Neural Diffusion Intensity Models for Point Process Data

Xinlong Du[1] **Harsha Honnappa**[1] **Vinayak Rao**[2]

[1]Edwardson School of Industrial Engineering, Purdue University
[2]Department of Statistics, Purdue University

Abstract

Cox processes model overdispersed point process data via a latent stochastic intensity, but both estimation of the nonparametric intensity model and posterior inference over intensity paths are generally intractable, relying on computationally-heavy MCMC methods. We introduce Neural Diffusion Intensity Models, a variational inference framework for Cox processes driven by neural stochastic differential equations (SDEs). Our key theoretical result, based on enlargement of filtrations, shows that conditioning on point process observations preserves the diffusion structure of the latent intensity with an explicit drift correction. This provides guidance on how to choose an appropriate variational family so that ELBO maximization coincides with maximum likelihood estimation under sufficient model capacity. We provide an amortized encoder architecture that maps variable-length event sequences to posterior intensity paths by simulating the drift-corrected SDE, replacing repeated MCMC runs with a single forward SDE simulation. Experiments on synthetic and real-world data demonstrate accurate recovery of latent intensity dynamics and posterior paths, with orders-of-magnitude speedups over MCMC-based methods.

1 INTRODUCTION

Point process data arise in many scientific and industrial applications, including neural spike trains [Amemori and Ishii, 2001], social interactions [Perry and Wolfe, 2013], finance and insurance models [Björk, 2021] and queuing models [Brémaud, 1981]. Such data are sparse, irregular, and often exhibit substantial *over-dispersion* (high variance to mean ratio) relative to simple inhomogeneous Poisson models [Jongbloed and Koole, 2001, Avramidis and L'Ecuyer, 2005]. As an example, in Section 4.4, we will consider a dataset of incoming calls from a large U.S. bank call center, and Figure 1 shows that for this dataset, 10-minute arrival counts are strongly over-dispersed. A natural and expressive framework is to model these data as Cox processes, wherein event arrivals follow a Poisson process with a latent *stochastic* intensity [Daley and Vere-Jones, 2003, Chapter 6]. The randomness of the intensity function helps explain the greater between-realization variability that is characteristic of over-dispersed point processes.

In this paper, we model the intensity process as a Markov diffusion that solves an unknown stochastic differential equation (SDE). Such a point process model is commonly referred to as a diffusion-driven Cox process (DDCP) [Gonçalves et al., 2024]. The Markovianity of the intensity process is a common requirement in many applications; for instance, [Zhang et al., 2014] propose a Cox-Ingersoll-Ross (CIR) diffusion model of the intensity of the U.S. bank data in Figure 1. Modeling via an SDE allows one to explicitly encode dynamic mechanisms, such as mean reversion or state-dependent volatility, giving an interpretable description of the event rate evolution.

Learning a DDCP amounts to estimating the coefficients of the latent SDE from point process observations. This is a challenging problem, involving intractable infinite-dimensional integrals over intensity paths and requiring computationally-intensive Markov chain Monte Carlo (MCMC) methods [Gonçalves et al., 2024]. These challenges persist even after the model is learned: computing the posterior of the latent intensity conditioned on a point process observation requires expensive MCMC simulations.

In this work, we introduce **Neural Diffusion Intensity Models**, a nonparametric variational inference framework for Cox processes driven by latent SDE intensity processes. Our approach combines three key ideas:

1. **Neural SDE priors.** We parametrize the drift of the latent SDE with a neural network, making a flexible generative model for intensity paths.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1448–1473.

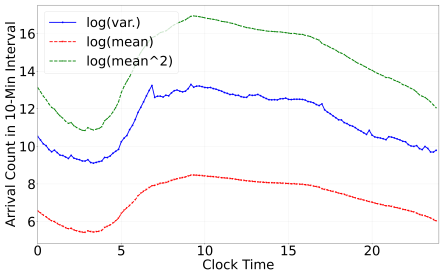


Figure 1: The variance curve exceeds the mean curve, indicating strong over-dispersion of the arrival point process, in the US Bank data from [SEE Lab, Technion, 2024].

2. **Posterior characterization.** Using tools from *enlargement of filtrations* (EoF), we show that conditioning on point process observations introduces a score-like drift correction while preserving the diffusion structure of the latent intensity. That is, the posterior intensity process remains a diffusion with the same diffusion coefficient and a modified score-type drift. This establishes a conjugacy at the level of path measures.
3. **Path-space amortized inference.** Rather than computing the posterior drift correction anew for each newly observed point process, we exploit the above result to replace the point process realization-dependent drift correction with a trainable neural network. This way, posterior inference is amortized: given point process observations, posterior intensity paths can be sampled by forward simulating the neural drift-corrected SDE.

Our architecture has the flavor of a variational autoencoder (VAE). The drift neural network is a decoder from Brownian paths to point process observations via the latent intensity, while the drift correction neural network forms an encoder mapping from point process observations to posterior distributions over intensity paths. We train both neural networks by maximizing the evidence lower bound (ELBO). For an expressive enough neural network, our approach approximates the true posterior arbitrarily well. This is in contrast to works that use, for example, a linear SDE approximation to the true posterior, as we discuss next.

1.1 RELATED WORK

There have been numerous prior works on point process models with stochastic intensity functions. A large number of these model the intensity function as a transformed realization of a smooth Gaussian process prior, and carry out posterior inference via MCMC [Møller et al., 1998, Adams et al., 2009], variational inference [Donner and Opper, 2018, Aglietti et al., 2019], or approaches like INLA [Simpson et al., 2016]. Here, the latent intensity primarily serves as a smoothing mechanism, and these Gaussian process models lack the capacity to represent mechanistic dynamics in the way that a drift term in an SDE can.

Closer to our work is Duncker et al. [2019] and Jaiswal et al. [2021], both of whom consider DDCPs, and attempt to flexibly learn the drift-function of the underlying SDE using variational Bayes. In the former, the authors make a Gaussian approximation to the posterior along the lines of Archambeau et al. [2007] (see below): however this introduces a variational gap. [Jaiswal et al., 2021] proposed a variational inference scheme for DDCPs, using a stochastic partial differential equation satisfied by the smoothing posterior density as a heuristic to identify the posterior diffusion. The present paper builds on this line of work by providing a rigorous EoF-based structural characterization of the posterior path measure (not considered in [Jaiswal et al., 2021]), and establishes that a sufficiently expressive variational family will contain the true posterior. Furthermore, these works do not consider amortized inference, necessitating a fresh inference run for each new observation.

A different approach is that taken by neural temporal point processes (neural TPPs) and related models [Du et al., 2016, Shchur et al., 2021, Zuo et al., 2020, Mei and Eisner, 2017, Shchur et al., 2020]: these model marked point processes autoregressively, predicting the time and mark of the next event in a sequence using neural network architectures like RNNs, GRUs or LSTMs. Works like APP-VAE [Mehrasa et al., 2019], VEPP [Pan et al., 2020], and [Shibue and Iwata, 2024] apply variational-autoencoder ideas to point processes by introducing latent variables at the level of inter-arrival times. These are flexible predictive models, but they do not aim to recover a latent diffusion intensity or quantify its uncertainty. Similarly, neural TPPs such as neural Hawkes and Transformer Hawkes models [Du et al., 2016, Mei and Eisner, 2017, Zuo et al., 2020, Shchur et al., 2021, 2020] parameterize the history-conditional law of future events directly, whereas our objective is to learn prior diffusion dynamics and perform posterior inference over latent intensity paths.

A separate recent line of research uses diffusion-model ideas directly on point process observations. ADD-THIN [Lüdke et al., 2023] and diffusion-excursion models [Hasan et al., 2023] also connect diffusions and point processes, but operate on event patterns or path-event constructions rather than on posteriors of a latent stochastic intensity diffusion.

Moving our focus from point processes to SDEs, there is an extensive literature on MCMC [Beskos et al., 2006, Gonçalves et al., 2024] and variational inference for continuous-time stochastic models. Variational SDE methods trace back to Archambeau et al. [2007], with later work learning posterior drifts for latent paths [Opper, 2019, Ryder et al., 2018]. These methods do not treat point process observations, identify the exact posterior structure induced by a Cox likelihood, or amortize posterior inference across event sequences. Latent Neural SDEs for generative mod-

eling have also been studied [Kidger et al., 2021, Li et al., 2020], but not for point process observations generated by a latent diffusion intensity.

Opper's formulation of variational inference as stochastic optimal control [Opper, 2019] is closely related to [Tzen and Raginsky, 2019], who derive the Föllmer drift for terminal-state conditioning. Since Cox process likelihood's depend on the full intensity trajectory, our EoF result is the corresponding path-measure analogue and identifies the full posterior path measure rather than just the marginal posterior distribution.

Discrete-time amortized inference for dynamical latent variables has been studied by Krishnan et al. [2017] and Marino et al. [2018]. Our work is the continuous-time point-process analogue, with EoF giving the structural reason to condition on the full event sequence.

2 PROBLEM FORMULATION

We formulate our model on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P}^\star)$, where $\mathbb{P}^\star$ denotes the unknown ground-truth data-generating measure. We observe event data as a simple point process [Daley and Vere-Jones, 2003, Chapter 1] $X = \{\tau_i\}_{i=1}^K$ on $[0,T]$, with $0 < \tau_1 < \cdots < \tau_K \leq T$ and K a random variable giving the number of events. We occasionally also use X to denote the counting measure, or the counting process $X_{0:T}$ where $X_t := \sum_i \mathbf{1}_{\{\tau_i \leq t\}}$. We let $\mu_{\text{gt}} = \mathbb{P}^\star \circ X^{-1}$ denote the ground-truth distribution over event sequences, and let $\mathcal{X} := \cup_{k\in\mathbb{N}} \mathbb{R}^k$ represent the sample space for X.

The filtration $(\mathcal{F}_t)$ is assumed to carry both the observed events X and a latent intensity process $Z = (Z_t)_{t\in[0,T]}$, which is $(\mathcal{F}_t)$-adapted but not directly observed. Let $\mathcal{Z} := C([0,T]; \mathbb{R}^d)$ represent the sample space of $\mathbb{R}^d$-valued continuous functions for Z.

2.1 NEURAL DIFFUSION INTENSITY MODEL

We focus on one-dimensional simple point processes, noting that our framework can be extended to multivariate point processes by replacing the scalar intensity process with a vector-valued SDE[1] on $[0,T]$. A rigorous multivariate posterior characterization is given in Appendix D. Our model class $\{P_\theta, \theta \in \Theta\}$ consists of Cox processes (doubly stochastic Poisson processes), defined by the hierarchical structure $P_\theta(X) := \int_{\mathcal{Z}} dP_\theta(X, Z)$ where $P_\theta(X, Z) = P(X|Z)P_\theta(Z)$, $P(X|Z)$ and $P_\theta(Z)$ are path measures and $\Theta \subseteq \mathbb{R}^p$, $p > 0$. Conditioned on a latent intensity trajectory $Z = (Z_t)_{t\in[0,T]}$, the observation X is an inhomogeneous Poisson process with rate $(Z_t)_{t\in[0,T]}$. The prior measure $P_\theta(Z)$ over intensities is induced by the Markov diffusion solution of the SDE. We focus on one-dimensional simple point processes X, though our framework can be extended to multivariate point processes by replacing the scalar intensity process with a vector-valued diffusion[2] on $[0,T]$. A rigorous multivariate posterior characterization is given in Appendix D. Our model class $\{P_\theta, \theta \in \Theta\}$ consists of Cox processes or doubly stochastic Poisson processes. Specifically, the marginal probability of the observations is $P_\theta(X) := \int_{\mathcal{Z}} dP_\theta(X, Z)$ where the joint distribution factors as $P_\theta(X, Z) = P(X|Z)P_\theta(Z)$, $P(X|Z)$, with $Z = (Z_t)_{t\in[0,T]}$ denoting the latent intensity process. Conditioned on Z, the observations X form an inhomogeneous Poisson process with intensity $(Z_t)_{t\in[0,T]}$, so that $P(X|Z)$ is the corresponding Poisson likelihood. The prior measure $P_\theta(Z)$ over intensities is induced by the solution of the SDE

$$dZ_t = b_\theta(Z_t, t)\, dt + \sigma(Z_t, t)\, dB_t, \qquad Z_0 = z. \tag{1}$$

Here B_t is a Brownian motion adapted to $(\mathcal{F}_t)_{t\in[0,T]}$, while

- b_θ is an unknown drift function, parametrized by a p-dimensional parameter θ. In this work, we will directly represent the drift function with a neural network, so that θ denotes the network parameters. Although finite-dimensional, a high-capacity parametrization serves as a flexible, effectively nonparametric model for the drift.
- σ is a known diffusion coefficient (e.g., fixed as a constant or set by the CIR model).

We assume that the drift and diffusion coefficients satisfy the usual conditions consistent for a strong solution of the SDE to exist; see [Øksendal, 2003, Theorem 5.2.1]. Then, the joint measure $P_\theta(X, Z)$ is well-defined on $\mathcal{X} \times \mathcal{Z}$ provided $Z_t \geq 0$ a.s., making it a valid intensity function. This construction yields an infinite mixture model [Lindsay, 1995a] of Poisson processes [Møller et al., 1998], with mixing distribution given by the SDE prior $P_\theta(Z)$. See Section B for the measure-theoretic desiderata for the model.

2.2 MAXIMUM MARGINAL LIKELIHOOD

With this setup, learning $P_\theta(X)$ reduces to estimating the drift parameter θ. We do this by maximizing the expected log-likelihood under the ground truth distribution μ_{gt}:

$$\sup_{\theta\in\Theta} \mathbb{E}_{\mu_{\text{gt}}}[\log P_\theta(X)] = \sup_{\theta\in\Theta} \mathbb{E}_{\mu_{\text{gt}}}\left[\log \int P(X|Z)P_\theta(dZ)\right].$$

For a flexible enough model class of the drift, this is effectively a nonparametric maximum likelihood estimation

[1] This would model either multiple event streams that may share a common multivariate intensity, or independently evolving ones.

[2] This would model either multiple event streams that may share a common multivariate intensity, or independently evolving ones.

(NPMLE) problem [Lindsay, 1995b]. We assume that the model class $\{P_\theta, \theta \in \Theta\}$ is *well-specified* in the sense that $\mathrm{KL}(\mu_{\text{gt}} \| P_\theta) < \infty$ for some $\theta \in \Theta$. This ensures that $\mathbb{E}_{\mu_{\text{gt}}}[\log P_\theta(X)]$ is well-defined and finite, and that the maximum likelihood objective is not trivially $-\infty$.

In practice, μ_{gt} is unknown and we assume access to n i.i.d. observations of the data, i.e., $\{X^i\}_{i=1}^n$, with $X^i \overset{\text{i.i.d.}}{\sim} \mu_{\text{gt}}$, leading to the empirical objective

$$\sup_{\theta \in \Theta} \mathbb{E}_{\hat{\mu}} \left[\log \int P(X|Z) dP_\theta(Z) \right] \tag{2}$$

where $\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \delta_{X^i}$ is the empirical measure.

Solving this maximization problem however involves a structural difficulty: with a few special exceptions, the integral on the RHS is intractable. This has been classically addressed in the mixture modeling literature by resorting to an expectation-maximization (EM) type algorithm; see Appendix E.1 and [Lindsay, 1995a, Chapter 3 and 6] for finite dimensional settings.

In our setting, the E-step requires sampling from the posterior path measure $P_\theta(Z|X)$, typically via MCMC, while the M-step updates the prior measure (with parameters θ). This can be computationally expensive, since every gradient step requires new MCMC samples from the path posterior distribution. Furthermore, the EM algorithm will only learn the prior dynamics, and posterior inference for a newly observed point process will require additional MCMC at test time. Since the posterior is a path space measure, such MCMC sampling can be challenging, especially in settings where repeated or real-time inference is needed.

To bypass this computational burden, we introduce an amortized approximation to the posterior. We model the posterior drift correction with a neural network u_β, and the posterior intensity process as the solution of the SDE

$$\begin{aligned} dZ_t = &\Big[b_\theta(Z_t, t) + \mathbf{1}_{\{t \le T'\}} \sigma(Z_t, t) u_\beta\left(Z_t, t, T', N_{0:T'}\right) \Big] dt \\ &+ \sigma(Z_t, t)\, dB_t. \end{aligned} \tag{3}$$

and treat the path measure induced by this SDE as the variational family $Q_\phi(Z|X)$, where $\phi = (\theta, \beta)$.

This construction preserves the diffusion coefficient, while introducing an observation-dependent drift correction through u_β. Once b_θ and u_β are learned, posterior inference for any observation reduces to simulating sample paths from Eq. (3). Posterior inference becomes fully amortized.

2.3 ENLARGEMENT OF FILTRATIONS

At this point, this structure imposed by our variational approximation remains a heuristic. There are two fundamental questions to consider regarding the true posterior process:

i. (**Structure**) Does the posterior intensity, conditioned on a point process observation, remain a diffusion with an SDE representation? And if so:

ii. (**Correction**) How does its drift relate to the prior drift?

Answering the first question justifies the validity of using path measures induced by Markov diffusions (as solutions of SDEs) for the approximate posteriors, and answering the second sheds light on the correct functional form of the posterior drift correction.

Both questions can be answered using tools from Enlargement of Filtrations (EoF) [Jeanblanc et al., 2009, Jacod, 2006], wherein one studies how semimartingale decompositions change when the filtration is enlarged to include additional information, typically information about future events. In our setting, the enlargement corresponds to conditioning on the observed point process trajectory $X = N_{0:T'}$, where $T' \le T$. Formally, we pass from the original filtration $(\mathcal{F}_t)$ to the "enlarged" filtration $\mathcal{G}_t = \mathcal{F}_t \vee \sigma(N_{0:T'})$, so that the future event times up to T' are known at all earlier times. This is the continuous-time analogue of *smoothing* in a hidden Markov model (HMM), where the latent state at time t is inferred given the entire observation sequence rather than only the observations up to time t; here, the role of the full observation sequence is played by the complete point process trajectory $X_{0:T'}$.

Under Jacod's absolute continuity condition (Theorem C.1), one can apply the general semimartingale decomposition theorem (Theorem. C.3) to obtain the following result[3]:

Theorem 2.1. *Fix $T' \le T$ and let $X_{0:T'}$ be a one-dimensional point process on $[0, T']$. Suppose the prior intensity process satisfies*

$$dZ_t = b(Z_t, t)\, dt + \sigma(Z_t, t)\, dB_t, \quad Z_0 = z_0.$$

Assume that, for each realized observation X, the function $h_X(z, t) := h(z, t, T', X)$ is strictly positive and differentiable in z on the interior of the state space visited by Z, and that the process $p_t^X = h_X(Z_t, t)$ satisfies

$$d\langle B, p^X \rangle_t = \sigma(Z_t, t) \partial_z h_X(Z_t, t)\, dt$$

on each interval between successive event times. A sufficient condition is that h_X is piecewise $C^{2,1}$ in (z, t) with locally bounded derivatives. Then, conditioned on the observed event times $X = (0 < \tau_1 < \cdots < \tau_{N_{T'}} \le T')$, the posterior intensity process admits the SDE representation

$$\begin{aligned} dZ_t = &\left[b(Z_t, t) + \mathbf{1}_{\{t \le T'\}} \sigma(Z_t, t)^2 \frac{\partial}{\partial z} \log h(Z_t, t, T', X) \right] dt \\ &+ \sigma(Z_t, t)\, d\tilde{B}_t, \quad Z_0 = z_0, \end{aligned}$$

[3] Proof in Appendix C.2.5

where $\tilde{B}$ is a Brownian motion with respect to $\mathcal{G}_t$ and

$$h(z,t,T',X) := \mathbb{E}\left[\exp\left(-\int_t^{T'} Z_s\,ds\right)\prod_{t<\tau_i\le T'} Z_{\tau_i}\middle| Z_t = z\right], \quad (4)$$

with the conditional expectation taken under the prior law of Z. Thus, h is the prior predictive likelihood of the unobserved future portion of the point-process realization, viewed as a function of the current latent state. h is also commonly referred to as a Doob's h-function.

Theorem 2.1 answers the two structural questions:

i. **Diffusion structure is preserved.** Under conditioning on the point process realization, the intensity remains a semi-martingale diffusion with the same diffusion coefficient. Only the drift is modified.

ii. **Posterior correction is a score-type correction.** The additional drift term $\sigma^2\partial_z \log h$ takes the form of a *score function*: it is the gradient of the log-density of future observations with respect to the current state, evaluated under the prior.

The posterior correction is analogous to the classifier guidance term in conditional diffusion models [Dhariwal and Nichol, 2021], where the score of a classifier steers the prior toward the conditioned distribution. Comparing Theorem 2.1 with the variational SDE (Eq. (3)) introduced earlier, we see that the amortized correction u_β is intended to approximate the quantity $\sigma(Z_t,t)\partial_z \log h(Z_t,t,T',X)$. The theorem therefore provides a principled, structured target for which variational inference methodology can be applied to learn. In this sense, EoF does not merely justify the SDE ansatz, but precisely identifies the functional object that amortization must approximate. When u_β can represent the target $\sigma\partial_z \log h$ arbitrarily well, the variational family contains the true posterior, and the variational gap vanishes.

3 VARIATIONAL METHODOLOGY

Consider the empirical NPMLE objective (2), with a single observation X, so that $\hat{\mu} = \delta_X$. For any probability measure Q equivalent to P_θ, i.e, $Q \sim P_\theta$, we write $\log P_\theta(X) = \log \int P(X|Z)\frac{dP_\theta(Z)}{dQ(Z)}\,dQ(Z)$. Applying Jensen's inequality gives the standard evidence lower bound (ELBO)

$$\log P_\theta(X) \ge \mathbb{E}_{Q(Z)}[\log P(X|Z)] - \mathbb{KL}[Q(Z)||P_\theta(Z)]\,. \quad (5)$$

This bound is tight when $Q = P_\theta(Z|X)$.

Recall the posterior drift correction $u_\beta(\cdot)$ indexed by β from Eq. (3), and let $\phi := (\theta,\beta)$. Then ELBO takes the following specific form:

Theorem 3.1. *Suppose the prior model $\{P_\theta(Z)\}$ is specified by Eq.* (1)*, and the variational family $\{Q_\phi\}$ is specified by Eq.* (3)*, then the evidence lower bound for the log-likelihood $\mathcal{L}(\theta,\beta;X)$ reduces to*

$$\mathbb{E}_{Q_\phi}\left[\log P(X|Z) - \frac{1}{2}\int_0^{T'} u_\beta^2(Z_t,t,T',X)\,dt\right]. \quad (6)$$

We direct readers to Appendix. F.1 for the proof of this.

To train our model, we jointly optimize the empirical ELBO,

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\theta,\beta;X^i), \quad (7)$$

with respect to θ and β, learning a generative prior over intensity trajectories and an amortized approximation to the posterior SDE. Once trained, posterior inference for a new point pattern X' reduces to simulating the SDE in Eq. (3), avoiding the need for repeated MCMC sampling.

3.1 IDENTIFIABILITY

The preceding sections are only concerned with maximizing the evidence lower-bound. However, there is no guarantee that the learned drift that optimizes the model fit is unique in describing the true latent dynamics. The Cox observation law identifies the latent random intensity measure through its Laplace functional, and, because the paths are continuous, it identifies the latent intensity-path law. With fixed diffusion coefficient and initial condition, the drift is then identifiable on the region visited by the latent process, up to the corresponding occupation measure. This is formally stated in the two propositions below. Proofs are given in Appendix A.

Proposition 3.2 (Identifiability through the Cox Laplace functional)**.** *Let X and X' be Cox processes on $[0,T]$ with random intensity measures $\Lambda(dt) = Z_t\,dt$ and $\Lambda'(dt) = Z'_t\,dt$, where Z and Z' have continuous nonnegative paths. If X and X' agree in distribution, then Λ and Λ' agree in distribution. Consequently, the induced path laws of Z and Z' agree.*

Proposition 3.3 (Drift identifiability w.r.t. the occupation measure)**.** *Let P^b and $P^{\tilde{b}}$ denote the laws on $C([0,T];\mathbb{R}^d)$ of weak solutions to*

$$dZ_t = b(Z_t,t)\,dt + \sigma(Z_t,t)\,dB_t, \qquad Z_0 = z_0,$$

and

$$dZ_t = \tilde{b}(Z_t,t)\,dt + \sigma(Z_t,t)\,dB_t, \qquad Z_0 = z_0,$$

with the same initial condition and diffusion coefficient. Assume that the coordinate process is a special semimartingale

under both laws and that the drift integrals are well-defined. If $P^b = P^{\tilde{b}} =: P$, then

$$b(Z_t, t) = \tilde{b}(Z_t, t) \qquad (P \otimes dt)\text{-a.e.}$$

Equivalently, $b = \tilde{b}$ μ_P-a.e. for the occupation measure

$$\mu_P(A) := \mathbb{E}_P \left[\int_0^T \mathbf{1}_{\{(Z_t,t)\in A\}}\, dt \right].$$

This identifiability result is what makes drift-level interpretation meaningful: under sufficient assumptions, the learned drift uniquely determines a latent diffusion law. If distinct drifts could induce the same Cox process observation law, then explanations based on the fitted drift would reflect a modeling choice rather than information identifiable from the data.

3.2 OPTIMIZATION STRATEGY

To optimize the ELBO in Eq. (7), we must first evaluate the gradients $\partial_\theta \mathcal{L}(\theta, \beta; X)$ and $\partial_\beta \mathcal{L}(\theta, \beta; X)$ associated with any point process observation X in Eq. (6). Under standard regularity conditions, differentiation can be interchanged with expectation, so the gradients are expectations of functionals of the simulated posterior path and its parameter sensitivities $J_t^\theta := (\partial_\theta Z)_t$ and $J_t^\beta := (\partial_\beta Z)_t$. We simulate the posterior SDE and sensitivity equations with Euler-Maruyama using common Brownian increments, average over m Brownian paths and mini-batched point process observations, and update (θ, β) with Adam. Full formulas are in Appendix F.2.

3.3 POSTERIOR CORRECTION ARCHITECTURES

The prior drift b_θ only depends only on (Z_t, t), so it can be parameterized by an MLP. On the other hand, the posterior correction is more complicated because Theorem 2.1 shows that its target is

$$u^\star(z, t, T', X) = \sigma(z, t)\partial_z \log h(z, t, T', X),$$

which depends on the current state, time, conditioning horizon, and a variable-sized set of future event times. We therefore use a Deep Sets encoder [Zaheer et al., 2017] to pool event-level features before producing the drift correction, which is discussed alongside with other optional architecture below.

3.3.1 Options of Posterior Correction Architecture

The target posterior correction in Eq. (3) is

$$u^\star(z, t, T', X) = \sigma(z, t)\partial_z \log h(z, t, T', X),$$

so an encoder must map (z, t, T') and the variable-sized future-event set $\{\tau_i : t < \tau_i \le T'\}$ to a drift correction.

Remaining-count. The simplest summary is the number of remaining events,

$$r_t(X) := X_{T'} - X_t,$$

and the correction can be parameterized as

$$u_\beta(z, t, T', X) = \sigma(z, t)\mathrm{MLP}_\beta(z, t, T', r_t(X)).$$

This is computationally cheap but discards event-time information, so it is generally misspecified except for special prior models.

Deep Sets. To encode relative event locations while handling variable-length inputs, we use a Deep Sets architecture inspired by [Zaheer et al., 2017]:

$$u_\beta(z, t, T', X) = \sigma(z, t)\rho_\beta \left(z, t, T', \sum_{t<\tau_i\le T'} \psi_\beta(z, \tau_i - t, T' - \tau_i, \Delta_i) \right),$$

where $\Delta_i = \tau_i - \tau_{i-1}$ and ρ_β, ψ_β are MLPs.

Hard-coded constraints. The h-function implies a jump in $\partial_z \log h$ at event times:

$$\lim_{s\nearrow\tau} \partial_z \log h - \lim_{s\searrow\tau} \partial_z \log h = \frac{1}{Z_\tau}.$$

Explicitly adding a hard constraint such as $r_t(X)/Z_t$ can therefore improve training efficiency by encoding part of the known structure.

4 NUMERICAL EXPERIMENTS

In this section, we evaluate prior recovery, amortized posterior inference, and running speed against an EM/MCMC baseline on synthetic and real datasets.

4.1 PRIOR RECOVERY

We begin with a synthetic experiment in which the ground-truth intensity process follows Cox-Ingersoll-Ross (CIR) dynamics,

$$dZ_t = 0.3(80 - Z_t)\, dt + \sqrt{Z_t}\, dB_t, \quad Z_0 = 5. \tag{8}$$

Using this model, we generate 256 Cox process observations on $[0, 4]$, each from an independently sampled intensity path. The SDEs are simulated using the Euler-Maruyama scheme, with 100 even steps. Monte Carlo gradient estimates are computed using 10 i.i.d. Brownian paths (discretized) (details in appendix F.2). The model is trained for 100 epochs with a mini-batch size of 32 and a learning rate of 0.005 for both the prior drift network b_θ and the control network

u_β. To stabilize optimization, gradients are clipped to have an L_2 norm of at most 5. During training, the true prior dynamics are not accessible.

Fig. 2 compares the learned drift with the ground truth.

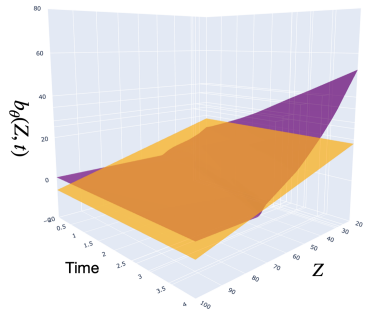


Figure 2: Comparing learned prior drift $b_\theta(Z_t, t)$ (purple) to the ground truth drift $\tilde{b}(Z_t, t) = 0.3(80 - Z_t)$ (yellow).

Note that the nonparametric drift is mainly learned on the state/time region visited the most by the data-generating diffusion. To check this is not a fundamental failure, we also ran a parametrized CIR experiment, where the only difference is that b_θ is parametrized as $b_\theta(z, t) = \theta_1(\theta_2 - z)$ instead of a full-size neural network. Within this correctly specified class, we recover the ground-truth drift; see Fig 11.

Generative performance is shown in Fig. 3: 128 samples from the learned drift match the empirical means and standard deviations of 128 samples from the true drift despite the local drift mismatch in Fig. 2.

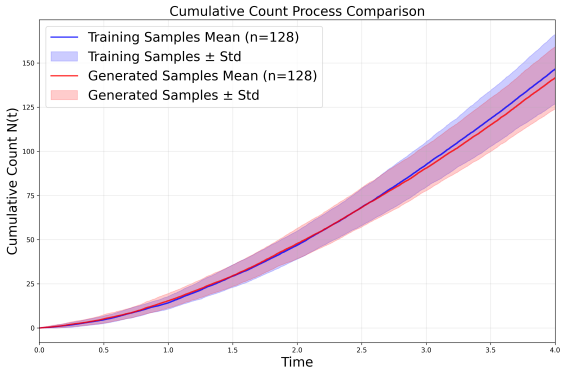


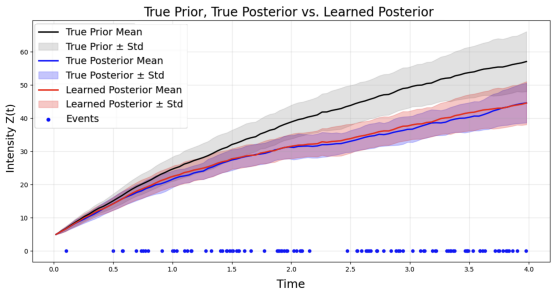
Figure 3: The learned data samples (red) compared to the true data samples (blue).

4.2 AMORTIZED POSTERIOR INFERENCE

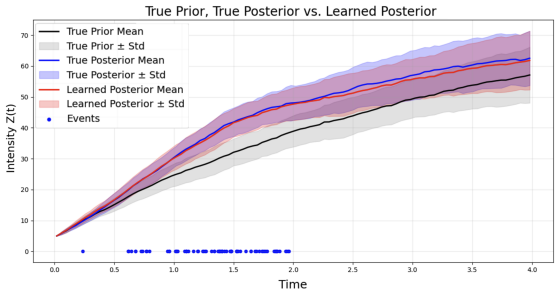
To assess posterior approximation quality, we compare the learned variational posterior with a high-fidelity MCMC approximation[4] using a burn-in of 10,000 iterations. Fig. 4 shows two representative inference scenarios.

In Fig. 4a, the posterior is conditioned on the test sample with the fewest events over $[0, T]$; despite the rarity of such samples in training, the variational posterior tracks the MCMC trajectories closely.

Fig. 4b shows partial observations on $[0, T/2]$ and posterior inference over $[T/2, T]$, where agreement with MCMC indicates that u_β uses the observed events effectively.



(a) Posterior Inference with complete data $X_{0:T}$.



(b) Posterior Inference with partially observed data $X_{0:T/2}$.

Figure 4: The learned posterior sample paths (red) are generated using the learned (b_θ, u_β) pair, and the baseline "true" posterior sample paths are generated using extensive MCMC simulations (blue).

DOES AMORTIZATION OVERFIT?: A natural question concerns the number of point process realizations (and the observation intervals) necessary for the amortized posterior drift correction $X \mapsto u_\beta(z, t, T', X)$ to reliably generalize. Even if the prior b_θ is learned well, an overfitted u_β may still yield unreliable posterior predictions for new point process observations. We use the empirical 2-Wasserstein distance under an L^2 path metric on the simulation grid; lower train than test distance indicates overfitting of the amortized correction.

Fig. 5 shows a train-test gap for $n \leq 8$, indicating overfitting of u_β, and the gap vanishes by $n \geq 16$.

[4]The MCMC posterior paths are generated with the standard Metropolis-Hasting algorithm, with proposal paths generated using the prior.

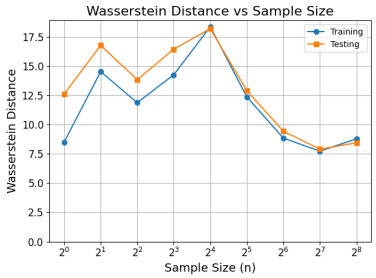


Figure 5: Train vs. test Wasserstein distance between amortized posterior samples and high-fidelity MCMC posterior samples as a function of the training sample size n. A train-test gap indicates overfitting of the amortized correction $u_\beta(\cdots, X)$, which vanishes as n goes beyond 16.

4.3 COMPARISON TO EM

We next compare against an approximate EM baseline (Algorithm 1) that alternates MCMC posterior approximation with gradient-based likelihood maximization.

LEARNING THE MODEL (PRIOR): We first evaluate prior recovery using the expected L^2 deviation between learned and ground-truth intensity paths: $\mathbb{E}\left[\int_0^T \left(Z_t^\theta - Z_t^{\text{gt}}\right)^2 dt\right]$, smaller values indicate better recovery. Using the same set of Brownian noises[5] and a five-hour budget[6], Table 1 shows that VI is comparable to, or better than, EM while also learning the amortized correction u_β.

True drift $\tilde{b}$	EM (ℓ^2 loss)	VI (ℓ^2 loss)
$0.3(80 - Z_t)$	6.159	5.792
$0.3(80-Z_t)-5t$	6.721	4.989

Table 1: Comparison of the learned prior quality in terms of ℓ^2 deviation of the intensity paths.

INFERENCE TIME: We compare posterior inference efficiency by matching posterior predictive log-likelihood,

$$\mathbb{E}_{X\sim\mu_{\text{gt}}}\left[\mathbb{E}_{Z\sim Q(\cdot|X)}\left[\sum_{T'<\tau_i\le T}\log Z_{\tau_i} - \int_{T'}^{T} Z_t\, dt\right]\right], \tag{9}$$

and measuring time to generate 32 posterior paths for each of 128 test observations. VI samples by simulating Eq. (3); EM uses a Metropolis-Hastings posterior sampler, where the number of MCMC steps is chosen such that the posterior quality matches that of the VI posteriors.

Tables 2 and 3 show one-to-two-order-of-magnitude speedups for VI at comparable predictive log-likelihood.

[5]The same Brownian noises are used in generating the intensity paths for all the trained drifts, which ensures they are compared w.r.t. to the same ω in the sample space.

[6]On an M3 MacBook Air with 8GB of RAM.

$\tilde{b}(Z_t, t) = 0.3(80 - Z_t)$

Horizon	Likelihood	MCMC	VI
$[0, T]$	398.0	17m 38.5s	39.9s
$[T/4, T]$	370.1	12m 43.8s	9.6s
$[T/2, T]$	288.4	14m 3.3s	19.6s
$[3T/4, T]$	162.6	15m 49.8s	29.7s

Table 2: Comparison of computation time used for posterior inference on the testing dataset. The data is generated with the ground truth drift $\tilde{b}(Z_t, t) = 0.3(80 - Z_t)$.

$\tilde{b}(Z_t, t) = 0.3(80 - Z_t) - 5t$

Horizon	Likelihood	MCMC	VI
$[0, T]$	249.9	16m 17.7s	38.2s

Table 3: Comparison of computation time with time-dependent $\tilde{b}(Z_t, t) = 0.3(80 - Z_t) - 5t$.

4.4 US BANK DATASET

Finally, we evaluate the method on call records from a large U.S.-based bank call center[7]. The call counts exhibit strong time-of-day variation and over-dispersion [Brown et al., 2005], motivating a Cox-process model. The data records arrivals per minute; we spread calls evenly within each minute, use 128 Mondays for training, and thin arrivals with probability 0.001 for training stability.

The model uses the same optimization procedure and five-hour training budget as the synthetic experiments. Since the true prior is unknown, Fig. 6 compares fitted-model samples with all training observations. The fitted mean captures the main temporal pattern, including peak periods, while variance mismatch suggests that learning the diffusion coefficient σ_α is an important next extension.

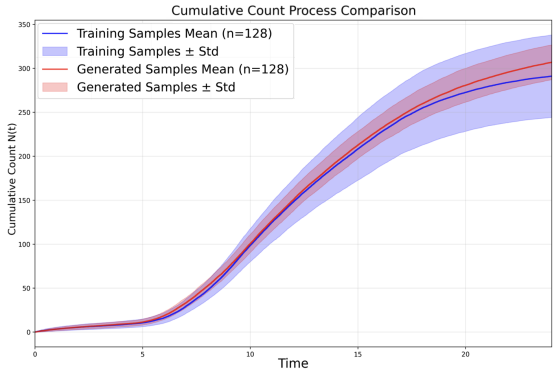


Figure 6: The learned data samples (red) compared to the true data samples (blue).

[7]https://seelab.net.technion.ac.il/data/; see Appendix H.

4.5 NEURAL SPIKES DATASET

We include an additional real-data experiment using the neural spike-train data from the Allen Visual Coding Neuropixels dataset[8]. Previous experiments only tested posterior recovery on synthetic CIR datasets, this experiment extends that to a real-world dataset.

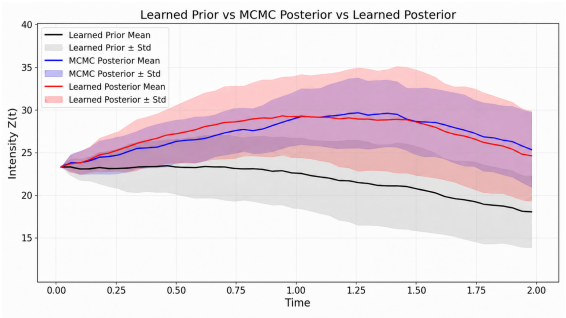


Figure 7: Neural spike-train experiment. The model is evaluated on a high-count spike-train observation. The learned amortized posterior matches the sample-wise MCMC posterior closely in their mean behavior, while the remaining variance mismatch may be attributed to using a fixed CIR diffusion during training.

The main qualitative conclusion is that the same inference framework can be used without changing the model or optimization procedure. At the same time, we still observe variance mismatch as the diffusion term is treated as fixed during the training process. We hope to extend to more flexible, learnable diffusion term in future works.

The mutivariate experiments are included in Appendix I showing that the same amortized posterior-correction idea extends beyond the scalar point-process setting.

5 CONCLUSION

We introduce Neural Diffusion Intensity Models, a variational framework for Cox process models with neural SDE intensities. Our key theoretical contribution leverages enlargement of filtrations to show that conditioning on point process observations preserves the diffusion structure of the latent intensity with an explicit drift correction, establishing a path-measure analogue of a conjugacy between our neural SDE priors and the Poisson likelihood. When the variational correction can represent the exact posterior correction, the variational gap vanishes and ELBO maximization coincides with likelihood maximization. Experiments show accurate prior recovery and orders-of-magnitude faster posterior inference than MCMC-based EM, with the same framework extending naturally to richer diffusion-intensity models.

[8]Dataset available at `https://allensdk.readthedocs.io/en/latest/visual_coding_neuropixels.html`.

Acknowledgements VR acknowledges support from the National Science Foundation through grant DMS-2503118. HH and XD are supported by the National Science Foundation through grant CAREER/CMMI-2143752.

References

Ryan Prescott Adams, Iain Murray, and David J.C. MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 9–16, 2009.

Virginia Aglietti, Edwin V Bonilla, Theodoros Damoulas, and Sally Cripps. Structured variational inference in continuous cox process models. *Advances in Neural Information Processing Systems*, 32, 2019.

Ken-ichi Amemori and Shin Ishii. Gaussian process approach to spiking neurons for inhomogeneous poisson inputs. *Neural computation*, 13(12):2763–2797, 2001.

Cedric Archambeau, Dan Cornford, Manfred Opper, and John Shawe-Taylor. Gaussian process approximations of stochastic differential equations. In *Gaussian Processes in Practice*, volume 1 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2007. URL `https://proceedings.mlr.press/v1/archambeau07a.html`.

A.N. Avramidis and P. L'Ecuyer. Modeling and simulation of call centers. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 9 pp.–, 2005. doi: 10.1109/WSC.2005.1574247.

Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):333–382, 2006.

Tomas Björk. *Point Processes and Jump Diffusions: An Introduction with Finance Applications*. Cambridge University Press, 2021. ISBN 978-1-316-51867-0. doi: 10.1017/9781009002127.

Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer Series in Statistics. Springer-Verlag, New York, 1981. ISBN 978-0-387-90536-5.

Lawrence Brown, Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn, and Linda Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005. doi: 10.1198/016214504000001808.

D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen, London, 1966.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer, New York, 2nd edition, 2003.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

Christian Donner and Manfred Opper. Efficient bayesian inference of sigmoidal gaussian cox processes. *Journal of Machine Learning Research*, 19(67):1–34, 2018.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564, 2016.

Lea Duncker, Gergo Bohner, Julien Boussard, and Maneesh Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *International conference on machine learning*, pages 1726–1734. PMLR, 2019.

Paul Glasserman. *Gradient Estimation Via Perturbation Analysis*, volume 116 of *The Springer International Series in Engineering and Computer Science*. Kluwer Academic Publishers, Boston, 1991. ISBN 978-0-7923-9095-4.

Emmanuel Gobet and Rémi Munos. Sensitivity analysis using Itô–Malliavin calculus and martingales, and application to stochastic optimal control. *SIAM Journal on Control and Optimization*, 43(5):1676–1713, 2005. doi: 10.1137/S0363012902419059.

Flávio B. Gonçalves, Krzysztof G. Łatuszyński, and Gareth O. Roberts. Exact bayesian inference for diffusion-driven cox processes. *Journal of the American Statistical Association*, 119(547):1882–1894, 2024. doi: 10.1080/01621459.2023.2223791. URL https://doi.org/10.1080/01621459.2023.2223791.

Karen Grigorian and Robert A. Jarrow. Enlargement of filtrations: An exposition of core ideas with financial examples, 2023. URL https://arxiv.org/abs/2303.03573.

Ali Hasan, Yu Chen, Yuting Ng, Mohamed Abdelghani, Anderson Schneider, and Vahid Tarokh. Inference and sampling of point processes from diffusion excursions. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 839–848, 2023.

Jean Jacod. Grossissement initial, hypothese (H') et theoreme de Girsanov. In *Grossissements de filtrations: exemples et applications: Séminaire de Calcul Stochastique 1982/83 Université Paris VI*, pages 15–35. Springer, 2006.

Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2 edition, 2003.

Prateek Jaiswal, Harsha Honnappa, and Vinayak Rao. Variational inference for diffusion modulated Cox processes. https://openreview.net/forum?id=snaT4xewUfX, 2021. Unpublished manuscript.

Monique Jeanblanc, Marc Yor, and Marc Chesney. *Mathematical Methods for Financial Markets*. Springer Finance. Springer, 2009.

Geurt Jongbloed and Ger Koole. Managing uncertainty in call centres using poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17 (4):307–318, 2001. doi: https://doi.org/10.1002/asmb.444. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.444.

Olav Kallenberg. *Random Measures, Theory and Applications*, volume 77 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2017.

Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, third edition, 2021. ISBN 978-3-030-61870-4. doi: 10.1007/978-3-030-61871-1.

Patrick Kidger, James Foster, Xuechen Li, and Terry J Lyons. Neural SDEs as infinite-dimensional GANs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5453–5463. PMLR, 2021. URL https://proceedings.mlr.press/v139/kidger21b.html.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2101–2109, 2017.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3870–3882. PMLR, 2020. URL https://proceedings.mlr.press/v108/li20i.html.

Bruce G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, Hayward, CA, 1995a. ISBN 0940600323, 9780940600324.

Bruce G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and American Statistical Association, Hayward, CA, 1995b.

David Lüdke, Marin Biloš, Oleksandr Shchur, Marten Lienen, and Stephan Günnemann. Add and thin: Diffusion for temporal point processes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Joseph Marino, Milan Cvitkovic, and Yisong Yue. A general method for amortizing variational filtering. In *Advances in Neural Information Processing Systems*, pages 7868–7879, 2018.

Nazanin Mehrasa, Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. A variational auto-encoder model for stochastic point processes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3165–3174, 2019.

Hongyuan Mei and Jason Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1612.09328.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998. doi: 10.1111/1467-9469.00115.

Bernt Øksendal. Stochastic differential equations. In *Stochastic differential equations: an introduction with applications*, pages 38–50. Springer, 2003.

Manfred Opper. Variational inference for stochastic differential equations. *Annalen der Physik*, 531(3):1800233, 2019. doi: https://doi.org/10.1002/andp.201800233. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.201800233.

Zhen Pan, Zhenya Huang, Defu Lian, and Enhong Chen. A variational point process model for social event sequences. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 173–180, 2020.

Patrick O. Perry and Patrick J. Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):821–849, 2013.

Tom Ryder, Andrew Golightly, A. Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4423–4432. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ryder18a.html.

SEE Lab, Technion. SEE lab dataset. https://seelab.net.technion.ac.il/data/, 2024. Accessed: 2025.

Oleksandr Shchur, Nicholas Gao, Marin Biloš, and Stephan Günnemann. Fast and flexible temporal point processes with triangular maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural temporal point processes: A review. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4585–4593, 2021.

Ryohei Shibue and Tomoharu Iwata. Marked point process variational autoencoder with applications to unsorted spiking activities. *PLOS Computational Biology*, 20(12): e1012620, 2024.

Daniel Simpson, Janine Baerbel Illian, Finn Lindgren, Sigrunn H Sørbye, and Havard Rue. Going off grid: Computationally efficient inference for log-gaussian cox processes. *Biometrika*, 103(1):49–70, 2016.

Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3084–3114. PMLR, 2019. URL https://proceedings.mlr.press/v99/tzen19a.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pages 3391–3401, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Xiaowei Zhang, L. Jeff Hong, and Jiheng Zhang. Scaling and modeling of call center arrivals. In *Proceedings of the 2014 Winter Simulation Conference*, pages 476–485, 2014. doi: 10.1109/WSC.2014.7019913.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11692–11702. PMLR, 2020. URL https://proceedings.mlr.press/v119/zuo20a.html.

Neural Diffusion Intensity Models for Point Process Data (Supplementary Material)

Xinlong Du[1] **Harsha Honnappa**[1] **Vinayak Rao**[2]

[1]Edwardson School of Industrial Engineering, Purdue University
[2]Department of Statistics, Purdue University

A PROOF OF IDENTIFIABILITY STATEMENTS

A.1 PROOF OF PROPOSITION. 3.2

Proof. For any continuous nonnegative test function f, the Cox Laplace functional is

$$\begin{aligned} L_X(f) &= \mathbb{E}\left[\exp\left(-\int_0^T f(t)X(dt)\right)\right] \\ &= \mathbb{E}\left[\exp\left(-\int_0^T (1-e^{-f(t)})\Lambda(dt)\right)\right]. \end{aligned}$$

Equality in law of X and X' gives equality of these functionals for all such f. As f varies, $1-e^{-f}$ ranges over continuous functions with values in $[0,1)$. This still determines the Laplace transform of $\Lambda(g)$ for every continuous nonnegative g: for small $s>0$, sg lies in this range, and uniqueness of one-dimensional Laplace transforms extends equality to all $s \geq 0$. Applying the same argument to finite linear combinations of test functions, or equivalently using the Laplace criterion for random measures [Kallenberg, 2017, Corollary 2.3], yields equality in distribution of Λ and Λ'. Since continuous functions that agree Lebesgue-a.e. agree everywhere, the map $z \mapsto z(t)\,dt$ is injective on continuous nonnegative paths, giving equality of the induced laws of Z and Z'. □

A.2 PROOF OF PROPOSITION. 3.3

Proof. On canonical path space, the same coordinate process admits under P the two decompositions

$$Z_t = Z_0 + \int_0^t b(Z_s,s)\,ds + M_t, \qquad Z_t = Z_0 + \int_0^t \tilde{b}(Z_s,s)\,ds + \tilde{M}_t,$$

where the finite-variation terms are predictable and $M, \tilde{M}$ are local martingales. By uniqueness of the canonical decomposition of a special semimartingale [Jacod and Shiryaev, 2003, Chapter II, 2c],

$$\int_0^t \{b(Z_s,s) - \tilde{b}(Z_s,s)\}\,ds = 0 \qquad \text{for all } t \in [0,T], \quad P\text{-a.s.}$$

Thus $b(Z_t,t) = \tilde{b}(Z_t,t)$ for $(P \otimes dt)$-a.e. (ω,t), equivalently μ_P-a.e. □

B MEASURE-THEORETIC FOUNDATIONS OF THE NEURAL SDE COX MODEL

We provide precise conditions under which the path measures in the Neural SDE Cox model (Section 2.1) are well-defined. Throughout, let $\mathcal{Z} := C([0,T], \mathbb{R}_{\geq 0})$ denote the space of non-negative continuous paths on $[0,T]$ equipped with the sup-norm topology and its Borel σ-algebra $\mathcal{B}(\mathcal{Z})$, and let $\mathcal{X} = \cup_{k\in\mathbb{N}} \mathbb{R}^k$ denote the space of locally finite counting measures on $[0,T]$ equipped with the vague topology and its Borel σ-algebra $\mathcal{B}(\mathcal{X})$.

Existence and uniqueness of the SDE solution. We require the drift and diffusion coefficients of the SDE (1) to satisfy the following standard conditions:

(**A1**) *Global Lipschitz continuity.* There exists a constant $K > 0$ such that for all $x, y \in \mathbb{R}$ and $t \in [0,T]$,

$$|b_\theta(x,t) - b_\theta(y,t)| + |\sigma(x,t) - \sigma(y,t)| \leq K|x-y|.$$

(**A2**) *Linear growth.* There exists a constant $C > 0$ such that for all $x \in \mathbb{R}$ and $t \in [0,T]$,

$$|b_\theta(x,t)| + |\sigma(x,t)| \leq C(1+|x|).$$

(**A3**) *Joint measurability.* The functions $b_\theta(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ are jointly Borel measurable in (x,t).

Under (**A1**)–(**A3**), the SDE (1) admits a unique strong solution for each $\theta \in \Theta$ [Øksendal, 2003, Theorem 5.2.1]. This solution induces a well-defined probability measure $P_\theta(Z)$ on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ via the pushforward of the Wiener measure through the Itô map.

Non-negativity of the intensity process. Since Z serves as an intensity, we additionally require:

(**A4**) *Non-negativity.* $Z_t \geq 0$ almost surely for all $t \in [0,T]$.

This condition is not implied by (**A1**)–(**A3**) and must be enforced through the model structure. For instance, if $\sigma(z,t) = \sigma_0\sqrt{z}$ (as in the CIR model), the Feller condition $2b_\theta(0,t) \geq \sigma_0^2$ for all t ensures that the boundary at zero is unattainable. Alternatively, one may model $\log Z_t$ as the SDE solution and recover Z_t by exponentiation, guaranteeing positivity. The global-Lipschitz assumptions above are sufficient conditions used for a clean well-definedness statement. They do not cover the square-root CIR diffusion coefficient used in the synthetic experiments. Instead, the CIR case is covered by the standard CIR existence and boundary theory under the usual Feller-type positivity condition. Thus, the appendix should be read as giving one regular sufficient regime, with the CIR experiments falling under a separate classical non-Lipschitz but well-posed regime.

The Poisson kernel as a Markov kernel. Given a non-negative intensity path $Z \in \mathcal{Z}$, the conditional law $P(X \mid Z)$ is that of an inhomogeneous Poisson process with rate function $(Z_t)_{t\in[0,T]}$. Its Radon–Nikodym derivative with respect to the unit-rate Poisson measure P_0 is

$$\frac{dP(X \mid Z)}{dP_0} = \exp\left(\int_0^T \log Z_t \, dN_t - \int_0^T (Z_t - 1)\, dt\right), \tag{10}$$

where X_t denotes the counting process associated with X. We require:

(**A5**) *Integrability.* $\int_0^T Z_t \, dt < \infty$ almost surely.

(**A6**) *Strict positivity at event times.* $P_\theta(Z_t > 0 \text{ for all } t \in [0,T]) = 1$.

Condition (**A5**) is automatically satisfied when Z has continuous paths on the compact interval $[0,T]$, but we state it explicitly for completeness. Condition (**A6**) ensures that $\log Z_t$ in (10) is well-defined at event times; it is implied by the Feller condition in the CIR setting or by the log-normal construction. Under (**A5**)–(**A6**), the map $Z \mapsto P(X \mid Z)$ defines a Markov kernel from $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, since the Poisson likelihood (10) is a measurable functional of the intensity path.

Well-definedness of the joint and marginal measures. Under conditions (**A1**)–(**A6**), the joint measure

$$P_\theta(X, Z) = P(X \mid Z)\, P_\theta(Z)$$

is well-defined on the product σ-algebra $\mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{Z})$ by the standard Markov kernel construction [Kallenberg, 2021, Chapter 3]. The marginal observation law

$$P_\theta(X) = \int_{\mathcal{Z}} P(X \mid Z)\, dP_\theta(Z)$$

is then a well-defined probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.

Regularity in θ. For gradient-based optimization via the pathwise ("differentiate then simulate") method [Glasserman, 1991, Gobet and Munos, 2005], we further require:

(**A7**) *Differentiability in parameters.* The map $\theta \mapsto b_\theta(x, t)$ is continuously differentiable for all (x, t), and $\nabla_\theta b_\theta$ satisfies conditions (**A1**)–(**A2**) uniformly in θ over compact subsets of Θ.

Under (**A7**), the augmented system of SDEs for $(Z_t, \partial_\theta Z_t, \partial_\beta Z_t)$ admits a unique strong solution, ensuring that the pathwise sensitivity processes are well-defined and that the resulting Monte Carlo gradient estimators are unbiased (up to time-discretization error).

C ENLARGEMENT OF FILTRATIONS

In this section, we will work with a *filtered probability space* $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ where the probability measure $\mathbb{P}$ encodes our "beliefs", and the filtration—collection of σ-algebras—$\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ represents the information that's available to us at any time. A stochastic process $(X_t)_{t\geq 0}$ is said to be *adapted* to $\mathbb{F}$ if X_t is $\mathcal{F}_t$-measurable for all t. That is, the value of X_t can be determined exactly by the information available at time t. The mathematical framework that studies how stochastic processes behave under such changes in information is called Enlargement of Filtrations, which has been studied extensively in the probability literature; see [Grigorian and Jarrow, 2023, Jacod, 2006, Jeanblanc et al., 2009].

C.1 MARTINGALES

Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space, and let M_t be an $\mathbb{F}$-adapted stochastic process with $\mathbb{E}\,|M_t| < \infty$ for all t. If

$$\mathbb{E}[M_t|\mathcal{F}_s] = M_s$$

for all $0 \leq s < t$, then M_t is said to be an $\mathbb{F}$-martingale. Intuitively, a martingale represents a "fair game": given the current information, its expected future value is just the present value. A process X_t is a *semimartingale* if it can be decomposed as

$$X_t = M_t + A_t$$

where M_t is a local martingale (which behaves like a martingale up to random stopping times)[1] and A_t is an $\mathbb{F}$-adapted finite-variation process. Semimartingales form the largest class of stochastic processes for which we can talk about Itô calculus.

C.2 ENLARGING THE FILTRATION

X_t may be a martingale with respect to some filtration $\mathbb{F}$, but not necessarily when $\mathbb{F}$ is "enlarged" with extra information ζ. Describing X_t in the presence of this information is the goal of Enlargement of Filtrations.

[1] A process M_t is a local martingale if there exists an increasing sequence of stopping times $\tau_n \nearrow \infty$ such that each stopped process $M_{t\wedge\tau_n}$ is a martingale.

C.2.1 Setup

There are two main types of "enlargement":

- **Initial enlargement:** all secret information is available at time 0.
- **Progressive enlargement:** information is gradually revealed over time.

We focus on the first case, which will be related to our variational problem. Let $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ be the original filtration, and let ζ be an $\mathcal{F}$-measurable random variable taking values in some measurable space E. We define the *enlarged filtration* $\mathbb{G} = (\mathcal{G}_t)_{t\geq 0}$ by

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(\zeta),$$

where $\vee$ denotes the smallest σ-algebra containing elements from $\mathcal{F}_t$ and $\sigma(\zeta)$. Intuitively, $\mathbb{G}$ describes a world in which the value of ζ is known from the start—one has "insider" information about ζ at time 0.

C.2.2 The $\mathcal{H}$ and $\mathcal{H}'$ Hypotheses

The question is how martingales behave under a filtration enlargement.

If every $\mathbb{F}$-martingale remains a $\mathbb{G}$-martingale, we say that $\mathbb{F}$ is *immersed* in $\mathbb{G}$, or that the pair $(\mathbb{F}, \mathbb{G})$ satisfies the $\mathcal{H}$-**hypothesis**. This is quite a strong condition—informally, it means that the extra information carried by ζ does not interfere with the "fairness" of any $\mathbb{F}$-martingale.

A weaker but more flexible condition is the $\mathcal{H}'$**-hypothesis**: every $\mathbb{F}$-martingale remains a $\mathbb{G}$-*semimartingale*. Under this assumption, Itô calculus remains valid, but an additional drift term may appear when expressing the $\mathbb{F}$-martingale in the $\mathbb{G}$ filtration. This is what's going under the hood when we enlarge the filtration with the observed point process $X_{0:T'}$.

C.2.3 Jacod's condition

Among several sufficient conditions ensuring the $\mathcal{H}'$-hypothesis, *Jacod's condition* is often the most convenient in practice. The following statements (theorem and lemma) are adapted from [Grigorian and Jarrow, 2023].

Theorem C.1 (Jacod's condition)**.** *Let ζ be a random element in a standard Borel space $(E, \mathcal{E})$, and let $Q_t(\omega, dx)$ be the regular conditional distribution of ζ given $\mathcal{F}_t$ for all $t \geq 0$. Suppose that there exists a positive σ-finite[2] measure η on $(E, \mathcal{E})$ such that*

$$Q_t(\omega, dx) \ll \eta(dx) \text{ a.s., } \forall t \geq 0.$$

Then $\mathcal{H}'$ holds.

Just knowing that a process remains a semimartingale is usually not enough, we typically would like the semimartingale decomposition in the enlarged filtration[3] $\mathbb{F}^{\sigma(\zeta)}$. To do so, we need the following lemma and theorem.

Lemma C.2. *Under $\mathcal{J}$, there exists a nonnegative $\mathcal{O}(\mathbb{F}) \times \mathcal{B}(\bar{\mathbb{R}})$-measurable function*

$$\Omega \times \mathbb{R}_+ \times \bar{\mathbb{R}} \ni (\omega, t, x) \mapsto p_t(\omega, x)$$

cadlag in t such that

(i) $Q_t(\omega, dx) = p_t(\omega, x)\eta(dx)$ *for every* $t \geq 0$,

(ii) for each $x \in \bar{\mathbb{R}}$, *the process* $(p_t(x))_{t\geq 0}$ *is an* $\mathbb{F}$*-martingale,*

(iii) $p(x) > 0$ *and* $p_-(x) > 0$ *on*[4] $[\![0, \zeta^x[\![$ *and* $p(x) = 0$ *on* $[\![\zeta^x, \infty[\![$, *where* $\zeta^x := \inf\{t : p_-(x) = 0\}$.

Theorem C.3. *Suppose that the random element ζ satisfies $\mathcal{J}$. If X is an $\mathbb{F}$-local martingale, the process $\tilde{X}$ defined as*

$$\tilde{X}_t := X_t - \int_0^t \frac{1}{p_{s-}(\zeta)} d\langle X, p(u)\rangle_s^{\mathbb{F}}|_{u=\zeta}, \quad t \leq T \quad (*)$$

is an $\mathbb{F}^{\sigma(\zeta)}$-local martingale.

[2] A measure is called σ-finite if it can be written as a countable union of finite measures.

[3] We use $\mathbb{F}^{\sigma(\zeta)}$ to denote the right-continuous version of the enlarged filtration $\mathbb{F} \vee \sigma(\zeta)$.

[4] The double square bracket is used for intervals with random boundary points.

C.2.4 Example: Brownian bridge via initial enlargement

Consider a standard Brownian motion $B = (B_t)_{t\geq 0}$ on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ with its natural filtration ($\mathcal{F}_t = \sigma(X_s, 0 \leq s \leq t)$). Suppose we enlarge $\mathbb{F}$ with the information given by $\zeta := B_T$, i.e.,

$$\mathcal{G}_t = \mathcal{F}_t \vee \sigma(B_T), \qquad \mathbb{G} = (\mathcal{G}_t)_{t\geq 0}.$$

As always, $\mathbb{G}$ is assumed to be right-continuous by taking

$$\mathcal{G}_t^+ = \bigcap_{s>t} \mathcal{G}_s.$$

Step 1: Verifying Jacod's condition. For $t < T$, the conditional law of B_T given $\mathcal{F}_t$ is Gaussian:

$$B_T \mid \mathcal{F}_t \sim \mathcal{N}\big(B_t,\ T-t\big).$$

Hence there exists a dominating measure (the Lebesgue measure) $\eta(dx) = dx$. $Q_t(\omega, dx)$ has a density with respect to Lebesgue measure: $Q_t(\omega, dx) \ = \ p_t(\omega, x)\, dx$, with

$$p_t(x) = \frac{1}{\sqrt{2\pi(T-t)}} \exp\Big(- \frac{(x-B_t)^2}{2(T-t)}\Big), \quad 0 \leq t < T.$$

Thus $Q_t(\cdot, dx) \ll \eta(dx)$ a.s. for each $t < T$, so Jacod's condition holds on $[0, t)$. (On $[T, \infty)$, the conditional law is degenerate at B_T.)

Step 2: Computing the semimartingale decomposition. We apply Theorem C.3 with $X = B$. To do so, we first compute the $\mathbb{F}$-predictable covariation $\langle B, p(x)\rangle^{\mathbb{F}}$. By Itô's formula, we can get

$$dp_t(x) = \partial_{B_t} p_t(x)\, dB_t \ = \ \frac{x - B_t}{T-t}\, p_t(x)\, dB_t, \qquad t < T.$$

Hence the predictable covariation with B is just

$$d\big\langle B, p(x)\big\rangle_t^{\mathbb{F}} = \frac{x - B_t}{T-t}\, p_t(x)\, dt, \qquad 0 \leq t < T.$$

Plugging this into $(*)$ with $x = \zeta = B_T$, we obtain from Theorem C.3 that the process

$$\tilde{B}_t := B_t - \int_0^t \frac{B_T - B_s}{T-s}\, ds, \qquad 0 \leq t < T$$

is a $\mathbb{G}$-local martingale. Since B is continuous with quadratic variation $[B]_t = t$, it follows that $[\tilde{B}]_t = t$ as well, and by Lévy's characterization $\tilde{B}$ is a $\mathbb{G}$-Brownian motion on $[0, T)$.

Equivalently, B admits the $\mathbb{G}$-semimartingale (indeed, SDE) representation

$$dB_t = d\tilde{B}_t \ + \ \frac{B_T - B_t}{T-t}\, dt, \qquad 0 \leq t < T,$$

where $\tilde{B}$ is a $\mathbb{G}$-Brownian motion. This is precisely the *Brownian bridge* drift toward the terminal pinning value B_T.

C.2.5 Proof of Theorem 2.1: Intensity process under initial enlargement

We give the proof for the scalar case; Appendix D gives the vector-valued analogue. Let $X = N_{0:T'}$ denote the realized point-process path, and let η be the law of a unit-rate Poisson process on $[0, T']$. Conditional on a latent path Z, the likelihood of a deterministic counting path x relative to η is

$$L_{T'}(x; Z) = \exp\left(\int_0^{T'} \log Z_s\, x(ds) - \int_0^{T'} (Z_s - 1) ds \right).$$

Hence the conditional law $Q_t(\omega, dx) := \mathbb{P}(X \in dx \mid \mathcal{F}_t)(\omega)$ is absolutely continuous with respect to $\eta(dx)$, with density process

$$p_t(x) = \mathbb{E}\left[L_{T'}(x; Z) \mid \mathcal{F}_t\right].$$

This verifies Jacod's condition under the standing positivity and integrability assumptions. By the Markov property, for $t \le T'$ we may factor

$$p_t(x) = A_t(x) H_x(Z_t, t),$$

where $A_t(x)$ depends only on the observed part of x up to time t and has finite variation, while

$$H_x(z,t) = \mathbb{E}\left[\exp\left(\int_t^{T'} \log Z_s \, x(ds) - \int_t^{T'} (Z_s - 1) ds\right) \middle| Z_t = z\right].$$

Up to the multiplicative factor $e^{T'-t}$, which is independent of z, this is exactly

$$h(z,t,T',x) = \mathbb{E}\left[\exp\left(-\int_t^{T'} Z_s ds\right) \prod_{t<\tau_i \le T'} Z_{\tau_i} \middle| Z_t = z\right].$$

Therefore $\partial_z \log H_x = \partial_z \log h$. Since $A_t(x)$ is finite variation, the continuous martingale part of $p_t(x)$ comes from $H_x(Z_t, t)$. Applying Itô's formula gives

$$dp_t(x)^c = p_{t-}(x)\, \sigma(Z_t, t) \partial_z \log h(Z_t, t, T', x)\, dB_t,$$

and hence

$$d\langle B, p(x)\rangle_t = p_{t-}(x)\, \sigma(Z_t, t) \partial_z \log h(Z_t, t, T', x)\, dt.$$

The semimartingale decomposition theorem for initial enlargement then yields

$$\widetilde{B}_t = B_t - \int_0^{t \wedge T'} \sigma(Z_s, s) \partial_z \log h(Z_s, s, T', X) ds,$$

which is a Brownian motion in the enlarged filtration. Equivalently,

$$dB_t = d\widetilde{B}_t + \mathbf{1}_{\{t \le T'\}} \sigma(Z_t, t) \partial_z \log h(Z_t, t, T', X) dt.$$

Substituting this identity into the prior SDE

$$dZ_t = b(Z_t, t) dt + \sigma(Z_t, t) dB_t$$

gives

$$dZ_t = \left[b(Z_t, t) + \mathbf{1}_{\{t \le T'\}} \sigma(Z_t, t)^2 \partial_z \log h(Z_t, t, T', X)\right] dt + \sigma(Z_t, t) d\widetilde{B}_t,$$

which proves Theorem 2.1.

D MULTIVARIATE ENLARGEMENT-OF-FILTRATIONS RESULT

We now state the multivariate analogue of Theorem 2.1. Let $Z_t \in \mathbb{R}^\ell$ solve

$$dZ_t = b(Z_t, t) dt + \Sigma(Z_t, t) dB_t,$$

where B is p-dimensional Brownian motion, $\Sigma(z,t) \in \mathbb{R}^{\ell \times p}$, and $a(z,t) := \Sigma(z,t)\Sigma(z,t)^\top$. We observe d counting processes $X^1, \dots, X^d$ on $[0, T']$. Conditional on the latent path Z, the channels are independent inhomogeneous Poisson processes with intensities $\lambda_i(Z_t)$, $i = 1, \dots, d$, where the link functions $\lambda_i : \mathbb{R}^\ell \to (0, \infty)$ are known. This allows dependence across observed channels through the shared latent diffusion, while preserving conditional independence given Z.

For a realized multichannel observation $X = (X^1, \dots, X^d)$, with $X^i = \sum_{k=1}^{K_i} \delta_{\tau_k^i}$, define

$$h_X(z,t) := \mathbb{E}\left[\exp\left(-\sum_{i=1}^d \int_t^{T'} \lambda_i(Z_s) ds\right) \prod_{i=1}^d \prod_{t<\tau_k^i \le T'} \lambda_i(Z_{\tau_k^i}) \middle| Z_t = z\right].$$

This is the prior predictive likelihood of the future multichannel observation, viewed as a function of the current latent state.

Theorem D.1 (Multivariate EoF posterior diffusion)**.** *Assume Jacod's condition holds for the initial enlargement by $X = N_{0:T'}$, the likelihood terms above are integrable and strictly positive, and h_X is piecewise $C^{2,1}$ in (z,t) between consecutive event times across all channels, with locally bounded spatial derivatives and $\Sigma(Z_t,t)^\top \nabla_z \log h_X(Z_t,t)$ locally square-integrable along the prior path. Then, in the enlarged filtration $\mathcal{G}_t = \mathcal{F}_t \vee \sigma(N_{0:T'})$, the posterior latent process satisfies*

$$dZ_t = \left[b(Z_t,t) + \mathbf{1}_{\{t\le T'\}} a(Z_t,t)\nabla_z \log h_X(Z_t,t)\right] dt + \Sigma(Z_t,t) d\widetilde{B}_t,$$

where $\widetilde{B}$ is a p-dimensional Brownian motion with respect to $(\mathcal{G}_t)$.

Proof. Let η be the law of d independent unit-rate Poisson processes on $[0,T']$. For a deterministic multichannel counting path x, the conditional likelihood relative to η is

$$L_{T'}(x;Z) = \exp\left(\sum_{i=1}^{d}\int_0^{T'} \log\lambda_i(Z_s)x^i(ds) - \sum_{i=1}^{d}\int_0^{T'}(\lambda_i(Z_s)-1)ds\right).$$

Thus $Q_t(\omega,dx) = \mathbb{P}(X\in dx \mid \mathcal{F}_t)(\omega)$ admits the density $p_t(x) = \mathbb{E}[L_{T'}(x;Z)\mid \mathcal{F}_t]$, so Jacod's condition holds. Factoring the likelihood at time t gives $p_t(x) = A_t(x)H_x(Z_t,t)$, where $A_t(x)$ is finite variation and $H_x = e^{d(T'-t)}h_x$. Since the exponential factor is independent of z, $\nabla_z \log H_x = \nabla_z \log h_x$. Applying Itô's formula to $H_x(Z_t,t)$ between observed event times gives the continuous martingale part

$$dp_t(x)^c = p_{t-}(x)\{\Sigma(Z_t,t)^\top \nabla_z \log h_x(Z_t,t)\}^\top dB_t.$$

Therefore

$$d\langle B, p(x)\rangle_t = p_{t-}(x)\Sigma(Z_t,t)^\top \nabla_z \log h_x(Z_t,t)dt.$$

Jacod's decomposition applied to the Brownian motion yields

$$\widetilde{B}_t = B_t - \int_0^{t\wedge T'} \Sigma(Z_s,s)^\top \nabla_z \log h_X(Z_s,s)ds,$$

which is Brownian in the enlarged filtration. Substituting

$$dB_t = d\widetilde{B}_t + \mathbf{1}_{\{t\le T'\}}\Sigma(Z_t,t)^\top \nabla_z \log h_X(Z_t,t)dt$$

into the prior SDE gives the claimed posterior SDE with drift correction $\Sigma\Sigma^\top \nabla_z \log h_X = a\nabla_z \log h_X$. □

Remark D.2. A natural multivariate variational family is therefore

$$dZ_t = \left[b_\theta(Z_t,t) + \mathbf{1}_{\{t\le T'\}} a(Z_t,t)u_\beta(Z_t,t,N_{0:T'})\right] dt + \Sigma(Z_t,t)dB_t,$$

where u_β approximates $\nabla_z \log h_X$. If instead one parameterizes the correction as Σu_β, then the target becomes $u_\beta^\star = \Sigma^\top \nabla_z \log h_X$, assuming the dimensions are compatible.

E EXPECTATION MAXIMIZATION AND GRADIENT COMPUTATIONS

In this section, we introduce the general expectation-maximization scheme. Then, we look at an approximate version as applied to our problem setting, with a focus on gradient computations in the M-step.

E.1 THE EXPECTATION-MAXIMIZATION ALGORITHM

First, we restate the variational lower bound in Eq. (5) for the log-likelihood:

$$\log P_\theta(X) \ge \mathbb{E}_{Q(Z)}[\log P(X|Z)] - \mathbb{KL}[Q(Z)||P_\theta(Z)]\,. \tag{11}$$

The expectation-maximization scheme maximizes this lower bound by iterating between two steps:

- **Expectation**-step
  Find $Q(Z)$ that tightens the inequality in Eq. (11)

- **Maximization**-step
  Maximize the right-hand side of Eq. (11) with respect to θ.

One can show that the $Q(Z)$ that tightens the inequality is exactly the true posterior, i.e.,

$$Q(Z) = P_\theta(Z|X).$$

This procedure is shown in detail in Algorithm 1.

Algorithm 1: Expectation-Maximization

Input: n observed samples $\{X^i\}_{i=1}^n$
Initialize with arbitrary $\theta^{(0)}$
for $t \leftarrow 0$ **to** ∞ **do**
  Expectation: Compute for all X^i
  $Q^i(Z) = P_{\theta^{(t)}}(Z|X^i)$;
  Maximization:
  $\theta^{(t+1)} \in \arg\max_\theta \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E}_{Q^i(Z)}\left[\log P(X^i|Z)\right] - \mathbb{KL}\left[Q^i(Z)||P_\theta(Z)\right]\right)$;
return Q^∞;

E.2 SAME ALGORITHM, ADAPTED TO OUR PROBLEM

In our problem setting (and many others), it is hard to compute the exact posterior $Q^i(Z) = P_{\theta^{(t)}}(Z|X^i)$ for each sample X^i in the **E**-step, so we may resort to approximations of the true posterior. In particular, we generate approximate sample paths from $P_\theta^{(t)}(Z|X^i)$ using Metropolis-Hastings type MCMC algorithm.

In the **M**-step, finding the argmax is difficult since the expectation over the approximate posterior does not have a closed-form expression. Instead, we perform a few gradient steps with respect to θ, which requires Monte-Carlo estimates of the gradients of right-hand side of Eq. (11). The next section explains how to compute these gradient estimates.

E.3 GRADIENT COMPUTATIONS

Starting with Eq. (11), observe that

$$\partial_\theta \,\text{ELBO} = -\partial_\theta \mathbb{E}_{Q(Z)}\left[\log \frac{dQ(Z)}{dP_\theta(Z)}\right] = -\partial_\theta \mathbb{E}_{Q(Z)}\left[\log \frac{dP^*(Z)}{dP_\theta(Z)}\right],$$

where $P^*(Z)$ is the path measure induced by the SDE

$$dZ_t = \sigma(Z_t, t)\, dB_t.$$

This identity holds because, during the M-step, the variational distribution $Q(Z)$ is fixed and does not depend on θ. Furthermore, the data likelihood term $\log P(X|Z)$ is also independent of θ.

Assuming regularity conditions (as before), we may write

$$\partial_\theta \,\text{ELBO} = -\partial_\theta \mathbb{E}_{Q(Z)}\left[\int_0^T \frac{b_\theta}{\sigma^2}\, dZ_t - \frac{1}{2}\int_0^T \frac{b_\theta^2}{\sigma^2}\, dt\right] = \mathbb{E}_{Q(Z)}\left[\int_0^T \frac{b_\theta\, \partial_\theta b_\theta}{\sigma^2}\, dt - \int_0^T \frac{\partial_\theta b_\theta}{\sigma^2}\, dZ_t\right].$$

To obtain Monte Carlo estimates of this gradient, we substitute discretized increments $dZ_{t_i} \approx (Z_{t_{i+1}} - Z_{t_i})$ and evaluate the resulting expression w.r.t. sample paths $Z \sim Q(Z)$ generated in the E-step. Averaging over these trajectories and over batched samples yields an unbiased estimator of the gradient (up to discretization error).

F VARIATIONAL INFERENCE GRADIENT COMPUTATIONS

In Section 3.2, we give the high-level ideas of how gradients of the ELBO are computed. Here, we dive into the details of those gradient computations.

F.1 ELBO DERIVATION (PROOF OF THEOREM 3.1)

We begin by deriving the Evidence lower bound (ELBO) for the variational formulation as shown in Eq. (6). Starting from the general ELBO expression in Eq. (5), we replace the variational measure $Q(Z)$ with the conditional path measure $Q_\phi(Z|X)$ that is induced by the posterior SDE in Eq. (3). Recall the version of the Girsanov's theorem adapted to our notations:

Theorem F.1 (Girsanov's theorem)**.** *Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0,T]}, \mathbb{P})$ be a filtered probability space carrying a Brownian motion B under which the prior diffusion satisfies*

$$dZ_t = b_\theta(Z_t, t)dt + \sigma(Z_t, t)dB_t, \qquad Z_0 = z_0.$$

Let u be progressively measurable and suppose the exponential martingale

$$M_T = \exp\left(\int_0^T u_s\, dB_s - \frac{1}{2}\int_0^T \|u_s\|^2 ds\right)$$

is a true martingale. Define $\mathbb{Q}$ by $d\mathbb{Q}/d\mathbb{P}|_{\mathcal{F}_T} = M_T$. Then

$$\widetilde{B}_t := B_t - \int_0^t u_s ds$$

is a $\mathbb{Q}$-Brownian motion, and the same coordinate process Z satisfies under $\mathbb{Q}$

$$dZ_t = (b_\theta(Z_t, t) + \sigma(Z_t, t)u_t)\, dt + \sigma(Z_t, t)d\widetilde{B}_t.$$

Moreover,

$$\mathbb{KL}[\mathbb{Q}\|\mathbb{P}] = \frac{1}{2}\mathbb{E}_\mathbb{Q}\int_0^T \|u_s\|^2 ds.$$

By the theorem, with $u_t = u_\beta(Z_t, t, T', X)\mathbf{1}_{\{t\le T'\}}$, the Radon–Nikodym derivative between the variational posterior path measure and the prior path measure is

$$\frac{dQ_\phi(Z \mid X)}{dP_\theta(Z)} = \exp\left(\int_0^{T'} u_\beta(Z_t, t, T', X)dB_t^P - \frac{1}{2}\int_0^{T'} u_\beta^2(Z_t, t, T', X)dt\right),$$

where B^P is the Brownian motion under the prior measure. Under Q_ϕ, $B_t^P = \widetilde{B}_t + \int_0^t u_\beta(Z_s, s, T', X)ds$, so

$$\mathbb{KL}[Q_\phi(Z \mid X)\|P_\theta(Z)] = \mathbb{E}_{Q_\phi}\left[\frac{1}{2}\int_0^{T'} u_\beta^2(Z_t, t, T', X)dt\right].$$

Substituting this into the general ELBO yields Eq. (6).

F.2 GRADIENT COMPUTATIONS

We now turn to the computation of gradients of the ELBO with respect to the model parameters θ and β. Under suitable regularity conditions, we may interchange differentiation and expectation (e.g., by Fubini's theorem), leading to the following pathwise gradient representations:

$$\partial_\theta \text{ELBO} = \mathbb{E}\left[\sum_{0<\tau_i\le T'} \frac{\partial_\theta Z_{\tau_i}}{Z_{\tau_i}} - \int_0^{T'} \partial_\theta Z_t\, dt - \int_0^{T'} u_\beta \frac{\partial u_\beta}{\partial Z_t}\partial_\theta Z_t\, dt\right],$$

$$\partial_\beta \text{ELBO} = \mathbb{E}\left[\sum_{0<\tau_i\le T'} \frac{\partial_\beta Z_{\tau_i}}{Z_{\tau_i}} - \int_0^{T'} \partial_\beta Z_t\, dt - \int_0^{T'} u_\beta \left(\frac{\partial u_\beta}{\partial Z_t}\partial_\beta Z_t + \partial_\beta u_\beta\right) dt\right].$$

The expectations above are taken with respect to the joint law of Z_t and its parameter sensitivities. These evolve according to the following set of SDEs. Z_t satisfies Eq. (3), while the sensitivities with respect to θ and β satisfy

$$d\partial_\theta Z_t = \left[\partial_\theta b_\theta + \frac{\partial b_\theta}{\partial Z_t}\partial_\theta Z_t + \sigma\frac{\partial u_\beta}{\partial Z_t}\partial_\theta Z_t + \frac{\partial \sigma}{\partial Z_t}u_\beta \partial_\theta Z_t\right] dt + \frac{\partial \sigma}{\partial Z_t}\partial_\theta Z_t\, dB_t, \quad \partial_\theta Z_0 = \mathbf{0}, \tag{12}$$

$$d\partial_\beta Z_t = \left[\frac{\partial b_\theta}{\partial Z_t}\partial_\beta Z_t + \sigma\frac{\partial u_\beta}{\partial Z_t}\partial_\beta Z_t + \sigma\partial_\beta u_\beta + \frac{\partial \sigma}{\partial Z_t}u_\beta \partial_\beta Z_t\right] dt + \frac{\partial \sigma}{\partial Z_t}\partial_\beta Z_t\, dB_t, \quad \partial_\beta Z_0 = \mathbf{0}. \tag{13}$$

Moreover, because the posterior correction in Eq. (3) is only in effect before T', every drift term involving u_β or its derivatives is also multiplied by $\mathbf{1}_{\{t\le T'\}}$.

To obtain Monte Carlo estimates of the gradients, we sample m i.i.d. discretized Brownian paths. For each path, simulate Z_t, $\partial_\theta Z_t$, and $\partial_\beta Z_t$ using a shared Brownian path B_t with Euler-Maruyama discretization. Substituting these simulated trajectories into the expressions above yields unbiased estimators of the gradients, the remaining bias comes from the Euler scheme.

G DATA-GENERATING PROCEDURES FOR SYNTHETIC COX SAMPLES

We describe two ways to generate event times conditioned on a simulated intensity path $Z_{0:T}$. The first is the thinning procedure used in our experiments; the second is the small-Δt binned approximation.

Algorithm 2: Exact thinning conditional on an intensity path

Input: Nonnegative intensity path $Z_{0:T}$
Output: Event times $X = \{\tau_i\}$
Set $Z_{\max} = \sup_{t\in[0,T]} Z_t$ (in simulations, the supremum is approximated on the Euler grid or on the chosen interpolation of the path).;
Generate candidate times $\{s_j\}$ from a homogeneous Poisson process on $[0, T]$ with constant rate $Z_{\max}$.;
For each candidate time s_j, keep it independently with probability $Z_{s_j}/Z_{\max}$ and discard it otherwise.;
Return the retained candidate times in increasing order.;

Conditional on the path Z, Algorithm 2 produces an inhomogeneous Poisson process with rate Z_t. It is exact at the level of the represented continuous path; the only numerical approximation in the implementation comes from simulating and evaluating the latent SDE path on a finite grid.

Algorithm 3: Binned Poisson approximation

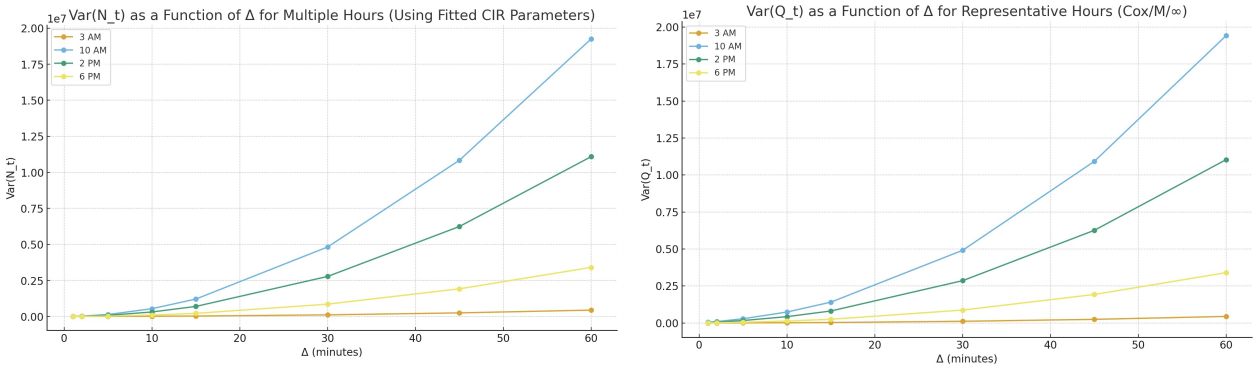
Input: Grid $0 = t_0 < t_1 < \cdots < t_M = T$, simulated intensities Z_{t_j}
Output: Approximate event times $X = \{\tau_i\}$
for $j = 0, \ldots, M-1$ **do**
  Sample $\Delta N_j \sim \text{Poisson}(Z_{t_j}\Delta t_j)$, where $\Delta t_j = t_{j+1} - t_j$.;
  Place ΔN_j event times independently and uniformly in $[t_j, t_{j+1})$.;
Return all sampled event times in increasing order.;

Algorithm 3 corresponds to replacing Z_t by a piecewise-constant approximation over each grid cell. More accurately, one may use mean $\int_{t_j}^{t_{j+1}} Z_s ds$ if the integral is available or numerically approximated. The method is fast and accurate when the grid size is small, while Algorithm 2 is the procedure used for the synthetic experiments reported in the paper.

H ABOUT THE BANK DATASET

The dataset [SEE Lab, Technion, 2024] originates from an undisclosed U.S. bank. With up to 300,000 calls per day, it provides a rich collection of information, including call time, call duration, and various categorical and continuous data. For our analysis, we focus solely on the initial arrival times of the calls. Accordingly, the dataset records the number of calls occurring within each minute of the day. Given the large volume of calls, we first evenly distribute calls within each minute and then apply independent thinning to each event with probability $p = 0.001$.

H.1 INFERENCE WITH OVERDISPERSED DATA



(a) Variance of event counts as a function of Δ, with CIR intensity model.

(b) Variance of Cox$/M/\infty$ queue length as a function of Δ.

Figure 8: Overdispersion affects downstream inferences.

We now describe a simple experiment demonstrating overdispersion in the U.S. Bank dataset. As Figure 1 shows, the estimated mean and variance of arrival counts over 10-minute intervals differ by more than an order of magnitude. The ratio of the variance to the mean is called the index of dispersion [Cox and Lewis, 1966]. Under a nonhomogeneous Poisson process, the index of dispersion would be uniformly equal to 1 by definition, so the discrepancy observed in the dataset is direct evidence that the data-generating process is not a simple Poisson point process.

Overdispersion has immediate implications for downstream inference. For a nonhomogeneous Poisson process with deterministic intensity (λ_t), the variance of the count $X(t) := \sum_{i\geq 1} \mathbb{1}_{\tau_i \leq t}$ over an interval $[t, t+\Delta)$ satisfies

$$\mathrm{Var}_\mu(N(t+\Delta) - N(t)) = \mathbb{E}_\mu[N(t+\Delta) - N(t)] \approx \lambda(t)\Delta.$$

That is, variance grows linearly in Δ at the same rate as the mean. For a Cox process with stochastic intensity (Z_t) satisfying $\mathbb{E}[Z(t)] = \lambda(t)$, the law of total variance gives

$$\begin{aligned}\mathrm{Var}_\mu(N(t+\Delta) - N(t)) &= \mathbb{E}_Z\left[\mathrm{Var}_N\left(N(t+\Delta) - N(t) \mid Z\right)\right] + \mathrm{Var}_Z\left(\mathbb{E}_N\left[N(t+\Delta) - N(t) \mid Z\right]\right) \\ &\approx \lambda(t)\Delta + \Delta^2\,\mathrm{Var}_Z(Z(t)).\end{aligned}$$

The additional term $\Delta^2\,\mathrm{Var}_Z(Z(t))$, absent in the Poisson case, reflects the randomness of the latent intensity and causes the variance to grow quadratically in Δ.

Figure 8a plots the empirical variance of arrival counts from the U.S. Bank dataset as a function of Δ, at four representative hours: 3AM, 10AM, 2PM, and 6PM. The variance grows nonlinearly, consistent with the Cox process prediction and inconsistent with a Poisson model with deterministic intensity, for which conditional and unconditional count variances cannot exceed the mean in this way. We also fit a CIR diffusion intensity model to each time period. Figure 8b then plots the variance of the queue length in a Cox$/M/\infty$ queue driven by the fitted Cox process. The nonlinear growth persists in the queue length as well, illustrating that overdispersion propagates into downstream inference: a model that does not explicitly represent the stochastic intensity will systematically misrepresent the variability of quantities that depend on it.

I MULTIVARIATE SYNTHETIC EXPERIMENTS

We evaluate the multivariate extension of the method on a two-dimensional Cox process. The goal of this experiment is to see whether the proposed framework can learn a vector-valued latent intensity process and perform posterior inference when the latent coordinates are statistically correlated. Conditional on the correlated latent path $(Z_t = (Z_t^1, Z_t^2)^\top)$, we generate two inhomogeneous Poisson processes independently. The ground truth drift is

$$b(z,t) = \begin{pmatrix} 0.3(80 - z_1) \\ 0.45(55 - z_2)) \end{pmatrix},$$

and the covariance matrix is fixed to be

$$\Sigma(z,t) = \begin{pmatrix} \sqrt{z_1} & 0 \\ \rho\sqrt{z_2} & \sqrt{1-\rho^2}\sqrt{z_2} \end{pmatrix}, \quad \rho = 0.65.$$

Hence the two latent intensity coordinates have instantaneous correlation $(\rho = 0.65)$, even though their drifts are decoupled. This construction gives a simple but nontrivial test case: the model must recover marginal mean-reverting dynamics for each coordinate while also representing the cross-coordinate dependence induced by the covariance matrix.

Figure 9 compares samples generated using the learned multivariate prior against samples generated from the ground-truth correlated CIR prior. The learned model captures the main marginal behavior of both intensity coordinates, including their different long-run levels.

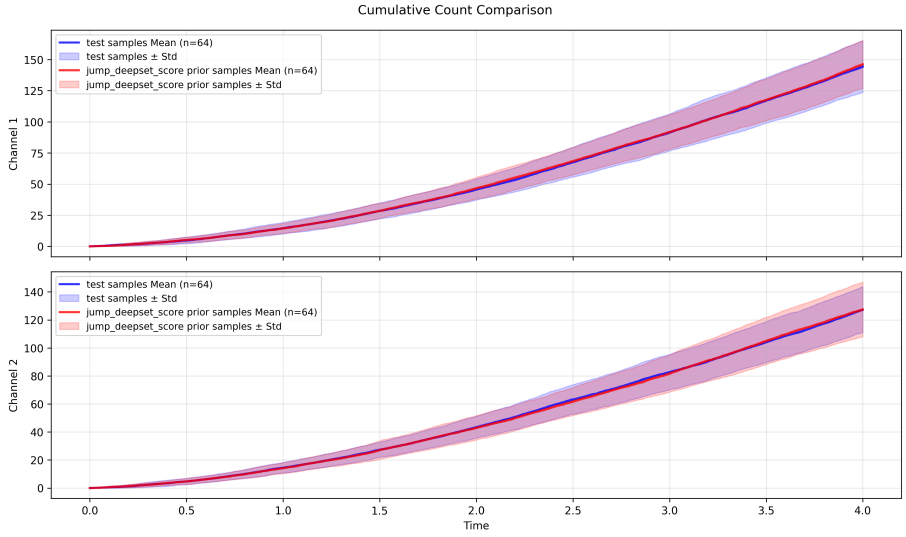


Figure 9: Multivariate synthetic prior recovery. The learned model is compared with samples generated from the ground-truth multivariate diffusion-driven Cox process. Agreement of the channel-wise sample statistics indicates that the learned neural SDE prior captures the main joint generative structure of the multichannel observations.

Figure 10 shows posterior inference for the same multivariate setting. The posterior samples generated from the learned drift-corrected SDE respond to the observed event patterns in both channels: higher observed activity in a channel pushes the corresponding latent intensity upward, while the correlation structure allows information from one channel to influence posterior uncertainty in the other.

Overall, this experiment supports the claim that the enlargement-of-filtrations characterization and the variational SDE construction extend naturally beyond the scalar case. The example verifies that the proposed method can handle multichannel observations and correlated latent stochastic intensities.

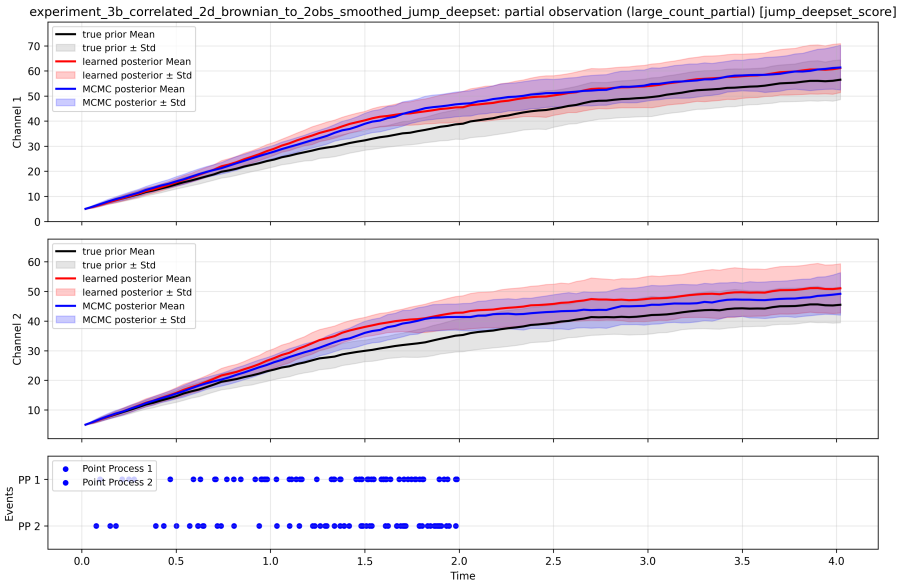


Figure 10: Multivariate posterior inference. The learned amortized posterior is evaluated on a multichannel observation from a multivariate CIR-type synthetic model. The posterior samples track the event information in the observed channels, illustrating the multivariate analogue of the scalar posterior-correction mechanism described by Theorem D.1.

J ADDITIONAL FIGURES

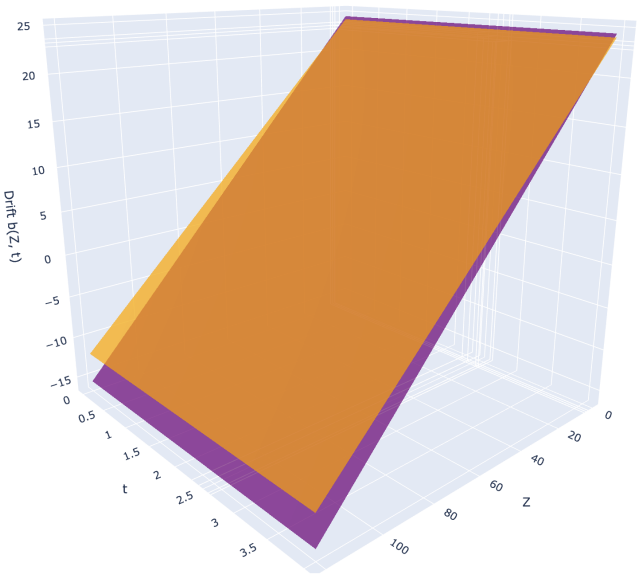


Figure 11