STREAMING INFERENCE FOR INFINITE NON-STATIONARY CLUSTERING

Rylan Schaeffer
Computer Science
Stanford University
rschaef@stanford.edu

Gabrielle Kaili-May Liu
Brain and Cognitive Sciences
Massachusetts Institute of Technology
gkml@mit.edu

Yilun Du
Electrical Engineering & Computer Science
Massachusetts Institute of Technology
yilundu@mit.edu

Scott Linderman
Statistics
Stanford University
scott.linderman@stanford.edu

Ila Rani Fiete
Brain and Cognitive Sciences
Massachusetts Institute of Technology
fiete@mit.edu

ABSTRACT

Learning from a continuous stream of non-stationary data in an unsupervised manner is arguably one of the most common and most challenging settings facing intelligent agents. Here, we attack learning under all three conditions (unsupervised, streaming, non-stationary) in the context of clustering, also known as mixture modeling. We introduce a novel clustering algorithm that endows mixture models with the ability to create new clusters online, as demanded by the data, in a probabilistic, time-varying, and principled manner. To achieve this, we first define a novel stochastic process called the Dynamical Chinese Restaurant Process (Dynamical CRP), which is a non-exchangeable distribution over partitions of a set. Next, we show that the Dynamical CRP provides a non-stationary prior over cluster assignments and yields an efficient streaming variational inference algorithm. We conclude with experiments showing that the Dynamical CRP can be applied on diverse synthetic and real data with Gaussian and non-Gaussian likelihoods.

1 INTRODUCTION

Biological intelligence operates in a radically different data regime than most artificial intelligence. In particular, biological intelligence must contend with data that is (i) unsupervised, (ii) streaming, and (iii) non-stationary, either as a consequence of the agent, its environment, or both. One goal of lifelong learning is to make artificial intelligence significantly more capable in this data regime, and accomplishing that goal requires asking and answering how agents in this data regime ought to approach learning.

Here, we consider the specific unsupervised problem of clustering, also known as mixture modeling. Clustering is a ubiquitous and important problem in its own right, with widespread applications, but clustering can also serve as a sub-goal in service of other goals. For instance, an agent in a partially observable world may wish to cluster sensory observations into world states to then use for spatial navigation or reinforcement learning. We specifically consider an agent who receives a single stream of observations from non-stationary clusters, with no ability to revisit past observations, but who must nonetheless identify the clusters and assign observations to them. In this data regime, the number of clusters is unknown and theoretically could be unbounded, and so the agent must use a clustering algorithm capable of growing in representational capacity as more observations are encountered.

In this paper, we define a novel distribution over partitions of a set that we call the Dynamical Chinese Restaurant Process (Dynamical CRP), due to its relationship with the Chinese Restaurant Process (CRP) (Ferguson, 1973; Blackwell & MacQueen, 1973; Antoniak, 1974). We then show how the Dynamical CRP can be used as a prior over cluster assignments in a manner that yields an efficient streaming clustering algorithm designed for non-stationary data. Starting with synthetic Gaussian and non-Gaussian data, and moving to more sophisticated real data including simultaneous

localization and mapping (SLAM) and self-supervised visual representation learning, we show that streaming inference using the Dynamical CRP achieves comparable or better performance than many common baselines, especially when the data is non-stationary.

2 BACKGROUND

2.1 NOTATION

We consider a single time series of D-dimensional observable variables $o_{1:N}$ ($o_n \in \mathbb{R}^D$) occurring at known times $t_{1:N}$, each corresponding to some latent cluster assignment variables $c_{1:N}$ (i.e. $c_n \in \{1, 2, ...\}$), where $\cdot_{1:N}$ denotes the sequence $(\cdot_1, \cdot_2, ..., \cdot_N)$. Our goal is to infer the latent cluster assignments $c_{1:N}$. Each cluster may have corresponding variables $\{\phi_c\}_{c=1}^C$ (e.g., per-cluster means and covariances) that we might also wish to infer. In the non-stationary setting, the clusters may change over time in a manner that we shall specify.

2.2 INFINITE CLUSTERING VIA THE CHINESE RESTAURANT PROCESS

In most clustering problems, the number of clusters is unknown and bounded only by the number of observations. Consequently, a useful clustering algorithm should be capable of (a) adding clusters as necessitated by the data, (b) generating predictions of future likely clusters, and (c) changing learnt representations of clusters over time. To meet the first two desiderata, many clustering algorithms use the Chinese Restaurant Process (CRP) or its related Dirichlet Process (Ferguson, 1973; Antoniak, 1974; Neal, 2000; Blei & Jordan, 2006; Kulis & Jordan, 2012). The CRP is a single-parameter ($\alpha > 0$) stochastic process that defines a discrete distribution over partitions of a set, making it an applicable prior for cluster assignments. The name CRP arises from a story of a sequence of customers (observations) arriving at a restaurant with an infinite number of tables (clusters), each table with infinite capacity. The first customer c_1 sits at the first table, and each subsequent customer c_n sits either at an unoccupied table with probability proportional to α or joins an occupied table with probability proportional to the number of preceding customers at that table. Denoting the number of non-empty tables after the first $n \in \mathbb{Z}^+$ customers $C_n \stackrel{\text{def}}{=} \max(c_1, ..., c_n)$, CRP($\alpha$) defines a conditional distribution for the nth customer c_n given the preceding customers $c_{<n}$:

$$p^{CRP}(c_n = c | c_{<n}, \alpha) \quad \propto \begin{cases} \sum_{n'<n} \mathbb{I}(c_{n'} = c) & \text{if } 1 \leq c \leq C_{n-1} \\ \alpha & \text{if } c = C_{n-1} + 1 \end{cases} \tag{1}$$

An example application of the CRP is task-free continual learning (Lee et al., 2020). However, the CRP is ill-suited to streaming data because the CRP's conditional form requires knowing the entire history of cluster assignments; Schaeffer et al. (2021) showed the CRP can be adapted for streaming data by rewriting the CRP in a recursive form:

$$p^{CRP}(c_n = c | \alpha) \propto \sum_{n'<n} p(c_{n'} = c | \alpha) + \alpha\, p(C_{n-1} = c - 1) \tag{2}$$

The intuition is that if many observations come from cluster c, then the next observation is also likely to come from cluster c, and the probability of more clusters should grow with the number of observations, giving the CRP the capacity to create an "infinite" number of clusters. A similar approach can be applied to feature models (Schaeffer et al., 2022).

2.3 NON-STATIONARY VARIANTS OF THE CHINESE RESTAURANT PROCESS

We say that observations are stationary if the joint distribution of $(o_n, o_{n+1}, ..., o_{n+p})$ equals the joint distribution of $(o_{n+\eta}, o_{n+\eta+1}, ..., o_{n+\eta+p})$ for all n, p, η. Although the CRP is widely used, the CRP has two properties which are inappropriate for non-stationary data. First, the CRP is exchangeable, meaning permuting the order of the data does not affect the probability of the resulting partition. Second, the CRP is consistent, meaning marginalizing out any observation is the same as if the observation never existed. To handle non-stationary data, Zhu et al. (2005) defined the time-sensitive CRP (tsCRP) by introducing exponential decay:

$$p^{tsCRP}(c_n = c | c_{<n}, \alpha) \propto \begin{cases} \sum_{n'<n} \exp((t_n - t_{n'})/\tau) \mathbb{I}(c_{n'} = c) & \text{if } 1 \leq c \leq C_{n-1} \\ \alpha & \text{if } c = C_{n-1} + 1 \end{cases} \tag{3}$$

Blei & Frazier (2011) later defined the distance-dependent Chinese Restaurant Process (ddCRP), which assigns customers to other customers in a possibly cyclic directed graph. While flexible, the ddCRP is impractical for streaming inference because observations can be assigned to future observations, and time/space complexities must be quadratic in the number of observations because the pairwise relationships have no structure and thus must all be remembered.

3 METHODS

3.1 DESIDERATA

Our goal is to define an efficient streaming inference algorithm for infinite non-stationary clustering. To do this, we define a novel stochastic process over partitions of a set called the **Dynamical CRP** to use as a prior over cluster assignments. The Dynamical CRP is designed with the following goals:

- Like the CRP, the Dynamical CRP can create "infinite" clusters (albeit upper bounded by the number of observations) and can generate predictions of future likely clusters.
- Unlike the CRP, the Dynamical CRP does not assume the observations are i.i.d., exchangeable or consistent, meaning the Dynamical CRP can model non-stationary data.
- Unlike the tsCRP, the Dynamical CRP does not restrict the influence of observation times to exponential decay and can therefore capture a richer class of temporal relationships.
- Unlike the ddCRP, the Dynamical CRP admits an efficient streaming inference algorithm, which is critical for practical use by agents with finite memory.

The Dynamical CRP thus sits in a "Goldilocks" zone: more powerful than the CRP or tsCRP, but less powerful than the ddCRP so as to still permit efficient streaming inference.

3.2 HIGH LEVEL IDEA

At the heart of the CRP are the "table occupancies" $N_c^{CRP}(t_n) \stackrel{\text{def}}{=} \sum_{n'<n} \mathbb{I}(c_n = c)\mathbb{I}(t_{n'} \leq t)$, which are the sufficient statistics of the stochastic process. The Dynamical CRP embeds those table occupancies in a dynamical system to evolve endogenously. By choosing or learning dynamics appropriate for a particular task, the Dynamical CRP gains rich time-dependent priors for cluster assignments.

3.3 DEFINITION

Let $\mathcal{H}$ be a Hilbert space and $\tilde{N}(t) \in \mathcal{H}$ contain both the "pseudo" table occupancies $N_c(t)$ and any desired higher-order temporal derivatives. Fix a linear dynamical system $\ell : \tilde{N} \to \tilde{N}$ and increment the c_n-th pseudo table occupancy N_{c_n} at time t_n by 1. As before, define $C_n \stackrel{\text{def}}{=} \max(c_1, ..., c_n)$. The Dynamical CRP, denoted $D\text{-}CRP(\ell, \alpha)$, is defined as the conditional distribution:

$$p^{D\text{-}CRP}(c_n = c | c_{<n}, t_{\leq n}, \ell, \alpha) \propto \begin{cases} N_c(t_n) & \text{if } 1 \leq c \leq C_{n-1} \\ \alpha & \text{if } c = C_{n-1} + 1 \end{cases} \tag{4}$$

Like the CRP, each customer increments a table's occupancy count, but unlike the CRP, the tables' occupancies can now change endogenously. We next show the flexibility that the Dynamical CRP provides.

3.4 EXAMPLES

For the following examples, let $\Delta \stackrel{\text{def}}{=} t_n - t_{n-1}$ be the elapsed time between two sequential observations.

Stationary Dynamics: Define $\ell(\tilde{N}) \stackrel{\text{def}}{=} \partial_t \tilde{N}(t)$ with initial conditions $N_c(0) = 0$. Then the Dynamical CRP assumes the data is stationary and simplifies to the CRP (Fig. 1, Time Function: $\Theta(\Delta)$).

Exponential Dynamics: Define $\ell(\tilde{N}) \stackrel{\text{def}}{=} \tau \partial_t \tilde{N}(t) + \tilde{N}(t)$. Then the relevance of previous customers (observations) decays exponentially with elapsed time, and the Dynamical CRP simplifies to the time-sensitive CRP (Fig. 1, Time Function: $\exp(-\Delta)$).

Oscillatory Dynamics: Suppose we want cluster assignments to be periodic on a particular timescale. For instance, dawn and dusk have visually similar light, but crepuscular animals need to distinguish them; similarly, fall and spring have similar day durations and temperatures, but migratory and hibernating/aestivating animals need to distinguish them. By defining the dynamics as a linear second order differential equation $\ell(\tilde{N}) \stackrel{\text{def}}{=} \partial_t^2 \tilde{N}(t) + \omega_0^2 \tilde{N}(t)$, the Dynamical CRP creates oscillatory table assignments (Fig. 1, Time Function: $\cos(\Delta)$).

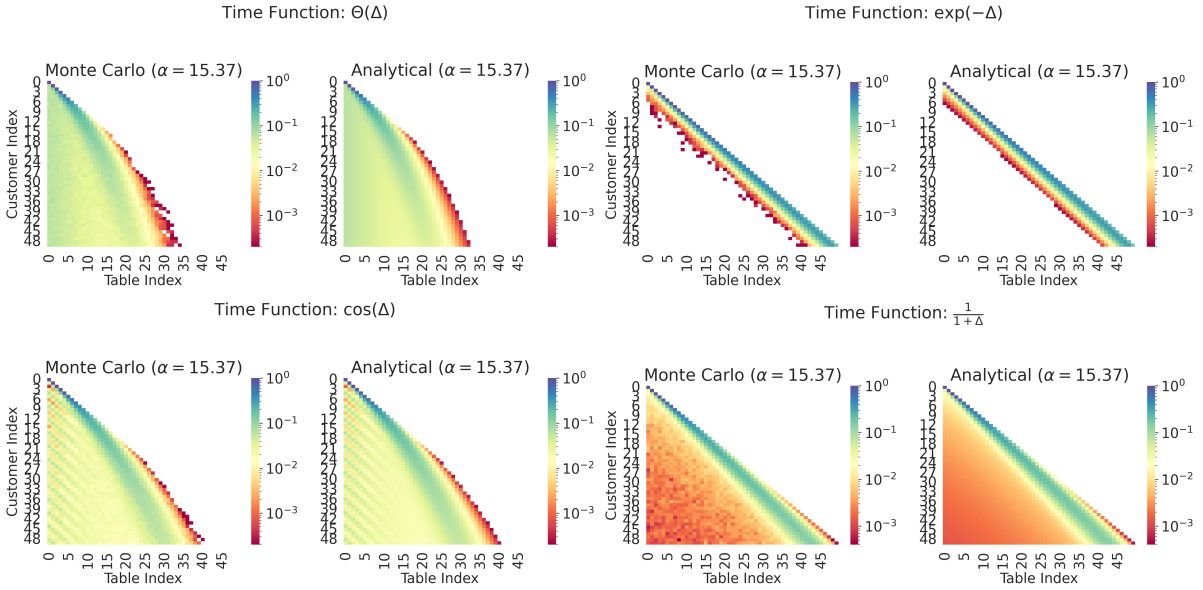

Figure 1: **The Dynamical Chinese Restaurant Process under 4 different dynamics.** Elapsed time is denoted by $\Delta \stackrel{\text{def}}{=} t_n - t_{n-1}$. Under different dynamics, the Dynamical CRP produces the CRP (Time Function: $\Theta(\Delta)$), the time-sensitive CRP (Time Function: $\exp(-\Delta)$), and new distributions over partitions of a set including sinusoidal (Time Function: $\cos(\Delta)$) and hyperbolic (Time Function: $1/(1+\Delta)$). Columns 1 and 3 are Monte Carlo samples; Columns 2 and 4 are our analytical recursion.

Hyperbolic Dynamics: Hyperbolic discounting is commonly used in reinforcement learning and observed across species including humans, monkeys, and rats (Sozou, 1998; Fedus et al., 2019). The Dynamical CRP also enables hyperbolic clustering (Fig. 1, Time Function: $1/(1+\Delta)$).

Power Law Dynamics: Some non-stationary clustering methods e.g. Dinari & Freifeld (2022) assume that the relevance of previous observations should fall as a power law of elapsed time. The Dynamical CRP can approximate power laws (Not shown, Time Function: $b^{-\lambda\Delta}$).

3.5 Generative Model

We now define the generative model for the streaming data $o_{1:N}$, observed at known times $t_{1:N}$, using the Dynamical CRP as a prior over cluster assignments $c_{1:N}$, and then turn to discussing how to perform streaming inference.

$$\begin{aligned} c_{1:N}|t_{1:N} &\sim D\text{-}CRP(\ell, \alpha) \\ \phi_k &\sim_{i.i.d.} p(\phi) \\ o_n|c_n, \{\phi_k\}_{k=1}^{\infty} &\sim p(o_n; \phi_{c_n}) \end{aligned} \tag{5}$$

3.6 Streaming Inference

Our approach will be to first show that the Dynamical CRP can be expressed in a recursive form designed for streaming inference, similar to the CRP, and then use this recursive form to define a variational family for streaming inference.

3.6.1 Recursive Form of the Dynamical CRP

As with the CRP, the Dynamical CRP's conditional distribution renders each cluster assignment c_n dependent on the entire history of previous cluster assignments $c_{<n}$. This handicaps its applicability to streaming data. We overcome this handicap by converting the conditional distribution to a marginal distribution by taking the average over all possible histories of cluster assignments (termed *sample paths* in the stochastic processes literature). Omitting $t_{1:N}$ for

brevity,

$$\begin{aligned} p(c_n = c|\ell, \alpha) &= \mathbb{E}_{p(c_n|\ell,\alpha)}[\mathbb{I}(c_n = c)] \\ &= \mathbb{E}_{p(c_{<n}|\ell,\alpha)}[\mathbb{E}_{p(c_n|c_{<n},\ell,\alpha)}[\mathbb{I}(c_n = c)]] \\ &= \mathbb{E}_{p(c_{<n}|\ell,\alpha)}[p(c_n = c|c_{<n}, \ell, \alpha)] \end{aligned}$$

Substituting the Dynamical CRP's conditional distribution and taking a first-order Taylor series approximation of the expectation yields:

$$\begin{aligned} p(c_n = c|\ell, \alpha) &= \mathbb{E}_{p(c_{<n}|\ell,\alpha)}\left[\frac{N_c(t_n)}{\alpha + \sum_c N_c(t_n)} + \frac{\alpha}{\alpha + \sum_c N_c(t_n)}\mathbb{I}(c = C_{n-1} + 1)\right] \\ &\approx \frac{\mathbb{E}[N_c(t_n)]}{\alpha + \mathbb{E}[\sum_c N_c(t_n)]} + \frac{\alpha}{\alpha + \mathbb{E}[\sum_c N_c(t_n)]} p(C_{n-1} = c - 1) \end{aligned}$$

Abusing notation slightly, we can write $N_c(t) = \sum_{n':t_{n'}<t} \ell(\mathbb{I}(c_{n'} = c), t_{n'}, t)$, where $\ell(\cdot, t_{n'}, t)$ means advancing the dynamical system from time $t_{n'}$ to time t. Because both the dynamics ℓ and the expectation are linear operators, the two commute and the expectation can be pulled inside:

$$\mathbb{E}[N_c(t)] = \sum_{n':t_{n'}<t} \ell(\mathbb{E}[\mathbb{I}(c_{n'} = c)], t_{n'}, t) = \sum_{n':t_{n'}<t} \ell(p(c_{n'} = c|\ell, \alpha), t_{n'}, t)$$

Together, this yields the recursive form of the Dynamical CRP:

$$p(c_n = c|\ell, \alpha) \propto \sum_{n':t_{n'}<t} \ell(p(c_{n'} = c|\ell, \alpha), t_{n'}, t_n) + \alpha\, p(C_{n-1} = c - 1) \tag{6}$$

The Dynamical CRP's recursive form (Eqn. 6) has similar intuition to the CRP's recursive form (Eqn. 2): previous cluster assignments influence the current cluster assignment, and clusters can appear with new observations. The key modification is that previous probability masses can now change over time. We confirm the correctness of Eqn. 6 by comparing the analytical expression to 5000 Monte Carlo samples drawn from the Dynamical CRP's conditional distribution over $\alpha \in \{1.1, 10.78, 15.37, 30.91\}$ and with step, exponential, sinusoidal, and hyperbolic dynamics (Fig. 1); visually, the analytical and Monte Carlo plots display excellent agreement. Quantitatively, the mean squared error between the analytical expression for all $p(c_n|\ell, \alpha)$ and the Monte Carlo estimates fall approximately as a power law in the number of Monte Carlo samples (Fig. 2) for all α values. This supports our claim that the recursive form of the Dynamical CRP is highly accurate for the stochastic process.

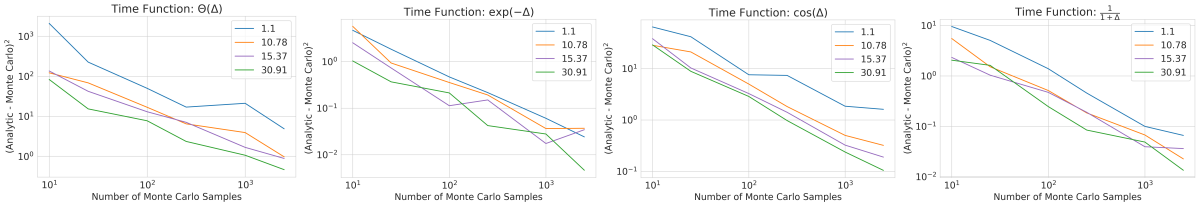

Figure 2: **Mean-Squared Error between analytical expression for $p(c_n|\ell, \alpha)$ and a Monte Carlo estimate, under 4 dynamics.** Over a wide range of α values, the mean-squared error between our analytical expression and Monte Carlo estimates falls approximately as a power law, showing the exactness of Eqn. 6.

3.6.2 Streaming Inference via Recursive Form of Dynamical CRP

To perform streaming inference, we start by considering the streaming evidence lower bound (ELBO) with variational parameters θ_n:

$$\begin{aligned} p(o_n|o_{<n}) &\geq \mathbb{E}_{q(c_n,\{\phi\}|o_{\leq n};\theta_n)}[\log p(o_n, c_n, \{\phi\}|o_{<n})] + H[q(c_n, \{\phi\}, o_{\leq n})] \\ &= \mathbb{E}_{q(c_n,\{\phi\}|o_{\leq n};\theta_n)}[\log p(o_n|c_n, \{\phi\}, o_{<n}) + \log p(c_n, \{\phi\}|o_{<n})] + H[q(c_n, \{\phi\}|o_{\leq n}; \theta_n)] \end{aligned}$$

However, computing this ELBO is tricky because the filtering prior $p(c_n, \{\phi\}|o_{<n})$ is unknown. Using the recursive form of the Dynamical CRP as inspiration, we replace the filtering prior with an approximate filtering prior:

$$q(c_n|o_{<n}) \overset{\text{def}}{\propto} \sum_{n'<n} \ell(q(c_{n'} = c|o_{\leq n'}, \ell, \alpha), t_{n'}, t_n) + \alpha\, q(C_{n-1} = c-1|o_{\leq n-1})$$

$$q(\{\phi\}|o_{<n}) \overset{\text{def}}{=} \prod_k q(\phi_k|o_{\leq n-1})$$

$$q(c_n, \{\phi\}|o_{<n}) \overset{\text{def}}{=} q(c_n|o_{<n}) q(\{\phi\}|o_{<n})$$

Substituting the approximate filtering prior yields an approximate filtering evidence lower bound that we maximize:

$$\mathbb{E}_{q(c_n,\{\phi\}|o_{\leq n};\theta_n)}[\log p(o_n|c_n, \{\phi\}, o_{<n}) + \log q(c_n, \{\phi\}|o_{<n})] + H[q(c_n, \{\phi\}|o_{\leq n}; \theta_n)] \tag{7}$$

4 Experimental Results

4.1 Room Clustering for Simultaneous Localization and Mapping

As a motivating example, we consider an agent moving spatially through an unfamiliar environment with the goal of clustering observations into states, perhaps to use for planning or reinforcement learning. This problem is motivated by the field of simultaneous localization and mapping (SLAM) (Rosen et al., 2021), in which an agent learns a map of its environment as well as its location within said environment; a common approach is to cluster sensory observations into rooms that can then be used to efficiently plan (Fairfield et al., 2010; Klukas et al., 2022).

We procedurally generated 1000 environments with multiple rooms, each containing a variable number of randomly placed objects (called "landmarks" in the SLAM literature), and then generated a single trajectory through each environment. At each position along the trajectory, a landmark is either visible or not, determined by a top-down limited field of view; the observations are thus high-dimensional binary vectors. We used a product-of-independent-Bernoullis likelihood for expressing whether each landmark is visible from a given position; for simplicity, the likelihood does not take into account position or velocity. As D-CRP takes a single trajectory through each novel environment, D-CRP aggregates sensory landmarks (black diamonds) into visually-distinguishable unique clusters (rooms) and visually-indistinguishable non-unique clusters (hallways, orange) (Fig. 3).

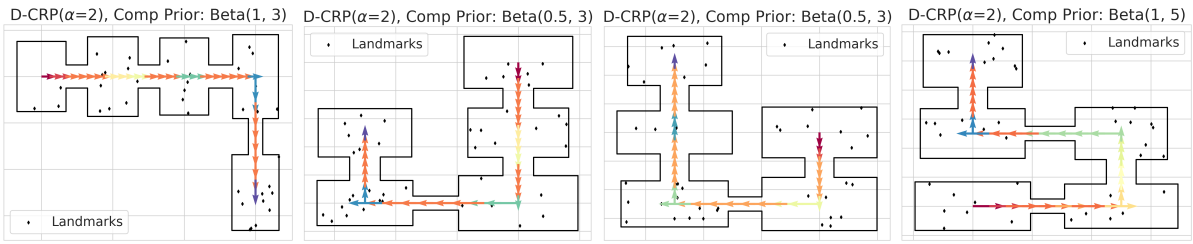

Figure 3: **Clusters inferred by Dynamical CRP in a 2D spatial navigation task.** Each color in each environment represents a unique cluster, inferred from visible landmarks (black diamonds) encountered along a single trajectory. The Dynamical CRP aggregates visually-distinguishable rooms (various colors) into distinct clusters and visually-identical hallways into the same cluster (orange). Each observation is a high dimensional binary vector, specifying which landmarks are (not) visible from a position, as determined by a top-down viewpoint with a limited field of view.

To emphasize how time provides a useful inductive bias, we tested D-CRP against R-CRP (Schaeffer et al., 2021); the two algorithms are identical save for R-CRP's assumption that the observations are exchangable. Across a wide range of hyperparameters (α for the stochastic processes, τ for D-CRP's dynamics, and Beta priors for cluster parameters), we found the D-CRP with exponentially decaying dynamics outperforms R-CRP as measured by normalized mutual information between true cluster labels and inferred cluster labels (Fig. 4).

D-CRP outperforms R-CRP because D-CRP's dynamics provides an inductive bias that trajectories are temporally and spatially smooth. This inductive bias is useful because it discounts clusters seen long ago (unless the sensory evidence is highly compelling) and because it preferentially allocates two sequential observations to the same cluster.

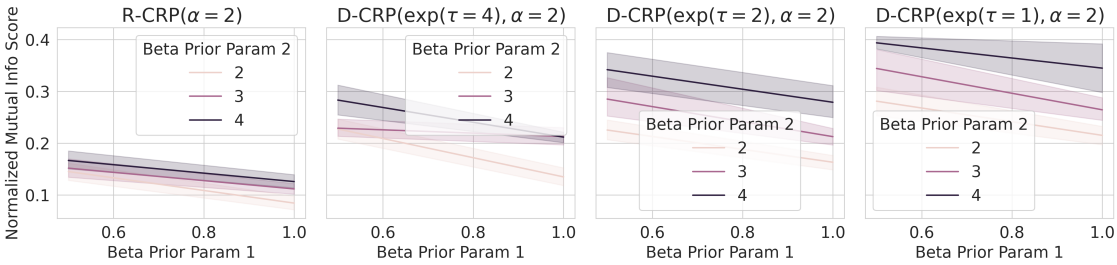

Figure 4: **Dynamical CRP outperforms Recursive CRP across a wide range of hyperparameters.** D-CRP's dynamics provide a useful inductive bias that trajectories are temporally smooth, enabling it to outperform R-CRP which assumes that observations are exchangeable.

4.2 Synthetic Mixture of Gaussians

Following previous work (Kulis & Jordan, 2012), we moved to synthetic mixtures of Gaussians. We generated 3600 datasets, each containing 1000 observations, by placing a Gaussian prior on the cluster means $p(\phi) = \mathcal{N}(0, \rho^2 I)$ and sweeping over dynamics, alpha, signal-to-noise ratios, and the number of dimensions. We compared Dynamical CRP against seven baseline inference algorithms; three are streaming and four are not. The non-streaming algorithms have unfettered access to all observations and therefore serve as upper bounds on performance; any comparison against these non-streaming baselines maximally disfavors our method. The baselines are:

- Collapsed Gibbs Sampling (non-streaming) (Neal, 2000).
- Variational Bayes Dirichlet-Process Gaussian Mixture Model (non-streaming) (Blei & Jordan, 2006), implemented in scikit-learn (Pedregosa et al., 2011).
- K-Means (both streaming and non-streaming variants) (MacQueen, 1967; Lloyd, 1982). For both variants, K-Means is given the ground-truth number of clusters.
- DP-Means (both streaming and non-streaming variants) (Kulis & Jordan, 2012; Broderick et al., 2013b).
- Recursive-CRP (streaming) (Schaeffer et al., 2021).

The baseline algorithms are all designed for stationary data, so we started our comparison with stationary dynamics, i.e. $\ell(\tilde{N}) = \partial_t \tilde{N}(t)$, but we also considered other dynamics (exponential and hyperbolic). We measured the performance of each algorithm by the normalized mutual information between the inferred cluster assignments and the true cluster assignments, implemented in scikit-learn (Pedregosa et al., 2011). We found that Dynamical CRP was competitive on stationary data (Fig. 5A), but excelled on non-stationary data (Fig. 5BC).

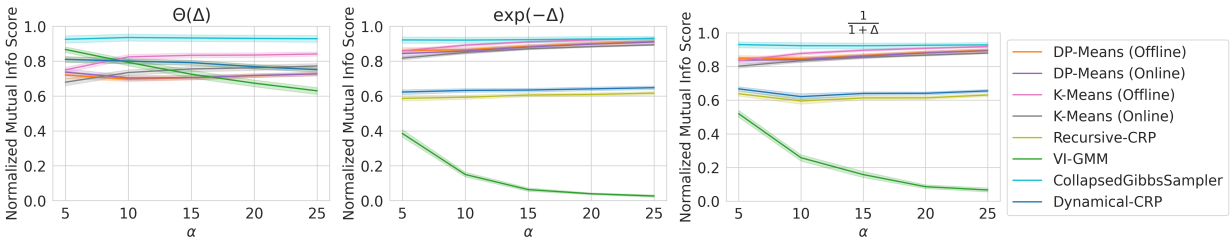

Figure 5: **Normalized mutual information between true cluster assignments and inferred cluster assignments in Gaussian Mixture Models under 3 different dynamics.**

We additionally plotted the number of clusters inferred by each algorithm. We found that D-CRP was often well within the correct order of magnitude, growing appropriately with the concentration hyperparameter α (Fig. 6).

To explore how clusters are created as observations are received, we visualized when the Dynamical CRP creates clusters by plotting the ratio of the number of inferred clusters to the total number of true clusters as a function of the number of observations, dividing by the total number of true clusters in that set of observations. We found that Dynamical CRP creates clusters over time, as necessitated by the data (Fig. 7).

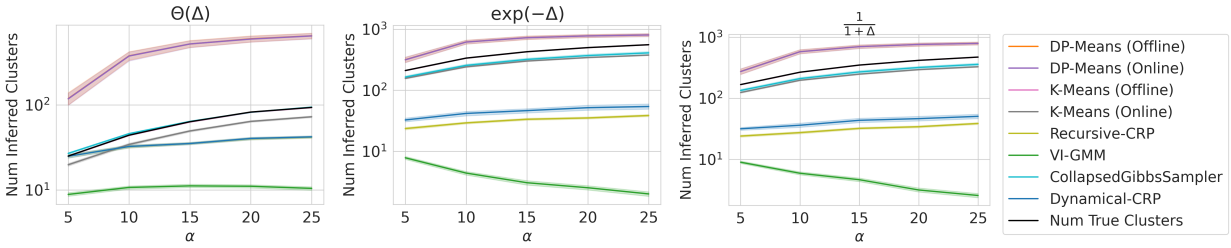

Figure 6: **Dynamical CRP recovers close to the correct number of clusters under 3 different dynamics.**

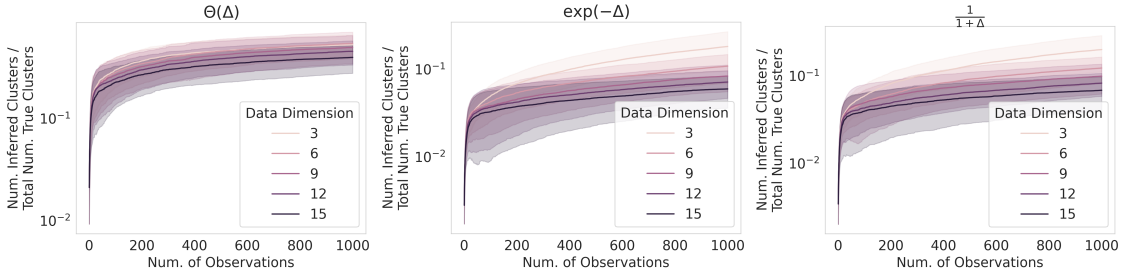

Figure 7: **Dynamical CRP creates clusters over time, as necessitated by incoming data.**

We also investigated how Dynamical CRP performs under different signal-to-noise (defined as the ratio of means covariance prefactor ρ to likelihood covariance prefactor σ) regimes; we found that Dynamical CRP displays better performance the higher the SNR (Fig. 8).

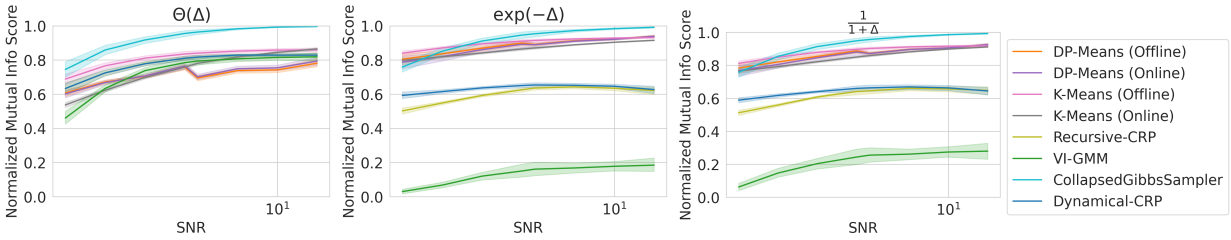

Figure 8: **Dynamical CRP displays better performance when the data has a higher signal-to-noise ratio.**

4.3 Synthetic Mixture of von Mises-Fisher

To demonstrate that the Dynamical CRP is not limited to Gaussian mixture models in Euclidean space, we turned to von Mises-Fisher mixture models on the surface of hyperspheres. We made this particular choice because one future line of work we are excited by involves combining deep learning with Bayesian nonparametrics for lifelong learning, and recent advances in self-supervised representation learning constrain deep neural network representations to the surface of hyperspheres (Chen et al., 2020; Grill et al., 2020; Caron et al., 2021). As with the mixture of Gaussians, we generated 3600 datasets, with 1000 observations each, from the generative model, with a uniform prior on the cluster directions $p(\phi) = \mathcal{VMF}(\kappa = 0)$ and sweeping over dynamics, alpha, signal-to-noise ratios, and the number of dimensions. Most previous baselines were designed for Gaussian likelihoods, meaning only the Recursive-CRP could be used. We again plotted the number of clusters inferred by each algorithm. We found that D-CRP often outperformed R-CRP (Fig. 9) and was often well within the correct order of magnitude, growing appropriately with the concentration hyperparameter α (Fig. 10).

4.4 Self-Supervised Visual Representation Learning

We returned to self-supervised representation learning on the surface of the hypersphere, albeit in a simplified form. Caron et al. (2021) recently proposed a self-supervised algorithm called SwAV for learning visual features that

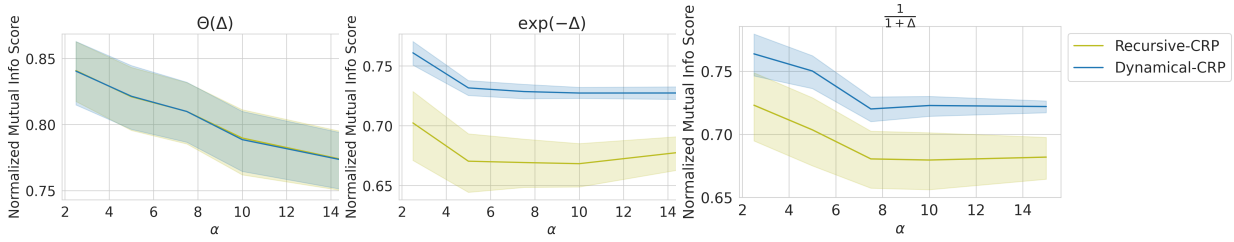

Figure 9: **Normalized mutual information between true cluster assignments and inferred cluster assignments in von Mises-Fisher Mixture Models under 3 different dynamics.**

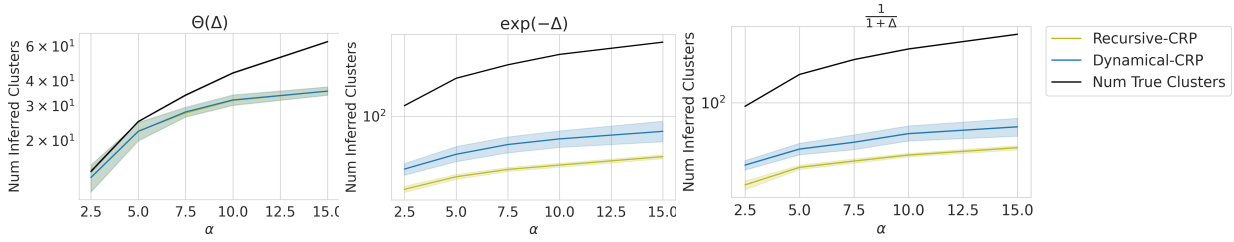

Figure 10: **Dynamical CRP recovers close to the correct number of clusters under 3 different dynamics.**

achieved state-of-the-art finetuning performance on ImageNet (Deng et al., 2009). At its heart, SwAV relies on clustering on the hypersphere to learn useful visual features; however, SwAV uses a fixed, finite number of cluster centroids (called "prototypes") that limit its applicability to streaming, nonstationary data. As a first step toward testing whether D-CRP can be used for high-dimensional clustering on the hypersphere, we tested how well D-CRP can extract the ImageNet classes from the pretrained SwAV embeddings in an unsupervised manner in a single pass. To do this, we extracted 10,000 ImageNet validation data embeddings from pretrained SwAV and then clustered the embeddings on the 128-dimensional hypersphere using Dynamical CRP. To introduce non-statinoarity, we followed Dinari & Freifeld (2022), in that we introduced what they call "recurring concept drift"; specifically, this means that observations are sampled from a small subset of the total ImageNet classes, where the subset changes over time as old classes are removed and new classes are added. This emulates what might happen naturally e.g. if a bird is seen, other birds are likely to be seen soon after.

We found that over a wide range of α and signal-to-noise (defined as the von-Mises Fisher likelihood κ) values, Dynamical CRP displayed high normalized mutual information between the inferred cluster labels and the ground-truth ImageNet classes (Deng et al., 2009) (Fig. 11). Note that here, α and κ are unknown hyperparameters that we swept; the high performance of Dynamical CRP across many values speaks to its robustness.

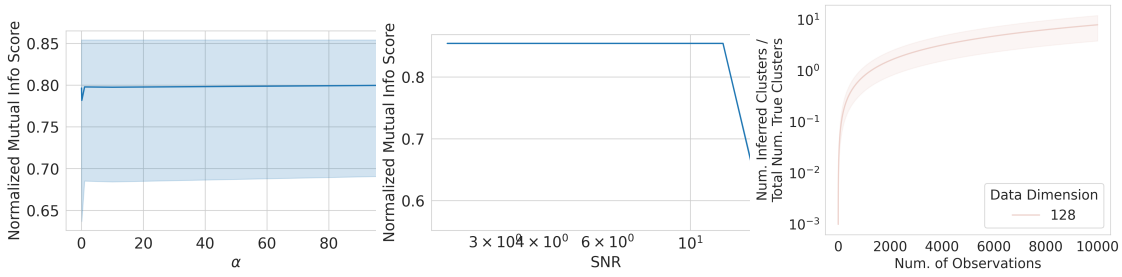

Figure 11: **Dynamical CRP capably clusters SwAV (Caron et al., 2021) embeddings of ImageNet validation data (Deng et al., 2009).**

5 Related Work

Streaming inference for "infinite" probabilistic models has a rich history (Broderick et al., 2013a; Lin, 2013; Campbell et al., 2015; Tank et al., 2015). On the specific subtopic of non-stationary clustering, there are many great recent and relevant papers. Dinari & Freifeld (2022) introduces a streaming augmented-space sampler extension of Chang & Fisher III (2013), with a heuristic of power-law decay as a function of elapsed time; in comparison, the Dynamic CRP uses variational inference in lieu of sampling and permits more general temporal structure. Saad & Mansinghka (2018) present a non-streaming temporally-reweighted CRP model, in which each cluster maintains a recent window of assigned observations, and new observations are compared to each cluster's product-of-Student-T distributions centered at the clusters' recent assigned observations. In comparison, D-CRP does not require one-hot cluster assignments and admits a streaming inference algorithm, enabling use on streaming data.

6 Limitations

One key limitation of our work is that the Dynamical CRP and the baseline algorithms use fixed embeddings of the data, rather than learning embeddings of the data. Future work should strive to combine these methods with representation learning.

7 Discussion

In this paper, we attack unsupervised learning on streaming non-stationary data in the specific setting of mixture modeling. We propose a novel stochastic process that defines a non-exchangable distribution over partitions of a set: the Dynamical Chinese Restaurant Process. We show that the Dynamical CRP provides a bespoke non-stationary prior over cluster assignments and is amenable to an efficient streaming variational inference algorithm. We then demonstrate on both synthetic and real data, with Gaussian and non-Gaussian likelihoods, that the Dynamical CRP provides a powerful clustering algorithm for non-stationary streaming unsupervised data.

REFERENCES

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics*, 2(6):1152–1174, November 1974. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176342871. URL https://projecteuclid.org/euclid.aos/1176342871. Publisher: Institute of Mathematical Statistics.

Matthew James Beal. Variational Algorithms for Approximate Bayesian Inference. *PhD thesis, Gatsby Computational Neuroscience Unit, UCL*, pp. 281, 2003.

David Blackwell and James B. MacQueen. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1(2):353–355, March 1973. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176342372. Publisher: Institute of Mathematical Statistics.

David M Blei and Peter I. Frazier. Distance Dependent Chinese Restaurant Processes. *Journal of Machine Learning Research*, 12:28, 2011.

David M. Blei and Michael I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1): 121–143, March 2006. ISSN 1936-0975. doi: 10.1214/06-BA104. URL http://projecteuclid.org/euclid.ba/1340371077.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming Variational Bayes. *Neural Information Processing Systems*, pp. 9, 2013a.

Tamara Broderick, Brian Kulis, and Michael I Jordan. MAD-Bayes: MAP-based Asymptotic Derivations from Bayes. *International Conference on Machine Learning*, pp. 9, 2013b.

Trevor Campbell, Julian Straub, John W Fisher Iii, and Jonathan P How. Streaming, Distributed Variational Inference for Bayesian Nonparametrics. *Neural Information Processing Systems*, pp. 9, 2015.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *arXiv:2006.09882 [cs]*, January 2021. URL http://arxiv.org/abs/2006.09882. arXiv: 2006.09882.

Jason Chang and John W Fisher III. Parallel Sampling of DP Mixture Models using Sub-Cluster Splits. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/bca82e41ee7b0833588399b1fcd177c7-Abstract.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2002.05709. arXiv: 2002.05709.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Or Dinari and Oren Freifeld. Sampling in Dirichlet Process Mixture Models for Clustering Streaming Data. *arXiv:2202.13312 [cs, stat]*, February 2022. URL http://arxiv.org/abs/2202.13312. arXiv: 2202.13312.

Nathaniel Fairfield, David Wettergreen, and George Kantor. Segmented SLAM in three-dimensional environments. *Journal of Field Robotics*, 27(1):85–103, 2010. ISSN 1556-4967. doi: 10.1002/rob.20320. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20320. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.20320.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic Discounting and Learning over Multiple Horizons. *arXiv:1902.06865 [cs, stat]*, February 2019. URL http://arxiv.org/abs/1902.06865. arXiv: 1902.06865.

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, March 1973. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176342360. Publisher: Institute of Mathematical Statistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv:2006.07733 [cs, stat]*, September 2020. URL http://arxiv.org/abs/2006.07733. arXiv: 2006.07733.

Mirko Klukas, Sugandha Sharma, YiLun Du, Tomas Lozano-Perez, Leslie Kaelbling, and Ila Fiete. Fragmented Spatial Maps from Surprisal: State Abstraction and Efficient Planning. Technical report, bioRxiv, January 2022. URL https://www.biorxiv.org/content/10.1101/2021.10.29.466499v2. Section: New Results Type: article.

Brian Kulis and Michael I Jordan. Revisiting k-means: New Algorithms via Bayesian Nonparametrics. *International Conference on Machine Learning*, pp. 8, 2012.

Soochan Lee, Junsoo Ha, Dongsu Zhang, and Gunhee Kim. A Neural Dirichlet Process Mixture Model for Task-Free Continual Learning. *arXiv:2001.00689 [cs, stat]*, January 2020. URL http://arxiv.org/abs/2001.00689. arXiv: 2001.00689.

Dahua Lin. Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation. *Neural Information Processing Systems*, pp. 9, 2013.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982. ISSN 1557-9654. doi: 10.1109/TIT.1982.1056489. Conference Name: IEEE Transactions on Information Theory.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967. Issue: 14.

Radford M Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, pp. 20, 2000.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

David M. Rosen, Kevin J. Doherty, Antonio Teran Espinoza, and John J. Leonard. Advances in Inference and Representation for Simultaneous Localization and Mapping. *arXiv:2103.05041 [cs]*, March 2021. URL http://arxiv.org/abs/2103.05041. arXiv: 2103.05041.

Feras A Saad and Vikash K Mansinghka. Temporally-Reweighted Chinese Restaurant Process Mixtures for Clustering, Imputing, and Forecasting Multivariate Time Series. *AISTATS*, pp. 10, 2018.

Rylan Schaeffer, Blake Bordelon, Mikail Khona, Weiwei Pan, and Ila Rani Fiete. Efficient Online Inference for Nonparametric Mixture Models. *Uncertainty in Artificial Intelligence*, pp. 10, 2021.

Rylan Schaeffer, Yilun Du, Gabrielle Kaili-May Liu, and Ila Rani Fiete. Streaming Inference for Infinite Feature Models. *Proceedings of the 39th International Conference on Machine Learning*, pp. 22, 2022.

P. D. Sozou. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409):2015–2020, October 1998. doi: 10.1098/rspb.1998.0534. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1998.0534. Publisher: Royal Society.

Alex Tank, Nicholas J Foti, and Emily B Fox. Streaming Variational Inference for Bayesian Nonparametric Mixture Models. *AISTATS*, pp. 9, 2015.

Martin J. Wainwright and Michael Irwin Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc, 2008. ISBN 978-1-60198-184-4. Google-Books-ID: zp5Mo3VsJbgC.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Time-Sensitive Dirichlet Process Mixture Models. pp. 14, 2005.

A Coordinate Ascent Variational Inference Parameter Updates

This section contains derivations for Coordinate Ascent Variational Inference (CAVI). Specifically, the following derivations show how to update variational parameters in different generative models with various likelihoods.

A.1 Closed-Form Expression for Variational Parameters in the Exponential Family

In the following three subsections, we use the following fact from Beal (2003); Wainwright & Jordan (2008): if a distribution p and its variational approximation q are both in the exponential family, then the optimal variational parameters ζ_i that correspond to the variational distribution over variable W_i are the solution to

$$\log q(W_i; \zeta_i) = \mathbb{E}_{q(W_{-i})}[\log p(W, X|\theta)] \tag{8}$$

This means that when optimizing the parameters for one variable, we can replace all other variables with their expectations under the variational distribution and then solve for that one variable's variational parameters.

A.2 CAVI for Multivariate Gaussian Likelihood

Mean field family:

$$\begin{aligned} q(c_n, \{\phi\}|o_{\leq n}) &\overset{\text{def}}{=} q(c_n|o_{\leq n}; \{\pi_{nc}\}) \prod_{k=1}^{C_n} q(\phi_{nc}|o_{\leq n}; \mu_{nc}, \Sigma_{nc}) \\ q(c_n|o_{\leq n}; \{\pi_{nc}\}) &\overset{\text{def}}{=} Categorical(\pi_n) \\ q(\phi_{nc}|o_{\leq n}; \mu_{nc}, \Sigma_{nc}) &\overset{\text{def}}{=} \mathcal{N}(\mu_{nc}, \Sigma_{nc}) \end{aligned}$$

where $\theta_n \overset{\text{def}}{=} \{\pi_{nc}\}_k \cup \{\mu_{nc}\}_k \cup \{\Sigma_{nc}\}_k$ are our variational parameters for the nth observation. The mixture weights' parameters π_n will be determined by solving the following:

$$\log q(c_n|o_{\leq n}; \pi_n) = \mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})]$$

The left-hand side (LHS) is:

$$\log q(c_n|o_{\leq n}; \pi_n) = \sum_k \mathbb{I}(c_n = k) \log \pi_{nc}$$

Dropping terms that don't include c_n, the right-hand side (RHS) contains two relevant terms:

$$\begin{aligned} &\mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})] \\ &= \mathbb{E}_{q(\{\phi_{nc}\})}[\log p(c_n|o_{<n}) + \log p(o_n|c_n, \{\phi_{nc}\})] \\ &= \log q(c_n|o_{<n}) + \mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n|c_n, \{\phi_{nc}\})] \end{aligned}$$

The first term is determined by the Dynamical CRP prior:

$$\log q(c_n|o_{<n}) = \sum_k \mathbb{I}(c_n = k) \log q(c_n = k|o_{<n})$$

The second term is given by:

$$\begin{aligned} &\mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n|c_n, \{\phi_{nc}\})] \\ &= \mathbb{E}_{q(\{\phi_{nc}\})}\Big[\sum_k -\frac{1}{2\sigma_o^2}||o_n - \phi_{nc}||^2 \, \mathbb{I}(c_n = k)\Big] \\ &= \sum_k -\frac{1}{2\sigma_o^2}(o_n^T o_n - 2o_n^T \mu_{nc} + \mathrm{Tr}[\Sigma_{nc} + \mu_{nc}\mu_{nc}^T]) \, \mathbb{I}(c_n = k) \end{aligned}$$

Comparing the simplified left-hand and right-hand sides, and solving for the variational parameter of the probability of the lth cluster π_{nl}:

$$\pi_{nl} \propto \exp\Big(\log q(c_n = l|o_{<n}) - \frac{1}{2\sigma_o^2} o_n^T o_n + \frac{1}{\sigma_o^2} o_n^T \mu_{nl} - \frac{1}{2\sigma_o^2} \operatorname{Tr}[\Sigma_{nl} + \mu_{nl}\mu_{nl}^T]\Big)$$

For the lth cluster centroid's variational parameters, we want to solve:

$$\log q(\phi_{nl}|o_{\leq n}; \mu_{nl}, \Sigma_{nl}) = \mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})]$$

The LHS is:

$$\begin{aligned}&\log q(\phi_{nl}|o_{\leq n}; \mu_{nl}, \Sigma_{nl})\\ &= -\frac{1}{2}(\phi_{nl} - \mu_{nl})^T \Sigma_{nl}^{-1}(\phi_{nl} - \mu_{nl})\end{aligned}$$

The RHS is:

$$\mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})] = \mathbb{E}_q[\log q(\phi_{nl}|o_{<n}) + \log p(o_n|c_n, \{\phi_{nc}\}, o_{<n})]$$

where the first RHS term is:

$$\mathbb{E}_q[\log q(\phi_{nl}|o_{<n})] = -\frac{1}{2}(\phi_{nl} - \mu_{n-1,l})^T \Sigma_{n-1,l}^{-1}(\phi_{nl} - \mu_{n-1,l})$$

and the second RHS term is:

$$\begin{aligned}\mathbb{E}_q[\log p(o_n|c_n, \{\phi_{nc}\}, o_{<n})] &= \mathbb{E}_q\Big[\sum_k -\frac{1}{2\sigma_o^2}(o_n - \phi_{nc})^T(o_n - \phi_{nc})\mathbb{I}(c_n = k)\Big]\\ &= -\frac{1}{2\sigma_o^2}(o_n - \phi_{nl})^T(o_n - \phi_{nl})\,\pi_{nl}\end{aligned}$$

Setting the LHS and RHS equal and isolating terms quadratic in ϕ_{nl} allows us to solve for the variational covariance parameter:

$$\begin{aligned}\phi_{nl}^T \Sigma_{nl}^{-1} \phi_{nl} &= \phi_{nl}^T \Sigma_{n-1,l}^{-1} \phi_{nl} + \phi_{nl}^T\big(\frac{\pi_{nl}}{\sigma_0^2} I\big)\phi_{nl}\\ \Sigma_{nl} &= \Big(\Sigma_{n-1,l}^{-1} + \frac{\pi_{nl}}{\sigma_0^2} I\Big)^{-1}\end{aligned}$$

Isolating terms linear in ϕ_{nl} allows us to solve for the variational mean parameter:

$$\begin{aligned}\phi_{nl}^T \Sigma_{nl}^{-1} \mu_{nl} &= \phi_{nl}^T \Sigma_{n-1,l}^{-1}\,\mu_{n-1,l} + \phi_{nl}^T(\frac{\pi_{nl}}{\sigma_o^2} I) o_n\\ \mu_{nl} &= \Sigma_{nl}\Big(\Sigma_{n-1,l}^{-1}\,\mu_{n-1,l} + \frac{\pi_{nl}}{\sigma_o^2} o_n\Big)\end{aligned}$$

A.3 CAVI FOR (PRODUCT OF) BERNOULLI LIKELIHOODS

Mean-field family:

$$\begin{aligned}q(c_n, \{\phi\}|o_{\leq n}) &\stackrel{\text{def}}{=} q(c_n|o_{\leq n}; \pi_n) \prod_{c=1}^{C_n} q(\phi_{nc}|o_{\leq n}; \gamma_{nc}, \beta_{nc})\\ q(c_n|o_{\leq n}; \{\pi_n\}) &\stackrel{\text{def}}{=} Categorical(\pi_n)\\ q(\phi_{nc}|o_{\leq n}; \gamma_{nc}, \beta_{nc}) &\stackrel{\text{def}}{=} \prod_c \prod_l Beta(\gamma_{ncl}, \beta_{ncl})\end{aligned}$$

where $\theta_n \stackrel{\text{def}}{=} \{\pi_{nc}\}_c \cup \{\gamma_{ncl}\} \cup \{\kappa_{ncl}\}$ are our variational parameters for the nth observation. The mixture weights' parameters π_n will be determined by solving the following:

$$\log q(c_n|o_{\leq n}; \pi_n) = \mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})]$$

The left-hand side (LHS) is:

$$\log q(c_n|o_{\leq n}; \pi_n) = \sum_c \mathbb{I}(c_n = c) \log \pi_{nc}$$

Dropping terms that don't include c_n, the right-hand side (RHS) contains two relevant terms:

$$\begin{aligned}\mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})] &= \mathbb{E}_q[\log p(c_n|o_{<n}) + \log p(o_n|c_n, \{\phi_{nc}\})] \\ &= \log q(c_n|o_{<n}) + \mathbb{E}_q[\log p(o_n|c_n, \{\phi_{nc}\})]\end{aligned}$$

The first term is determined by the Dynamical CRP prior:

$$\log q(c_n|o_{<n}) = \sum_c \mathbb{I}(c_n = c) \log q(c_n = c|o_{<n})$$

The second term is given by:

$$\begin{aligned}&\mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n|c_n, \{\phi_{nc}\})] \\ &= \mathbb{E}_{q(\{\phi_{nc}\})}\left[\log \prod_c \Big(\prod_l \phi_{ncl}^{x_{nl}}(1-\phi_{ncl})^{1-x_{nl}}\Big)^{\mathbb{I}(c_n=c)}\right] \\ &= \mathbb{E}_{q(\{\phi_{nc}\})}\left[\sum_c \mathbb{I}(c_n = c) \sum_l \log\Big(\phi_{ncl}^{x_{nl}}(1-\phi_{ncl})^{1-x_{nl}}\Big)\right] \\ &= \sum_c \mathbb{I}(c_n = c) \sum_l \Big(x_{nl}\mathbb{E}_{q(\phi_{ncl})}[\log \phi_{ncl}] + (1-x_{nl})\mathbb{E}_{q(\phi_{ncl})}[\log(1-\phi_{ncl})]\Big) \\ &= \sum_c \mathbb{I}(c_n = c) \sum_l \Big(x_{nl}(\psi(\gamma_{ncl}) - \psi(\gamma_{ncl}+\beta_{ncl})) + (1-x_{nl})(\psi(\beta_{ncl}) - \psi(\gamma_{ncl}+\beta_{ncl}))\Big)\end{aligned}$$

where $\psi(x) \stackrel{\text{def}}{=} \frac{d}{dx}\log\Gamma(x)$ is the digamma function. Comparing the simplified left-hand and right-hand sides, and solving for the variational parameter of the probability of the lth cluster π_{nl}:

$$\begin{aligned}\pi_{nc} \propto \exp\Bigg(&\log q(c_n = c|o_{<n}) + \sum_l \Big(x_{nl}(\psi(\gamma_{ncl}) - \psi(\gamma_{ncl}+\beta_{ncl})) \\ &+ (1-x_{nl})(\psi(\beta_{ncl}) - \psi(\gamma_{ncl}+\beta_{ncl}))\Big)\Bigg)\end{aligned}$$

For the c-th cluster's variational parameters, we want to solve:

$$\log q(\phi_{ncl}|o_{\leq n}; \gamma_{ncl}, \beta_{ncl}) = \mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})]$$

The LHS is:

$$\begin{aligned}\log q(\phi_{ncl}|o_{\leq n}; \gamma_{ncl}, \beta_{ncl}) &= \log Beta(\phi_{ncl}; \gamma_{ncl}, \beta_{ncl}) \\ &= (\gamma_{ncl}-1)\log(\phi_{ncl}) + (\beta_{ncl}-1)\log(1-\phi_{ncl})\end{aligned}$$

Dropping terms that don't contain ϕ_{ncl}, the RHS is:

$$\begin{aligned}&\mathbb{E}_{q(-\phi_{ncl})}[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})] \\ &= \mathbb{E}_{q(-\phi_{ncl})}\left[\sum_{c'} \mathbb{I}(c_n = c') \sum_{l'} \log\Big(\phi_{ncl'}^{x_{nl'}}(1-\phi_{ncl'})^{1-x_{nl'}}\Big)\right] + \mathbb{E}_{q(-\phi_{ncl})}\Big[q(\phi_{ncl}|o_{<n})\Big] \\ &= \pi_{nc}\Big(x_{nl}\log(\phi_{ncl}) + (1-x_{nl})\log(1-\phi_{ncl})\Big) + (\gamma_{n-1,cl}-1)\log(\phi_{ncl}) + (\beta_{n-1,cl}-1)\log(1-\phi_{ncl})\end{aligned}$$

Grouping terms and setting equal:

$$\begin{aligned}\gamma_{ncl} - 1 &= \pi_{nc} x_{nl} + \gamma_{n-1,cl} - 1 \\ \gamma_{ncl} &= \pi_{nc} x_{nl} + \gamma_{n-1,cl} \\ \beta_{ncl} - 1 &= \pi_{nc}(1 - x_{nl}) + \beta_{n-1,cl} - 1 \\ \beta_{ncl} &= \pi_{nc}(1 - x_{nl}) + \beta_{n-1,cl}\end{aligned}$$

A.4 CAVI FOR VON-MISES-FISHER LIKELIHOOD

Mean field family:

$$\begin{aligned}q(c_n, \{\phi\}|o_{\leq n}) &\stackrel{\text{def}}{=} q(c_n|o_{\leq n}; \{\pi_{nc}\}) \prod_{k=1}^{C_n} q(\phi_{nc}|o_{\leq n}; \mu_{nc}, \Sigma_{nc}) \\ q(c_n|o_{\leq n}; \{\pi_{nc}\}) &\stackrel{\text{def}}{=} Categorical(\pi_n) \\ q(\phi_{nc}|o_{\leq n}; \mu_{nc}, \kappa_{nc}) &\stackrel{\text{def}}{=} \mathcal{VMF}(\mu_{nc}, \kappa_{nc})\end{aligned}$$

where $\theta_n \stackrel{\text{def}}{=} \{\pi_{nc}\}_k \cup \{\mu_{nc}\}_k \cup \{\kappa_{nc}\}_k$ are our variational parameters for the nth observation. The mixture weights' parameters π_n will be determined by solving the following:

$$\log q(c_n|o_{\leq n}; \pi_n) = \mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})]$$

The left-hand side (LHS) is:

$$\log q(c_n|o_{\leq n}; \pi_n) = \sum_k \mathbb{I}(c_n = k) \log \pi_{nc}$$

Dropping terms that don't include c_n, the right-hand side (RHS) contains two relevant terms:

$$\begin{aligned}\mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})] &= \mathbb{E}_q[\log p(c_n|o_{<n}) + \log p(o_n|c_n, \{\phi_{nc}\})] \\ &= \log q(c_n|o_{<n}) + \mathbb{E}_q[\log p(o_n|c_n, \{\phi_{nc}\})]\end{aligned}$$

The first term is determined by the Dynamical CRP prior:

$$\log q(c_n|o_{<n}) = \sum_k \mathbb{I}(c_n = k) \log q(c_n = k|o_{<n})$$

The second term is given by:

$$\begin{aligned}\mathbb{E}_{q(\{\phi_{nc}\})}[\log p(o_n|c_n, \{\phi_{nc}\})] &= \mathbb{E}_{q(\{\phi_{nc}\})}\Big[\sum_k \frac{1}{\sigma_o^2} \phi_{nc}^T o_n \, \mathbb{I}(c_n = k)\Big] \\ &= \sum_k \frac{1}{\sigma_o^2} \left(\frac{I'_{D/2-1}(\kappa)}{I_{D/2-1}(\kappa)} - \frac{D/2-1}{\kappa}\right) \mu_{nc}^T o_n \, \mathbb{I}(c_n = k)\end{aligned}$$

Comparing the simplified left-hand and right-hand sides, and solving for the variational parameter of the probability of the lth cluster π_{nl}:

$$\pi_{nl} \propto \exp\left(\log q(c_n = l|o_{<n}) + \frac{1}{\sigma_o^2}\left(\frac{I'_{D/2-1}(\kappa)}{I_{D/2-1}(\kappa)} - \frac{D/2-1}{\kappa}\right)\mu_{nc}^T o_n\right)$$

For the lth cluster centroid's variational parameters, we want to solve:

$$\log q(\phi_{nl}|o_{\leq n}; \mu_{nl}, \kappa_{nl}) = \mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})]$$

The LHS is:

$$\log q(\phi_{nl}|o_{\leq n}; \mu_{nl}, \kappa_{nl}) = \kappa_{nl}\, \mu_{nl}^T\, \phi_{nl}$$

The RHS is:

$$\mathbb{E}_q[\log p(o_n, c_n, \{\phi_{nc}\}|o_{<n})] = \mathbb{E}_q[\log q(\phi_{nl}|o_{<n}) + \log p(o_n|c_n, \{\phi_{nc}\}|o_{<n})]$$

where the first RHS term is:

$$\mathbb{E}_q[\log q(\phi_{nl}|o_{<n})] = \kappa_{n-1,l}\, \mu_{n-1,l}^T\, \phi_{nl}$$

and the second RHS term is:

$$\mathbb{E}_q[\log p(o_n|c_n, \{\phi_{nc}\}, o_{<n})] = \mathbb{E}_q\Big[\sum_k \frac{1}{\sigma_o^2}\, \phi_{nc}^T\, o_n\, \mathbb{I}(c_n = k)\Big] = \frac{1}{\sigma_o^2}\phi_{nl}^T\, o_n\, \pi_{nl}$$

Setting the LHS and RHS equal:

$$\kappa_{nl}\, \mu_{nl} = \kappa_{n-1,l}\, \mu_{n-1,l} + \frac{1}{\sigma_o^2}\pi_{nl}\, o_n$$

The two variational parameters are separately recoverable by setting μ_{nl} equal to the unit direction of the right hand side $RHS \stackrel{\text{def}}{=} \kappa_{n-1,l}\, \mu_{n-1,l} + \frac{1}{\sigma_o^2}\pi_{nl}\, o_n$ and by setting κ_{nl} equal to the magnitude of the right hand side:

$$\kappa_{nl} = ||RHS||_2 \qquad \mu_{nl} = \frac{RHS}{||RHS||_2}$$