STREAMING INFERENCE FOR INFINITE NON-STATIONARY CLUSTERING

Anonymous authors
Paper under double-blind review

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 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; then, 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 preliminary experiments showing that the Dynamical CRP can be applied on diverse data.

1 INTRODUCTION

Biological intelligence operates in a radically different data regime than (most) artificial intelligence. 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 open-endedness 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.

In this paper, 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 an intermediary 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 planning or reinforcement learning. In this paper, we specifically consider an agent that receives a single stream of observations from non-stationary clusters, with no ability to revisit past observations, but must nonetheless make progress towards grouping together "similar" observations. In this data regime, the number of clusters (i.e. the number of mixture components) is unknown and unbounded, and so the agent must use a clustering algorithm capable of growing in representational capacity as more observations are seen.

In this workshop 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 classic Chinese Restaurant Process (CRP). We then show how the Dynamical CRP can be used as a prior over cluster assignments in a way that yields an efficient streaming inference algorithm designed for non-stationary data. Finally, we conclude with preliminary experiments showing its applicability.

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 nonstationary setting, the clusters may change over time in a manner that we shall specify.

2.2 Infinite Clustering via the Chinese Restaurant Process

In the non-stationary streaming data regime, the number of clusters is unknown and unbounded. 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 appropriate 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 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 $c_{<n}$:

$$p^{CRP}(c_n = c|c_{<n}, \alpha) \quad \propto \begin{cases} \sum_{n'<n} \mathbb{I}(c_{n'} = c) & \text{if } 1 \leq c \leq C_{n-1} \\ \alpha & \text{if } c = C_{n-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Recent work showed that the CRP is useful for 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 that 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 two-part intuition is that (i) if many observations come from cluster c, then the next observation is also likely to come from cluster c, and (ii) the probability of more clusters should grow with the number of observations, giving the CRP the capacity to create an "infinite" number of clusters.

2.3 Non-stationary Variants of the Chinese Restaurant Process

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:

$$p^{tsCRP}(c_n = c|c_{<n}, \alpha) \propto \begin{cases} \sum_{n'<n} \exp((t_n - t_{n'})/\tau)\mathbb{I}(c_{n'} = c) & \text{if } 1 \leq c \leq C_{n-1} \\ \alpha & \text{if } c = C_{n-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

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 extremely flexible, the ddCRP is impractical for streaming inference because (i) observations can be assigned to future observations, and (ii) the time and space complexity 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 the next likely clusters
- Unlike the CRP, the Dynamical CRP does not assume the observations are 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

The heart of the CRP are the "table occupancies" $N_c(t) \stackrel{\text{def}}{=} \sum_{n' \leq n} \mathbb{I}(c_n = c)\mathbb{I}(t_n \leq t)$, 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 "psuedo" 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 table 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:

$$p^{D\text{-}CRP}(c_n = c | c_{<n}, t_{\leq n}, \ell, \alpha) \propto \begin{cases} N_c(t_n) & \text{if } 1 \leq c \leq C_{n-1} \\ \alpha & \text{if } c = C_{n-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

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

3.4 EXAMPLES

Step Clustering: If we define the dynamics as a perfect integrator $\ell(\tilde{N}) \stackrel{\text{def}}{=} \partial_t \tilde{N}(t)$ with initial conditions $N_c(0) = 0$, the Dynamical CRP simplifies to the CRP (Fig. 1 Time Function: $\Theta(\Delta)$).

Exponential Clustering: If we define the dynamics as a leaky integrator $\ell(\tilde{N}) \stackrel{\text{def}}{=} \partial_t \tilde{N}(t) + \tilde{N}(t)$, 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 Clustering: 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 temperatures, but migratory and hibernating/aestivating animals need to distinguish them. By defining the dynamics as a linear second order differential equation $\ell(\tilde{N}) \stackrel{\text{def}}{=} \partial_t^2 \tilde{N}(t) + \partial_t \tilde{N}(t) + \tilde{N}(t)$, the Dynamical CRP creates oscillatory table assignments (Fig. 1 Time Function: $\cos(\Delta)$).

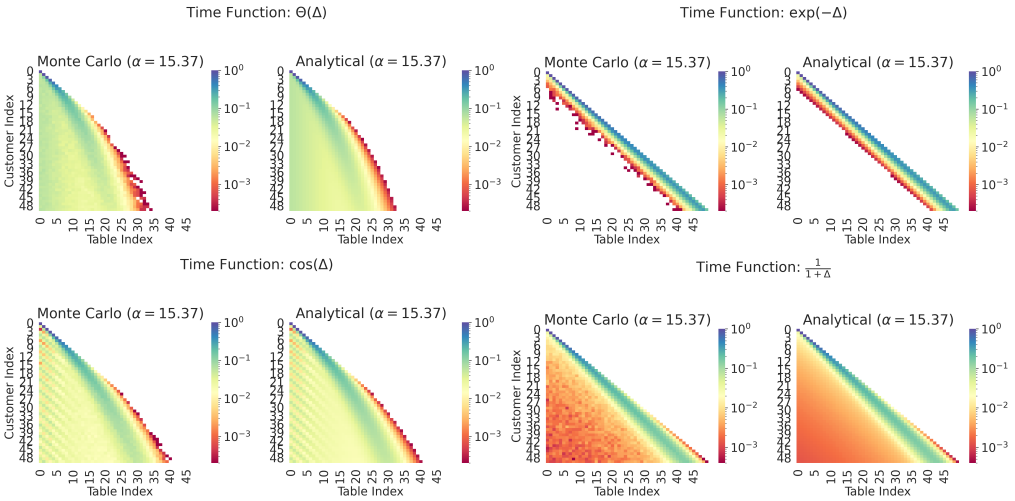

Figure 1: The Dynamical Chinese Restaurant Process under 4 different dynamics. $\Delta \stackrel{\text{def}}{=} t_n - t_{n-1}$ denotes elapsed time. Under different dynamics, the Dynamical CRP produces the CRP (Time Function: $\Theta(\Delta)$), the time-sensitive CRP (Time Function: $\exp(-\Delta)$) 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 Clustering: Hyperbolic discounting is observed across species including humans, monkeys, and rats, and can be explained as exponential discounting with an uncertain timescale (Sozou, 1998). The Dynamical CRP also enables hyperbolic clustering (Fig. 1 Time Function: $1/(1+\Delta)$).

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 $c_{1:N}$:

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

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, similarly to the CRP in (Schaeffer et al., 2021), and then use this recursive form as inspiration for a variational family to use for streaming variational 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 by converting the conditional distribution to a marginal distribution by taking the average over all possible sample paths. Omitting $t_{1:N}$ for brevity,

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

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

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

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 following recursion:

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

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 falls approximately as a power law in the number of Monte Carlo samples (Fig. 2) for all α values. This supports our claim that recursive form of the Dynamical CRP is exact for the stochastic process.

3.6.2 Streaming Inference via Recursive Form of Dynamical CRP

To perform streaming inference, we start by considering the streaming evidence lower bound:

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

with variational parameters θ_n. However, computing this ELBO is tricky because the filtering prior $p(c_n, \{\phi\}|o_{<n})$ is unknown, so we replace the filtering prior with an approximate filtering prior:

$$q(c_n|o_{<n}) \overset{\text{def}}{\propto} \sum_{n':t_{n'}<t} \ell(q(c_{n'} = c|o_{\leq n'}, \ell, \alpha), t_{n'}, t) + \alpha\, q(C_{n-1} = c - 1|o_{\leq n-1})$$

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

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

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

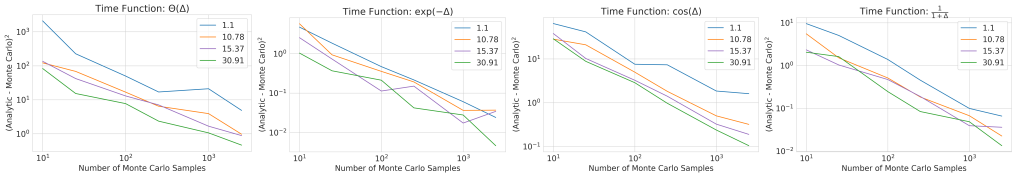

Figure 2: **Mean-Squared Error between analytical expression for $p(c_n|\ell, \alpha)$ and a Monte Carlo estimate.** Over a wide range of α values, the mean-squared error between our analytical expression and Monte Carlo estimates falls approximately as a power law, showing the exactness of Eqn. 6.

4 EXPERIMENTAL RESULTS

4.1 ROOM CLUSTERING FOR SIMULTANEOUS LOCALIZATION AND MAPPING

Clustering is useful not just in its own right, but also as a sub-task for other tasks. For instance, an agent may wish to cluster sensory observations into states for use in planning or reinforcement learning. As a demonstration, we turn to the well-studied problem of simultaneous localization and mapping (SLAM) (Rosen et al., 2021), in which an agent must both learn a map of its environment as well as its location within that environment. One common approach is to learn hierarchically (Fairfield et al., 2010; Klukas et al., 2022) by clustering sensory observations into rooms that can then be used to efficiently plan by planning at the level of rooms instead of at the level of granular positions. We procedurally generated environments with multiple rooms, each containing a variable number of (binary) sensory observations commonly called "landmarks" in the SLAM literature, and then applied D-CRP to a single trajectory through each novel environment. We implemented inference in this setting using a product-of-independent-Bernoullis likelihood for expressing whether each landmark is visible from a given position; the model does not take into account position or velocity information. We found the D-CRP with exponentially decaying dynamics excels at this task (Fig 3). 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). These rooms could then serve as a high-level representation for hierarchical spatial navigation. The D-CRP's excels because its exponentially decaying dynamics gives the D-CRP prior knowledge that trajectories are (typically) temporally and spatially smooth, meaning the D-CRP has a strong prior towards allocating two sequential observations to the same cluster even if the two sensory observations differ significantly.

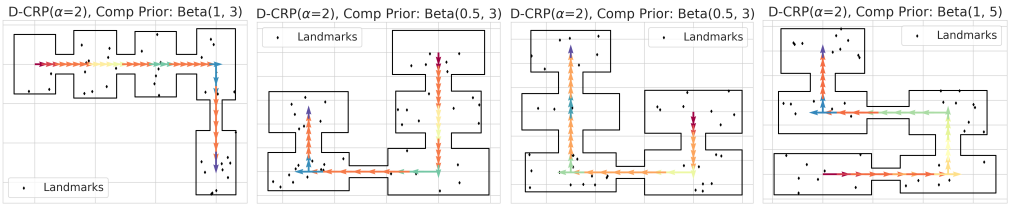

Figure 3: **Clusters inferred by Dynamical CRP in a 2D spatial navigation task.** Each color in each environment represents a unique cluster, inferred from visible landmarks (black diamonds) encountered along a single trajectory. The Dynamical CRP aggregates visually-distinguishable rooms into distinct clusters and visually-identical hallways into the same cluster (orange).

4.2 SYNTHETIC MIXTURE OF MULTIVARIATE GAUSSIANS DISTRIBUTIONS

Following previous work e.g. Kulis & Jordan (2012), we tested our inference algorithm on synthetic mixture of Gaussians data. We sampled 1000 observations from Gaussian mixture models with a Dynamical CRP prior over cluster assignments, assuming known cluster variances, sweeping dynamics, alpha, signal-to-noise ratio and ambient dimension. We found that our streaming inference algorithm excelled at recovering the cluster assignments, as measured by normalized mutual information between the inferred cluster assignments and the true cluster assignments (Fig 4).

We also tested how well our streaming inference algorithm performs at recovering the cluster assignments as a function of signal-to-noise, again considering four different types of dynamics and measuring performance with the normalized mutual information between true cluster assignments and inferred cluster assignments. We generally found that performance increases with higher signal-to-noise (Fig. 5), although not always, meriting further investigation.

5 DISCUSSION

Ultimately, we ran out of time for the experimental results we wanted, but we intend to continue working on this and will hopefully have better results come April!

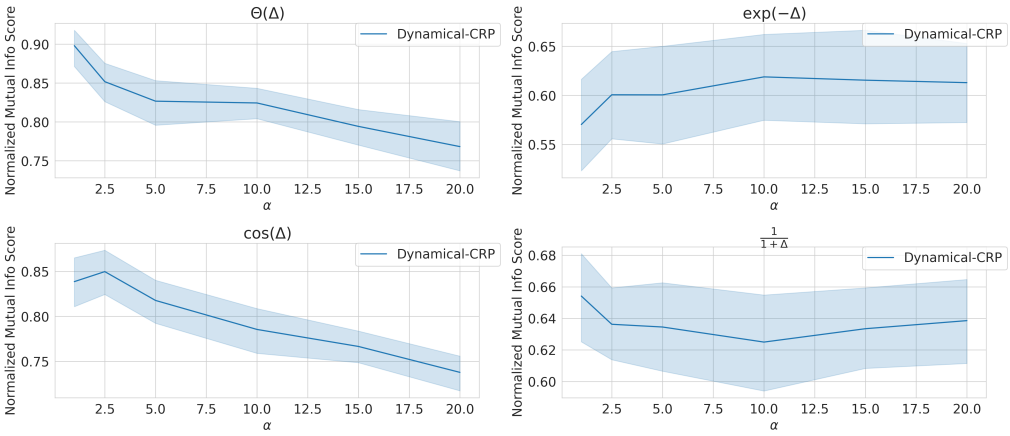

Figure 4: **Normalized mutual information between true cluster assignments and inferred cluster assignments in 4 different dynamics, sweeping α.** Higher α means a harder inference problem.

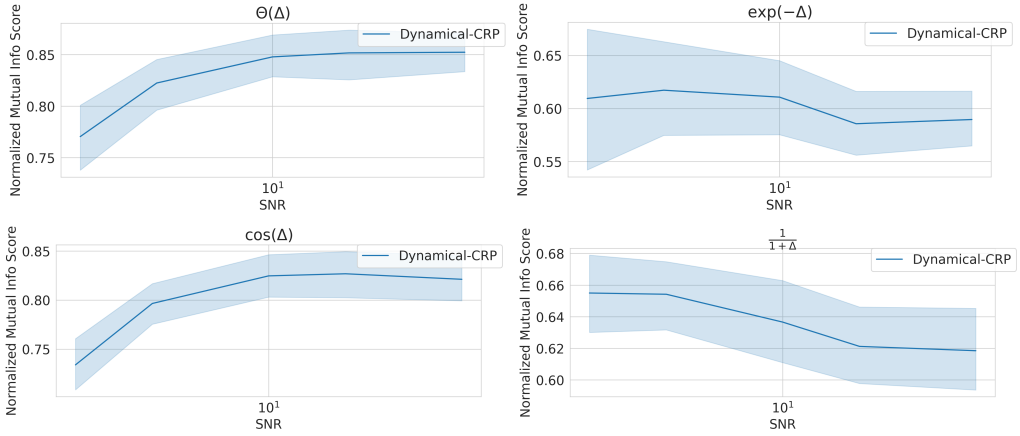

Figure 5: **Normalized mutual information between true cluster assignments and inferred cluster assignments in 4 different dynamics, sweeping signal-to-noise (SNR).** Higher SNR means an easier inference problem.

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 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.

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.

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.

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.

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

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.

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.

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.

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

A.1 CLOSED-FORM EXPRESSION FOR VARIATIONAL PARAMETERS IN THE EXPONENTIAL FAMILY

In the following three subsections, we use the following fact 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 parameters.

A.2 CAVI FOR MULTIVARIATE GAUSSIAN LIKELIHOOD

Mean field family:

$$q(c_n, \{\phi\}|o_{\leq n}) \stackrel{\text{def}}{=} q(c_n|o_{\leq n}; \{\pi_{nc}\}) \prod_{k=1}^{C_n} q(\phi_{nc}|o_{\leq n}; \mu_{nc}, \Sigma_{nc})$$

$$q(c_n|o_{\leq n}; \{\pi_{nc}\}) \stackrel{\text{def}}{=} Categorical(\pi_n)$$

$$q(\phi_{nc}|o_{\leq n}; \mu_{nc}, \Sigma_{nc}) \stackrel{\text{def}}{=} \mathcal{N}(\mu_{nc}, \Sigma_{nc})$$

where $\theta_n \stackrel{\text{def}}{=} \{\pi_{nc}\}_k \cup \{\mu_{nc}\}_k \cup \{\Sigma_{nc}\}_k$ 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}:

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

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:

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

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 RNCRP 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:

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