Published at ProbML 2026

Identifiability, Fisher Information, and Amortized Inference for Heterogeneous Diffusion from Discrete-Time Noisy Observations

Yeo Zhen Yuan

Department of Statistics and Data Science. National University of Singapore.

Correspondence to zhenyuan.yeo13@sps.nus.edu.sg

Abstract

Estimating diffusion coefficients from discrete-time noisy particle trajectories is a core problem in single-particle tracking, yet its statistical foundations remain incomplete. We establish exact identifiability structure for single-species and heterogeneous-population models, derive closed-form Fisher information bounds, and show how these results jointly characterize the feasibility regime for practical inference in terms of the normalized diffusion scale $\alpha = 2D\Delta t/\sigma^2$ and effective particle occupancy $\beta = \rho\sigma^2$. A key finding is that one-step increment distributions alone cannot separate the diffusion coefficient from localization noise, but temporal autocorrelation structure resolves this ambiguity without additional calibration. For heterogeneous populations, identifiability holds under a variance separation condition, and Fisher information for rare components degrades quadratically with mixture weight, setting a fundamental limit on what unlabeled trajectory data can recover. We use these theoretical results to derive and validate an amortized inference architecture: time-averaging aggregation, factored posterior parameterization, and pretraining objective each follow directly from the theory, and on simulated data the trained posterior respects the κ-level-set non-identifiability and recovers a visibly two-mode density under mixture label ambiguity.

1. Introduction

Estimating the diffusion coefficient D from single-particle tracking (SPT) data is central to quantitative biophysics: D reports on molecular size, membrane viscosity, confinement, and active transport (Manzo and Garcia-Parajo, 2015; Shen et al., 2017). The standard model, which we take as our scope, observes a Brownian particle at discrete frame intervals Δt with additive isotropic Gaussian localization noise of variance σ^2, arising from photon shot noise and fitting errors of the optical point-spread function (PSF); extensions to anomalous or confined diffusion are discussed in Section 6. Classical mean-squared displacement (MSD) analysis (Qian et al., 1991; Saxton and Jacobson, 1997) provides a consistent estimator; the MSD curve's variance structure and fit-length optimization under localization noise have been analyzed by Michalet (2010), and optimal linear MSD variants derived by Vestergaard et al. (2014). However, MSD estimators are suboptimal at finite sample sizes, their variance is poorly characterized for heterogeneous populations, and they require trajectory linking, a preprocessing step that accumulates errors at high particle densities.

Modern SPT experiments present two deeper difficulties that MSD analysis is not designed to address. First, many experiments involve multiple co-existing molecular species with distinct diffusion

behaviors in spatially varying environments: inference for such heterogeneous populations has been approached via variational Bayes (Persson et al., 2013), hidden Markov models (Monnier et al., 2015; Slator and Burroughs, 2018), and Bayesian nonparametrics (Calderon and Arora, 2009), but the fundamental statistical limits of these problems have not been derived. Second, modern high-throughput imaging produces thousands of image sequences per experiment, so inference methods must scale without per-dataset likelihood evaluation. Amortized and simulation-based inference address this issue (Cranmer et al., 2020; Lueckmann et al., 2021; Papamakarios et al., 2018). Deep learning has also been applied to SPT for diffusion mode classification (Granik et al., 2019) and coefficient regression (Kowalek et al., 2019; Thapa et al., 2018). However, these approaches are rarely grounded in or validated against information-theoretic limits.

This paper closes both gaps simultaneously. We derive the complete identifiability structure of the noisy diffusion model, provide Fisher information bounds for single-species and mixture settings, construct a feasibility phase diagram, and build an amortized inference architecture whose structural choices are justified by theory. We ran experiments to test these theory guided models.

Contributions. Several of the building blocks employed in this work are established: Gaussian increment laws under additive localization noise are classical, and the identifiability of finite Gaussian scale mixtures follows from standard results (Teicher, 1963; Yakowitz and Spragins, 1968). Our contribution lies in the synthesis of these components and in extending them to the heterogeneous setting. Specifically:

1. We establish the identifiability structure of the noisy diffusion model as a formal theorem, demonstrating that one-step increments collapse the parameter pair into the scalar $\kappa = D\Delta t + \sigma^2$, while temporal correlations restore joint identifiability of (D, σ^2) (Appendix C). A detailed comparison with prior single-species treatments is given in Appendix A.

2. We provide the closed-form Fisher information $\mathcal{I}(D) = d\Delta t^2/(2\kappa^2)$ and characterize its three asymptotic regimes in terms of $\alpha = 2D\Delta t/\sigma^2$ (Proposition 1), identifying the information-rate-optimal frame interval $\Delta t^* = \sigma^2/D$.

3. For K-species mixtures with piecewise-constant diffusion fields, we establish identifiability under a variance separation condition, derive the per-component Fisher information, and obtain a quadratic information penalty w_j^2 for rare components (Theorem 3, Proposition 5, and Corollary 6). To the best of our knowledge, these mixture-level results do not appear in prior work.

4. We identify the two dimensionless parameters (α, β) that govern practical feasibility and characterize the resulting failure regimes as direct corollaries of the theorems (Section 3); an illustrative phase diagram is given in Appendix B.

5. We design an amortized inference network whose architectural components, time-averaging aggregation, lag-1 encoder feature, posterior family, and pretraining target, are each motivated by the preceding theoretical results (Table 1). On simulated data, the network respects the predicted κ-level-set non-identifiability on a probe grid and recovers a visibly two-mode posterior under mixture label ambiguity, with the NLL and Bayes MSE rankings tracking the expressivity of each posterior family (Section 4). The correspondence between theoretical results and architectural choices is, to our knowledge, novel.

Organization. Section 2 treats the single-species model. Section 3 extends to heterogeneous mixtures and identifies the two dimensionless parameters (α, β) that govern practical feasibility, with an illustrative phase diagram given in Appendix B. Section 4 presents the architecture and experiments. A detailed comparison with prior single-species treatments, together with a broader discussion of related work, is provided in Appendix A. Formal identifiability proofs for the single-species model are deferred to Appendix C, the remaining proofs appear in Appendix D, and the four posterior families used in Section 4 are specified in Appendix E.

2. Single-Species Model

2.1. Observation Model

Let $d \geq 1$ be the spatial dimension. A single particle evolves in an interval, Δt, as

$$X_{n+1} = X_n + \sqrt{2D\Delta t}\,\xi_n, \quad \xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0,\, I_d)\,, \tag{1}$$

and is observed with additive localization noise, σ^2,

$$Y_n = X_n + \varepsilon_n, \quad \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}\big(0,\, \sigma^2 I_d\big)\,, \tag{2}$$

with $\{\xi_n\}$ and $\{\varepsilon_n\}$ independent. The observed increment is $\Delta Y_n := Y_{n+1} - Y_n$.

2.2. Identifiability

The first structural fact is that the one-step increment distribution does not distinguish D from localization variance once they appear through the single combination

$$\kappa = D\Delta t + \sigma^2. \tag{3}$$

Hence marginal increments alone identify only an effective variance scale. The missing information is purely temporal. In fact, the lag-1 increment covariance satisfies

$$\mathrm{Cov}(\Delta Y_n, \Delta Y_{n+1}) = -\sigma^2 I_d, \tag{4}$$

so the localization variance can be recovered directly from the minimal first-order temporal statistic, after which D follows from the lag-0 covariance (cf. Berglund, 2010, eq. (8), $R = 0$). This decomposition cleanly separates marginal information, which determines κ, from temporal information, which determines σ^2. Figure 5 in Appendix C illustrates the collapse: all points on a given contour of κ produce identical one-step increment distributions and are separable only through (4). A formal statement and proof are given there.

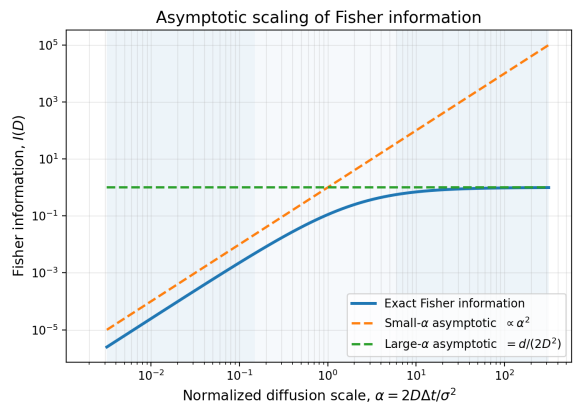

Figure 1. Fisher information $\mathcal{I}(D)$ versus $\alpha = 2D\Delta t/\sigma^2$ on log-log axes ($D = 1$, $d = 2$). The dashed line is the high-α plateau $d/(2D^2)$. Three regimes are marked.

2.3. Fisher Information

Proposition 1 (Fisher information, single species). *Assume σ^2 known. The Fisher information for D from a single increment $\Delta Y_n \in \mathbb{R}^d$ is*

$$\mathcal{I}(D) = \frac{d\,\Delta t^2}{2\kappa^2}, \qquad \kappa = D\Delta t + \sigma^2. \tag{5}$$

For N independent increments, $\mathcal{I}_N(D) = N\mathcal{I}(D)$, and $\mathrm{Var}(\widehat{D}) \geq 2\kappa^2/(Nd\Delta t^2)$.

Proof. See Appendix D. □

For use in the mixture analysis below, we define the *single-species Fisher information benchmark* for a scale κ_j as

$$\mathcal{I}_{\text{single}}(\kappa_j) \;:=\; \frac{\Delta t^2}{2\kappa_j^2}, \tag{6}$$

which is the one-dimensional ($d = 1$) specialization of (5): the Fisher information that would be available for a single Brownian species at scale κ_j in the absence of any mixture.

In terms of the normalized diffusion scale

$$\alpha \;:=\; \frac{2D\Delta t}{\sigma^2}, \tag{7}$$

equation (5) becomes $\mathcal{I}(D) = \frac{d}{2D^2}\left(\frac{\alpha}{\alpha+2}\right)^2$, revealing two asymptotic limits and a transition region, illustrated in Figure 1. When $\alpha \ll 1$ (localization dominated), $\mathcal{I}(D) \approx d\Delta t^2/(2\sigma^4)$ and information grows quadratically with Δt. When $\alpha \gg 1$ (motion dominated), $\mathcal{I}(D) \approx d/(2D^2)$ and information saturates. The information rate $\mathcal{I}(D)/\Delta t$ is maximized at $\Delta t^* = \sigma^2/D$, equivalently $\alpha = 2$, which gives a principled criterion for frame-rate selection.

3. Heterogeneous Populations and Mixture Models

3.1. Model Setup

Let $\Omega \subset \mathbb{R}^d$ be the spatial domain of the experiment. Partition Ω into M non-overlapping cells $\{C_m\}_{m=1}^M$. Each particle carries a permanent species label $k \in \{1, \ldots, K\}$; a particle of species k in cell C_m has diffusion coefficient $D_{km} > 0$ and follows (1) and (2) with $D = D_{km}$. A randomly drawn particle has species k with probability π_k and occupies cell C_m with stationary probability ρ_m, independently. Define

$$\kappa_{km} := D_{km}\Delta t + \sigma^2, \quad w_{km} := \pi_k \rho_m, \quad \sum_{k,m} w_{km} = 1. \tag{8}$$

The two-index (k, m) notation is structural: it mirrors the species-plus-cell decomposition typical of fluorescence microscopy. No spatial correlation between cells is assumed, and species labels and cell occupations are taken independent. Statistically, therefore, the increment law depends only on the unordered flat collection of components, and we collapse (k, m) to a single index $j \in \{1, \ldots, J\}$ with $J = KM$ when no weight structure is assumed, writing w_j, κ_j, D_j. Recovering the product structure $w_{km} = \pi_k \rho_m$ (and the rectangular array $\{D_{km}\}$) requires additional information beyond the increment distribution alone, as discussed in the remark below.

3.2. Identifiability

The increment of a randomly drawn particle satisfies

$$\Delta Y_n \sim \sum_{j=1}^{J} w_j \, \mathcal{N}(0,\, 2\kappa_j I_d)\,, \tag{9}$$

a finite isotropic Gaussian scale mixture.

Assumption 2 (Variance separation). *All $\kappa_1, \ldots, \kappa_J$ are pairwise distinct.*

Theorem 3 (Mixture identifiability). *Assume σ^2 known.*

1. *Under Assumption 2, the collection $\{(w_j, \kappa_j)\}_{j=1}^J$, and hence $\{D_j\}$, is uniquely identifiable from (9) up to label permutation (Teicher, 1963; Yakowitz and Spragins, 1968).*

2. *If two components share the same κ_j, only their pooled weight is identifiable; no estimator can resolve them from increment data alone.*

3. *The weighted mean $\bar{\kappa} = \sum_j w_j \kappa_j$ is always identifiable: $\bar{\kappa} = \frac{1}{2d} \operatorname{tr} \operatorname{Cov}(\Delta Y_n)$.*

Proof. See Appendix D. □

Remark 4 (Structured weight tensors). *When $w_{km} = \pi_k \rho_m$, recovering the full array $\{D_{km}\}$ and the factorization requires additional structure beyond increment data alone, such as labeled trajectories or spatial co-localization information.*

3.3. Fisher Information for Mixture Components

Proposition 5 (Fisher information, K-species mixture). *Assume σ^2 known and $d = 1$. The Fisher information for D_j from one increment is*

$$\mathcal{I}(D_j) = w_j^2 \, \Delta t^2 \, \mathcal{R}_j(\Theta), \tag{10}$$

where $\mathcal{R}_j(\Theta) = \int [\partial_\kappa \phi(\delta; \kappa_j)]^2 / p(\delta; \Theta) \, d\delta$ is a mixture penalty factor. Moreover,

$$\mathcal{I}(D_j) \leq \frac{w_j \, \Delta t^2}{2\kappa_j^2}, \tag{11}$$

with equality if and only if $J = 1$.

Proof. See Appendix D. □

Corollary 6 (Quadratic penalty for rare components). *The ratio of mixture to single-species Fisher information obeys*

$$\frac{\mathcal{I}(D_j)}{\mathcal{I}_{\text{single}}(\kappa_j)} = \frac{\mathcal{I}(D_j)}{\Delta t^2 / (2\kappa_j^2)} \leq w_j. \tag{12}$$

The absolute information itself is

$$\mathcal{I}(D_j) = w_j^2 \, \Delta t^2 \, \mathcal{R}_j(\Theta), \tag{13}$$

which contains the explicit prefactor w_j^2 from (10). Rare components are therefore penalized twice: once through the weight-scaled score magnitude (a factor w_j in the numerator of the score), and once through the mixture denominator $p(\delta; \Theta)$, which for $p \gg w_j \phi(\delta; \kappa_j)$ dilutes the jth component's contribution. The upper bound (11) shows that the relative *loss versus the single-species benchmark is at most w_j, so the* absolute *Fisher information for a rare component scales no faster than $w_j \cdot \mathcal{I}_{\text{single}}(\kappa_j)$, and its leading prefactor is quadratic in w_j. This is a fundamental limit, independent of the estimation algorithm.*

To quantify how much information is lost in practice, we evaluate the relative Fisher information $\mathcal{I}(D_j)/\mathcal{I}_{\text{single}}(\kappa_j)$ numerically for a two-component mixture ($J = 2$) across the joint space of mixture weight w_j and scale ratio $\kappa_{\text{other}}/\kappa_j$, comparing exact quadrature against Monte Carlo estimates at two sample sizes. Figure 2 shows that the upper bound w_j from Corollary 6 is never tight away from full scale separation, and that the additional loss due to scale overlap is most severe near the degenerate line $\kappa_{\text{other}} = \kappa_j$ where condition (Assumption 2) fails.

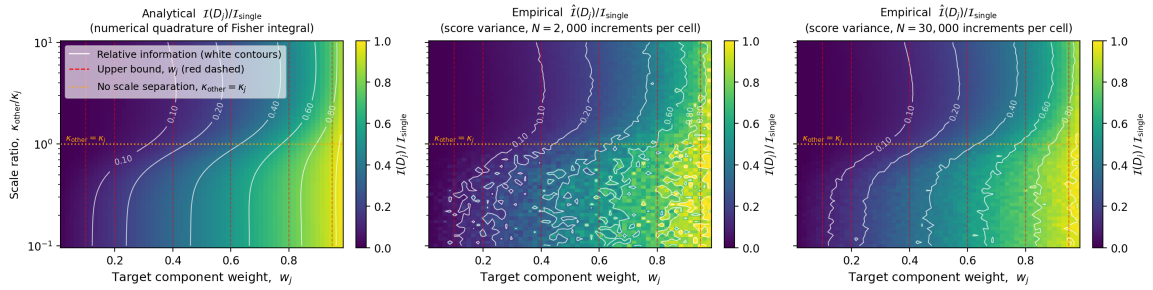

Figure 2. Relative Fisher information $\mathcal{I}(D_j)/\mathcal{I}_{\text{single}}(\kappa_j)$ for a target diffusion component in a two-component isotropic Gaussian scale mixture ($J = 2$, $d = 1$, $\kappa_j = 1$, $\Delta t = 1$), as a function of the target mixture weight w_j and the scale ratio $\kappa_{\text{other}}/\kappa_j$. The single-species benchmark is $\mathcal{I}_{\text{single}}(\kappa_j) = \Delta t^2/(2\kappa_j^2)$, defined in (6). Left: exact values obtained by numerical quadrature of the Fisher integral in Proposition 5. Center and right: Monte Carlo estimates via score variance averaging over $N = 2{,}000$ and $N = 30{,}000$ mixture increments per cell, respectively, confirming convergence of the empirical estimator to the analytical surface. Red dashed vertical lines show the upper bound w_j (Corollary 6), white contours trace the exact relative information. The orange dotted horizontal line marks $\kappa_{\text{other}} = \kappa_j$, the degenerate case of zero scale separation where Assumption 2 fails and the information collapses regardless of w_j. The gap between the white and red contours quantifies the additional information loss due to scale overlap beyond the mixture-weight penalty alone.

Operating regimes. Two dimensionless quantities govern practical feasibility in fluorescence microscopy, where many particles contribute to the same diffraction-limited image through the PSF: the normalized diffusion scale α from (7), and the effective particle occupancy

$$\beta \;:=\; \rho\,\sigma^2, \tag{14}$$

where ρ is the expected particle density and β counts overlapping PSF footprints per localization radius. Two failure modes follow directly from the theorems above. *Motion blur* occurs for large α: the effective scale κ becomes dominated by motion, so Proposition 1 gives $\mathcal{I}(D) \propto \kappa^{-2} \to 0$. *Overlap* occurs for large β: multiple emitters contribute to the same PSF region, effectively reducing usable component weights and triggering the quadratic w_j^2 penalty of Corollary 6. Both predictions are consequences of the theorems and are independent of any particular estimator. An illustrative phase diagram visualizing the joint feasible region, together with the simulation protocol used to produce it, is given in Appendix B.

4. Amortized Inference Architecture

4.1. Theory-Driven Design

Each architectural and training choice is grounded in one of the theoretical results derived above. Table 1 summarizes the connections. Three are worth emphasizing. First, time-averaging of per-frame embeddings is justified by the stationarity and 1-dependence of observed increments established in Proposition 7: because the increment process is strictly stationary and Gaussian with a single non-trivial lag, the sample moments S_0 and S_1 defined below are sufficient for the second-order structure, and their time-averaging is a lossless reduction at the level of second-order statistics. Second, pre-training on $\bar{\kappa}$ regression is motivated by Theorem 3 (part 3): the mixture-averaged effective scale $\bar{\kappa} = \sum_j w_j \kappa_j$ is always identifiable from the second moment alone, providing a well-posed and unconditionally supervised warmup target before end-to-end training on the full posterior. Third, wide posteriors in degenerate regimes are due to the non-identifiability established in Theorem 3 (part 2): when Assumption 2 is violated, no estimator can resolve the individual D_j, and the posterior mass should spread over the unresolvable manifold.

Theoretical result	**Architectural consequence**
Stationary, 1-dependent increments (Proposition 7)	Time-averaging of frame-pair embeddings in Stage 1
Non-identifiability under separation failure (Theorem 3, part 2)	Wide posteriors in degenerate regimes are correct
Partial identifiability of $\bar{\kappa}$ (Theorem 3, part 3)	Pretraining Stage 1 on $\bar{\kappa}$ regression
Quadratic Fisher penalty w_j^2 (Proposition 5, Corollary 6)	Importance-weighted simulation budget for rare components
Cramér–Rao bound (11)	Hard floor on posterior variance; misspecification diagnostic
Lag-1 covariance identifies σ^2 (Proposition 7)	Lag-1 inner product included as an encoder feature

Table 1. Theoretical grounding for each architectural choice.

4.2. Architecture

Rather than operating on raw image sequences, which require a full convolutional forward model, we work directly with hand-crafted sufficient statistics of the observed increments. This choice is justified by Proposition 7 and Theorem 3 (part 3): the second moment and the lag-1 covariance together capture all identifiable information about D and σ^2 from the increment process. Further training details are in Appendix E.

Posterior families. We compare four families for $q_\phi(\log D \mid x)$, each modeling the posterior over $\log D$ to enforce positivity. A **log-normal** (LogNormal) posterior serves as the unimodal baseline (Kingma and Welling, 2013). A **mixture of log-normals** (MoLN) extends this to two components (Goodfellow et al., 2021), providing the minimal representational capacity needed for the bimodal posteriors predicted by Theorem 3 (part 2). A **log-Student-t** (StudentT) posterior is included as a heavy-tailed alternative (Jaakkola and Jordan, 1998; Futami et al., 2017). Finally, a **neural spline flow** (NSF) provides a fully flexible normalizing flow baseline (Durkan et al., 2019; Papamakarios et al., 2021; Cranmer et al., 2020). Full mathematical definitions of all four families are given in Appendix E.

5. Empirical Validation

We validate the architecture in three stages. We first verify that the trained posterior faithfully represents the identifiability structure established in Theorem 9, by probing how the posterior varies across level sets of $\kappa = D\Delta t + \sigma^2$ (Section 5.1). We then evaluate two regimes predicted by the theory: confounding from unknown σ^2 (Section 5.2), and posterior bimodality from mixture non-identifiability (Section 5.3). Full per-architecture numerical results and extended discussion are deferred to Appendix F.

5.1. Identifiability Consistency of the Amortized Posterior

Before evaluating performance, we verify that the trained posterior respects the non-identifiability structure of Theorem 9 (part 1): the law of ΔY_n depends on (D, σ^2) only through $\kappa = D\Delta t + \sigma^2$, so a posterior built from marginal features must be constant along level sets of κ. We probe this on a 7×7 grid of (D, σ^2) pairs with diagonal $\kappa = 2$, upper-right $\kappa > 2$, and lower-left $\kappa < 2$, using a

marginal-only feature set $(S_0, \log S_0)$; the full grid specification is in Appendix F. The MoLN and LogNormal posteriors on this grid are shown in Figures 6 and 7. Histograms are near-identical along each anti-diagonal of constant κ, while the true-D marker slides across a stationary shape: neither model resolves D within a level set, as required. Interestingly, the two families differ in the shape they assign to the Bayes posterior $p(D \mid \kappa)$: LogNormal forces a right-skewed unimodal density onto what is essentially a plateau, while MoLN covers the support more faithfully. All subsequent experiments restore the lag-1 feature S_1, which by Proposition 7 identifies σ^2 and lifts the κ-manifold confounding analyzed here.

5.2. Experiment A: unknown σ^2

With $D \sim \text{LogUniform}[0.5, 4.0]$, $\sigma^2 \sim \text{LogUniform}[0.2, 2.0]$, and $\Delta t = 0.1$ fixed, Theorem 9 (part 1) predicts that D is not identifiable from the marginal increment distribution: only κ is. At the test point shown in Figure 3, all four architectures produce broad right-skewed posteriors with Bayes MSE in a tight band (0.838–0.865); the four families are separated primarily by NLL and median calibration, not by MSE. NSF attains the best NLL (0.631) and near-ideal 50% coverage (0.49); MoLN is close behind (NLL = 0.706, $\text{Cov}_{50} = 0.48$) and shows a mild secondary shoulder at higher D, indicating that the manifold effect of the consistency check (Section 5.1) leaves a faint imprint on a single test posterior, with the two competing branches being low D with high σ^2 and high D with low σ^2. LogNormal and StudentT are structurally unimodal and show median undercoverage ($\text{Cov}_{50} \approx 0.39$) together with 90% coverage at the upper end (≈ 0.95), the joint signature of mild over-dispersion at the median with tails that over-cover. This experiment evaluates the downstream consequence of the (D, σ^2) confounding on the marginal posterior of D; a direct test of joint (D, σ^2) recovery via the lag-1 feature S_1 is flagged as future work in Section 6. Full per-architecture results are in Appendix F.2.

5.3. Experiment B: $K = 2$ mixture

With $D_A \sim \text{LogUniform}[0.4, 1.2]$, $D_B = 2.0$ fixed, equal mixing weights, and $\sigma^2 = 1.0$ known, Theorem 3 (part 2) combined with label ambiguity predicts a bimodal posterior over D_A: one mode near the true value, and a secondary mode induced by the D_B alternative explanation. At the test point shown in Figure 10, MoLN is the only architecture that produces a visibly two-mode posterior, with a primary mode in the D_A-prior support and a clear secondary mass at higher D. NSF concentrates a sharp peak in the secondary region with broad mass below, achieving the best NLL (2.355); MoLN is close in NLL (2.522) and attains the lowest Bayes MSE among finite-MSE families (0.726 versus 0.735 for LogNormal and 0.841 for NSF). LogNormal incurs a catastrophic NLL (78.165) because its unimodal density must compromise between D_A and D_B on each test dataset, assigning near-zero density at the true D_A for a substantial fraction of them. StudentT has divergent Bayes MSE under the mixture likelihood: its heavy tail produces an unbounded second moment in this regime, a more severe failure than the merely large MSE of the other unimodal families. Central-quantile coverage is poor across all four architectures ($\text{Cov}_{50} \approx 0.23$–$0.26$, $\text{Cov}_{90} \approx 0.51$–$0.58$); this is expected for bimodal posteriors measured by central intervals, and reflects a limitation of the diagnostic rather than of the architectures. A highest posterior-density coverage variant would separate MoLN from the unimodal families more sharply. Per-architecture results are in Appendix F.3.

5.4. Summary

The consistency check in Section 5.1 verifies that the amortized posterior respects the κ-manifold non-identifiability of Theorem 9 and exposes a posterior-family expressivity gap between LogNormal and MoLN. Exp A shows that identifiability failure from unknown nuisance parameters forces wide posteriors with separation between families dominated by NLL and median calibration rather than

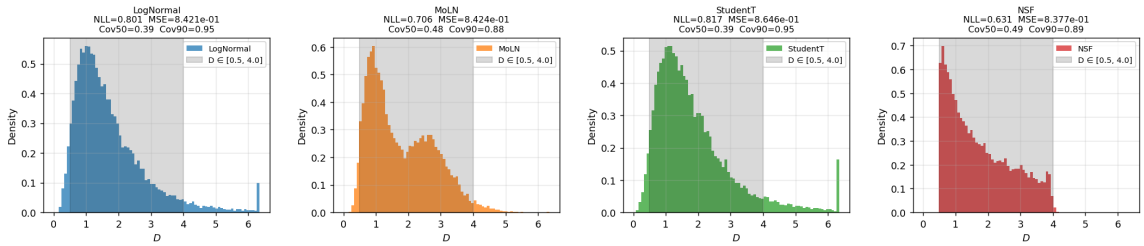

Figure 3. Posterior samples for all four architectures under Exp A (top row, unknown σ^2). In Exp A, all architectures produce broad right-skewed posteriors consistent with the D-σ^2 non-identifiability of Theorem 9 (part 1). For the parametric families (LogNormal, MoLN, StudentT) the plotted curves can equivalently be read as analytical posterior densities; for NSF the kernel density of samples is shown.

MSE, and that controlled-tail families (LogNormal, NSF) outperform heavy-tailed alternatives (StudentT). Exp B shows that mixture non-identifiability induces bimodality captured visibly only by MoLN, while NSF reaches the best NLL by concentrating a sharp secondary mode and LogNormal incurs catastrophic NLL from mode-forcing. The Cramér–Rao bound (11) provides a lower bound on posterior variance; the variances observed above the bound in Exp A are the expected consequence of the D-σ^2 confounding identified in Theorem 9 (part 1).

6. Conclusion

We have derived the complete identifiability and information-theoretic structure of diffusion coefficient estimation from discrete-time noisy observations, for both single-species and heterogeneous-population settings. The central results are three. First, marginal increment distributions collapse (D, σ^2) into a single scalar $\kappa = D\Delta t + \sigma^2$; the lag-1 autocovariance is the unique minimal temporal statistic that resolves this degeneracy without external calibration. Second, the Fisher information $\mathcal{I}(D) = d\Delta t^2/(2\kappa^2)$ has three qualitatively distinct regimes as a function of $\alpha = 2D\Delta t/\sigma^2$, with an optimal frame interval $\Delta t^* = \sigma^2/D$ that maximizes information rate. Third, in J-component mixtures with unknown labels, Fisher information for component j is bounded above by $w_j \Delta t^2/(2\kappa_j^2)$, establishing a quadratic information penalty for rare components that is a fundamental limit independent of the estimation algorithm. While we establish a quadratic Fisher information penalty w_j^2 for rare mixture components, it remains open whether this bound is achievable or whether the irregular mixture likelihood creates a fundamental gap between the Cramér–Rao bound and any estimator.

Empirically, the amortized network respects the κ-level-set non-identifiability on a 7×7 probe grid (Section 5.1), produces broad posteriors with a tight MSE band (0.838–0.865) under unknown σ^2 (Exp A), and recovers a visibly two-mode posterior under mixture label ambiguity only with MoLN, while LogNormal incurs catastrophic NLL (78.165) and StudentT diverges in MSE (Exp B).

Together, the theoretical framework and the amortized network provide both a quantitative guide for experimental design in fluorescence microscopy and a scalable tool for diffusion inference that is grounded in, and validated against, exact statistical limits.

Acknowledgments

This work was supported by the National University of Singapore (NUS) and the Institute for Digital Molecular Analytics and Science (IDMxS) under WBS A-8000517-06-00.

References

Andrew J Berglund. Statistics of camera-based single-particle tracking. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 82(1 Pt 1):011917, July 2010.

Christopher P Calderon and Karunesh Arora. Extracting kinetic and stationary distribution information from short MD trajectories via a collection of surrogate diffusion models. *J. Chem. Theory Comput.*, 5(1):47–58, January 2009.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proc. Natl. Acad. Sci. U. S. A.*, 117(48):30055–30062, December 2020.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *arXiv [stat.ML]*, June 2019.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. *arXiv [stat.ML]*, pages 813–822, October 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. https://mitpress.mit.edu/9780262035613/deep-learning/, December 2021. Accessed: 2026-3-21.

Naor Granik, Lucien E Weiss, Elias Nehme, Maayan Levin, Michael Chein, Eran Perlson, Yael Roichman, and Yoav Shechtman. Single-particle diffusion characterization by deep learning. *Biophys. J.*, 117(2):185–192, July 2019.

Tommi S Jaakkola and Michael I Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, pages 163–173. Springer Netherlands, Dordrecht, 1998.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv [stat.ML]*, December 2013.

Patrycja Kowalek, Hanna Loch-Olszewska, and Janusz Szwabiński. Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. *Phys. Rev. E.*, 100 (3-1):032410, September 2019.

Jan-Matthis Lueckmann, Jan Boelts, David S Greenberg, Pedro J Gonçalves, and Jakob H Macke. Benchmarking simulation-based inference. *arXiv [stat.ML]*, pages 343–351, January 2021.

Carlo Manzo and Maria F Garcia-Parajo. A review of progress in single particle tracking: from methods to biophysical insights. *Rep. Prog. Phys.*, 78(12):124601, December 2015.

Xavier Michalet. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 82(4 Pt 1):041914, October 2010.

Nilah Monnier, Zachary Barry, Hye Yoon Park, Kuan-Chung Su, Zachary Katz, Brian P English, Arkajit Dey, Keyao Pan, Iain M Cheeseman, Robert H Singer, and Mark Bathe. Inferring transient particle transport dynamics in live cells. *Nat. Methods*, 12(9):838–840, September 2015.

George Papamakarios, David C Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. *arXiv [stat.ML]*, pages 837–848, May 2018.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22:1–64, 2021.

Fredrik Persson, Martin Lindén, Cecilia Unoson, and Johan Elf. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods*, 10(3):265–269, March 2013.

H Qian, M P Sheetz, and E L Elson. Single particle tracking. analysis of diffusion and flow in two-dimensional systems. *Biophys. J.*, 60(4):910–921, October 1991.

M J Saxton and K Jacobson. Single-particle tracking: applications to membrane dynamics. *Annu. Rev. Biophys. Biomol. Struct.*, 26(1):373–399, 1997.

Hao Shen, Lawrence J Tauzin, Rashad Baiyasi, Wenxiao Wang, Nicholas Moringo, Bo Shuang, and Christy F Landes. Single particle tracking: From theory to biophysical applications. *Chem. Rev.*, 117(11):7331–7376, June 2017.

Paddy J Slator and Nigel J Burroughs. A hidden markov model for detecting confinement in single-particle tracking trajectories. *Biophys. J.*, 115(9):1741–1754, November 2018.

Henry Teicher. Identifiability of finite mixtures. *Ann. Math. Stat.*, 34(4):1265–1269, December 1963.

Samudrajit Thapa, Michael A Lomholt, Jens Krog, Andrey G Cherstvy, and Ralf Metzler. Bayesian analysis of single-particle tracking data using the nested-sampling algorithm: maximum-likelihood model selection applied to stochastic-diffusivity data. *Phys. Chem. Chem. Phys.*, 20(46):29018–29037, November 2018.

Christian L Vestergaard, Paul C Blainey, and Henrik Flyvbjerg. Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 89(2): 022726, February 2014.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *Ann. Math. Stat.*, 39(1):209–214, February 1968.

A. Relation to Prior Work and Limitations

The single-species portion of our analysis connects to two parallel branches of the SPT literature: the spectral maximum-likelihood treatment of Berglund (2010) and the MSD-based analysis of Michalet (2010). We make the correspondence explicit here.

Increment covariance and the lag-1 identity. Berglund (2010) derives the full covariance of observed displacements for a camera-based tracking experiment with motion blur coefficient $R \in [0, 1/4]$ (his eq. (8)). The lag-1 identity (4) used throughout our paper is the $R = 0$ (delta-pulse illumination, or equivalently ideal snapshot sampling) specialization of that formula. Earlier work in the SPT literature had noted pieces of this structure: static localization noise contributes to the lag-0 MSD, and motion blur provides corrections; Berglund (2010) unifies these into a single 1-dependent Gaussian framework. Our contribution in this part is to reframe the lag-1 identity as the basis of an identifiability theorem (Theorem 9) rather than as an ingredient of a moment estimator or a spectral MLE.

Single-species Fisher information. Proposition 1, stating $\mathcal{I}(D) = d\Delta t^2/(2\kappa^2)$, is equivalent in the $R = 0$ limit to the diagonal element I_D^{11} of the spectral Fisher matrix derived in Berglund (2010, eqs. (20)–(21)). The two derivations differ in style: Berglund works in the Fourier basis using the circulant approximation of the Toeplitz covariance, which is asymptotically equivalent to the exact likelihood for large N; we work directly from the per-increment Gaussian log-likelihood, which is exact for any N but slightly more verbose. We keep our derivation in the paper (Appendix D) for self-containedness and because it is the form used in the mixture proofs.

MSD-based analyses. A parallel line of work, represented by Michalet (2010), analyzes the same physical model through the mean-squared-displacement (MSD) curve rather than the increment likelihood. His offset $\epsilon = 4\sigma^2$ (his eq. (14)) is the σ^2-contribution to our lag-0 covariance $\operatorname{tr}\operatorname{Cov}(\Delta Y_n) = 2d(D\Delta t + \sigma^2)$ at $d = 2$, and his reduced localization error $x = \sigma^2/(D\Delta t)$ (his eq. (20)) is exactly $2/\alpha$ in our notation, so the dimensionless quantity governing statistical difficulty is the same across the two analyses. The difference is the level at which the analysis is conducted: Michalet (2010) derives the variance and covariance of the MSD curve (his eqs. (D13), (F13)) and studies the estimator-level relative error $\sigma_{\widehat{D}}/D$ as a function of the number of fitting points p and trajectory length N, arriving at the empirical optimum $p_{\min} \approx \lfloor 2 + 2.7\, x^{1/2} \rfloor$ (his eq. (30)). We work one level up, at the Fisher information of the increment likelihood itself, which is estimator-independent and yields the Cramér–Rao bound as a lower bound that any MSD-based estimator must obey. Michalet (2010) himself notes, in a "Note added," that the maximum-likelihood treatment of Berglund (2010) is a distinct formalism from his MSD analysis; our Proposition 1 and the identifiability theorems of Appendix C belong to that maximum-likelihood branch, not the MSD branch.

Contributions beyond prior work. The mixture-level results, to the best of our knowledge, do not appear in the existing literature. Theorem 3 (identifiability of K-species mixtures under variance separation), Proposition 5 (per-component Fisher information with the mixture penalty $\mathcal{R}_j(\Theta)$), Corollary 6 (the quadratic w_j^2 penalty for rare components), and the theory-to-architecture mapping in Table 1 do not appear in Berglund (2010) or related single-species treatments. Classical finite-mixture identifiability results (Teicher, 1963; Yakowitz and Spragins, 1968) are used as ingredients but do not themselves yield the per-component Fisher bound.

Limitations and future work. Our theorems assume Brownian motion with additive isotropic Gaussian localization noise; anomalous diffusion, confined motion, and state-switching dynamics lie outside their scope. We assume increments are already extracted, which in dense-image regimes requires trajectory linking and is its own source of error. The empirical evaluation uses simulated data only, chosen so that the true parameters are known and comparisons against the Cramér–Rao floor are meaningful; validation on real SPT datasets with ground-truth calibrations is natural next

work. Finally, although the lag-1 feature S_1 identifies σ^2 (Proposition 7), in Exp A we evaluate only the marginal posterior of D under unknown σ^2; a direct test of joint (D, σ^2) recovery using S_1 is an immediate follow-up.

B. Feasibility Phase Diagram

The results of Sections 2 and 3 characterize inference in the trajectory model. In fluorescence microscopy, however, many particles contribute to the same diffraction-limited image through the PSF, making inference substantially harder. The two failure-mode predictions stated in the "Operating regimes" paragraph of Section 3 (motion blur for large α, overlap for large β) follow directly from Proposition 1 and Corollary 6. The precise location of the boundary between feasible and infeasible regions, however, depends on the forward model and estimator and is therefore simulation-driven rather than predicted by the theorems.

Simulation protocol. For each grid point (α, β) we generate synthetic image sequences from a forward model consisting of Brownian dynamics (1), Gaussian PSF image formation, and Poisson photon noise. We estimate D using the moment-based maximum likelihood estimator implied by Theorem 9 (part 3),

$$\widehat{D} \;=\; \frac{1}{\Delta t}\left(\frac{1}{2d}\,\mathrm{tr}\,\widehat{\mathrm{Cov}}(\Delta Y_n) - \sigma^2\right), \tag{15}$$

with σ^2 known, and report the relative error $E = |\widehat{D} - D|/D$ averaged over replicates. Figure 4 is illustrative: it is produced by a stylized parametric error surface and is intended to visualize the qualitative regime structure, not to locate a precise transition boundary.

The diagram in Figure 4 translates directly into experimental design: for a given D and imaging system σ, the feasible region specifies the admissible range of frame intervals and particle densities. The feasible region is widest near $\alpha = 2$, consistent with the optimal frame interval $\Delta t^* = \sigma^2/D$ from Section 2.3.

C. Single-Species Identifiability

This appendix makes precise the claim used in the main text: marginal increment statistics identify only the collapsed scale $\kappa = D\Delta t + \sigma^2$, whereas first-order temporal correlation recovers σ^2 and then D. At the end, Figure 5 visualizes the identifiability structure: contours of constant $\kappa = D\Delta t + \sigma^2$ define equivalence classes that are indistinguishable from one-step increments and are resolved only by the lag-1 covariance (4).

C.1. Increment process and covariance structure

We first write the observed increment explicitly:

$$\Delta Y_n = Y_{n+1} - Y_n = \sqrt{2D\Delta t}\,\xi_n + \varepsilon_{n+1} - \varepsilon_n. \tag{16}$$

This representation exposes two distinct sources of randomness: the latent Brownian displacement and the differenced localization noise.

Proposition 7 (Increment covariance structure). *Under (1) and (2), the process $(\Delta Y_n)_{n\geq 0}$ is a stationary centered Gaussian process with*

$$\Delta Y_n \sim \mathcal{N}\big(0,\, 2(D\Delta t + \sigma^2)I_d\big)\,, \tag{17}$$

$$\mathrm{Cov}(\Delta Y_n, \Delta Y_{n+1}) = -\sigma^2 I_d, \tag{18}$$

$$\mathrm{Cov}(\Delta Y_n, \Delta Y_{n+h}) = 0, \qquad |h| \geq 2. \tag{19}$$

In particular, (ΔY_n) is 1-dependent and is not independent unless $\sigma^2 = 0$.

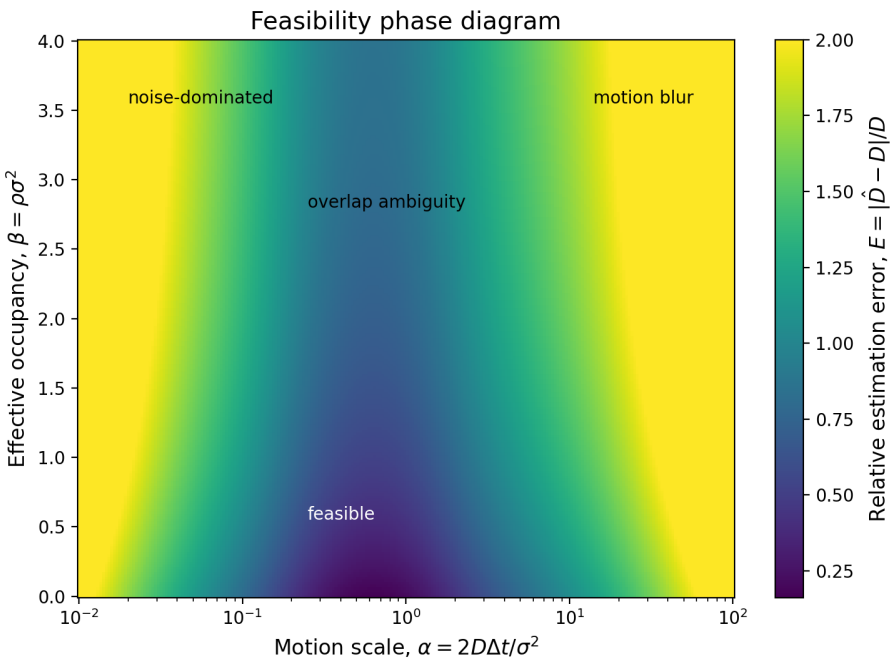

Figure 4. Feasibility phase diagram (illustrative) for diffusion inference, displaying the relative estimation error $E = |\hat{D} - D|/D$ over the two-dimensional parameter space defined by the normalized diffusion scale $\alpha = 2D\Delta t/\sigma^2$ and the effective particle occupancy $\beta = \rho\sigma^2$. Three distinct failure regimes are visible. At small α, particle displacements between frames are dominated by localization noise, consistent with the $\mathcal{I}(D) \approx d\Delta t^2/(2\sigma^4)$ scaling of Proposition 1. At large α, motion blur reduces the per-increment information as $\mathcal{I}(D) \propto \kappa^{-2}$ with $\kappa = D\Delta t + \sigma^2 \approx D\Delta t$. At large β, high particle density produces PSF overlap, inducing mixture ambiguity and the quadratic Fisher information penalty of Corollary 6. A connected feasible region of low error persists near $\alpha \sim 1$ and $\beta \lesssim 1$, defining the practical operating regime for reliable diffusion coefficient recovery. The precise location of the boundary depends on the estimator and forward model and is not a prediction of the theorems above.

Proof. Equation (16) shows that ΔY_n is a linear combination of independent centered Gaussian variables, hence it is centered Gaussian. Its covariance is

$$\mathrm{Cov}(\Delta Y_n) = 2D\Delta t\, I_d + \mathrm{Cov}(\varepsilon_{n+1} - \varepsilon_n) = 2D\Delta t\, I_d + 2\sigma^2 I_d = 2(D\Delta t + \sigma^2) I_d, \tag{20}$$

which proves (17).

For the lag-1 covariance, write

$$\Delta Y_{n+1} = \sqrt{2D\Delta t}\, \xi_{n+1} + \varepsilon_{n+2} - \varepsilon_{n+1}. \tag{21}$$

Every cross term between (16) and this expression vanishes by independence except the shared localization-noise variable ε_{n+1}. Therefore

$$\mathrm{Cov}(\Delta Y_n, \Delta Y_{n+1}) = \mathrm{Cov}(\varepsilon_{n+1} - \varepsilon_n,\, \varepsilon_{n+2} - \varepsilon_{n+1}) = -\mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_{n+1}) = -\sigma^2 I_d, \tag{22}$$

which proves (18).

If $|h| \geq 2$, then ΔY_n and ΔY_{n+h} depend on disjoint families of independent Gaussian variables, so their covariance is zero. This gives (19). Since the underlying innovations are identically distributed across time, the process is stationary. The 1-dependence claim follows from vanishing covariance at lags of magnitude at least 2, and the failure of independence for $\sigma^2 > 0$ follows from the nonzero lag-1 covariance. □

Remark 8. *The proposition yields a particularly clean interpretation. The lag-0 covariance aggregates physical motion and localization noise through the single collapsed parameter $\kappa = D\Delta t + \sigma^2$, while the lag-1 covariance isolates the localization-noise contribution exactly.*

C.2. Identifiability theorem

We now state the formal identifiability result used in the main text.

Theorem 9 (Single-species identifiability). *Let $\kappa := D\Delta t + \sigma^2$. Then:*

1. *The marginal law of a one-step increment depends on (D, σ^2) only through κ. Consequently, if σ^2 is unknown, D is not identifiable from the one-step increment distribution alone.*
2. *If σ^2 is known, then D is identifiable from one-step increments via $D = (\kappa - \sigma^2)/\Delta t$.*
3. *If the full increment process is observed, then (D, σ^2) are jointly identifiable. Indeed,*

$$\sigma^2 I_d = -\mathrm{Cov}(\Delta Y_n, \Delta Y_{n+1}), \qquad D = \frac{1}{\Delta t}\left(\frac{1}{2d}\mathrm{tr}\,\mathrm{Cov}(\Delta Y_n) - \sigma^2\right). \tag{23}$$

Proof. Part (1) follows directly from Proposition 7. The one-step increment law is Gaussian with covariance $2\kappa I_d$, so any two parameter pairs (D, σ^2) and (D', σ'^2) satisfying $D\Delta t + \sigma^2 = D'\Delta t + \sigma'^2$ produce exactly the same one-step increment distribution. Therefore no estimator based only on that marginal law can distinguish such pairs.

Part (2) is immediate once σ^2 is known, because the collapsed scale κ is identified by the one-step covariance and then $D = (\kappa - \sigma^2)/\Delta t$.

For part (3), Proposition 7 gives the identities

$$\mathrm{Cov}(\Delta Y_n) = 2(D\Delta t + \sigma^2)I_d, \qquad \mathrm{Cov}(\Delta Y_n, \Delta Y_{n+1}) = -\sigma^2 I_d. \tag{24}$$

The second identity identifies σ^2 uniquely. Substituting this value into the first identity identifies $D\Delta t$, and hence D. The recovery formulas in (23) follow by taking traces. □

Corollary 10 (Minimal temporal statistic). *Within the Gaussian observation model (1) and (2), first-order temporal dependence is sufficient to restore joint identifiability of (D, σ^2) once marginal increments are known. No higher lag is needed, because all lag covariances beyond 1 vanish.*

Proof. By Proposition 7, the only nonzero off-diagonal covariance of (ΔY_n) occurs at lag 1 and equals $-\sigma^2 I_d$. Hence lag 1 recovers σ^2, while the lag-0 covariance recovers $\kappa = D\Delta t + \sigma^2$. All higher lags carry no additional second-order information, since their covariances are zero. □

Remark 11. *This corollary is the reason the main text keeps only (4). It is the minimal temporal signature needed to separate physical diffusion from localization uncertainty.*

C.3. Illustration of Parameter Collapse

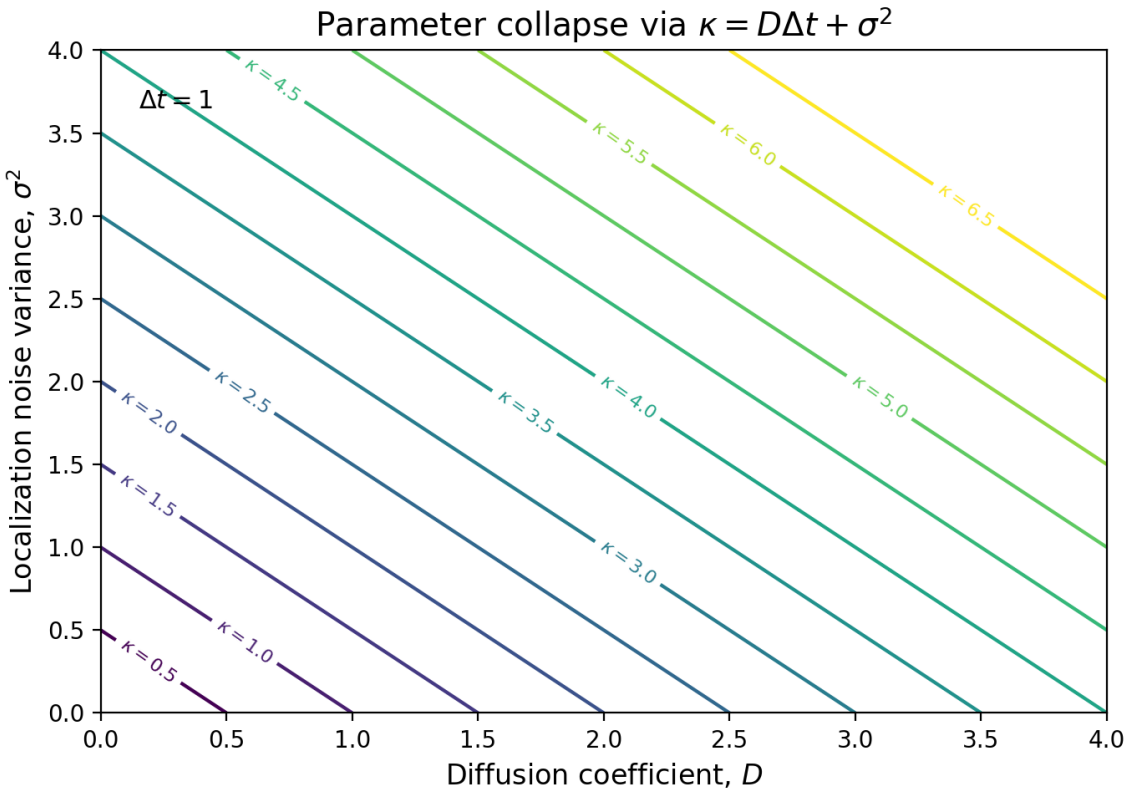

Figure 5. Parameter collapse. Contours of $\kappa = D\Delta t + \sigma^2$ in the (D, σ^2) plane ($\Delta t = 1$). Points on the same contour produce identical one-step increment distributions; they are distinguishable only through the lag-1 covariance.

D. Proofs for Fisher Information and Mixture Results

Proof of Proposition 1

The log-likelihood for one observation $\delta = \Delta Y_n \in \mathbb{R}^d$ is

$$\ell(D; \delta) = -\frac{d}{2}\log(4\pi\kappa) - \frac{\|\delta\|^2}{4\kappa}, \qquad \kappa = D\Delta t + \sigma^2. \tag{25}$$

Since $\partial_D \kappa = \Delta t$, the score is

$$\partial_D \ell(D; \delta) = \Delta t \left(-\frac{d}{2\kappa} + \frac{\|\delta\|^2}{4\kappa^2} \right). \tag{26}$$

Because $\delta \sim \mathcal{N}(0,\, 2\kappa I_d)$, we have $\|\delta\|^2/(2\kappa) \sim \chi^2_d$, and therefore

$$\mathrm{Var}(\|\delta\|^2) = (2\kappa)^2 \mathrm{Var}(\chi^2_d) = 4\kappa^2 \cdot 2d = 8d\kappa^2. \tag{27}$$

The deterministic term in the score has zero variance, so

$$\mathcal{I}(D) = \mathrm{Var}(\partial_D \ell) = \Delta t^2 \cdot \frac{1}{16\kappa^4} \cdot 8d\kappa^2 = \frac{d\Delta t^2}{2\kappa^2}. \tag{28}$$

For independent increments, Fisher information adds linearly, giving $\mathcal{I}_N(D) = N\mathcal{I}(D)$. The Cramér-Rao bound yields $\mathrm{Var}(\widehat{D}) \geq 1/\mathcal{I}_N(D) = 2\kappa^2/(Nd\Delta t^2)$ for any unbiased estimator.

Proof of Theorem 3

For part (1), by isotropy it suffices to consider the one-dimensional radial scale parameterization. The characteristic function of the increment law is

$$\varphi(\omega) = \sum_{j=1}^{J} w_j \exp(-\kappa_j \omega^2). \tag{29}$$

Under Assumption 2, the functions $\exp(-\kappa_j\omega^2)$ are linearly independent, so the finite mixture representation is unique up to permutation by classical identifiability results for Gaussian mixtures (Teicher, 1963; Yakowitz and Spragins, 1968). Since σ^2 is known, each diffusion coefficient is recovered as $D_j = (\kappa_j - \sigma^2)/\Delta t$.

For part (2), if two components coincide, they produce the same Gaussian density, so their individual weights enter the mixture only through their sum.

For part (3), the covariance of a mean-zero finite mixture is the weight-averaged covariance:

$$\mathrm{Cov}(\Delta Y_n) = \sum_{j=1}^{J} w_j \cdot 2\kappa_j I_d = 2\bar{\kappa} I_d, \qquad \bar{\kappa} := \sum_{j=1}^{J} w_j \kappa_j. \tag{30}$$

Taking traces gives $\bar{\kappa} = \frac{1}{2d}\mathrm{tr}\,\mathrm{Cov}(\Delta Y_n)$, proving identifiability of the weighted mean scale.

Proof of Proposition 5

Let

$$p(\delta;\Theta) = \sum_{\ell=1}^{J} w_\ell\, \phi(\delta;\kappa_\ell), \qquad \phi(\delta;\kappa) = (4\pi\kappa)^{-1/2} \exp\!\left(-\frac{\delta^2}{4\kappa}\right). \tag{31}$$

Only the jth component depends on κ_j, so

$$\partial_{\kappa_j} p(\delta;\Theta) = w_j\, \partial_\kappa \phi(\delta;\kappa_j). \tag{32}$$

By the chain rule, $\partial_{D_j} = \Delta t\, \partial_{\kappa_j}$, hence

$$\partial_{D_j} \log p(\delta;\Theta) = \Delta t\, \frac{w_j\, \partial_\kappa \phi(\delta;\kappa_j)}{p(\delta;\Theta)}. \tag{33}$$

Squaring and integrating against $p(\delta;\Theta)$ gives

$$\mathcal{I}(D_j) = \Delta t^2 w_j^2 \int \frac{[\partial_\kappa \phi(\delta;\kappa_j)]^2}{p(\delta;\Theta)}\, d\delta = w_j^2 \Delta t^2 \mathcal{R}_j(\Theta), \tag{34}$$

which proves (10).

For the upper bound, observe that $p(\delta;\Theta) \geq w_j \phi(\delta;\kappa_j)$ pointwise. Therefore

$$\mathcal{R}_j(\Theta) \leq \frac{1}{w_j} \int \frac{[\partial_\kappa \phi(\delta;\kappa_j)]^2}{\phi(\delta;\kappa_j)}\, d\delta. \tag{35}$$

The remaining integral is the Fisher information of the Gaussian scale parameter κ for the density $\phi(\delta;\kappa)$, which equals $1/(2\kappa_j^2)$. Consequently,

$$\mathcal{I}(D_j) \leq w_j^2 \Delta t^2 \cdot \frac{1}{w_j} \cdot \frac{1}{2\kappa_j^2} = \frac{w_j \Delta t^2}{2\kappa_j^2}. \tag{36}$$

Equality holds only if the pointwise inequality $p(\delta;\Theta) \geq w_j \phi(\delta;\kappa_j)$ is an equality almost everywhere, which requires the absence of all other components. Hence $J = 1$.

Proof of Corollary 6

Starting from (11), divide both sides by the single-species information level $\mathcal{I}_{\text{single}}(\kappa_j) = \Delta t^2/(2\kappa_j^2)$ to obtain

$$\frac{\mathcal{I}(D_j)}{\mathcal{I}_{\text{single}}(\kappa_j)} \leq w_j. \tag{37}$$

Since (10) already contains an explicit prefactor w_j^2, the result shows that rare components are penalized twice: once by the direct scarcity of samples from that component, and once by the mixture denominator that dilutes score information. This establishes the quadratic penalty interpretation.

E. Posterior families and their training details.

We compare four tractable families for $q_\phi(\log D \mid x)$, all trained by minimizing (43). Each family models the posterior over $\log D$ rather than D directly, which naturally enforces positivity and stabilizes gradients across the wide prior range.

Log-normal (LogNormal). The network outputs $(\mu_\phi(x),\, s_\phi(x)) \in \mathbb{R} \times \mathbb{R}_{>0}$ and sets

$$q_\phi(\log D \mid x) = \mathcal{N}\big(\log D;\, \mu_\phi(x),\, s_\phi(x)^2\big)\,, \tag{38}$$

so $D \mid x$ follows a log-normal distribution with posterior mean $\exp\big(\mu_\phi + s_\phi^2/2\big)$ and posterior variance $\exp(2\mu_\phi + s_\phi^2)(\exp(s_\phi^2) - 1)$. The log-normal is the canonical unimodal positive-support family in variational inference (Kingma and Welling, 2013) and serves as the single-component baseline.

Mixture of log-normals (MoLN). The network outputs mixture logits, means, and log-scales for a two-component mixture:

$$q_\phi(\log D \mid x) = \sum_{c=1}^{2} w_c(x)\, \mathcal{N}\big(\log D;\, \mu_c(x),\, s_c(x)^2\big)\,, \tag{39}$$

where $w_1(x)+w_2(x) = 1$, $w_c(x) > 0$, obtained via softmax. Finite mixtures of Gaussians are universal approximators of smooth densities on $\mathbb{R}$ (Goodfellow et al., 2021), and the two-component variant is the minimal extension capable of representing the bimodal posteriors predicted by Theorem 3 (part 2).

Log-Student-t (StudentT). The network outputs location $\mu_\phi(x)$, scale $s_\phi(x) > 0$, and degrees of freedom $\nu_\phi(x) > 2$:

$$q_\phi(\log D \mid x) = t_{\nu_\phi(x)}(\log D;\, \mu_\phi(x),\, s_\phi(x))\,, \tag{40}$$

where $t_\nu(\cdot; \mu, s)$ denotes the location-scale Student-t density. The constraint $\nu > 2$ ensures a finite variance. Heavy-tailed variational families of this form have been proposed to reduce underestimation of posterior uncertainty in variational inference (Jaakkola and Jordan, 1998; Futami et al., 2017).

Neural spline flow (NSF). A normalizing flow that transforms a standard Gaussian base density on $\log D$ through a composition of monotone rational-quadratic spline layers (Durkan et al., 2019):

$$q_\phi(\log D \mid x) = p_{\text{base}}\Big(T_\phi^{-1}(\log D;\, x)\Big) \left| \frac{\partial T_\phi^{-1}}{\partial \log D} \right|, \tag{41}$$

where $T_\phi(\,\cdot\,;\, x)$ is the spline-based invertible transform conditioned on x. NSF is the most expressive family in our comparison and can in principle approximate any smooth unimodal or multimodal density. Normalizing flows have been used as posterior approximators in simulation-based inference (Papamakarios et al., 2021; Cranmer et al., 2020).

Feature extraction. Given N observed increments $\{\Delta Y_n\}_{n=1}^N$ in d dimensions, we compute four scalar statistics:

$$S_0 = \frac{1}{N}\sum_n \|\Delta Y_n\|^2, \quad S_1 = \frac{1}{N-1}\sum_n \langle \Delta Y_n, \Delta Y_{n+1}\rangle, \quad \log S_0, \quad \log|S_1|. \tag{42}$$

By Proposition 7, S_0 converges to $2d\bar{\kappa}$ and S_1 converges to $-d\sigma^2$, so together they identify both $\bar{\kappa}$ and σ^2 when σ^2 is unknown, and reduce to a sufficient pair for D when σ^2 is known. The log-transformed versions stabilize the feature scale across the wide prior range. We stack these four scalars into a feature vector $x := \big(S_0, S_1, \log S_0, \log|S_1|\big) \in \mathbb{R}^4$, which serves as the input to the posterior head.

Training details and loss function. We assume increments are already extracted from trajectories; linking errors in dense-image regimes are a separate preprocessing issue outside the scope of our theorems and of the experiments reported. All models are trained by minimizing the negative log-posterior on simulated pairs $(x^{(i)}, \log D^{(i)})$, where $x^{(i)} \in \mathbb{R}^4$ is the feature vector defined above, computed from the i-th simulated dataset:

$$\mathcal{L}(\phi) = -\frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} \log q_\phi\Big(\log D^{(i)} \mid x^{(i)}\Big), \tag{43}$$

using AdamW with cosine annealing, $N_{\text{train}} = 120{,}000$, and $N = 500$ increments per simulated dataset.

F. Details of Empirical Validation

This appendix provides the full specification of the experiments reported in Section 5, including grid parameters for the consistency check, the interpretation of the cell-by-cell structure of the 7×7 posterior grids, and per-architecture numerical results for Exp A and Exp B.

F.1. Consistency check: grid specification and interpretation

Grid setup. The consistency check of Section 5.1 uses a 7×7 grid of (D, σ^2) pairs with $\Delta t = 0.5$ fixed. Rows correspond to $\sigma^2 \in \{0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75\}$ and columns to $D \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5\}$, chosen so that the main diagonal satisfies $\kappa = D\Delta t + \sigma^2 = 2$ exactly, the upper-right triangle satisfies $\kappa > 2$, and the lower-left triangle satisfies $\kappa < 2$. For each cell we draw 10 independent realizations of $N = 400$ increments, pass them through the encoder with features restricted to $(S_0, \log S_0)$, and sample 400 posterior draws per realization (4000 samples per cell in total). The true-D marker in each cell is the column value.

Interpreting the grid. Three observations confirm that the amortized posterior respects the identifiability structure predicted by Theorem 9 (part 1). First, histograms are nearly identical along each anti-diagonal: cells sharing a common κ produce the same posterior shape even though the underlying D varies from 0.5 to 3.5, and the red true-D line slides across an essentially stationary histogram. Second, the posterior shifts smoothly across anti-diagonals with increasing κ. Third, the observed shapes match the Bayes posterior $p(D \mid \kappa)$ induced by the training prior $D \sim \text{Uniform}[D_{\min}, D_{\max}]$ and $\sigma^2 \sim \text{Uniform}[\sigma^2_{\min}, \sigma^2_{\max}]$, restricted to the feasible region $\sigma^2 = \kappa - D\Delta t \in [\sigma^2_{\min}, \sigma^2_{\max}]$. Cells near the corners of the grid, where the prior support on D is narrow, show sharp concentration; interior cells, where the prior support is wide, show plateau-shaped posteriors. For example, at $\kappa = 0.5$ (bottom-left corner) the feasible D-interval is narrow and the posterior concentrates near $D \approx 0.5$; at $\kappa = 2$ (main diagonal) the feasible interval spans almost the full prior and the posterior is close to the marginal D-prior.

Expressivity gap between LogNormal and MoLN. Both architectures satisfy the anti-diagonal constancy test, so neither is leaking D through a non-κ statistic. They differ in the shape they can assign to the Bayes posterior on a κ level set. For interior κ values the target density is close to uniform on the feasible D-interval, and a single log-normal head is structurally incapable of representing a plateau: the LogNormal posterior approximates it with a right-skewed unimodal density, systematically over-weighting smaller D. The MoLN posterior, with two components, covers the support more faithfully, although a mild bimodal artifact is visible on the widest interior cells (where the support is wide enough that the two components split to cover the left and right thirds, leaving a shallow dip in the middle). This shape gap is a posterior-family limitation, not an identifiability failure, and it motivates a broader comparison of posterior families in Exp A and Exp B below.

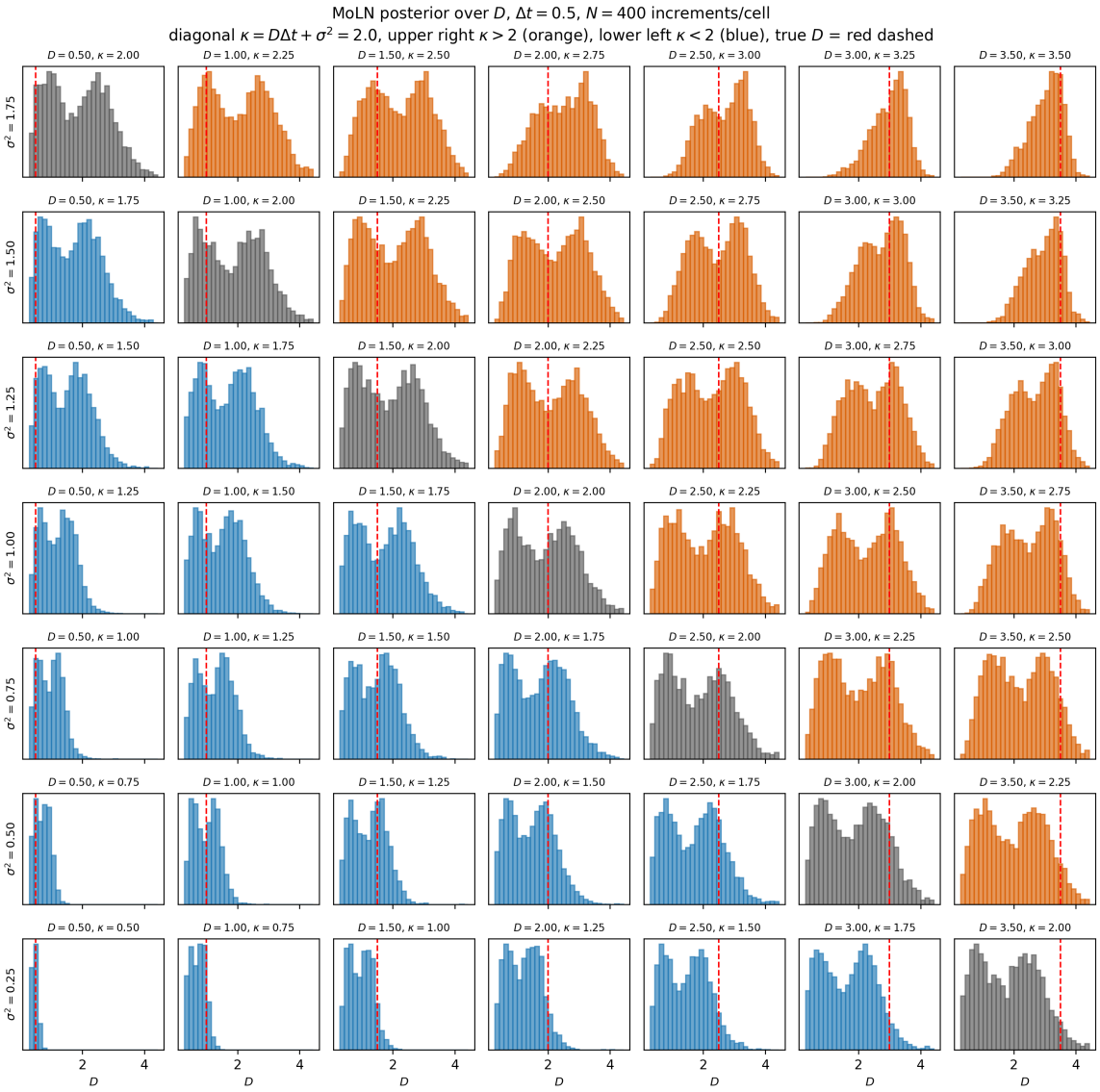

Figure 6. MoLN posterior on the 7×7 identifiability-consistency grid (Section 5.1, Appendix F.1). Each cell shows the posterior $q_\phi(D \mid x)$ for a distinct (D, σ^2) pair, with $\Delta t = 0.5$ fixed and features restricted to $(S_0, \log S_0)$. The main diagonal satisfies $\kappa = 2$ (grey), the upper-right triangle $\kappa > 2$ (orange), the lower-left triangle $\kappa < 2$ (blue). The red dashed line marks the true D. Histograms are nearly identical along each anti-diagonal, confirming that the posterior depends on the data only through κ.

F.2. Experiment A: results

Evaluation is performed at the test point shown in Figure 3, with the full feature vector $x = (S_0, S_1, \log S_0, \log |S_1|)$ restored. All four architectures produce broad, right-skewed posteriors consistent with the (D, σ^2) confounding predicted by Theorem 9 (part 1). NSF achieves the lowest NLL (0.631), well-calibrated 90% coverage (0.89), and near-ideal 50% coverage (0.49); MoLN follows with NLL = 0.706, $\text{Cov}_{50} = 0.48$, and $\text{Cov}_{90} = 0.88$. LogNormal and StudentT yield 90% coverage at the upper end (both ≈0.95), but their 50% coverage falls short of the ideal (both ≈0.39), the joint signature of mild over-dispersion at the median with tails that over-cover. The StudentT posterior carries the largest test NLL (0.817) and the largest Bayes MSE (0.865) of the four families; its heavy

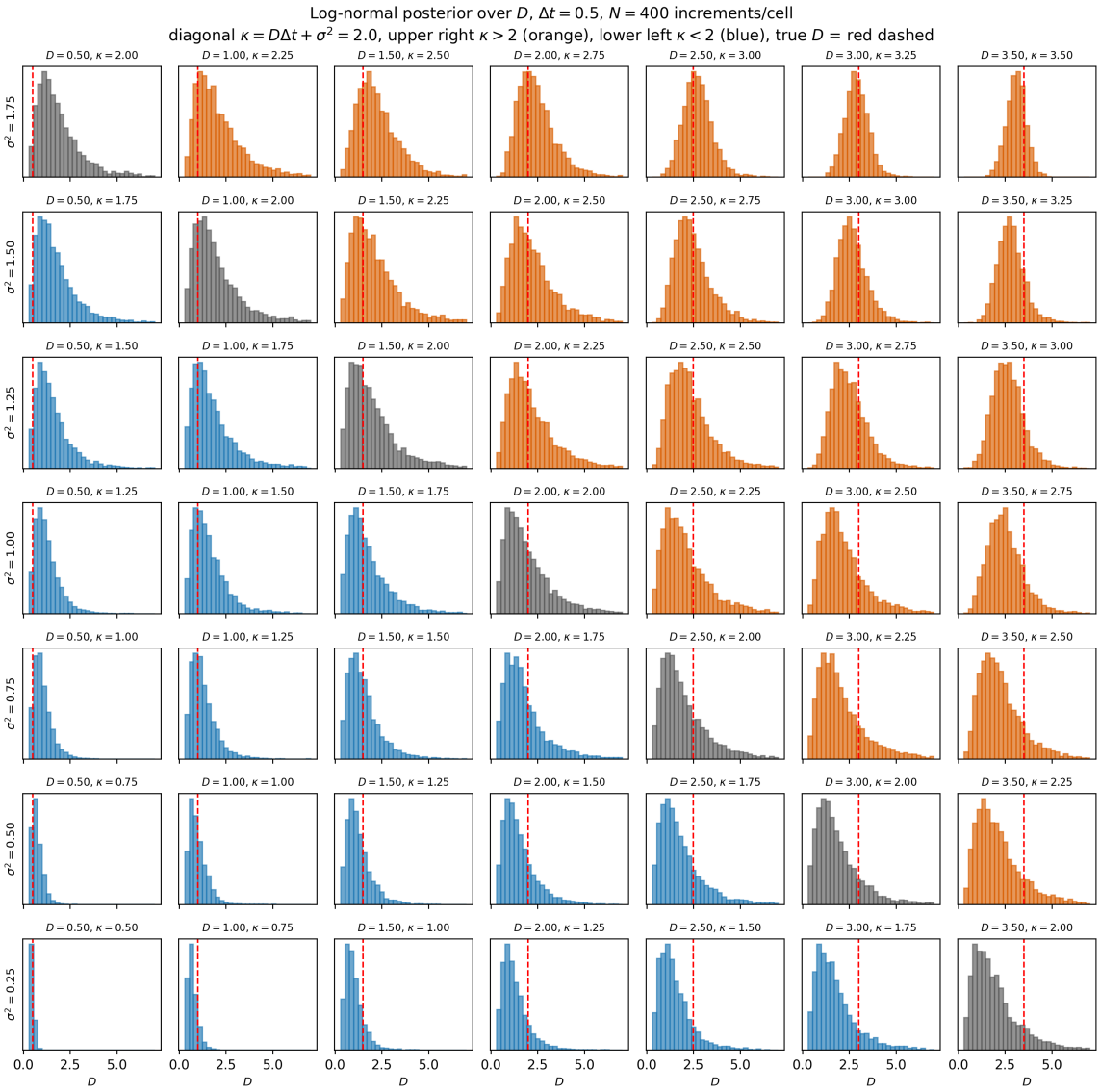

Figure 7. LogNormal posterior on the same 7×7 grid as Figure 6. Anti-diagonal constancy is preserved, so the LogNormal posterior also depends on the data only through κ. Compared to MoLN, the LogNormal posterior is structurally unable to represent plateau-shaped Bayes posteriors on interior κ level sets, approximating them with a right-skewed unimodal density. This expressivity gap is the reason MoLN (and NSF) are preferred in Exp A and Exp B.

tails assign non-negligible mass to large values of D, making it the least preferable family for this inference problem despite its theoretical flexibility, since the log-Student-t tail index interacts with the exponential tail of the Brownian likelihood to produce an amplified posterior tail that dominates any MSE-type metric even when the posterior mode is correctly placed.

Additionally, we also show the sigma^2 posteriors.

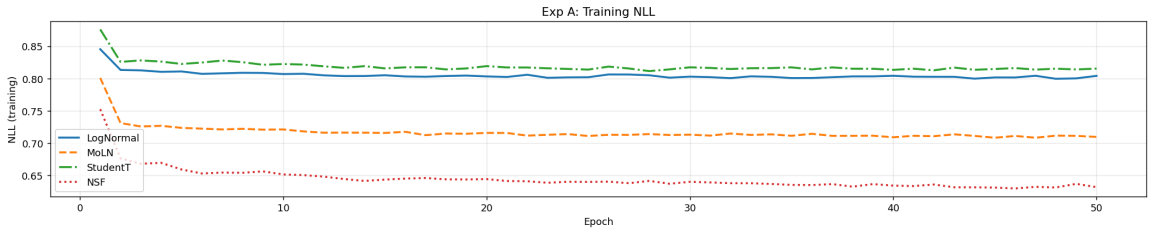

Figure 8. Training NLL trajectories for the four posterior families on Exp A across 50 epochs. NSF (red) reaches the lowest training NLL, MoLN (orange) is next, and LogNormal (blue) and StudentT (green) plateau at higher values; the training-time ordering tracks the test-time NLL reported in Figure 3.

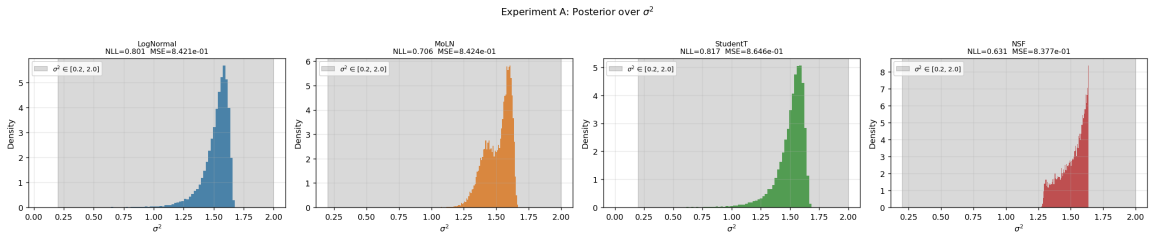

Figure 9. Marginal posteriors over σ^2 produced by the four architectures at the Exp A test point (σ^2-prior LogUniform$[0.2, 2.0]$, shaded region). All four families place posterior mass in the upper portion of the prior support; the recovered σ^2-posteriors reflect the lag-1 information accessible through the S_1 feature, complementing the marginal-of-D posteriors of Figure 3.

F.3. Experiment B: results

Evaluation is at the Exp B test point with D_A-prior LogUniform$[0.4, 1.2]$, $D_B = 2.0$ fixed, equal mixing weights, and $\sigma^2 = 1.0$ known. MoLN places its primary mode in the D_A-prior support and a secondary, broader mode at an intermediate location that captures the label-ambiguity mass associated with D_B; the secondary mode is displaced from the nominal value $D_B = 2.0$, consistent with the prior-weighted mixing of the D_B-explanation against the D_A-prior. The unimodal families (LogNormal, StudentT) collapse to a single mode in the D_A-prior support and miss the D_B-induced mass, with LogNormal incurring a catastrophic NLL (78.165) from mode-forcing on each test dataset and StudentT producing an unbounded Bayes MSE under the mixture likelihood. NSF places a sharp secondary spike with broad mass below it and achieves the best NLL (2.355); MoLN is close behind in NLL (2.522) and attains the lowest Bayes MSE among finite-MSE families (0.726 versus 0.735 for LogNormal and 0.841 for NSF). Central-quantile coverage is uniformly low across architectures ($\mathrm{Cov}_{50} \approx 0.23$–$0.26$, $\mathrm{Cov}_{90} \approx 0.51$–$0.58$), as predicted for bimodal posteriors measured by central intervals; the differences across families lie almost entirely in posterior shape and NLL, as expected from Corollary 6.

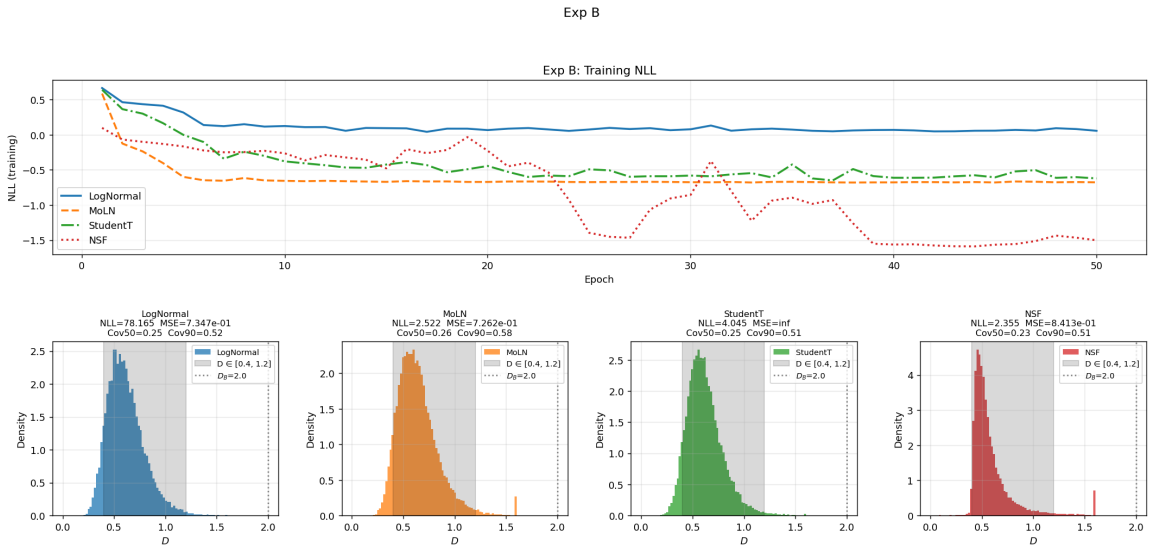

Figure 10. Top: training NLL trajectories for Exp B (K=2 mixture) over 50 epochs. Bottom: posterior densities at the Exp B test point, with $D_B = 2.0$ marked by a dotted line and the D_A-prior support shaded. MoLN is the only family producing a visibly two-mode density; NSF places a sharp secondary spike with broad mass below; LogNormal and StudentT collapse onto a single unimodal mode and miss the D_B-induced mass, with LogNormal incurring catastrophic NLL (78.165) and StudentT producing an unbounded Bayes MSE.

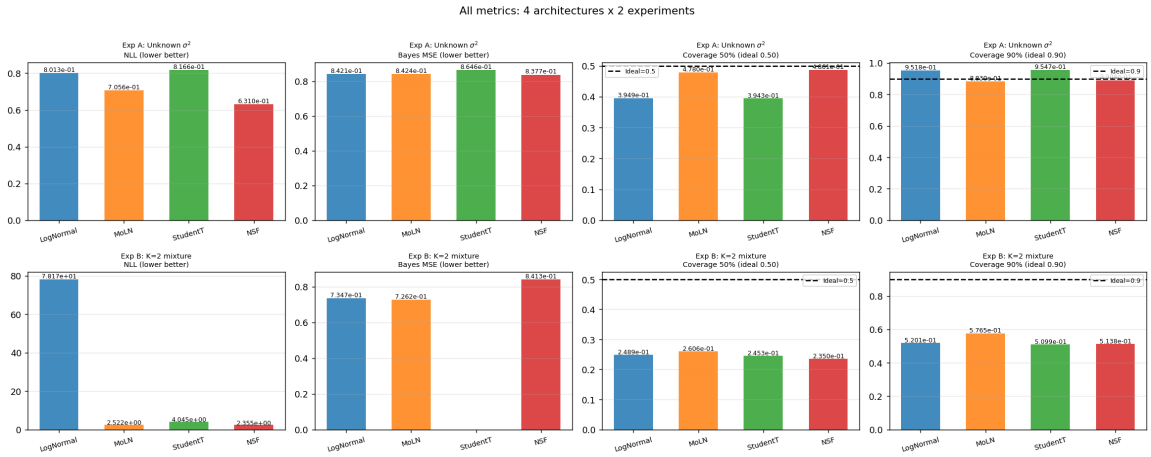

Figure 11. Aggregate metrics for the four architectures across Exp A (top row, unknown σ^2) and Exp B (K=2 mixture, bottom row). Columns from left to right: NLL, Bayes MSE, Cov_{50}, and Cov_{90}, with dashed lines marking ideal coverage. In Exp A, all architectures attain comparable Bayes MSE near 0.84, NSF achieves the lowest NLL (0.631) and best median calibration, and StudentT carries the largest NLL and MSE among the four. In Exp B, MoLN attains the lowest Bayes MSE among finite-MSE families (0.726), NSF reaches the lowest NLL (2.355), LogNormal incurs a catastrophic NLL (78.165) and StudentT diverges in MSE.