Non-exchangeability in Infinite Switching Linear Dynamical Systems

Victor Geadah ✉
Program in Applied and Computational Mathematics
Princeton University
victor.geadah@princeton.edu

Jonathan W. Pillow
Princeton Neuroscience Institute
Princeton University
pillow@princeton.edu

Abstract

Complex nonlinear time-series data can be effectively modeled by Switching Linear Dynamical System (SLDS) models. In trying to allow for unbounded complexity in the discrete modes, most approaches have focused on Dirichlet Process mixture models. Such non-parametric Bayesian models restrict the distribution over dynamical modes to be exchangeable, making it difficult to capture important temporally and spatially sequential dependencies. In this work, we address these concerns by developing the *non-exchangeable SLDS* (neSLD) model class effectively extending infinite-capacity SLDS models to capture non-exchangeable distributions over dynamical mode partitions. Importantly, from this non-exchangeability, we can learn transition probabilities with infinite capacity that depend on observations or on the continuous latent states. We leverage partial differential equations (PDE) in the modeling of latent sufficient statistics to provide a Markovian formulation and support efficient dynamical mode updates. Finally, we demonstrate the flexibility and expressivity of our model class on synthetic data.

1 Introduction

Partitioning experience into coherent clusters of observations is a guiding principle of learning. Models that can identify such modes of activity and allow for different continous dynamical behavior for each have found ubiquitous use in modern machine learning methods [1, 2, 3, 4]. In parallel, recent work in classical conditioning [5, 6] have focused on partitioning of observations into coherent clusters through generative mixture processes, using a Hierarchical Dirichlet process prior (HDP) over the latent causes [7]. Bridging such non-parametric Bayesian modeling techniques with discrete mode driven continuous latent-state models remains an open challenge [2]. Specifically, how to introduce recurrent connections to use current internal representations to guide updates of the underlying discrete modes remains unsolved. This is what this work focuses on, leveraging non-exchangeable infinite capacity partitioning processes [8].

1.1 Preliminaries

In this work, we focus on fully observable time-stamped data $\{(t_n, \boldsymbol{y}_n)\}_{n=1}^N$. We consider a (recurrent) *Switching Linear Dynamical System model* as the generative model for the data, which is defined by

$$z_{n+1} \sim P(z_{n+1}|z_n, \boldsymbol{x}_n) \tag{1}$$

$$\boldsymbol{x}_{n+1} = A^{(z_n)}\boldsymbol{x}_n + a^{(z_n)} + \boldsymbol{\epsilon}_n \tag{2}$$

at time $n \in \{0, \dots, N\}$, with discrete dynamical modes $z_n \in \{1, 2, \dots, K\}$ ($K \in \mathbb{N}$ an hyperparameter), continuous latent dynamics $\boldsymbol{x}_n$ with Gaussian noise $\epsilon_n \sim \mathcal{N}(0, \Sigma_x)$, and outputs as

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

linear Gaussian readout from those dynamics $\boldsymbol{y}_n = C\boldsymbol{x}_n + c + \boldsymbol{\omega}_n$ with $\boldsymbol{\omega}_n \sim \mathcal{N}(0, \Sigma_y)$. Conjugate matrix normal inverse Wishart (MNIW) priors are normally placed on the linear recurrent continuous dynamics and output parameters for Bayesian inference. The discrete modes z_n follow Markovian dynamics with a categorical distribution that can depend on z_n only (SLDS), or include a dependence on the previous continuous latent states x_{n-1} (rSLDS, Ref. [1]).

To allow infinite mode cardinality in such mode-driven architectures, most approaches have focused on non-parametric extensions based on Dirichlet Processes (DP) (e.g. [2]). DP mixture models provide a powerful random measure over clusterings [9, 10, 7], and in practice sequential sampling from a draw from a DP can be defined through the *Chinese Restaurant Process* (CRP, details in Appendix §4.1). However a key limitation of the CRP is that it induces a joint probability over cluster assignments that is invariant to the order of allocations. We call this property *exchangeability*. This enforces a strong and limited prior on distributions of partitions that can arise from this model. Furthermore, the purpose of recurrence is to actively control the dynamical mode transition, in such a way that fundamentally breaks exchangeability.

Ref. [8] offer an alternate formulation that bypasses this exchangeability property by considering cluster assignments *with each-other*. At a given time step n, this *distance-dependent CRP* (ddCRP, [8]) assigns time step $i \in [n]$ with $c_i \in [n]$ following

$$p(c_i = j | D, \alpha, \beta) \propto \begin{cases} f(D_{ij}; \beta) & \text{if } i \neq j \\ \alpha & \text{else} \end{cases} \tag{3}$$

with distance matrix D_{ij}, decay function $f(\cdot; \beta)$ and decay parameter $\beta > 0$. As we consider time-stamped data, we set $D_{ij} = t_i - t_j$ for $i \geq j$, and let $f(D_{ij}) = 0$ if $i < j$ to enforce sequentiality (no step is assign with future steps). We finally use the entire history $\boldsymbol{c}_{:n} = \{\boldsymbol{c}_{:n-1}, c_n\}$ of pairwise assignments to perform clustering, and denote z_i the cluster assignment of time step i,

$$\mathcal{Z} : \boldsymbol{c}_{:n} \mapsto \boldsymbol{z}_{:n} \tag{4}$$

and thereby setting z_n.

2 Non-exchangeable infinite-mode switching linear dynamics

The central idea being targeted is that of the exchangeability in the order of the discrete modes z_n, a challenging assumption in the modeling of complex time-stamped dependencies. To this end, we introduce *non-exchangeable switching linear dynamics* (neSLD[1]) models, which combine the distance-dependent CRP with the SLDS, and most especially, the rSLDS.

A reasonable first step towards extending the SLDS model to non-exchangeable and infinite capacity modes is to combine the dynamical mode assignments z_i from (3) and (4) with the continuous latent dynamics x_i in (2). Together this defines the generative model for a naive neSLD class (see Appendix Fig. 3**A**). One can introduce recurrence by parameterizing the decay function f and allowing the decay parameter to depend on the previous continuous states, thus introducing a dependency $c_n \leftarrow \boldsymbol{x}_{n+1}$. However, performing Bayesian inference in this model, while tractable, is of significant computational complexity. Indeed, we show in Appendix §4.2.1 that for appropriate choices of decay function f, we can leverage Pólya-gamma augmentation following [11, 1] to handle non-Gaussian factors emerging from recurrence. Unfortunately the resulting Gaussian augmentation grows quadratically with the number of steps considered, at each time step. Luckily as we'll see below, we can use sufficient statistics and recurrent dynamics to circumvent this problem.

2.1 Partial differential equations for neSLD modeling

To remedy the challenges arising out of a naive combination of the original ddCRP and SLDS models, notice that we can write the cluster allocations (eq. 3-4) directly through an *influence function*

$$\mathbf{w} : \mathcal{J} \times \mathbb{R}_+ \to \mathbb{R}_+, \quad \mathbf{w} : (j, t) \mapsto \sum_{\{i \,:\, t_i \leq t, z_i = j\}} f(t - t_i; \beta)$$

[1] To be read as *Nestled*.

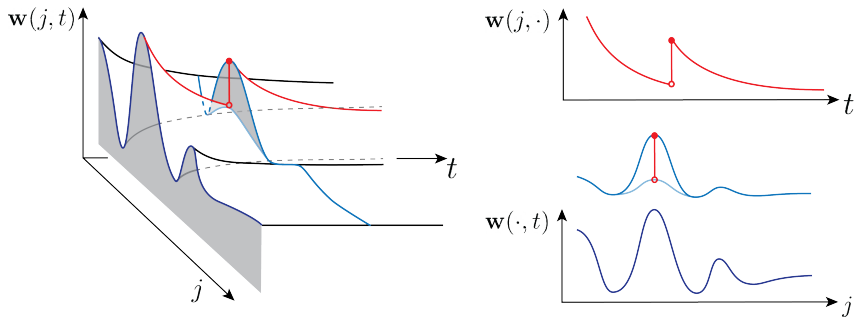

Figure 1: Modeling the influence function $\mathbf{w}(j,t)$ as a solution to the heat equation. (**top-right**) A sample trajectory of conventional choice-driven exponential decay. (**bottom-right**) Different probability profiles $\mathbf{w}(\cdot,t)$ for fixed t over U. Red bar indicates the increase in weight from a choice of mode j, which also increase the probability of "nearby" modes.

with distance function $f(\cdot;\beta)$ and decay parameter $\beta > 0$ such that

$$p(z_n = j|\boldsymbol{c}_{:n}, \alpha, \beta) = p(z_n = j|\mathbf{w}(\cdot,t_n), \alpha, \beta) \propto \begin{cases} \mathbf{w}(j,t_n;\beta) & \text{if } j \text{ in history} \\ \alpha & \text{else} \end{cases} \tag{5}$$

thereby making $\mathbf{w}(\cdot,t_n)$ a sufficient statistic, the use of which makes the entire process Markovian. Setting $f(x;\beta) = \exp(-\beta x)$ to be an exponentially decreasing function, we can rewrite $\mathbf{w}(j,t)$ above as a solution to the continuous time ODE

$$\dot{\mathbf{w}}(j,t) = -\beta \mathbf{w}(j,t) + \mathbf{1}_{\{z_n=j\}} \tag{6}$$

with càdlàg trajectories $\mathbf{w}(j,t)$ in time t, for $j \in \mathcal{J}$. The inputs $\mathbf{1}_{\{z_n=j\}} \in L^1$ represent a point bump in the influence function $\mathbf{w}$ at the mode $z_n \in \mathcal{J}$, increasing the weight of this mode for future time steps (see Fig. 1 for a visualization). As we show in Appendix §4.3, we can model the sufficient statistic to evolve according to a PDE, the heat equation. In doing so we obtain a compact form to express the time evolution, which satisfies (eq. 6), *and* how the trajectories $\mathbf{w}(j,\cdot)$ relate to one another for pairs in $\mathcal{J}$. Finally, treating $\mathbf{w}(\cdot,t)$ as a function over a continuous dynamic mode space allows to naturally restrain it to a discrete and infinite number of modes.

We implement such sufficient statistic $\mathbf{w}$ following heat-equation time dynamics through *finite difference methods*. First, let $\mathbf{w}(j+\Delta j, t_n+\Delta t) =: \boldsymbol{w}_{n+1}(j+1)$ be our discrete approximation. We then use a finite approximation of the derivative and central difference approximation to the second order spatial partial derivative to obtain the solution

$$\boldsymbol{w}_{n+1} = U\boldsymbol{w}_n, \quad U = \text{tridiag}(\beta, 1-2\beta, \beta)$$

where $\beta = \gamma\frac{\Delta t}{\Delta x^2}$. We impose $\Delta t \leq \frac{\Delta x^2}{4\gamma}$ as a general requirement for stability, and let Δx be adjusted accordingly given γ (model parameter) and Δt (data parameter). Inputs can be added to drive the system, including (1) the desired $\mathbf{1}_{z_n=j}$ adding self-reinforcement to the system, and (2) the past internal states encoded by a matrix $R \in \mathbb{R}^{J\times N_x}$ for recurrence. In all, the dynamics of the sufficient statistic $\boldsymbol{w}_n$ follow

$$\boldsymbol{w}_{n+1} = U\boldsymbol{w}_n + \kappa \mathbf{1}_{z_n=j} + R\boldsymbol{x}_n$$

with parameters of decay $\beta > 0$ and self-reinforcement $\kappa \in \mathbb{R}$. Dynamical modes z_n are then sampled according to (eq. 5), with final random transition parameter $\alpha > 0$.

3 Experiments and Results

We test the validity and performance of our neSLD model class on synthetic data. We consider the Synthetic NASCAR experiment used in [1], where the data consists of toy stock car trajectories on a NASCAR track. The true dynamics are sampled from a custom rSLDS model and rely on 4 modes z, but conceptually 2 states use similar linear dynamics (center dynamics during the turn). We use for comparison the SLDS and rSLDS models, and train all models via variational inference by maximizing the Evidence Lower Bound (ELBO) using Laplace-EM from [4].

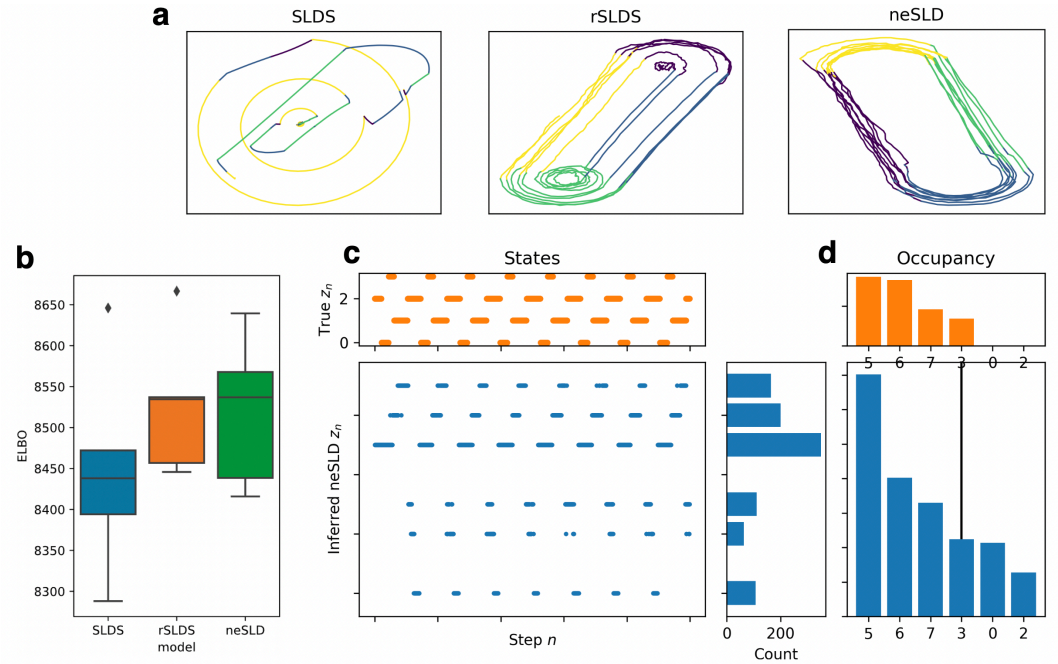

Figure 2: *NASCAR experimental results*. (**a**) Sample x_t dynamics from trained models. True data resembles the neSLD dynamics. (**b**) ELBO attained by each model, over 5 seeds. Higher is better. (**c**) True and inferred neSLD dynamical modes z_n. (**d**) Sorted histogram of z_n occupancy.

For a similar number of dynamical modes, we find the neSLD model to be able to match or exceed the rSLDS model in terms of ELBO (Fig. 2**b**), and find sample trajectories to provide a good qualitative fit to the true x_t state dynamics (Fig. 2**a**, other seeds in Appendix Fig. 5). Increasing the number of modes to $K = 8$, we find that it accurately inferred the active states (Fig. 2**c**), indeed occupying a lower number of states that the K prescribed (see occupancy curve Fig. 2**d**).

4 Conclusion

We note that the discrete and finite domain representation of $\mathbf{w}(\cdot, t)$ that we use does requires the practitioner to set the number of states K ahead of training. However such value serves only as an upper bound, as the model sequentially adds states as needed, and we showed that it infers the required number of truly occupied states. For an infinite mode domain while still maintaining some degree of finite representation for implementation purposes, one can turn to a truncated Fourier series expansion of $\mathbf{w}(\cdot, t)$. We refer to Appendix §4.3 for details on such a formalism and how external inputs would be treated. Future work would aim to integrate this framework into the neSLD model class, and test the performance on more realistic tasks.

In closing, in this work we introduce the non-exchangeable SLDS (neSLD) model class, extending the SLDS model class to allow for non-exchangeable priors over the discrete modes z_n with unbounded complexity. This non-exchangeability is the key characteristic making it possible to apply this non-parametric machinery to the the widely used rSLDS model [1]. We first discuss a tractable Bayesian inference framework for learning in fully connected, "naive", neSLD models with Pólya gamma augmentation. While this helps with tractability, it renders the process highly computationally expensive both at generation and inference. To mitigate these problems, we leverage PDE theory to derive a faster, semi-parametric formulation of sufficient statistics to the dynamical mode updates in the neSLD model. Finally, we implement and test the flexibility and expressivity of the neSLD model on a toy task compared to (r)SLDS baselines. We find that it matches performance of the true underlying model, and accurately infers the number of active states. In all, this demonstrates the use of PDE-informed modeling in latent generative models of spatio-temporal data and in their extensions to infinite capacity models.

References

[1] Scott W. Linderman, Andrew C. Miller, Ryan P. Adams, David M. Blei, Liam Paninski, and Matthew J. Johnson. Recurrent switching linear dynamical systems, 2016.

[2] Emily Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. Bayesian nonparametric inference of switching dynamic linear models. *IEEE Transactions on Signal Processing*, 59(4):1569–1585, apr 2011.

[3] Scott S. Bolkan, Iris R. Stone, Lucas Pinto, Zoe C. Ashwood, Jorge M. Iravedra Garcia, Alison L. Herman, Priyanka Singh, Akhil Bandi, Julia Cox, Christopher A. Zimmerman, Jounhong Ryan Cho, Ben Engelhard, Jonathan W. Pillow, and Ilana B. Witten. Opponent control of behavior by dorsomedial striatal pathways depends on task demands and internal state. *Nat Neurosci*, 25(3):345–357, mar 2022.

[4] David M. Zoltowski, Jonathan W. Pillow, and Scott W. Linderman. Unifying and generalizing models of neural dynamics during decision-making, 2020.

[5] Samuel J. Gershman and Yael Niv. Exploring a latent cause theory of classical conditioning. *Learn Behav Behavior*, 40(3):255–268, aug 2012.

[6] Samuel J Gershman, Kenneth A Norman, and Yael Niv. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50, oct 2015.

[7] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, dec 2006.

[8] David M. Blei and Peter I. Frazier. Distance Dependent Chinese Restaurant Processes. *Journal of Machine Learning Research*, 12(74):2461–2488, 2011.

[9] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2), mar 1973.

[10] Charles E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann. Statist.*, 2(6), nov 1974.

[11] Scott W. Linderman, Matthew J. Johnson, and Ryan P. Adams. Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3456–3464, 2015.

[12] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Appendix

4.1 Background

Dirichlet Processes (DP) mixture models provide a powerful random measure over clusterings. We refer to Ref. [9, 10] for a measure-theoretic treatment of DPs, and Ref. [7] for a machine-learning overview. They can be alternatively defined through the *Chinese Restaurant Process* (CRP), a process akin to the Polya Urn process. The analogy goes as follows : upon entering a restaurant, a customer i selects to sit at a table k with probability proportional to the number of people already sat at that table. With some fixed rate α, they may decide to start a new table. Put otherwise, for a new customer i, its table allocation z_i follows

$$p(z_i = k | z_{:i}, \alpha) \propto \begin{cases} n_k & \text{if } k \leq K \\ \alpha & \text{if } k = K+1 \end{cases} \tag{7}$$

with n_k the size of cluster $k \in [K]$. In this work, one should think of tables as clusters or *dynamical modes*, and the customers i as time-steps. It can be easily seen that this process induces a joint probability over cluster assignments that is invariant to the order of customers entering. We call this property *exchangeability*. This enforces a strong and limited prior on distributions of partitions that can arise from this model.

4.2 Modeling details

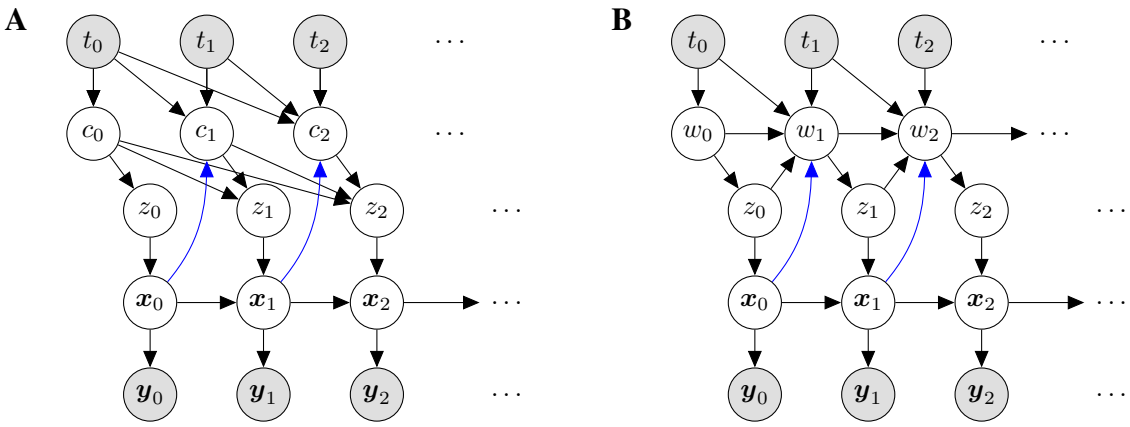

Figure 3: *Graphical models for the neSLD class*. We do not include model parameters. Optional recurrence indicated in blue. (**A**) Graphical model of the *naive* infinite non-exchangeable SLD (neSLD) model. (**B**) Markov neSLD model, with discrete approximation w_n of the sufficient statistic $\mathbf{w}(\cdot, t_n) \in L^1$.

We show in Figure 4**D** how we can effectively control the mode transitions $z_n \to z_{n+1}$ by leveraging the decay parameter $\beta > 0$ and the self reinforcement parameter κ.

4.2.1 Tractable Bayesian Inference in Naive neSLD models

To perform inference in the (decay-recurrent)-Naive neSLD model, we leverage message passing to perform Gibbs sampling. It revolves around the conditional density

$$p(\boldsymbol{x}_{1:N} | \boldsymbol{c}_{1:N}, \boldsymbol{z}_{1:N}, \{\boldsymbol{y}_{1:N}, \boldsymbol{t}_{1:N}\}) \propto \prod_{n=1}^{N} \psi(\boldsymbol{x}_{n-1}, \boldsymbol{x}_n, z_n) \psi(\boldsymbol{x}_{n-1}, c_n, \boldsymbol{t}_{:n}) \psi(\boldsymbol{x}_n, \boldsymbol{y}_n)$$

where $\psi(\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, c_{n+1})$ is the potential from the continuous recurrent dynamics, and $\psi(\boldsymbol{x}_n, y_n)$ the evidence potentials. The decay-recurrent connections introduce the dependencies captured in $\psi(\boldsymbol{x}_n, c_{n+1})$, which adds significant challenges for inference. Without it, in the standard SLDS, the potentials are all Gaussian, allowing for analytical integration.

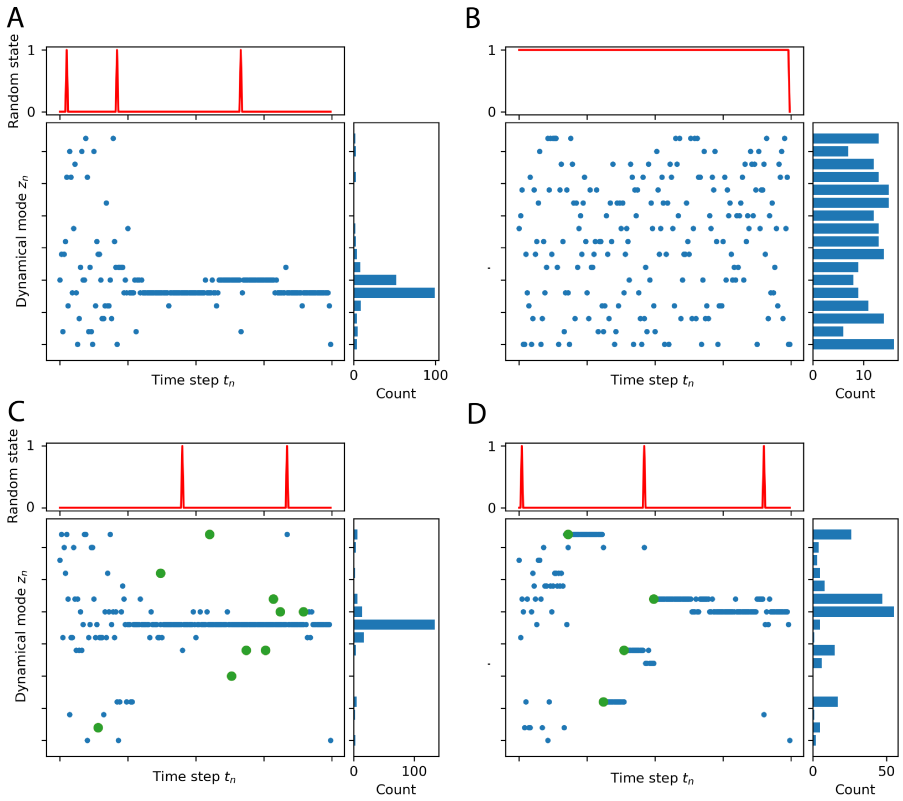

Figure 4: *Mode transition control*. The transition parameter $\alpha = 0.05$ unless stated otherwise. Top subplot indicates when a random transition was triggered, with probability α. (**A**) "Typical" mode transitions, with transients until one mode is preferentially reinforced. Reinforcing one mode also raises the probability of transitioning to nearby modes. (**B**) $\alpha = 1.0$, random transitions a.s. (**C**) Introduction of weak external inputs $\boldsymbol{u}$ of height $\kappa = 1.0$ (green dots). Qualitatively similar to **A**. (**D**) Strong external inputs $\boldsymbol{u}$ of height $\kappa = 100.0$ (green dots) dictate mode transitions.

Following [1], we leverage Pólya-gamma augmentation [11] to deal with the non-Gaussian factors. The key is that instead of performing a categorical choice over a pre-determined set of K dynamical modes, we perform an association, a categorical choice, with a previous time-step $j \in \{1, \ldots, t\}$. Because of the conceptual categorical similarities in updating, we find similarities in inference methodology with the rSLDS. This is where the choice of decay function f comes in, in enforcing that link. This non-Gaussian factor is

$$\begin{aligned}\psi(\boldsymbol{x}_{n-1}, c_n, \boldsymbol{t}_{:n}) = p(c_n|\boldsymbol{x}_{n-1}, \boldsymbol{t}_{:n}) &\propto \prod_{j=1}^{n} f(t_n - t_j; \beta(\boldsymbol{x}_{n-1}))^{\mathbb{I}[c_n=j]} \alpha^{\mathbb{I}[c_n=n]} \\ &= \alpha^{\mathbb{I}[c_n=n]} \prod_{j=1}^{n-1} \left(\frac{e^{[\boldsymbol{\nu}_n]_j}}{1+e^{[\boldsymbol{\nu}_n]_j}}\right)^{\mathbb{I}[c_n=j]}\end{aligned}$$

for $\boldsymbol{\nu}_n \in \mathbb{R}^{n-1}$, $[\boldsymbol{\nu}_n]_j \coloneqq \beta(\boldsymbol{x}_{n-1}) \cdot (t_n - t_j)$. We can leverage the following integral quantity

$$\frac{(e^{\nu})^a}{(1+e^{\nu})^b} = 2^{-b} e^{\kappa\nu} \int_0^\infty e^{-\omega\nu^2/2} p_{\mathrm{PG}}(\omega|b, 0)\mathrm{d}\omega \qquad b > 0, \kappa = a - \frac{b}{2}$$

to introduce auxiliary variables $\{\omega_j\}_{j=1}^n$ such that the conditional density $p(c_{n+1}|\boldsymbol{x}_n, \boldsymbol{t}_{:n+1}, \omega_n)$ becomes Gaussian

$$\psi(\boldsymbol{x}_n, c_{n+1}, \boldsymbol{t}_{:n+1}, \omega_n) \propto \alpha\, \mathcal{N}(\boldsymbol{\nu}_n \mid \Omega_n^{-1}\boldsymbol{\kappa}_n, \Omega_n^{-1})$$

where $\Omega_n = \mathrm{diag}(\omega_{1:t-1})$, and $[\boldsymbol{\kappa}_n]_j = \frac{1}{2}\mathbb{I}[c_n = j]$, $\boldsymbol{\kappa}_n \in \mathbb{R}^n$. With this augmentation, the required potentials are Gaussian and the integral can be calculated analytically. We refer to [8] for details on the handling of messages $m_{n\to n+1}(c_{n+1})$.

4.3 PDEs formalism and Fourier analysis

We consider a discrete set $\mathcal{J}$ of potentially infinite size, as we aim to learn the number of discrete modes from data. To this end, we treat $\mathbf{w} : U \times \mathbb{R}_+ \to \mathbb{R}$ as a function over both mode space $U \subset \mathbb{R}$ open set and time t. The mode space U encompasses $\mathcal{J}$, and we consider w.l.o.g. $U = (0, J)$, $J > 0$. We require the following modeling restrictions on $\mathbf{w}$:

1. For fixed $j \in U$, $\mathbf{w}(j, \cdot)$ has the continuous time evolution described in (6).
2. For a fixed time t, $\mathbf{w}(\cdot, t) \in L^1(U)$.
3. In the absence of external inputs, $\mathbf{w}(\cdot, t)$ should tend to a constant function as $t \to \infty$.

The second regularity condition is to ensure that $\mathbf{w}$ can be normalized, and the last condition can be interpreted as requiring that if there are no discrete mode allocations, no memory is held for the choices and we tend to make a uniform choice over possible clusters $j \in \mathcal{J}$.

We can synthesize the above conditions by imposing $\mathbf{w}$ to evolve according to the *heat equation*, drawing inspiration from physics-informed neural networks [12]. A function $u : U \times \mathbb{R}_+ \to \mathbb{R}$ is a solution of the heat equation if $\frac{\partial}{\partial t} u = \gamma \frac{\partial^2}{\partial x^2} u$, and any solution[2] u satisfies the three conditions listed above. Modeling $\mathbf{w}$ as a solution of the heat equation thus allows us to satisfy the conditions, and furthermore gives us a compact form to express both the time evolution *and* how the trajectories $\mathbf{w}(j, \cdot)$ relate to one another for pairs in $\mathcal{J}$. In particular, the function form of $\mathbf{w}(\cdot, t)$ over U allows for an arbitrary number of modes j. In implementations, we will be dealing with finite representations $w_n(j)$ of $\mathbf{w}(t_n, j)$.

4.3.1 Sufficient statistic dynamics over continuous domain with Fourier analysis

With a continuous domain, to nonetheless maintain a finite representation $\tilde{\mathbf{w}}$ of $\mathbf{w}$ for implementation purposes, we consider the truncated Fourier series

$$\tilde{\mathbf{w}}(x, t) = \sum_{k=-N}^{N} a_k(t) \cdot e^{ikx/J}$$

with coefficients $a_k(t) = \frac{1}{2\pi J} \int_U w(x, t) e^{-ikx/J} dx$. We denote $\tilde{\mathbf{w}}(\cdot, t_n) =: w_n$.

Finally, we consider un-normalized Gaussian input drives $g(x; \mu, \sigma, A) = Ae^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. The idea would be that at time n, a bump centered around $\mu = z_n$ would be added to the influence profile $\mathbf{w}(\cdot, t_n)$. See Figure 1 for a visualization. A fixed variance σ^2 embodies how one choice of z_n affects the "nearby" dynamical modes. A is further introduced to capture the strength of the drive; we usually take $A = 1$. Its Fourier transform is known

$$\mathcal{F}\{g(\cdot\,; \mu, \sigma)\}(\lambda) = \frac{\sigma e^{-i\mu\lambda}}{\sqrt{2\pi}} e^{-\frac{1}{2}(\lambda\sigma)^2} = \frac{\sigma e^{-i\mu\lambda}}{\sqrt{2\pi}} g(\lambda; \sigma^{-1}) =: G(\lambda; \sigma, \mu) \tag{8}$$

We state the dynamics of the coefficients a_k of w_n under these inputs.

Proposition 4.1. *Let* $\mathbf{w}$ *be a solution of the heat equation on* $U \subset \mathbb{R}$ *with point drives* $g(\cdot; z_n, \sigma) \in L^1(U)$ *at time* t_n*. The dynamics of the truncated Fourier series* w_n *of* $\mathbf{w}(\cdot, t_n)$ *have coefficients*

$$a_k(t_{n+1}) = \left(a_k(t_n) + J^{-1} G(k/J; \sigma, z_n)\right) e^{-\gamma \frac{k^2}{J^2}(t_{n+1} - t_n)}$$

for $t_n \in \{0, \ldots, T-1\}$*, with* $a_k(t_0) = G(\frac{k}{J}; \sigma, z_0)$.

Proof. From the definition of the Fourier transform $\hat{\mathbf{w}}$ of $\mathbf{w}$, it follows that $a_k(t) = \frac{1}{J} \hat{\mathbf{w}}(\frac{k}{J}, t)$. For $\mathbf{w}$ a solution of the heat equation, its Fourier transform satisfies

$$\hat{\mathbf{w}}(\lambda, t_{n+1}) = \hat{\mathbf{w}}(\lambda, t_n) e^{-\gamma\lambda^2(t_{n+1} - t_n)}$$

where in our case the previous state $\hat{\omega}(\lambda, t_n)$ in Fourier domain satisfies

$$\hat{\mathbf{w}}(\lambda, t_n) = \frac{1}{2\pi} \int_U \left(\lim_{t \uparrow t_n} \mathbf{w}(x, t) + g(x; z_n, \sigma) \right) e^{i\lambda x} dx = \lim_{t \uparrow t_n} \hat{\mathbf{w}}(\lambda, t) + \hat{g(\cdot; z_n, \sigma)}(\lambda)$$

[2] We consider initial conditions $\mathbf{w}(\cdot, 0) \equiv \mathbf{0}$ and Dirichlet boundary conditions $\mathbf{w}(x, \cdot) = 0$ on $x \in \partial U$.

such that

$$\hat{\mathbf{w}}(\lambda, t_{n+1}) = \left(\lim_{t \uparrow t_n} \hat{\mathbf{w}}(\lambda, t_n) + G(\lambda; \sigma, z_n) \right) e^{-\gamma \lambda^2 (t_{n+1} - t_n)}$$

with initial condition $w(\lambda, 0) \equiv \mathbf{0}$ and $w(\lambda, t_0) = \frac{1}{2\pi} e^{i\lambda z_0}$. Combining it all, we get that the coefficients of the truncated Fourier expansion follow

$$a_k(t_{n+1}) = \frac{1}{J} \hat{\mathbf{w}}\left(k/J, t_{n+1}\right) = \left(a_k(t_n) + \frac{1}{J} G\left(\frac{k}{J}; \sigma, z_n \right) \right) e^{-\gamma \frac{k^2}{J^2} (t_{n+1} - t_n)}$$

as desired. □

Alternate influence dynamics and associated Fourier coefficients

Instead of Gaussian bumps g, we can consider delta δ_{z_n} inputs. This yields

Proposition 4.2. *Let w be a solution of the heat equation on $U \subset \mathbb{R}$ with point drives $\delta_{z_n} \in L^1(U)$ at time t_n. The dynamics of the truncated Fourier series w_n of $\mathbf{w}(\cdot, t_n)$ have coefficients*

$$a_k(t_{n+1}) = \left(a_k(t_n) + \frac{1}{J} \hat{\delta_{z_n}}(k/J) \right) e^{-\gamma \frac{k^2}{J^2} (t_{n+1} - t_n)}$$

for $t_n \in \{0, \ldots, T-1\}$, with $a_k(t_0) = \hat{\delta_{z_0}}(\frac{k}{J})$.

Proof. From the definition of the Fourier transform $\hat{\mathbf{w}}$ of $\mathbf{w}$, it follows that $a_k(t) = \frac{1}{J} \hat{\mathbf{w}}(\frac{k}{J}, t)$. For $\mathbf{w}$ a solution of the heat equation, its Fourier transform satisfies

$$\hat{\mathbf{w}}(\lambda, t_{n+1}) = \hat{\omega}(\lambda, t_n) e^{-\gamma \lambda^2 (t_{n+1} - t_n)}$$

where in our case the previous state $\hat{\omega}(\lambda, t_n)$ in Fourier domain satisfies

$$\begin{aligned} \hat{\omega}(\lambda, t_n) = \frac{1}{2\pi} \int_U \left(\mathbf{w}(x, t_n) + \delta_{z_n}(x) \right) e^{i\lambda x} dx &= \hat{\mathbf{w}}(\lambda, t_n) + \frac{1}{2\pi} \int_U \delta_{z_n}(x) e^{i\lambda x} dx \\ &= \hat{\mathbf{w}}(\lambda, t_n) + \frac{1}{2\pi} e^{i\lambda z_n} \end{aligned}$$

such that

$$\hat{\mathbf{w}}(\lambda, t_{n+1}) = \left(\hat{\mathbf{w}}(\lambda, t_n) + \frac{1}{2\pi} e^{i\lambda z_n} \right) e^{-\gamma \lambda^2 (t_{n+1} - t_n)}$$

with initial condition $w(\lambda, 0) \equiv \mathbf{0}$ and $w(\lambda, t_0) = \frac{1}{2\pi} e^{i\lambda z_0}$. Combining it all, we get that the coefficients of the truncated Fourier expansion follow

$$\begin{aligned} a_k(t_{n+1}) = \frac{1}{J} \hat{\mathbf{w}}\left(k/J, t_{n+1}\right) &= \frac{1}{J} \left(\hat{\mathbf{w}}(k/J, t_n) + \frac{1}{2\pi} e^{i \frac{k}{J} z_n} \right) e^{-\gamma k^2 (t_{n+1} - t_n)/U^2} \\ &= \left(a_k(t_n) + \frac{1}{2\pi J} e^{i \frac{k}{J} z_n} \right) e^{-\gamma \frac{k^2}{J^2} (t_{n+1} - t_n)} \end{aligned}$$

as desired. □

4.4 Further results on the NASCAR experiment

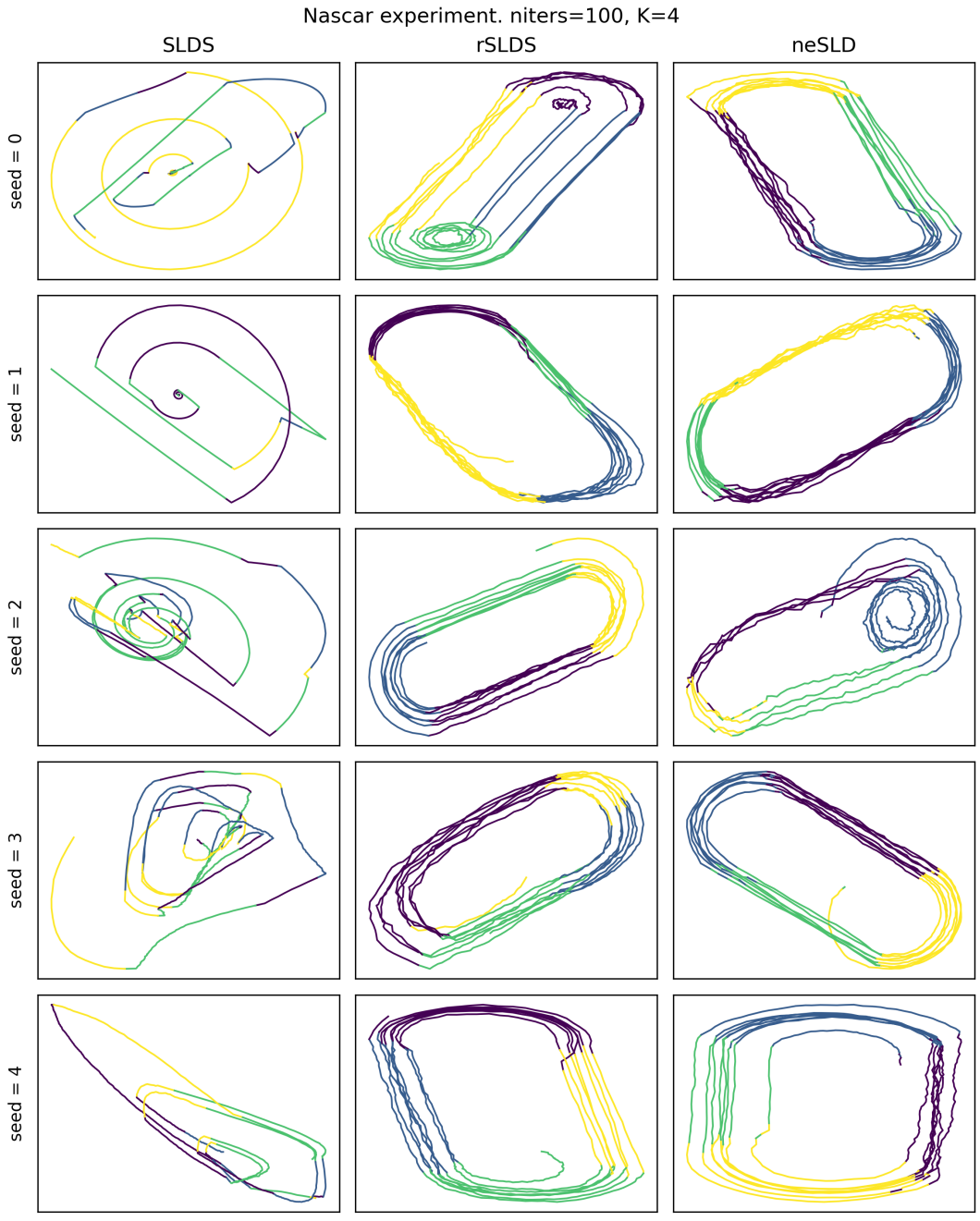

Figure 5: Sample continuous dynamics x_t for each seed for each model. Models were trained for 100 iterations, with $K = 4$ dynamical modes z_n.