus head)**.** *Fix a measurable feature map $\phi : \mathcal{X} \to \mathbb{R}^k$. Suppose the pushforward $P_X \circ \phi^{-1}$ admits a density (with respect to Lebesgue measure on $\mathbb{R}^k$) that is positive almost everywhere on some non-empty open set $U \subseteq \mathbb{R}^k$. Then the class $\mathcal{F}_\beta = \{x \mapsto \mathrm{softplus}(w^\top \phi(x)) : w \in \mathbb{R}^k\}$ is scale-separated with respect to $P_X$.*

*Proof.* Suppose $\mathrm{softplus}(w'^\top \phi(x)) = a\ \mathrm{softplus}(w^\top \phi(x))$ $P_X$-a.e., with $a > 0$. Define

$$F(z) \;=\; \mathrm{softplus}(w'^\top z) - a\ \mathrm{softplus}(w^\top z), \qquad z \in \mathbb{R}^k.$$

$F$ is real-analytic on $\mathbb{R}^k$ (as a finite combination of real-analytic functions). Let $U \subseteq \mathbb{R}^k$ be the non-empty open set on which the pushforward $P_X \circ \phi^{-1}$ has positive density a.e.; let $\mu := P_X \circ \phi^{-1}$ and let $Z := \{z \in \mathbb{R}^k : F(z) = 0\}$. The hypothesis is $\mu(\mathbb{R}^k \setminus Z) = 0$; combined with $\mu(U) > 0$ and the positive density of $\mu$ a.e. on $U$, this forces $\mathrm{Leb}(U \setminus Z) = 0$, i.e. $Z$ contains a Lebesgue-positive-measure subset of $U$. A nontrivial real-analytic function on a connected open subset of $\mathbb{R}^k$ has zero set of Lebesgue measure zero (Mityagin, 2020); since $Z$ has positive Lebesgue measure, $F$ must be identically zero on the connected domain $\mathbb{R}^k$.

Evaluating at $z = 0$: $\mathrm{softplus}(0) - a \cdot \mathrm{softplus}(0) = \log 2\,(1 - a) = 0$, so $a = 1$. Hence $\mathrm{softplus}(w'^\top z) = \mathrm{softplus}(w^\top z)$ for all $z \in \mathbb{R}^k$. Since softplus is strictly increasing (hence injective), $w'^\top z = w^\top z$ for all $z \in \mathbb{R}^k$, so $w' = w$. □

**Proof of Theorem 3.3 (per-sample softplus-affine identifiability).** The proof proceeds in three steps. The first two reuse the machinery already in this appendix; the third is the new contribution.

*Step 1: reduction to the product $c(\nu(x))\beta(x)$.* Univariate Student-$t$ density identifiability gives $\gamma_1(x) = \gamma_2(x)$, $\alpha_1(x) = \alpha_2(x)$, and $c(\nu_1(x))\beta_1(x) = c(\nu_2(x))\beta_2(x)$ $P_X$-a.e. (same step as in the proof of Theorem 3.2; the only change is that $\nu$ is now a measurable function rather than a global scalar).

*Step 2: real-analytic continuation from $P_X$-a.e. to all $z \in \mathbb{R}^k$.* Define

$$Q_{u,d,w,b}(z) \;:=\; \mathrm{softplus}(w^\top z + b) \cdot c\big(\mathrm{softplus}(u^\top z + d)\big), \qquad z \in \mathbb{R}^k.$$

$Q$ is real-analytic on $\mathbb{R}^k$ because $\mathrm{softplus}(r) = \log(1 + e^r)$ is real-analytic on $\mathbb{R}$, $\mathrm{softplus}(r) > 0$, and $c(t) = 1 + 1/t$ is analytic on $(0, \infty)$. The hypothesis $c(\nu_{u_1,d_1}(x))\beta_{w_1,b_1}(x) = c(\nu_{u_2,d_2}(x))\beta_{w_2,b_2}(x)$ $P_X$-a.e. implies $(Q_1 - Q_2)(\phi(x)) = 0$ $P_X$-a.e.; by the same density-positive-a.e.-on-$U$ argument as in the proof of Corollary B.2, the zero set of $Q_1 - Q_2$ has positive Lebesgue measure. A nontrivial real-analytic function on a connected open subset of $\mathbb{R}^k$ has zero set of Lebesgue measure zero (Mityagin, 2020), so $Q_1(z) = Q_2(z)$ for every $z \in \mathbb{R}^k$.

*Step 3: the softplus-affine product factorization lemma (new).*

**Lemma B.3** (Softplus-affine product factorization on $\mathbb{R}$)**.** *Let*

$$q_{a,b,r,d}(t) := \mathrm{softplus}(at + b)\, c\big(\mathrm{softplus}(rt + d)\big), \qquad t \in \mathbb{R},$$

*where* $\mathrm{softplus}(x) = \log(1 + e^x)$ *and* $c(s) = 1 + 1/s$ *on* $(0, \infty)$*. If* $(a, r) \neq (0, 0)$*, then the map* $(a, b, r, d) \mapsto q_{a,b,r,d}(\cdot)$ *is injective as a map into real-valued functions on* $\mathbb{R}$*, and every non-degenerate quadruple yields a non-constant function. If* $a = r = 0$*, then* $q_{0,b,0,d}(t) \equiv \mathrm{softplus}(b)\, c(\mathrm{softplus}(d))$*, so the only identifiable functional is the scalar value* $\mathrm{softplus}(b)c(\mathrm{softplus}(d))$*, and for each fixed* $C > 0$ *the level set* $\{(b, d) \in \mathbb{R}^2 : \mathrm{softplus}(b)c(\mathrm{softplus}(d)) = C\}$ *is the one-parameter curve* $\{(\mathrm{softplus}^{-1}(C/c(\mathrm{softplus}(d))), d) : d \in \mathbb{R}\}$.

*Proof.* Write $s(x) := \mathrm{softplus}(x)$ and $h(x) := c(s(x)) = 1 + 1/s(x)$. Note $s : \mathbb{R} \to (0, \infty)$ is strictly increasing and $h : \mathbb{R} \to (1, \infty)$ is strictly decreasing; both are injective with inverses $s^{-1}(y) = \log(e^y - 1)$ and $h^{-1}(K) = s^{-1}(1/(K - 1))$ for $K > 1$.

*Tail expansions.* As $X \to +\infty$,

$$s(X) = X + O(e^{-X}), \qquad h(X) = 1 + \tfrac{1}{X} + O(e^{-X}X^{-2}). \tag{2}$$

As $X \to -\infty$,

$$s(X) = e^X(1 + O(e^X)), \qquad h(X) = e^{-X}\big(1 + \tfrac{3}{2}e^X + O(e^{2X})\big). \tag{3}$$

These follow from the Taylor expansion of $\log(1+u)$ at $u=0$. All $O(|t|^{-2})$ remainders below are bounded and integrable on the corresponding tail; all exponential remainders are likewise bounded and integrable. The non-integrable $1/t$ corrections are always displayed explicitly.

We now give identifying limits in all sign cases. Throughout, $A(t) := at + b$ and $R(t) := rt + d$.

**Case 1:** $a > 0, r > 0$**.** As $t \to +\infty$, both $A(t), R(t) \to +\infty$, so by (2),

$$q_{a,b,r,d}(t) = (at+b)\big(1 + \tfrac{1}{rt+d}\big) + O(t^{-2}) = at + \big(b + \tfrac{a}{r}\big) + \tfrac{br-ad}{r^2 t} + O(t^{-2}).$$

As $t \to -\infty$, both $A(t), R(t) \to -\infty$, so by (3),

$$q_{a,b,r,d}(t) = e^{(a-r)t+b-d}\big(1 + O(e^{at}) + O(e^{rt})\big).$$

Define

$$\alpha := \lim_{t\to+\infty} \tfrac{q(t)}{t} = a, \quad \delta := \lim_{t\to-\infty} \tfrac{\log q(t)}{t} = a - r, \quad \rho := \alpha - \delta = r,$$

$$B_0 := \lim_{t\to+\infty} \big(q(t) - \alpha t\big) = b + \tfrac{\alpha}{\rho}, \quad b = B_0 - \tfrac{\alpha}{\rho},$$

$$B_1 := \lim_{t\to+\infty} t\big(q(t) - \alpha t - B_0\big) = \tfrac{b\rho - \alpha d}{\rho^2}, \quad d = \tfrac{b\rho - B_1\rho^2}{\alpha}.$$

*Slope-cancellation subcase* $a = r > 0$. Then $\delta = 0$, so $\rho = \alpha = a$ and no division by $a - r$ is required; $B_1 = (b-d)/a$ and $d = b - aB_1$.

**Case 2:** $a > 0, r = 0$**.** Put $K := h(d) > 1$. Then $q_{a,b,0,d}(t) = K\, s(at+b)$, so

$$q(t) = K(at+b) + O(e^{-at}) \ (t \to +\infty), \qquad q(t) = Ke^{at+b}(1 + O(e^{at})) \ (t \to -\infty).$$

Define $\alpha := \lim_{t\to-\infty} \log q(t)/t = a$ and $M := \lim_{t\to+\infty} q(t)/t = Ka$, giving $K = M/\alpha > 1$ and $d = h^{-1}(K)$. Then $L := \lim_{t\to-\infty} e^{-\alpha t} q(t) = Ke^b$, so $b = \log(L/K)$.

**Case 3:** $a > 0, r < 0$**.** As $t \to +\infty$, $A(t) \to +\infty$, $R(t) \to -\infty$, so $q(t) = e^{-rt-d}(at + b)(1 + O(e^{rt}) + O(e^{-at}/t))$. As $t \to -\infty$, $A(t) \to -\infty$, $R(t) \to +\infty$, so $q(t) = e^{at+b}\big(1 + \frac{1}{rt+d} + O(t^{-2}) + O(e^{at})\big)$. Define $\alpha := \lim_{t\to-\infty} \log q(t)/t = a$ and $b = \log(\lim_{t\to-\infty} e^{-\alpha t} q(t))$; $\eta := \lim_{t\to+\infty} \log q(t)/t = -r > 0$, so $r = -\eta$; and $L := \lim_{t\to+\infty} q(t)/(te^{-rt}) = ae^{-d}$, giving $d = \log a - \log L$.

**Case 4:** $a = 0, r > 0$**.** Put $B := s(b) > 0$. Then $q_{0,b,r,d}(t) = B\, h(rt+d)$, so $q(t) = B + \frac{B}{rt+d} + O(t^{-2})$ as $t \to +\infty$ and $q(t) = Be^{-rt-d}\big(1 + \frac{3}{2}e^{rt+d} + O(e^{2rt})\big)$ as $t \to -\infty$. Identify: $B = \lim_{t\to+\infty} q(t)$, $b = s^{-1}(B)$; $\rho := -\lim_{t\to-\infty} \log q(t)/t = r$; $L := \lim_{t\to-\infty} e^{\rho t} q(t) = Be^{-d}$, so $d = \log B - \log L$.

**Case 5:** $a = 0, r < 0$**.** Put $B := s(b) > 0$. Symmetric to Case 4 with the tails swapped: $q(t) = Be^{-rt-d}\big(1 + \frac{3}{2}e^{rt+d} + O(e^{2rt})\big)$ as $t \to +\infty$ and $q(t) = B + \frac{B}{rt+d} + O(t^{-2})$ as $t \to -\infty$. Identify: $B = \lim_{t\to-\infty} q(t)$, $b = s^{-1}(B)$; $\eta := \lim_{t\to+\infty} \log q(t)/t = -r > 0$, so $r = -\eta$; $L := \lim_{t\to+\infty} e^{rt} q(t) = Be^{-d}$, so $d = \log B - \log L$.

**Case 6:** $a < 0, r > 0$**.** As $t \to +\infty$, $A(t) \to -\infty$, $R(t) \to +\infty$, so $q(t) = e^{at+b}(1 + \frac{1}{rt+d} + O(t^{-2}) + O(e^{at}))$. As $t \to -\infty$, $A(t) \to +\infty$, $R(t) \to -\infty$, so $q(t) = e^{-rt-d}(at + b)(1 + O(e^{rt}) + O(e^{-at}/|t|))$. Identify: $\alpha := \lim_{t\to+\infty} \log q(t)/t = a < 0$, $b = \log(\lim_{t\to+\infty} e^{-\alpha t} q(t))$; $\rho := -\lim_{t\to-\infty} \log q(t)/t = r > 0$; and using $at + b \sim (-a)(-t)$ as $t \to -\infty$, $L := \lim_{t\to-\infty} q(t)/((-t)e^{-\rho t}) = (-a)e^{-d}$, so $d = \log(-a) - \log L$.

**Case 7:** $a < 0, r = 0$**.** Mirror of Case 2 with the linear/exponential tails reflected. Put $K := h(d) > 1$; $q_{a,b,0,d}(t) = K\, s(at+b)$. So $q(t) = Ke^{at+b}(1 + O(e^{at}))$ as $t \to +\infty$ and $q(t) = K(at+b) + O(e^{-at})$ as $t \to -\infty$. Define $\alpha := \lim_{t\to+\infty} \log q(t)/t = a$ (negative) and $M := \lim_{t\to-\infty} q(t)/t = Ka$ (also negative). Then $K = M/\alpha > 1$, $d = h^{-1}(K)$, and $L := \lim_{t\to+\infty} e^{-\alpha t} q(t) = Ke^b$, so $b = \log(L/K)$.

**Case 8:** $a < 0, r < 0$**.** Mirror of Case 1 with the linear tail at $-\infty$. As $t \to -\infty$, both $A(t), R(t) \to +\infty$, so $q(t) = at + (b + a/r) + \frac{br-ad}{r^2 t} + O(t^{-2})$. As $t \to +\infty$, both $A(t), R(t) \to -\infty$, so $q(t) =$

$e^{(a-r)t+b-d}(1+O(e^{at})+O(e^{rt}))$. Define $\alpha := \lim_{t\to-\infty} q(t)/t = a$, $\delta := \lim_{t\to+\infty} \log q(t)/t = a - r$, $\rho := \alpha - \delta = r$. Then $B_0 := \lim_{t\to-\infty}(q(t) - \alpha t) = b + \alpha/\rho$ gives $b$, and $B_1 := \lim_{t\to-\infty} t(q(t) - \alpha t - B_0) = (b\rho - \alpha d)/\rho^2$ gives $d$. The cancellation subcase $a = r < 0$ is handled by $\delta = 0$, $\rho = \alpha$, $d = b - \alpha B_1$.

**Row signatures and disjointness.** For each non-degenerate sign pattern $(\operatorname{sgn} a, \operatorname{sgn} r)$, the leading-order asymptotic type of $q(t)$ at $t \to +\infty$ and $t \to -\infty$ is summarized below. Notation: $P_1$ = polynomial of degree 1 in $t$ (positive linear leading); $C$ = positive constant; $\mathrm{EXP}_g(\rho)$ = exponential growth $\sim e^{\rho|t|}$ at the corresponding tail with rate $\rho > 0$; $\mathrm{EXP}_d(\rho)$ = exponential decay $\sim e^{-\rho|t|}$; $P_1 \cdot \mathrm{EXP}_g(\rho)$ = polynomial-times-exponential growth.

| Case | sgn$(a, r)$ | $t \to +\infty$ tail | $t \to -\infty$ tail |
|---|---|---|---|
| 1 | $(+,+)$ | $P_1$ slope $a$ | $\mathrm{EXP}_g(r-a)$ if $a<r$; $C$ if $a=r$; $\mathrm{EXP}_d(a-r)$ if $a>r$ |
| 2 | $(+,0)$ | $P_1$ slope $Ka$, $K>1$ | $\mathrm{EXP}_d(a)$ |
| 3 | $(+,-)$ | $P_1 \cdot \mathrm{EXP}_g(\lvert r\rvert)$ | $\mathrm{EXP}_d(a)$ |
| 4 | $(0,+)$ | $C=B>0$ | $\mathrm{EXP}_g(r)$ |
| 5 | $(0,-)$ | $\mathrm{EXP}_g(\lvert r\rvert)$ | $C=B>0$ |
| 6 | $(-,+)$ | $\mathrm{EXP}_d(\lvert a\rvert)$ | $P_1 \cdot \mathrm{EXP}_g(r)$ |
| 7 | $(-,0)$ | $\mathrm{EXP}_d(\lvert a\rvert)$ | $P_1$ slope $Ka<0$, $K>1$ |
| 8 | $(-,-)$ | $\mathrm{EXP}_g(a-r)$ if $\lvert a\rvert<\lvert r\rvert$; $C$ if $a=r$; $\mathrm{EXP}_d(r-a)$ if $\lvert a\rvert>\lvert r\rvert$ | $P_1$ slope $a$ |

(Sign check, Case 1 with $a < r$: $q(t) \sim e^{(a-r)t}$ as $t \to -\infty$; with $a - r < 0$ and $t \to -\infty$, the exponent $(a-r)t \to +\infty$, so $q$ grows; rate is $|a - r| = r - a > 0$. With $a > r$: $(a-r)t \to -\infty$, so $q$ decays at rate $|a - r| = a - r > 0$. Case 8 with $|a| < |r|$ (i.e., $a > r$ since both negative): as $t \to +\infty$, $(a-r)t \to +\infty$, $q$ grows at rate $a - r > 0$; with $|a| > |r|$ (i.e., $a < r$), $(a-r)t \to -\infty$, $q$ decays at rate $|a - r| = r - a > 0$. The rates $\rho$ in $\mathrm{EXP}_g(\rho)$ and $\mathrm{EXP}_d(\rho)$ are always positive in the table, by construction.)

The $(+\infty, -\infty)$-tail-type pair distinguishes Cases 3, 4, 5, 6 from each other and from all others by inspection of the table. Cases 1 vs. 2 share signature ($P_1$ on $+\infty$, exponential decay on $-\infty$); Cases 7 vs. 8 likewise. All other pairs differ in either $+\infty$ or $-\infty$ type. Furthermore, every nondegenerate quadruple yields a nonconstant function $q$ (each row in the table has at least one tail with linear or exponential growth/decay), so no nondegenerate quadruple can produce the same function as the constant degenerate family $\{q_{0,b,0,d} \equiv \mathrm{softplus}(b)c(\mathrm{softplus}(d))\}$. We work out the Case 1 vs. Case 2 overlap below; the reflected Case 7 vs. Case 8 follows by the substitution $t \mapsto -t$, which sends $q_{a,b,r,d}(t)$ to $q_{-a,b,-r,d}(t)$ and swaps the two tails, so any Case-7-vs-Case-8 collision pulls back to a Case-2-vs-Case-1 collision under sign flips of $(a, r)$, contradicting the same algebraic identity below. The Case 2 vs. Case 1 overlap: a function with a Case-2 representation $(a_0 > 0, r_0 = 0)$ matching a Case-1 representation $(a > 0, r > 0)$, i.e. $K\, s(a_0 t + b_0) \equiv s(at + b)\, h(rt + d)$ with $K = h(d_0) > 1$. Matching $\pm\infty$ leading orders forces $0 < r < a$, $a_0 = a - r$, and $K = a/(a - r) > 1$. The $-\infty$ leading coefficient gives $Ke^{b_0} = e^{b-d}$. At $+\infty$, the $r_0 = 0$ representation has no algebraic $1/t$ term: $K\, s(a_0 t + b_0) = at + Kb_0 + O(e^{-a_0 t})$. Comparing with the Case-1 expansion gives $Kb_0 = b + a/r$ and $br - ad = 0$, so $d = br/a$. Substituting into $b_0 + \log K = b - d$ then yields $b_0 = b(1 - r/a) - \log K$, hence $Kb_0 = b - K \log K$. The constant-term identity $Kb_0 = b + a/r$ would therefore require $-K \log K = a/r$, impossible because $K > 1$ makes the left side negative while $a/r > 0$. The reflection $t \mapsto -t$ rules out the only other potential overlap (Case 7 vs Case 8). Therefore the eight non-degenerate rows are mutually disjoint, and within each row the explicit functionals above pin $(a, b, r, d)$ uniquely.

**Degenerate $a = r = 0$ case.** Then $q_{0,b,0,d}(t) \equiv s(b)\, h(d) = \mathrm{softplus}(b)\, c(\mathrm{softplus}(d))$, constant in $t$. For fixed $C > 0$, since $h(d) > 1$ for every $d \in \mathbb{R}$, the equation $s(b)h(d) = C$ is equivalent to $s(b) = C/h(d) \in (0, \infty)$, hence $b = s^{-1}(C/h(d)) = \mathrm{softplus}^{-1}(C/c(\mathrm{softplus}(d)))$ for any $d \in \mathbb{R}$. The level set is precisely the one-parameter curve in the lemma statement. □

**Step 3 (continued): from line restrictions to the multivariate $(u, d, w, b)$.** Step 2 already gave $Q_1 \equiv Q_2$ on all of $\mathbb{R}^k$. Fix $z_0 \in \mathbb{R}^k$ and a direction $v \in \mathbb{R}^k$; restricting the identity to the line $z = z_0 + tv$ gives, with $a_i := w_i^\top v$, $b_i^v := w_i^\top z_0 + b_i$, $r_i := u_i^\top v$, $d_i^v := u_i^\top z_0 + d_i$, the univariate identity $\mathrm{softplus}(a_1 t + b_1^v)\, c(\mathrm{softplus}(r_1 t + d_1^v)) = \mathrm{softplus}(a_2 t + b_2^v)\, c(\mathrm{softplus}(r_2 t + d_2^v))$ for all $t \in \mathbb{R}$.

Define the linear subspaces $E_i := \{v \in \mathbb{R}^k : w_i^\top v = 0 \text{ and } u_i^\top v = 0\}$, $i = 1, 2$. Because $(u_i, w_i) \neq (0, 0)$, each $E_i$ is a proper linear subspace; therefore $D := \mathbb{R}^k \setminus (E_1 \cup E_2)$ is dense in $\mathbb{R}^k$. For every $v \in D$, both line restrictions are non-degenerate, so Lemma B.3 applies and yields $a_1 = a_2$ (i.e. $w_1^\top v = w_2^\top v$), $r_1 = r_2$ (i.e. $u_1^\top v = u_2^\top v$), $b_1^v = b_2^v$, and $d_1^v = d_2^v$ for every $z_0 \in \mathbb{R}^k$ and every $v \in D$. Since $D$ is dense and $v \mapsto w_i^\top v$, $v \mapsto u_i^\top v$ are continuous linear functionals on $\mathbb{R}^k$, the first two equalities extend to every $v \in \mathbb{R}^k$, hence $w_1 = w_2 =: w$ and $u_1 = u_2 =: u$. Taking any $v \in D$ and $z_0 = 0$ in the intercept equalities then gives $b_1 = b_2$ and $d_1 = d_2$. This proves Theorem 3.3. □

**Corollary B.4** (Biased nonconstant $\beta$-head)**.** *Set $u = 0$ in Theorem 3.3. Then $\nu_{0,d}(x) \equiv \mathrm{softplus}(d)$ is a global scalar, and the pair $(\mathrm{softplus}(d), \beta_{w,b})$ is identifiable iff $w \neq 0$.*

*Proof.* Set $u = 0$ in Theorem 3.3. The hypothesis $(u, w) \neq (0, 0)$ becomes $w \neq 0$. The conclusion $w_1 = w_2$, $b_1 = b_2$, $d_1 = d_2$ specializes to: a biased softplus-linear $\beta$-head with $w \neq 0$ and a global-$\nu_{0,d} = \mathrm{softplus}(d)$ scalar are jointly identifiable. The constant subfamily $w = 0$ remains the sharp obstruction. □

## C Full proofs: multivariate gauge

**Lemma C.1** (Multivariate Student-$t$ identifiability)**.** *The map $(\mu, \Lambda, \nu_t) \mapsto \text{mv-St}_t(\,\cdot\,; \mu, \Lambda, \nu_t)$ is injective on $\mathbb{R}^d \times \mathrm{SPD}(d) \times (0, \infty)$.*

*Proof.* The mv Student-$t$ density is

$$p(y; \mu, \Lambda, \nu_t) = \frac{\Gamma((\nu_t + d)/2)}{\Gamma(\nu_t/2)(\nu_t \pi)^{d/2} |\Lambda|^{1/2}} \Big(1 + \tfrac{(y-\mu)^\top \Lambda^{-1}(y-\mu)}{\nu_t}\Big)^{-(\nu_t + d)/2}.$$

We give a density-level argument that is valid for all $\nu_t \in (0, \infty)$ (and does not require existence of first or second moments). Suppose two parameter triples $(\mu_i, \Lambda_i, \nu_{t,i})$, $i = 1, 2$, induce identical densities $p_1 \equiv p_2$ on $\mathbb{R}^d$.

*Step 1: $\mu$ is identified.* For each fixed $(\Lambda, \nu_t)$, the density $y \mapsto p(y; \mu, \Lambda, \nu_t)$ is strictly decreasing in the Mahalanobis distance $(y - \mu)^\top \Lambda^{-1}(y - \mu)$ and attains its unique global maximum at $y = \mu$. Hence $p_1 \equiv p_2$ forces $\mu_1 = \mu_2 =: \mu$.

*Step 2: $\Lambda$ is identified up to a positive scalar.* Define $Q_i(z) := z^\top \Lambda_i^{-1} z$ for $z \in \mathbb{R}^d$. Because $p_i$ is a strictly decreasing function of $Q_i(y - \mu)$, each density level set $\{y : p_i(y) = \tau\}$ equals the ellipsoid $\{y : Q_i(y - \mu) = r_i(\tau)^2\}$ for a unique $r_i(\tau) > 0$ determined by $(\Lambda_i, \nu_{t,i})$, provided $0 < \tau < p_i(\mu)$ (the unique modal value, attained only at $y = \mu$, must be excluded so the radius is strictly positive). Pick any sub-modal level $\tau_0 \in (0, \min\{p_1(\mu), p_2(\mu)\})$; both densities take this value somewhere because $p_i$ is continuous and decays to 0 at infinity, so $r_1(\tau_0), r_2(\tau_0) > 0$ are well-defined. From $p_1 \equiv p_2$ we then have $\{y : Q_1(y - \mu) = r_1(\tau_0)^2\} = \{y : Q_2(y - \mu) = r_2(\tau_0)^2\}$. The following elementary fact then applies.

**Lemma C.2** (Proportional quadratic forms from equal level sets)**.** *Let $Q_1, Q_2 : \mathbb{R}^d \to \mathbb{R}$ be two positive definite quadratic forms. If there exist $r_1, r_2 > 0$ with $\{z : Q_1(z) = r_1^2\} = \{z : Q_2(z) = r_2^2\}$, then $Q_2 = (r_2/r_1)^2 \, Q_1$.*

*Proof of Lemma C.2.* Any nonzero $z \in \mathbb{R}^d$ scales to the first level set at $z_1 := r_1 z / \sqrt{Q_1(z)}$; by hypothesis $Q_2(z_1) = r_2^2$. Substituting and using homogeneity $Q_i(\lambda z) = \lambda^2 Q_i(z)$, $Q_2(z) = (r_2/r_1)^2 \, Q_1(z)$. At $z = 0$ both sides vanish. □

Apply Lemma C.2 with $Q_i(z) = z^\top \Lambda_i^{-1} z$: $\Lambda_2^{-1} = (r_2(\tau_0)/r_1(\tau_0))^2 \, \Lambda_1^{-1}$, i.e. $\Lambda_2 = \rho \, \Lambda_1$ with $\rho = (r_1(\tau_0)/r_2(\tau_0))^2 > 0$.

*Step 3: $\nu_t$ is identified from tail decay.* Choose any $u \in \mathbb{R}^d$ with $u^\top \Lambda_1^{-1} u = 1$ (such $u$ exists because $\Lambda_1^{-1}$ is SPD). Since $\Lambda_2^{-1} = \rho^{-1} \Lambda_1^{-1}$, we have $u^\top \Lambda_2^{-1} u = \rho^{-1}$. Along the ray $y = \mu + ru$ ($r \geq 0$),

$$p_i(\mu + ru) \;=\; C_i \Big(1 + \tfrac{r^2}{\nu_{t,i} \, q_i}\Big)^{-(\nu_{t,i} + d)/2}, \qquad q_1 = 1, \; q_2 = \rho,$$

with $C_i$ independent of $r$, where we used the explicit form of $p_i$ and the identities above. As $r \to \infty$, $\log p_i(\mu + ru) = -(\nu_{t,i} + d)\log r + O(1)$. Matching the leading tail coefficient across $p_1 \equiv p_2$ gives $(\nu_{t,1} + d) = (\nu_{t,2} + d)$, hence $\nu_{t,1} = \nu_{t,2} =: \nu_t$.

*Step 4: $\Lambda$ is identified.* With $\nu_t$ and $\mu$ fixed, the equality on the ray reduces to

$$C_1\big(1 + r^2/\nu_t\big)^{-(\nu_t+d)/2} = C_2\big(1 + r^2/(\rho\nu_t)\big)^{-(\nu_t+d)/2} \qquad \forall r \geq 0.$$

At $r = 0$ this gives $C_1 = C_2$; for any $r > 0$, strict monotonicity of $s \mapsto (1 + r^2/s)^{-(\nu_t+d)/2}$ in $s \in (0, \infty)$ forces $\rho = 1$. Hence $\Lambda_2 = \Lambda_1$.

Therefore the map $(\mu, \Lambda, \nu_t) \mapsto p$ is injective on $\mathbb{R}^d \times \mathrm{SPD}(d) \times (0, \infty)$. □

**Proof of Theorem 3.4.** Let $(\mu_0^{(1)}, \kappa^{(1)}, s^{(1)}, S^{(1)}, \nu_W^{(1)})$ and $(\mu_0^{(2)}, \kappa^{(2)}, s^{(2)}, S^{(2)}, \nu_W^{(2)})$ induce identical predictives $P_X$-a.e. The induced Student-$t$ parameter triples are $(\mu_0^{(i)}(x), \Lambda^{(i)}(x), \nu_t^{(i)})$, with $\Lambda^{(i)}(x) = c(\kappa^{(i)})s^{(i)}(x)S^{(i)}(x)/\nu_t^{(i)}$ and $\nu_t^{(i)} = \nu_W^{(i)} - d + 1$. By Lemma C.1, $\mu_0, \Lambda, \nu_t$ are identified $P_X$-a.e., hence $\nu_W$ is identified (from $\nu_t$).

Next, positive homogeneity of $g$ (degree 1) applied to $\Lambda(x) = c(\kappa)s(x)S(x)/\nu_t$ gives

$$g(\Lambda(x)) = c(\kappa)s(x)g(S(x))/\nu_t = c(\kappa)s(x)/\nu_t,$$

using $g(S(x)) = 1$ by the gauge-fixed factorization. Hence $c(\kappa^{(i)})s^{(i)}(x)/\nu_t^{(i)}$ is identified from $g(\Lambda(x))$ $P_X$-a.e., and $S^{(i)}(x) = \Lambda^{(i)}(x)\nu_t^{(i)}/(c(\kappa^{(i)})s^{(i)}(x))$ is identified $P_X$-a.e.

It remains to identify $\kappa$ and $s$ from the product $c(\kappa)s(x)$ (given $\nu_t$). This is the same scalar identifiability question as Theorem 3.2, now for the $\mathcal{F}_s$-class. Under scale-separation of $\mathcal{F}_s$, the argument of Theorem 3.2's proof applies verbatim: if $c(\kappa_1)s_1 = c(\kappa_2)s_2$ $P_X$-a.e., then $s_2 = as_1$ $P_X$-a.e. with $a = c(\kappa_1)/c(\kappa_2)$. Scale-separation forces $a = 1$ and $s_1 = s_2$, and $c$'s injectivity gives $\kappa_1 = \kappa_2$.

*Necessary direction (explicit construction).* If $\mathcal{F}_s$ is not scale-separated, then by definition there exist $s_1, s_2 \in \mathcal{F}_s$ and $a > 0$ with $a \neq 1$ such that $s_2 = a\, s_1$ $P_X$-a.e. Without loss of generality (swap and replace $a \to 1/a$ if needed), $a > 1$. Because $c : \mathbb{R}_+ \to (1, \infty)$ is a continuous bijection with $c(\nu) \to \infty$ as $\nu \to 0^+$, choose $\kappa_1 > 0$ with $c(\kappa_1) > a$; then $c(\kappa_1)/a \in (1, \infty)$ lies in the range of $c$, so there is a unique $\kappa_2 > 0$ with $c(\kappa_2) = c(\kappa_1)/a$. By construction, $c(\kappa_1)s_1(x) = c(\kappa_2)s_2(x)$ $P_X$-a.e., so the multivariate Student-$t$ scale matrices $\Lambda_1(x) = c(\kappa_1)s_1(x)S(x)/\nu_t$ and $\Lambda_2(x) = c(\kappa_2)s_2(x)S(x)/\nu_t$ coincide $P_X$-a.e. The two parameter tuples $(\mu_0, \kappa_1, s_1, S, \nu_W)$ and $(\mu_0, \kappa_2, s_2, S, \nu_W)$ thus produce identical mv-Student-$t$ predictives $P_X$-a.e., and any strictly proper score has the same population value at both. Hence $\kappa$ is not identifiable. □

**Proof of Theorem 3.5 (per-sample softplus-affine mv identifiability).** Let $(\mu_0^{(1)}, \kappa_{u_1,d_1}, s_{w_1,b_1}, S^{(1)}, \nu_W^{(1)})$ and $(\mu_0^{(2)}, \kappa_{u_2,d_2}, s_{w_2,b_2}, S^{(2)}, \nu_W^{(2)})$ induce identical mv predictives $P_X$-a.e., with $(u_i, w_i) \neq (0, 0)$ for $i = 1, 2$ and $\nu_W^{(i)}$ either global or measurable. By Lemma C.1, the triple $(\mu_0(x), \Lambda(x), \nu_t(x))$ is identified $P_X$-a.e., hence $\nu_W^{(i)}(x) = \nu_t^{(i)}(x) + d - 1$ pointwise. Positive homogeneity of $g$ applied to

$$\Lambda^{(i)}(x) \;=\; c(\kappa_{u_i,d_i}(x))\, s_{w_i,b_i}(x)\, S^{(i)}(x)/\nu_t^{(i)}(x)$$

together with $g(S^{(i)}(x)) = 1$ gives

$$g(\Lambda^{(i)}(x)) \;=\; c(\kappa_{u_i,d_i}(x))\, s_{w_i,b_i}(x)/\nu_t^{(i)}(x),$$

so the scalar product $c(\kappa_{u_i,d_i}(x))\, s_{w_i,b_i}(x)$ is identified $P_X$-a.e. (multiply by the identified $\nu_t^{(i)}(x)$). This is exactly the per-sample softplus-affine product factorization problem of Theorem 3.3, with $(\nu, \beta) \leftarrow (\kappa_{u,d}, s_{w,b})$ and the same fixed-$\phi$, full-dimensional thickness assumption. Apply Theorem 3.3 to recover $u_1 = u_2$, $d_1 = d_2$, $w_1 = w_2$, $b_1 = b_2$ uniquely. The shape then satisfies

$$S^{(i)}(x) \;=\; \Lambda^{(i)}(x)\, \nu_t^{(i)}(x)\, /\, [c(\kappa_{u,d}(x))\, s_{w,b}(x)],$$

identified $P_X$-a.e. The sharpness clause ($u = w = 0$ admits a one-parameter $(b, d)$-family for the constant product) is inherited from Theorem 3.3. The quotient-level caveat for raw $\hat{S}$ vs. normalized $S = \hat{S}/g(\hat{S})$ is identical to the global-$\kappa$ Theorem 3.4. □

# D EXTENDED EXPERIMENTS (DEFERRED FROM §4)

This appendix expands the empirical evidence summarized in the main §4. Notation and protocols carry over from there.

## D.1 SCALAR SYNTHETIC: PER-CELL NUMBERS

Table 5: Scalar synthetic amplitude CV (percent) at $n = 8000$, across twenty data-and-training seeds. Per-sample heads are compared on common probe inputs; the global head has the same CV at each input.

| Method | $\nu^* = 0.5$ | $\nu^* = 1$ | $\nu^* = 2$ | $\nu^* = 5$ |
|---|---|---|---|---|
| Ours | 3.96 | 5.34 | 4.19 | 2.65 |
| Vanilla ($\lambda = 0$) | 51.34 | 47.71 | 37.35 | 29.63 |
| Vanilla ($\lambda = 0.1$) | 34.18 | 30.73 | 35.41 | 32.59 |

All heads start with $\nu(x) = 1$ and train for 1000 epochs. At $n = 8000$, unregularized vanilla has mean pointwise CV 8.91–12.95 times that of our global head; regularized vanilla has CV 5.75–12.29 times larger. The regularized estimates have means between 0.017 and 0.036, showing why small absolute standard deviation alone is insufficient to assess stability. Both vanilla variants have mean test NLL approximately 0.02–0.03 nats below ours. Data and training seeds vary together in this synthetic suite; the common probe inputs align the evaluation, while the fixed-data comparisons in Sections D.7 and H isolate training randomness. Figure 8 plots the CVs, and Figure 9 gives a separate predictive-rescaling diagnostic.

Our mean estimate is approximately 1.2 at $\nu^* = 5$ under the original budget. This is a biased but concentrated estimate, not successful parameter recovery.

**Recovery diagnostics at $\nu^* = 5$.** For five data seeds at $n = 8000$, extending the learned-head fits from 1000 to 10000 Adam epochs raises mean $\hat{\nu}$ from 1.16 to 2.28, while mean test NLL rises from 1.091 to 1.152. Subsequent full-data L-BFGS refinement (up to 1000 iterations) gives $\hat{\nu} = 2.42 \pm 0.31$ and mean relative error 51.7%. The bias remains under the longer optimization budget.

In a separate diagnostic we fix $\gamma, \alpha$ to their generating values and fit only $\beta$ weights and $\nu$, using four initial values $\{0.5, 1, 5, 10\}$ per data seed and the same L-BFGS budget. At $n = 8000$, $\hat{\nu} = 4.5249 \pm 0.4816$ with mean relative error 10.82%; at $n = 80000$, $5.1405 \pm 0.2585$ with error 4.74%. Means and SDs are across five data seeds after averaging over the four initializations. Within each dataset, the largest difference between initialization endpoints is below $8.4 \cdot 10^{-5}$. This known-nuisance diagnostic separates initialization consistency from finite-data variation, but does not replace the learned-head experiment or establish faithful epistemic uncertainty.

## D.2 PER-SAMPLE $n$-SWEEP CONVERGENCE (THEOREM 3.3)

Table 6: Absolute parameter $L^2$ errors for the per-sample fits in Table 3. Mean $\pm$ sample SD across 20 seeds.

| Truth | $(u, d)$ | $(w, b)$ |
|---|---|---|
| Nonconstant | $0.468 \pm 0.147$ | $0.541 \pm 0.185$ |
| Constant $\nu$ | $0.807 \pm 0.182$ | $0.749 \pm 0.175$ |
| Both constant | $0.656 \pm 0.145$ | $0.733 \pm 0.129$ |

We run the per-sample experiment on $n \in \{500, 2000, 8000, 32000\}$ at a fixed 2000-epoch Adam budget, with 20 seeds, in two regimes: nonconstant $u^*, w^* \neq 0$ (Figure 3) and constant-$\nu$, $u^* = 0, w^* \neq 0$ (Figure 4). Both fitted heads initialize $\nu(x)$ at one. The global baseline retains an affine softplus $\beta$ head, including its learned bias, and is correctly specified in the constant-$\nu$ regime.

The regime names describe the generating heads. The per-sample model learns both slope vectors $u, w$ in every regime, including the two-constant truth in Section 4.4. The global baseline fixes $u = 0$

and learns $w$ and both intercepts. Product errors compare the full fitted function $c(\hat{\nu}(x))\hat{\beta}(x)$ with its generating function over the test inputs.

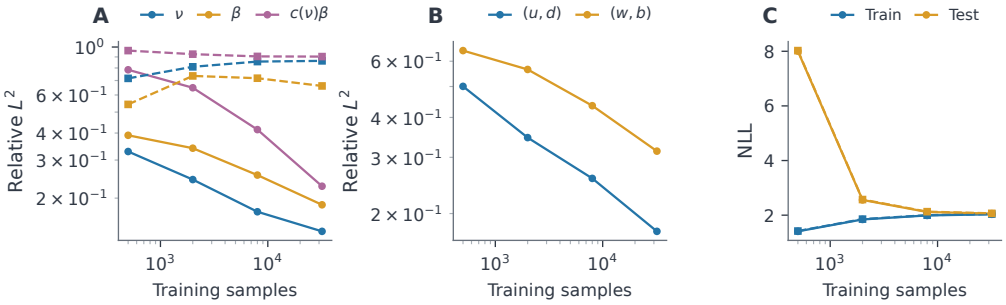


Figure 3: Recovery with nonconstant ground-truth heads ($u^*, w^* \neq 0$). Points show means across 20 seeds after 2000 Adam epochs at learning rate $3 \cdot 10^{-3}$. Solid lines denote the per-sample head; dashed lines denote the global-$\nu$ baseline. (A) Relative function errors. (B) Relative parameter errors of the per-sample head. (C) Train and test NLL. The per-sample head's recovery improves with sample size; the misspecified global head retains substantial error.

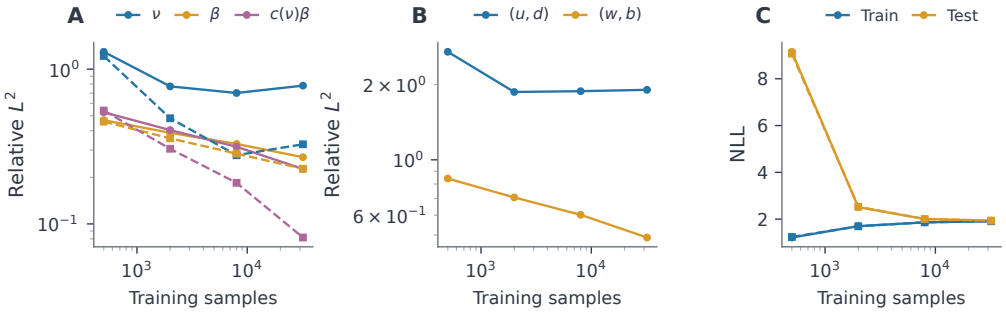


Figure 4: Recovery with constant ground-truth $\nu$ ($u^* = 0, w^* \neq 0$), using the same protocol and line styles as Figure 3. The global-$\nu$ baseline is correctly specified and recovers the functions more accurately. The per-sample head remains identifiable, but its $\nu$ error is not monotone over the tested sample sizes at this fixed training budget.

In the nonconstant regime, mean per-sample $\nu$ error falls from $0.329$ at $n = 500$ to $0.140$ at $n = 32000$. In the constant-$\nu$ regime it falls initially, then rises from $0.701$ at $n = 8000$ to $0.782$ at $n = 32000$. Since $w^* \neq 0$, this regime remains identifiable under Theorem 3.3: the observed finite-budget difficulty does not introduce a predictive equivalence. Absolute parameter errors are included in Table 6, because relative errors depend on the ground-truth parameter norms.

**Local conditioning of the softplus-affine head.** Predictive injectivity distinguishes parameter tuples, while local conditioning describes how sensitively the predictive distribution changes with those parameters. We evaluate the score-outer-product empirical Fisher for $(u, d, w, b)$ at three ground-truth configurations, with $\gamma$ and $\alpha$ fixed at their true values ($k = 3$, $n = 20{,}000$). The nonconstant and constant-$\nu$ configurations have condition numbers $5.13 \times 10^2$ and $1.21 \times 10^4$, respectively. At $u = w = 0$, the matrix is numerically singular, using an eigenvalue threshold of $10^{-10}\lambda_{\max}$; the maximum absolute score in the constant-product orbit direction is $6.66 \times 10^{-16}$. These measurements describe local conditioning at the specified truths. Cross-initialization stability is assessed by the separate training comparisons.

### D.3 SENSITIVITY TO OPTIMIZATION BUDGET AT $n = 32000$

Train and test NLL can be close even while parameters continue to move. We compare checkpoints from the same five trajectories at 2000 and 4000 Adam epochs. The table reports means $\pm$ sample SD; the separate sample-size sweep above uses twenty seeds.

Table 7: Paired budget comparison at $n = 32000$: the same five seeds at 2000 and 4000 epochs. Mean $\pm$ sample SD. The 20-seed sample-size comparison is reported separately.

| Truth | Head | Epoch | $\nu$ rel. $L^2$ | Product rel. $L^2$ | NLL |
|---|---|---|---|---|---|
| Nonconstant | Per-sample | 2000 | $0.147 \pm 0.035$ | $0.276 \pm 0.122$ | $2.061 \pm 0.015$ |
| Nonconstant | Per-sample | 4000 | $0.093 \pm 0.012$ | $0.186 \pm 0.136$ | $2.071 \pm 0.017$ |
| Nonconstant | Global $\nu$ | 2000 | $0.861 \pm 0.011$ | $0.908 \pm 0.006$ | $2.073 \pm 0.015$ |
| Nonconstant | Global $\nu$ | 4000 | $0.931 \pm 0.016$ | $0.876 \pm 0.010$ | $2.077 \pm 0.018$ |
| Constant $\nu$ | Per-sample | 2000 | $0.789 \pm 0.274$ | $0.212 \pm 0.107$ | $1.943 \pm 0.013$ |
| Constant $\nu$ | Per-sample | 4000 | $0.543 \pm 0.086$ | $0.142 \pm 0.037$ | $1.953 \pm 0.014$ |
| Constant $\nu$ | Global $\nu$ | 2000 | $0.336 \pm 0.033$ | $0.097 \pm 0.031$ | $1.942 \pm 0.013$ |
| Constant $\nu$ | Global $\nu$ | 4000 | $0.297 \pm 0.037$ | $0.089 \pm 0.033$ | $1.950 \pm 0.012$ |

Doubling the budget reduces the per-sample head's mean $\nu$ relative error from $0.147$ to $0.093$ in the nonconstant regime and from $0.789$ to $0.543$ in the constant-$\nu$ regime. Product errors also decrease, while mean test NLL rises by approximately $0.010$ nats in both cases. Thus parameter recovery continues to change at a budget where predictive fit changes little.

### D.4 MULTIVARIATE SYNTHETIC

**Setup.** $d = 3$ output, $\phi : \mathbb{R}^{10} \to \mathbb{R}^3$ Gaussian projection, mv-Student-$t$ generator with $\kappa^* \in \{0.5, 1, 2, 5\}$, $n \in \{2000, 8000\}$, 20 seeds. Three heads: **ours-gauge** (global $\kappa$, gauge-fixed $\Psi = s \cdot S$ with $\mathrm{tr}(S)/d = 1$, $s = \mathrm{softplus}(w^\top \phi)$); **ours-closed** (same but with $s = a \cdot \mathrm{softplus}(w^\top \phi)$, where the trainable scalar $a$ closes $\mathcal{F}_s$ and breaks scale-separation); **vanilla-mv** (per-sample $\kappa(x)$, free $s(x)$, free $S(x)$, no gauge).

The method comparison uses $\kappa(x) = 1$ initially and evaluates pointwise CV. A separate orbit diagnostic adds the free multiplier $a$: for fixed $w$, the scale-closed head admits a predictive-equivalence curve with invariant $\log a + \log c(\kappa)$ and slope $-1$ in log coordinates. To probe that curve, the diagnostic samples $a_{\mathrm{init}}, \kappa_{\mathrm{init}} \sim U[0.3, 3.0]$ per seed. This deliberately varied-amplitude protocol is distinct from the common-initialization method comparison.

**Diverse-init results ($n = 8000$).**

| $\kappa^*$ | fitted slope | std($\hat{\kappa}$) | std($\hat{a}$) | std($\log \hat{a} + \log c(\kappa)$) | mean NLL |
|---|---|---|---|---|---|
| 0.5 | $-0.907$ | 0.741 | 0.487 | **0.085** | 3.334 |
| 1.0 | $-0.848$ | 0.848 | 0.384 | **0.093** | 2.720 |
| 2.0 | $-0.808$ | 0.914 | 0.320 | **0.102** | 2.284 |
| 5.0 | $-0.773$ | 0.960 | 0.277 | **0.113** | 1.947 |

Fitted slopes lie between $-0.77$ and $-0.91$, near the predicted $-1$. The invariant-sum std is 2.6–$4.7\times$ smaller than the individual log-coordinate spreads, consistent with the scalar orbit invariant of Theorem 3.4. In the method comparison at $n = 8000$, vanilla-to-ours pointwise CV ratios are 2.03–6.09 at $d = 3$ and 2.57–7.96 at $d = 5$ (Figure 10). These synthetic comparisons vary both data and training seeds. Our mean NLL is higher by $0.074$–$0.123$ nats at $d = 3$ and $0.064$–$0.111$ at $d = 5$.

### D.5 UCI SIGNATURE SUITE

The UCI signature suite examines amplitude constraints and predictive rescaling on the 9-dataset DER subset of Amini et al. (2020), using 20 splits per dataset (5 for protein) from Hernández-Lobato & Adams (2015). The amplitude-constraint comparison uses one training seed per split. These likelihood diagnostics complement the pointwise, within-split stability comparison in Section F, under the real-feature protocol described there.

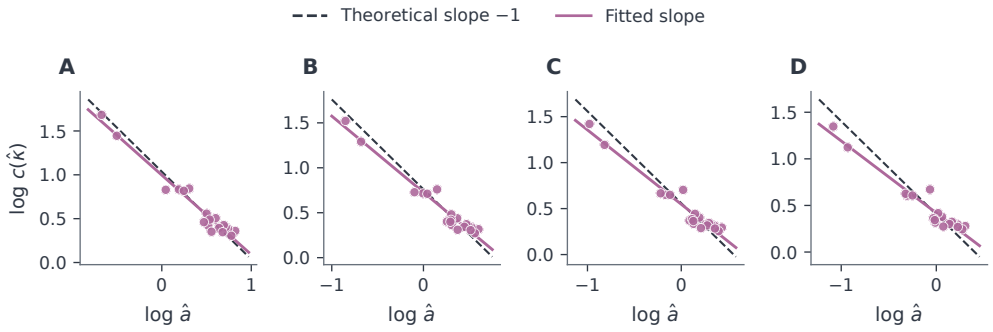


Figure 5: Multivariate scale-closed fits under diverse initialization: twenty fits per panel at $n = 8000$, with (A–D) $\kappa^* \in \{0.5, 1, 2, 5\}$. With $s = a\ \mathrm{softplus}(w^\top \phi)$, the fitted log-coordinate slopes range from $-0.77$ to $-0.91$; dashed lines indicate the theoretical slope $-1$ for compensating rescalings at fixed $w$. The standard deviation of $\log \hat{a} + \log c(\hat{\kappa})$ is 2.6–4.7× smaller than the individual log-coordinate spreads. These independently trained fits show alignment with the scaling direction, rather than exact equality of their pointwise predictives.

**Learnable vs. fixed amplitude.** We compare our global learnable $\nu$ with $\nu$ fixed at 1, using the same frozen Stage-2 trunk and remaining heads. The test-NLL difference satisfies $|\Delta\mathrm{NLL}| \leq 0.006$ on all 9 datasets. This fixes one amplitude while leaving the degrees-of-freedom head free; it is a different constraint from fixing their ratio.

**Effect of the evidence regularizer.** On UCI-protein, sweeping $\lambda \in \{0, 10^{-2}, 10^{-1}\}$ over five splits moves the mean fitted $\hat{\nu}$ toward zero as $\lambda$ increases. Mean NLL rises from 2.749 at $\lambda = 0$ to 2.812 at $\lambda = 0.1$. The near-zero amplitudes explain why an absolute standard deviation can be small even when relative dispersion is appreciable. The five-seed comparison in Table 15 therefore uses pointwise CV.

**Held-out NLL along the orbit.** Starting from vanilla DER fits on UCI-protein with $\lambda \in \{0, 0.1\}$ (five splits, three seeds, 80 training epochs), we freeze $(\gamma, \alpha)$ and replace $\hat{\nu}(x)$ by a common value $\nu'$ on a 21-point log grid in $[0.1, 10]$. The pointwise rescaling $\beta'(x) = \hat{\beta}(x)c(\hat{\nu}(x))/c(\nu')$ preserves the fitted predictive product at every input. Held-out NLL is flat along this path for every seed and both regularizer settings (Figure 6). The rescaled functions are evaluated directly; they need not remain in the fitted softplus head class.

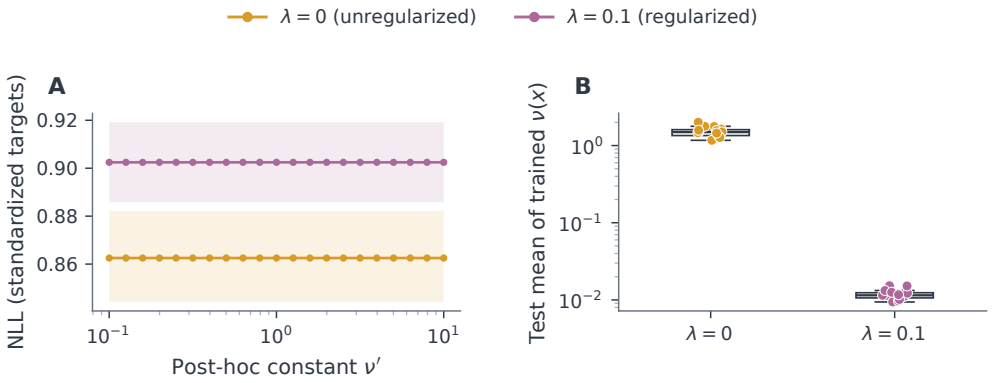


Figure 6: Likelihood and regularizer diagnostics on UCI-protein. (A) Pointwise rescaling preserves each fitted predictive and its NLL, here evaluated on standardized targets. Shaded bands show the 25–75 percentiles over fifteen split–seed fits. (B) Test-input means of the trained $\nu(x)$ heads. This diagnostic uses a separate three-seed protocol; the main method comparison uses the pointwise CVs in Table 15.

**Predictive comparison.** The five-seed NLL and calibration results in Sections F and G compare the complete methods, including unregularized vanilla DER. They show dataset-dependent predictive tradeoffs alongside the lower amplitude CV of our two-stage head.

### D.6 SCALE-CLOSED ORBIT DIAGNOSTIC

The scale-closed configuration adds a free positive multiplier $a$ to our scale-separated $\beta$ head. Its purpose is to visualize the equivalence predicted by the theory; it is not the vanilla DER baseline. Figure 7 uses ten runs with fixed data seed 0, $(a_{\text{init}}, \nu_{\text{init}}) \sim U[0.3, 3]$, and seed-dependent head weights and minibatch order, trained for 30 epochs. The fitted slope is close to the compensating-rescaling slope $-1$.

Allowing the data seed to vary gives the same qualitative pattern: a slope near $-1$ and an invariant sum with smaller dispersion than its summands. The numerical values change with the data realization. The common-initialization comparison against vanilla DER is reported separately in Section D.7.

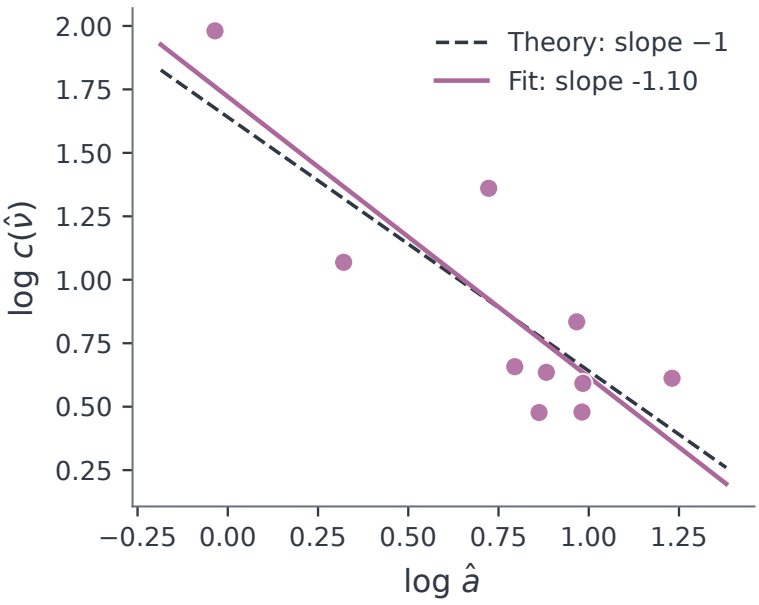


Figure 7: Scaling-direction diagnostic on CIFAR-rotation. Each point is an independently trained scale-closed fit. The theoretical and fitted slopes are $-1$ and $-1.10$, respectively. The standard deviation of $\log \hat{a} + \log c(\hat{\nu})$ is 0.54 times the sum of the two individual standard deviations. The fits align with the theoretical orbit direction; because the other head parameters are also learned, this plot does not assert exact predictive equality between fitted models.

### D.7 CIFAR-ROTATION: FROZEN-FEATURE STRESS TEST

We use frozen ImageNet ResNet-18 features to test the empirical stability of the heads. The finite CIFAR image set is augmented by a single rotation angle. Random rotation of this finite collection does not produce a full-dimensional feature density in $\mathbb{R}^{512}$. This comparison lies outside the theorem's density assumption. The rank, projected KDE and anisotropy diagnostics in Section I describe the observed feature geometry.

**Setup.** CIFAR-10 (Krizhevsky, 2009) images are rotated by $\theta \sim U[-\pi, \pi]$ and the target is $\theta$ in radians. The 512-dimensional pooled features come from a frozen ImageNet-pretrained ResNet-18 (He et al., 2016). We compare our global-$\nu$ head, unregularized vanilla DER and vanilla DER with $\lambda = 0.1$, using Adam at learning rate $10^{-3}$ for 1000 epochs. Each of three data realizations uses ten training seeds. Data and test inputs remain fixed within each ten-seed comparison; every method initializes $\nu(x) = 1$ and learns it thereafter.

The unregularized baseline separates the behavior of vanilla DER from that of its evidence regularizer. All three methods share the same frozen features. The separate predictive-family comparison in Section D.9 includes Gaussian MLE, Deep Ensembles and MC-Dropout.

Table 8: CIFAR-rotation at 1000 epochs. CVs are across ten training seeds at each fixed input, then averaged over test inputs; units are percent. NLL is mean ± sample SD across seeds. All heads start with $\nu(x) = 1$ and learn it.

| Data | Method | $\nu$ CV | Aleatoric CV | Epistemic CV | NLL |
|---|---|---|---|---|---|
| 1 | Ours | 1.474 | 7.998 | 7.380 | $1.969 \pm 0.009$ |
| 1 | Vanilla ($\lambda = 0$) | 10.064 | 7.608 | 12.896 | $1.974 \pm 0.011$ |
| 1 | Vanilla ($\lambda = 0.1$) | 7.530 | 16.527 | 16.809 | $2.088 \pm 0.006$ |
| 2 | Ours | 3.290 | 7.339 | 7.491 | $1.994 \pm 0.018$ |
| 2 | Vanilla ($\lambda = 0$) | 12.999 | 7.050 | 14.547 | $1.993 \pm 0.014$ |
| 2 | Vanilla ($\lambda = 0.1$) | 7.589 | 16.021 | 18.847 | $2.093 \pm 0.006$ |
| 3 | Ours | 1.793 | 7.572 | 7.239 | $1.987 \pm 0.012$ |
| 3 | Vanilla ($\lambda = 0$) | 13.222 | 5.989 | 13.820 | $1.982 \pm 0.013$ |
| 3 | Vanilla ($\lambda = 0.1$) | 6.947 | 11.414 | 14.478 | $2.088 \pm 0.004$ |

Our head has lower mean pointwise amplitude and epistemic-variance CV in all three realizations. This ordering concerns the mean over inputs: for the second realization, regularized vanilla has median amplitude CV $2.98\%$, below our global CV $3.29\%$, although its mean pointwise CV is $7.59\%$. Thus the mean-CV advantage does not imply a lower CV at every individual input. Relative to unregularized vanilla, our mean NLL changes by $-0.005$, $+0.002$ and $+0.005$ nats across the three realizations.

**References without image information.** On the same test inputs as the first data realization, we evaluate zero prediction, a constant Student-$t$ distribution fitted to the training angles, and the known generating distribution $U[-\pi, \pi]$ (Table 9). The constant Student-$t$ fit is selected by training NLL and approaches the Gaussian limit. Our mean RMSE is $1.812$, compared with $1.821$ for zero prediction, a reduction of approximately $0.50\%$. Our NLL is lower than the fitted constant Student-$t$ ($1.969$ versus $2.018$), while the known uniform distribution has lower NLL ($1.838$). The uniform reference uses the task's known bounded support and lies outside the Student-$t$ predictive family. These references show limited gains in conditional prediction, alongside the cross-seed stability differences reported above.

Table 9: CIFAR-rotation references on the same $10{,}000$ test inputs as data realization 1 in Table 8. Ours uses the ten fixed-evidence-initialization fits at $1000$ epochs (mean ± sample SD). The other predictors do not use image features.

| Predictor | RMSE | NLL |
|---|---|---|
| Ours | $1.812 \pm 0.004$ | $1.969 \pm 0.009$ |
| Zero point prediction | 1.821 | — |
| Fitted constant Student-$t$ | 1.821 | 2.018 |
| Known uniform distribution | 1.821 | 1.838 |

### D.8 CIFAR-ROTATION: FROZEN-FEATURE CALIBRATION

Table 10 reports coverage, interval width and analytic Student-$t$ CRPS for the same fits as the stability comparison. The scores assess the predictive distribution and complement the dispersion of its internal uncertainty components.

### D.9 MATCHED UNCERTAINTY BASELINES AND SEMANTIC SHIFT

This comparison evaluates predictive quality against Deep Ensembles (DE) and MC-Dropout (MCD). It uses a separate 30-epoch protocol from the $1000$-epoch stability comparison in Section D.7.

**Protocol.** All methods share frozen ImageNet ResNet-18 features of dimension $512$, the actual applied rotation angle as target, and one fixed data/rotation realization (data seed $0$). Five training

Table 10: CIFAR-rotation calibration for the same 1000-epoch fits as Table 8. Mean ± sample SD across ten seeds. Intervals and CRPS use the Student-$t$ predictive; widths and CRPS are in radians.

| Data | Method | $\text{PICP}_{90}$ | $\text{MPIW}_{90}$ | CRPS |
|---|---|---|---|---|
| 1 | Ours | $0.948 \pm 0.010$ | $5.791 \pm 0.139$ | $1.031 \pm 0.003$ |
| 1 | Vanilla ($\lambda = 0$) | $0.941 \pm 0.017$ | $5.792 \pm 0.265$ | $1.030 \pm 0.003$ |
| 1 | Vanilla ($\lambda = 0.1$) | $0.975 \pm 0.005$ | $8.208 \pm 0.447$ | $1.052 \pm 0.003$ |
| 2 | Ours | $0.937 \pm 0.012$ | $5.680 \pm 0.140$ | $1.042 \pm 0.002$ |
| 2 | Vanilla ($\lambda = 0$) | $0.938 \pm 0.022$ | $5.787 \pm 0.321$ | $1.042 \pm 0.005$ |
| 2 | Vanilla ($\lambda = 0.1$) | $0.978 \pm 0.006$ | $8.426 \pm 0.486$ | $1.061 \pm 0.003$ |
| 3 | Ours | $0.941 \pm 0.009$ | $5.710 \pm 0.135$ | $1.029 \pm 0.002$ |
| 3 | Vanilla ($\lambda = 0$) | $0.928 \pm 0.020$ | $5.677 \pm 0.295$ | $1.032 \pm 0.005$ |
| 3 | Vanilla ($\lambda = 0.1$) | $0.975 \pm 0.006$ | $8.290 \pm 0.323$ | $1.048 \pm 0.001$ |

seeds vary initialization and minibatch order. Each head trains for 30 epochs with batch size 256 and learning rate $10^{-3}$. Ours and the vanilla DER controls use one predictive pass; DE uses five heteroscedastic Gaussian members; MCD uses dropout probability 0.1 and 30 stochastic passes. The comparison is therefore not training- or inference-compute matched. Train and test images are disjoint, but the deterministic angle rule uses a matched schedule across splits rather than an independently reseeded test-angle realization. All reported standard deviations below are sample standard deviations across the five training seeds, not uncertainty across independent datasets.

**Metrics.** DER NLL and interval quantiles are exact for its Student-$t$ predictive. DE/MCD use Gaussian-mixture NLL and analytic CRPS, with interval quantiles obtained by deterministic CDF bisection. DER CRPS uses 512 seeded independent-pair Monte Carlo samples per test point. Differences of a few $10^{-3}$ in CRPS are therefore descriptive, not strong rankings across these estimators.

Table 11: Additional full CIFAR-10 ID comparison. Mean ± sample SD across five training seeds on one fixed data realization. NLL, CRPS and RMSE are lower-is-better; PICP and MPIW refer to the nominal 90% interval. Width and RMSE are in radians.

| Method | Passes | NLL | CRPS | $\text{PICP}_{90}$ | $\text{MPIW}_{90}$ | RMSE |
|---|---|---|---|---|---|---|
| Ours | 1 | $1.9761 \pm 0.0021$ | $1.0288 \pm 0.0029$ | $0.9461 \pm 0.0037$ | $5.7988 \pm 0.0884$ | $1.8024 \pm 0.0054$ |
| Vanilla, regularized | 1 | $2.0868 \pm 0.0132$ | $1.0532 \pm 0.0087$ | $0.9800 \pm 0.0073$ | $8.2849 \pm 0.5530$ | $1.8223 \pm 0.0146$ |
| Vanilla, unregularized | 1 | $1.9772 \pm 0.0169$ | $1.0296 \pm 0.0091$ | $0.9271 \pm 0.0365$ | $5.7295 \pm 0.4646$ | $1.8057 \pm 0.0038$ |
| Gaussian MLE | 1 | $1.9712 \pm 0.0036$ | $1.0289 \pm 0.0053$ | $0.9357 \pm 0.0373$ | $5.9808 \pm 0.5833$ | $1.7977 \pm 0.0067$ |
| Deep Ensemble | 5 | $1.9599 \pm 0.0019$ | $1.0241 \pm 0.0014$ | $0.9413 \pm 0.0114$ | $5.8739 \pm 0.1807$ | $1.7927 \pm 0.0006$ |
| MC-Dropout | 30 | $1.9789 \pm 0.0087$ | $1.0334 \pm 0.0061$ | $0.9625 \pm 0.0239$ | $6.1947 \pm 0.4684$ | $1.8038 \pm 0.0060$ |

DE has the lowest reported mean full-ID NLL, CRPS and RMSE. Ours has NLL 0.0162 higher than DE and 0.0028 lower than MCD at one predictive pass. This is competitive single-pass prediction.

**Semantic class holdout.** Training and ID evaluation use the six CIFAR-10 animal classes $\{2, 3, 4, 5, 6, 7\}$; OOD evaluation uses the four vehicle classes $\{0, 1, 8, 9\}$. Class labels determine membership only, never the rotation target. This is one internal class-holdout diagnostic, not a general OOD benchmark.

Table 12: Animal-class ID evaluation under the semantic protocol. Same aggregation and metric caveats as Table 11.

| Method | NLL | CRPS | $\text{PICP}_{90}$ | $\text{MPIW}_{90}$ |
|---|---|---|---|---|
| Ours | $2.0108 \pm 0.0070$ | $1.0422 \pm 0.0014$ | $0.9435 \pm 0.0064$ | $5.8412 \pm 0.0889$ |
| Vanilla, regularized | $2.1125 \pm 0.0026$ | $1.0632 \pm 0.0028$ | $0.9853 \pm 0.0043$ | $8.4013 \pm 0.4907$ |
| Vanilla, unregularized | $2.0115 \pm 0.0122$ | $1.0427 \pm 0.0025$ | $0.9269 \pm 0.0174$ | $5.7752 \pm 0.2223$ |
| Gaussian MLE | $1.9896 \pm 0.0088$ | $1.0428 \pm 0.0062$ | $0.9200 \pm 0.0284$ | $5.7701 \pm 0.3630$ |
| Deep Ensemble | $1.9855 \pm 0.0019$ | $1.0409 \pm 0.0015$ | $0.9250 \pm 0.0104$ | $5.7826 \pm 0.1206$ |
| MC-Dropout | $1.9885 \pm 0.0025$ | $1.0464 \pm 0.0018$ | $0.9328 \pm 0.0111$ | $5.7405 \pm 0.1197$ |

Table 13: Vehicle-class OOD evaluation. Mean ± sample SD across five training seeds. AUROC uses predictive variance to discriminate OOD from ID; OOD/ID variance is the ratio of their mean predictive variances. Spearman measures variance–error correlation.

*Predictive quality*

| Method | NLL | CRPS | $\text{PICP}_{90}$ | $\text{MPIW}_{90}$ |
|---|---|---|---|---|
| Ours | $2.0126 \pm 0.0044$ | $1.0492 \pm 0.0034$ | $0.9496 \pm 0.0087$ | $6.0502 \pm 0.1023$ |
| Vanilla, regularized | $2.1421 \pm 0.0057$ | $1.0772 \pm 0.0055$ | $0.9898 \pm 0.0027$ | $8.8984 \pm 0.4452$ |
| Vanilla, unregularized | $2.0151 \pm 0.0071$ | $1.0508 \pm 0.0056$ | $0.9391 \pm 0.0140$ | $6.0865 \pm 0.2431$ |
| Gaussian MLE | $2.0156 \pm 0.0023$ | $1.0511 \pm 0.0038$ | $0.9266 \pm 0.0309$ | $6.0191 \pm 0.4292$ |
| Deep Ensemble | $2.0153 \pm 0.0033$ | $1.0515 \pm 0.0027$ | $0.9310 \pm 0.0122$ | $6.0396 \pm 0.1520$ |
| MC-Dropout | $2.0054 \pm 0.0029$ | $1.0477 \pm 0.0027$ | $0.9462 \pm 0.0249$ | $5.9630 \pm 0.2733$ |

*OOD diagnostics*

| Method | AUROC | OOD/ID variance | Spearman |
|---|---|---|---|
| Ours | $0.5492 \pm 0.0064$ | $1.0543 \pm 0.0069$ | $0.2273 \pm 0.0094$ |
| Vanilla, regularized | $0.5555 \pm 0.0160$ | $1.1015 \pm 0.0357$ | $0.2479 \pm 0.0167$ |
| Vanilla, unregularized | $0.5571 \pm 0.0135$ | $1.0951 \pm 0.0259$ | $0.2321 \pm 0.0141$ |
| Gaussian MLE | $0.5505 \pm 0.0351$ | $1.0734 \pm 0.0605$ | $0.2286 \pm 0.0211$ |
| Deep Ensemble | $0.5539 \pm 0.0125$ | $1.0765 \pm 0.0216$ | $0.2248 \pm 0.0064$ |
| MC-Dropout | $0.5576 \pm 0.0523$ | $1.0698 \pm 0.0691$ | $0.2261 \pm 0.0256$ |

On animal-class ID, ours has higher NLL than both DE and MCD. On vehicle-class OOD, their predictive metrics are close, but all three variance AUROCs are about $0.55$, with OOD/ID variance ratios about $1.05$–$1.08$. None shows strong OOD discrimination in this diagnostic.

**Head-only inference.** A separate synchronized benchmark on one H200 measures the forward pass(es) and predictive mean/variance aggregation for $10{,}000$ device-resident feature vectors. It excludes training, feature extraction, transfer and metric evaluation. Each seed uses $20$ warmups and $100$ repetitions; mean ± sample SD of the five per-seed median latencies is $0.090 \pm 0.006$ ms for ours, $0.230 \pm 0.009$ ms for DE and $1.424 \pm 0.122$ ms for MCD on the full protocol. The semantic protocol gives $0.092 \pm 0.009$, $0.238 \pm 0.020$ and $1.393 \pm 0.048$ ms, respectively. These are head-only measurements, not end-to-end latency or training-cost comparisons.

**Practical use.** Freeze a pre-trained representation before training the evidential heads; use the anchored bias-free softplus class for global $\nu$, or the softplus-affine class under its stated feature condition; distinguish the nonconstant softplus-affine subclass with trainable biases from the both-constant exception. Feature diagnostics can flag obvious degeneracy but cannot certify the theorem assumptions. Predictive quality should still be evaluated separately from parameter identifiability.

## E Additional figures referenced from the main text

This appendix collects the synthetic pointwise-CV comparisons and the predictive-rescaling diagnostic.

## F UCI 5-seed aggregated NLL

Table 14 gives test-set NLL averaged over $5$ seeds $\times \{20, 5\}$ splits on the $9$-dataset DER subset of Amini et al. (2020), using the split protocol of Hernández-Lobato & Adams (2015). Lower is better; bold marks the lowest mean per row. Our two-stage head has lower mean pointwise amplitude CV than all three vanilla variants on all nine datasets (Table 15). Predictive NLL is dataset-dependent: relative to unregularized vanilla, ours is lower by $0.501$ nats on naval and higher by $0.065$, $0.024$ and $0.202$ nats on boston, protein and yacht, respectively.

**Protocols.** Arm A pretrains and then freezes a ReLU trunk. Its feature dimension exceeds the input dimension, so this deterministic map does not produce a full-dimensional feature density. We

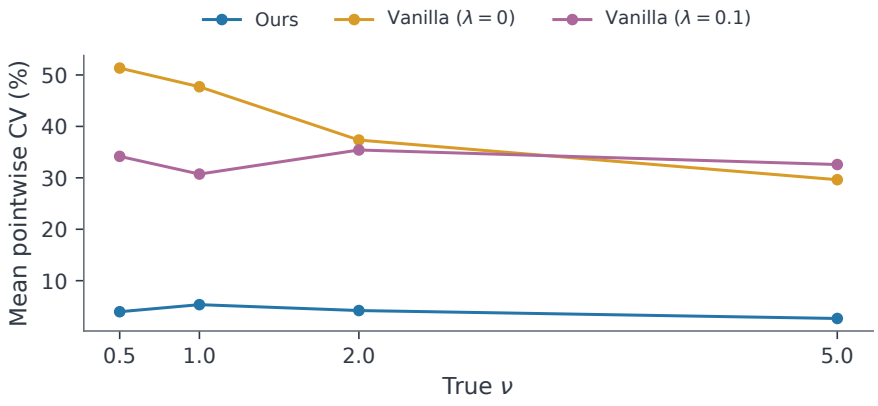


Figure 8: Scalar synthetic amplitude stability at $n = 8000$: four values of $\nu^*$, three methods and twenty data-and-training seeds. Points show mean pointwise CV on common probe inputs. All methods initialize $\nu(x) = 1$. Lower dispersion does not imply accurate recovery: at $\nu^* = 5$, our mean estimate is approximately 1.18.

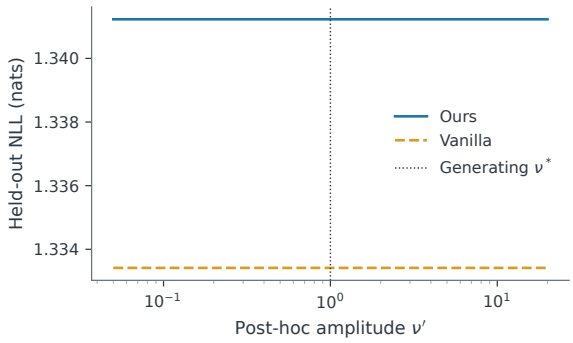


Figure 9: Held-out NLL under $\beta'(x) = \beta(x) \cdot c(\hat{\nu}_{\text{trained}})/c(\nu')$ at $\nu^* = 1$, $n = 8000$. For ours, this rescaling preserves $c(\nu)\beta(x)$ pointwise, so NLL is flat in $\nu'$ along the orbit (Proposition B.1). For per-sample vanilla DER, substituting a single $\nu'$ no longer preserves the predictive at every $x$, so the curve is a constrained projection.

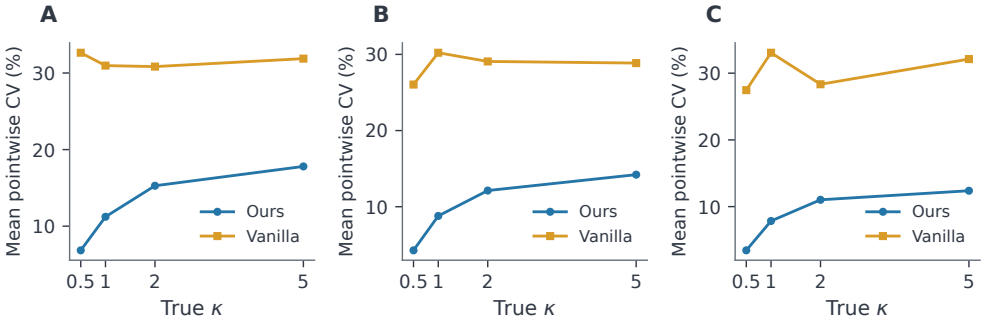


Figure 10: Multivariate synthetic amplitude stability at $n = 8000$, across twenty data-and-training seeds. Mean pointwise CV compares vanilla-mv with the global-$\kappa$ method at dimensions (A) $d = 2$, (B) $d = 3$ and (C) $d = 5$. Both initialize $\kappa = 1$. These are descriptive comparisons with the original global-head fits; the fixed-data, matched-update comparison is the QM9 experiment in Section H.

evaluate the resulting heads empirically, outside the density guarantee of Corollary B.2. Arm B jointly trains the trunk and heads. All DER methods start with $\nu(x) = 1$ and learn it. Stability is evaluated across the five seeds within each split, then averaged over inputs and splits. The original single-seed-per-split diagnostics in Section D.5 are reported separately.

Table 14: UCI test NLL in original units. Mean $\pm$ sample SD across five seeds and 20 splits (five for protein). DER heads start with $\nu = 1$ and learn it. Arm A uses two-stage training; Arm B trains jointly. Bold marks the lowest mean, without implying significance.

| Dataset | Ours A | Ours B | Vanilla $\lambda = 0$ | Vanilla $\lambda = .1$ | DER+MAE | Gaussian |
|---|---|---|---|---|---|---|
| Boston | $2.583 \pm 0.189$ | $\mathbf{2.504 \pm 0.169}$ | $2.518 \pm 0.172$ | $2.521 \pm 0.157$ | $2.522 \pm 0.158$ | $2.631 \pm 0.239$ |
| Concrete | $3.340 \pm 0.093$ | $3.325 \pm 0.098$ | $\mathbf{3.315 \pm 0.098}$ | $3.327 \pm 0.095$ | $3.328 \pm 0.094$ | $3.326 \pm 0.104$ |
| Energy | $2.055 \pm 0.141$ | $2.068 \pm 0.147$ | $2.053 \pm 0.145$ | $2.045 \pm 0.144$ | $2.045 \pm 0.144$ | $\mathbf{2.012 \pm 0.164}$ |
| Kin8nm | $-1.000 \pm 0.036$ | $-0.988 \pm 0.033$ | $-0.999 \pm 0.031$ | $-0.964 \pm 0.030$ | $-0.966 \pm 0.030$ | $\mathbf{-1.000 \pm 0.034}$ |
| Naval | $\mathbf{-4.688 \pm 0.181}$ | $-4.171 \pm 0.098$ | $-4.187 \pm 0.086$ | $-4.150 \pm 0.077$ | $-4.165 \pm 0.074$ | $-4.534 \pm 0.176$ |
| Power | $2.791 \pm 0.025$ | $2.790 \pm 0.024$ | $\mathbf{2.788 \pm 0.024}$ | $2.802 \pm 0.023$ | $2.802 \pm 0.023$ | $2.825 \pm 0.038$ |
| Protein | $2.778 \pm 0.018$ | $\mathbf{2.754 \pm 0.021}$ | $2.754 \pm 0.021$ | $2.807 \pm 0.025$ | $2.807 \pm 0.025$ | $2.884 \pm 0.023$ |
| Wine | $0.943 \pm 0.068$ | $\mathbf{0.936 \pm 0.075}$ | $0.937 \pm 0.078$ | $0.944 \pm 0.069$ | $0.944 \pm 0.069$ | $0.962 \pm 0.133$ |
| Yacht | $3.311 \pm 0.203$ | $3.200 \pm 0.224$ | $3.109 \pm 0.254$ | $3.109 \pm 0.243$ | $\mathbf{3.106 \pm 0.242}$ | $3.324 \pm 0.201$ |

Table 15: UCI amplitude stability. Within each split, CV is computed across five seeds at each fixed test input, then averaged over inputs and splits. Global heads give the same CV at every input. Values are percentages.

| Dataset | Ours A | Ours B | Vanilla $\lambda = 0$ | Vanilla $\lambda = .1$ | DER+MAE |
|---|---|---|---|---|---|
| Boston | 1.17 | 0.70 | 9.02 | 10.52 | 10.47 |
| Concrete | 0.79 | 0.91 | 8.88 | 17.16 | 17.20 |
| Energy | 0.84 | 0.69 | 5.29 | 15.39 | 15.52 |
| Kin8nm | 2.79 | 1.44 | 35.05 | 42.35 | 42.32 |
| Naval | 6.19 | 2.58 | 26.28 | 19.62 | 19.84 |
| Power | 1.69 | 1.02 | 20.23 | 24.03 | 23.95 |
| Protein | 10.32 | 8.00 | 33.10 | 46.95 | 44.85 |
| Wine | 1.41 | 1.27 | 14.47 | 27.74 | 27.78 |
| Yacht | 0.65 | 0.33 | 11.68 | 16.53 | 16.58 |

## G UCI-9 CALIBRATION AND CRPS

Alongside NLL (Section F), we evaluate analytic CRPS for each predictive distribution: Student-$t$ for DER methods and Gaussian for Gaussian MLE. ECE-90 $= |\text{coverage} - 0.9|$ uses the corresponding predictive intervals. Tables 16 and 17 aggregate five seeds over twenty splits per dataset (five for protein). CRPS is evaluated from the Student-$t$ distribution itself, without a Gaussian approximation.

With unregularized vanilla included, Arm A has the lowest mean ECE-90 on naval and yacht. No method dominates the predictive comparisons. Rankings use unrounded means and do not establish significance. These empirical calibration results are separate from injectivity: exact orbit motion preserves the entire predictive distribution, including coverage.

## H MULTIVARIATE REAL-DATA COMPARISONS

**QM9 multivariate regression with a jointly trained trunk.** We train SchNet (Schütt et al., 2017) jointly with each multivariate head to predict polarizability $\alpha$, HOMO and LUMO on QM9 (Ramakrishnan et al., 2014). The fixed split contains 110,000 training, 10,000 validation and 10,000 test molecules, and targets are standardized using training-set statistics. Ten seeds vary initialization and training order; $\kappa(x)$ starts at one for all three methods. Each run has 120 epochs and 103,200 optimizer updates. The uncoupled NIW baseline learns per-input $\kappa$ and a free scale matrix, with separate global predictive degrees of freedom. The coupled baseline follows the bounded $\nu_W$ head and Cholesky parameterization of Meinert & Lavin (2021), with $\nu_W = r\kappa$ and $r \simeq 6$, chosen to match the initial evidence strength and predictive degrees of freedom. Both NIW baselines optimize predictive NLL alone. Table 18 reports the 20, 60 and 120-epoch checkpoints.

Table 16: Analytic CRPS of the actual predictive distribution on UCI, in original target units. Means use the same five-seed split protocol as Table 14; the Student-$t$ formula is used for all DER methods. Bold marks the lowest unrounded mean.

| Dataset | Ours A | Vanilla $\lambda = 0$ | Vanilla $\lambda = .1$ | DER+MAE | Gaussian |
|---|---|---|---|---|---|
| Boston | 1.8556 | **1.7794** | 1.7912 | 1.7912 | 1.7953 |
| Concrete | 4.0212 | 3.9223 | 3.9616 | 3.9637 | **3.9184** |
| Energy | 1.3521 | 1.3610 | 1.3552 | 1.3547 | **1.2998** |
| Kin8nm | 0.0570 | 0.0570 | 0.0566 | **0.0564** | 0.0573 |
| Naval | **0.0015** | 0.0023 | 0.0024 | 0.0023 | 0.0016 |
| Power | 2.2442 | **2.2385** | 2.2448 | 2.2448 | 2.2522 |
| Protein | **2.5282** | 2.5605 | 2.6058 | 2.5988 | 2.5484 |
| Wine | 0.3493 | **0.3480** | 0.3510 | 0.3509 | 0.3480 |
| Yacht | 4.4642 | 4.0562 | 4.0491 | **4.0380** | 4.2898 |

Table 17: UCI coverage error $|\mathrm{PICP}_{90} - 0.9|$ for the same fits as Table 14. Lower is better; bold marks the lowest unrounded mean.

| Dataset | Ours A | Ours B | Vanilla $\lambda = 0$ | Vanilla $\lambda = .1$ | DER+MAE | Gaussian |
|---|---|---|---|---|---|---|
| Boston | 0.037 | 0.034 | **0.032** | 0.032 | 0.032 | 0.035 |
| Concrete | 0.028 | 0.029 | 0.030 | 0.028 | **0.028** | 0.029 |
| Energy | 0.054 | 0.060 | 0.061 | **0.049** | 0.050 | 0.056 |
| Kin8nm | 0.015 | 0.014 | 0.015 | 0.018 | 0.018 | **0.013** |
| Naval | **0.011** | 0.023 | 0.022 | 0.024 | 0.025 | 0.025 |
| Power | 0.010 | **0.009** | 0.009 | 0.025 | 0.024 | 0.014 |
| Protein | 0.011 | **0.008** | 0.009 | 0.021 | 0.022 | 0.027 |
| Wine | 0.027 | 0.032 | 0.032 | 0.022 | **0.022** | 0.026 |
| Yacht | **0.039** | 0.042 | 0.041 | 0.044 | 0.044 | 0.045 |

Table 18: Jointly trained QM9 models, ten seeds per method, with $\kappa(x) = 1$ initially. CVs are mean pointwise percentages; epistemic CV uses the covariance trace. NLL is mean $\pm$ sample SD.

| Epoch | Method | $\kappa$ CV | Epistemic CV | NLL |
|---|---|---|---|---|
| 20 | Ours | 1.553 | 27.515 | -2.319 $\pm$ 0.290 |
| 20 | Uncoupled NIW | 24.797 | 40.586 | -2.282 $\pm$ 0.193 |
| 20 | Coupled NIW ($r = 6$) | 0.496 | 26.234 | -2.312 $\pm$ 0.312 |
| 60 | Ours | 2.921 | 26.381 | -3.352 $\pm$ 0.180 |
| 60 | Uncoupled NIW | 23.785 | 41.887 | -3.383 $\pm$ 0.312 |
| 60 | Coupled NIW ($r = 6$) | 0.586 | 27.960 | -3.312 $\pm$ 0.397 |
| 120 | Ours | 3.074 | 27.581 | -4.689 $\pm$ 0.055 |
| 120 | Uncoupled NIW | 27.850 | 44.921 | -4.718 $\pm$ 0.052 |
| 120 | Coupled NIW ($r = 6$) | 0.819 | 31.783 | -4.681 $\pm$ 0.054 |

At the final checkpoint, the uncoupled-NIW-to-ours amplitude CV ratio is $9.06$, while the epistemic covariance-trace CV ratio is $1.63$. Coupled NIW has lower $\kappa$ CV than ours, while ours has $13.2\%$ lower epistemic-trace CV. Mean NLL for ours is $0.029$ nats higher than uncoupled NIW and $0.008$ nats lower than coupled NIW. The ordering depends on training budget: at 20 epochs the coupled baseline has lower CV for both quantities; ours has lower epistemic-trace CV at 60 and 120 epochs. All evaluated predictive degrees of freedom exceed two, so the covariance summaries are defined. These are comparisons of complete parameterizations under joint training.

**CIFAR-10 rotation regression with circular targets.** The label $(\sin\theta, \cos\theta) \in \mathbb{R}^2$ avoids the branch cut of scalar angle regression. Using frozen ResNet-18 features, 30 epochs and three data-and-training seeds, our global-$\kappa$ head has CV $0.61\%$, compared with mean pointwise CV $5.40\%$ for vanilla-mv; mean NLL is $1.933$ and $1.935$, respectively. Both initialize $\kappa(x) = 1$. This small auxiliary comparison varies data and training randomness together; the QM9 comparison above uses ten seeds on fixed data.

# I FEATURE-SUPPORT DIAGNOSTIC PROTOCOL

Corollary B.2 requires the pushforward $P_X \circ \phi^{-1}$ to admit a density with respect to Lebesgue measure on $\mathbb{R}^k$ that is positive almost everywhere on some non-empty open set $U \subseteq \mathbb{R}^k$. We report three finite-sample diagnostics on $\Phi = \{\phi(x_i)\}_{i=1}^N$. They describe observed feature geometry, not a verification of this population condition. A pass below means only that the chosen diagnostic thresholds are met.

**(i) Rank sufficiency.** Compute the centered feature covariance $\Sigma_\phi = \frac{1}{N-1}\sum_i(\phi(x_i) - \bar{\phi})(\phi(x_i) - \bar{\phi})^\top$ and its eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$. Define effective dimension $m^\star = \max\{k : \lambda_k/\lambda_1 \geq \tau_r\}$ with default $\tau_r = 10^{-3}$. The diagnostic uses $m^\star \geq 2$ to exclude rank-one samples. This is not sufficient for a density in the ambient feature space: even full covariance rank can occur on a curved lower-dimensional support.

**(ii) Local positivity.** Set the projection dimension to $m = \min(m^\star, 3)$. Project features onto the top-$m$ eigenvectors, whiten by dividing each axis by $\sqrt{\lambda_k}$, and evaluate a Gaussian kernel density estimate $\hat{p}_\phi$ with per-axis Silverman bandwidth $h_k = \sigma_k N^{-1/(m+4)}$ on a grid inside the whitened 2-ball $\{\|z\| \leq 2\}$. Report $\min \hat{p}_\phi$. The diagnostic threshold is $\min \hat{p}_\phi \geq \tau_p = 10^{-6}$. A Gaussian KDE is positive by construction; its values depend on projection, bandwidth, and grid resolution, so passing cannot establish positivity of the underlying ambient-space density.

**(iii) Anisotropy disclosure.** Report $\lambda_1/\lambda_{m^\star}$ as context. This does not gate pass/fail but documents how compressed the feature distribution is across its effective dimensions.

**Audit result for ResNet-18 on CIFAR-rotation.** On $N$=10,000 features (pooled; audit rerun per-seed for seeds $0, \ldots, 4$ gives the same pass verdict with $m^\star$ in a $\pm 2$ range of 397) we measure $m^\star = 397$, $\min \hat{p}_\phi = 9.94 \cdot 10^{-4}$, anisotropy $\lambda_1/\lambda_{m^\star} = 996$; the rank and projected-KDE thresholds pass, while anisotropy is descriptive. This is the setting used in Section D.7.

**Choice of thresholds.** We use $\tau_r = 10^{-3}$ as the numerical-rank threshold. Our KDE threshold $\tau_p = 10^{-6}$ is about four orders of magnitude below the observed mean density ($\overline{\hat{p}_\phi} \approx 2.18 \cdot 10^{-2}$).

**Threshold sensitivity.** On a $5 \times 5$ grid $\tau_r \in \{10^{-4} \ldots 10^{-2}\}$, $\tau_p \in \{10^{-8} \ldots 10^{-4}\}$, $m^\star$ varies from 512 to 118 ($510, 397, 232$ at $\tau_r = 3{\cdot}10^{-4}, 10^{-3}, 3{\cdot}10^{-3}$); the diagnostic thresholds are met in every cell.

The released code passes the three checks on a synthetic $\mathcal{N}(0, I_3)$ sample and fails check (i) on a rank-1 Gaussian (`audit_features.py`).

A finite-resolution projected KDE cannot rule out density zeros below its bandwidth or singularity in unobserved directions.

# J IMPLEMENTATION DETAILS

## J.1 OUTPUT PARAMETERIZATIONS

The tables specify the heads used in Section 4. Write $\sigma(t) = \mathrm{softplus}(t)$, $A_j(z) = w_j^\top z + b_j$ for a learned affine map, and $B_j(z) = w_j^\top z$ for a learned map without bias. Each subscript denotes a separate set of parameters. $M_j$ is a ReLU MLP with two hidden layers of width 50 and biases in every layer. The fixed output offsets are $\epsilon = 10^{-4}$ and $\eta = 10^{-3}$. All head parameters are learned unless stated otherwise.

For the scalar heads, $\gamma = H_\gamma(z)$ and $\alpha = \sigma(H_\alpha(z)) + 1.1$. The column $H$ specifies whether these maps are affine or MLPs. A symbol $\theta$ denotes a learned scalar shared across inputs.

Table 19: Scalar output heads. Together with the definitions above, each row specifies all four NIG outputs. Vanilla variants share the same head architecture.

| Experiment and method | $H$ | $\nu$ | $\beta$ |
|---|---|---|---|
| Scalar synthetic: ours | $M$ | $\sigma(\theta_\nu)$ | $\sigma(B_\beta(z))$ |
| Scalar synthetic: vanilla | $M$ | $\sigma(M_\nu(z)) + \epsilon$ | $\sigma(M_\beta(z)) + \epsilon$ |
| CIFAR / UCI: ours | $A$ | $\sigma(\theta_\nu)$ | $\sigma(B_\beta(z)) + \epsilon$ |
| CIFAR / UCI: vanilla | $A$ | $\sigma(A_\nu(z)) + \epsilon$ | $\sigma(A_\beta(z)) + \epsilon$ |
| Per-sample synthetic: per-sample | $M$ | $\sigma(A_\nu(z))$ | $\sigma(A_\beta(z))$ |
| Per-sample synthetic: global baseline | $M$ | $\sigma(\theta_\nu)$ | $\sigma(A_\beta(z))$ |

Both synthetic suites use a fixed Gaussian feature map. CIFAR uses the same frozen ResNet-18 features for all heads. UCI Arm A pretrains a trunk by Gaussian likelihood, then freezes only that trunk; its evidential heads remain trainable. Arm B and the vanilla variants train the ReLU trunk jointly with their heads. The UCI trunk has one hidden layer of width 50, or 100 for protein. The separate known-nuisance recovery diagnostic fixes $\gamma, \alpha$ to their generating functions.

The scalar-synthetic global head uses the exact bias-free class in Corollary B.2. The real-feature global heads add the fixed offset shown in Table 19; they have no learned bias in the scale preactivation and no free scale multiplier. The global baseline in the per-sample recovery study deliberately retains a learned scale bias. It is a different class. In the scale-closed orbit diagnostics, a learned positive multiplier is added to the scale output; these diagnostics are reported separately from the method comparisons.

For QM9, $z$ is the jointly learned SchNet embedding and $\mu_0 = A_\mu(z)$ is vector-valued. A network with one ReLU hidden layer of width 64 and biased affine layers outputs the lower-triangular entries. Let $L_+$ use diagonal entries $\sigma(\ell_{ii}) + \epsilon$, and let $L_{\exp}$ use $\exp(\ell_{ii})$; off-diagonal entries are unconstrained. Define $T(P) = P/[\mathrm{tr}(P)/d]$ and $q = \sigma(\theta_q) + \eta$. Here $q$ is the predictive degrees of freedom, so the NIW parameter is $\nu_W = q + d - 1$.

Table 20: QM9 output heads, with $d = 3$. All three methods jointly train SchNet and their heads. Coupled NIW uses the fixed ratio specified in Section H.

| Method | $\kappa$ | $\Psi$ | $\nu_W$ |
|---|---|---|---|
| Ours | $\sigma(\theta_\kappa)$ | $[\sigma(B_s(z)) + \epsilon]T(L_+L_+^\top)$ | $q + d - 1$ |
| Uncoupled NIW | $\sigma(A_\kappa(z)) + \epsilon$ | $[\sigma(A_s(z)) + \epsilon]L_+L_+^\top$ | $q + d - 1$ |
| Coupled NIW | $\nu_W/r$ | $\nu_W L_{\exp}L_{\exp}^\top$ | $d + 6 + 5\tanh(A_\nu(z)) + \eta$ |

No free global multiplier is present in the main method comparisons. The scalar evidence-regularized variants use the NIG penalty in Section 2.1 with coefficient 0.1; DER+MAE additionally uses $0.1|y - \gamma|$. Their unregularized counterparts and all three QM9 methods use predictive NLL alone. Freezing a representation does not establish the feature-density assumption; the real-data settings and the theorem-class synthetic settings are distinguished in their experiment descriptions.

**Compute.** Each training run uses a single H200 GPU. Frozen ResNet-18 features are computed once per data realization and shared across the corresponding CIFAR heads. Training budgets and seed counts are specified with each experiment.

**Scalar synthetic (Section 4.3).** Input $x \in \mathbb{R}^{10} \sim \mathcal{N}(0, I_{10})$; projection $\phi$ a fixed Gaussian matrix $\in \mathbb{R}^{10\times 3}$ seeded at 0. Ground-truth $a^*, b^* \in \mathbb{R}^3$ for $\gamma^*, \alpha^*$ and $w^* \in \mathbb{R}^3$ for $\beta^*$ drawn once at $w^*$-scale 3.0. Adam LR $10^{-3}$, 1000 epochs for the reported $n = 8000$ recovery/dispersion runs, full-batch, no weight decay, gradient-norm clip 1.0. $\nu_{\text{global}}$ initialized at $\text{softplus}^{-1}(1)$.

**Multivariate synthetic (Section D.4).** $d = 3$ output; $\nu_t^* = 4$; $S^*$ Cholesky-synthesized and trace-normalized to $\text{tr}(S^*)/d = 1$. Cholesky head for $\hat{S}$ with softplus diagonal. Diverse-init runs sample $a_{\text{init}}, \kappa_{\text{init}} \sim U[0.3, 3.0)$ with seed-tied initialization.

**UCI-9.** MLP trunk, 1 hidden layer of 50 ReLU units (Amini 2020 standard). Stage-1 pretrain: Gaussian NLL, 40 epochs, Adam $10^{-3}$. Stage-2 DER heads: Student-$t$ NLL, 40 epochs. Protein uses 100 hidden units. Standard 20-split protocol (Amini 2020), protein 5-split. All numbers reported in original units (un-standardized per the Amini 2020 reporting convention).

**Reproducibility.** The anonymous code supplement provides the implementations, experiment configurations, reproduction scripts and instructions needed to regenerate the reported comparisons and figures.

**Pointwise stability.** For a positive quantity $q_s(x)$ predicted by seed $s$, we report

$$\overline{\text{CV}}(q) = \frac{100}{m} \sum_{i=1}^{m} \frac{\text{sd}_s[q_s(x_i)]}{\text{mean}_s[q_s(x_i)]},$$

where sd is the sample standard deviation. The inputs $x_i$ are shared across seeds. For a global parameter, this reduces to its across-seed CV. On UCI, the statistic is computed within each split before averaging over splits. Epistemic stability uses $\beta/[\nu(\alpha - 1)]$ for NIG heads and the trace of the epistemic covariance for multivariate heads.

## K TWO-STAGE TRAINING DETAIL

Corollary B.2 requires a fixed $\phi$. Jointly training $\phi$ with $w$ enlarges the effective $\beta$-class to $\{x \mapsto \text{softplus}(w^\top \phi_\theta(x)) : w, \theta\}$, which can fail scale-separation. Our implementation pretrains a trunk by minimizing heteroscedastic Gaussian NLL (Stage 1) and then trains $(\gamma, \alpha, \beta, \nu_{\text{global}})$ on the precomputed features with the trunk fully frozen (Stage 2). A stop-gradient on the $\beta$-branch is not enough: the $\gamma$- and $\alpha$-heads still move the trunk.

*Remark* K.1 (Realizability bridge). Theorem 3.2 is a population statement. Under realizability, any strictly proper score has its unique population minimizer at the true predictive (Gneiting & Raftery, 2007); injectivity under scale-separation then yields $P_X$-a.e. recovery of $(\gamma, \nu, \alpha, \beta)$. Finite-sample $\hat{\nu}$ tightness is governed by the loss surface and sample size. Without realizability, the score-optimal predictive in the constrained class becomes the target.