Efficient Online Inference for Nonparametric Mixture Models

Rylan Schaeffer[1, 3, 4] Blake Bordelon[1] Mikail Khona[2, 3, 4] Weiwei Pan[1] Ila Rani Fiete[3, 4]

[1]School of Engineering and Applied Sciences, Harvard University
[2]Department of Physics, Massachusetts Institute of Technology
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[4]McGovern Institute for Brain Research, Massachusetts Institute of Technology

Abstract

Natural data are often well-described as belonging to latent clusters. When the number of clusters is unknown, Bayesian nonparametric (BNP) models can provide a flexible and powerful technique to model the data. However, algorithms for inference in nonparametric mixture models fail to meet two critical requirements for practical use: (1) that inference can be performed online, and (2) that inference is efficient in the large time/sample limit. In this work, we propose a novel Bayesian recursion to efficiently infer a posterior distribution over discrete latent variables from a sequence of observations in an online manner, assuming a Chinese Restaurant Process prior on the sequence of latent variables. Our recursive filter, which we call the Recursive Chinese Restaurant Process (R-CRP), has quasilinear average time complexity and logarithmic average space complexity in the total number of observations. We experimentally compare our filtering method against both online and offline inference algorithms including Markov chain Monte Carlo, variational approximations and DP-Means, and demonstrate that our inference algorithm achieves comparable or better performance for a fraction of the runtime.

1 INTRODUCTION

Since the introduction of the Hidden Markov Model [Baum and Petrie, 1966], latent-variable models have become a common starting point for modeling temporal data. When the sample space of latent variables is unknown (or growing over time), Bayesian nonparametric models (BNP; Hjort, N. et al. [2010]) provide a probabilistic framework for allowing a model to grow in complexity as more data are observed. One common BNP model, useful for mixture modeling, is the Chinese Restaurant Process (CRP; Blackwell and MacQueen [1973], Aldous [1985]) and its de Finetti mixing distribution, the Dirichlet Process (DP; Ferguson [1973], Antoniak [1974]). However, inference algorithms for these clustering models suffer from two key limitations. First, inference algorithms are formulated for the offline setting, in which the entire sequence of observations is available. Second, the algorithms' computational complexity scales poorly with sequence length. These limitations make practical use of these models in the streaming setting difficult, if not impossible.

To provide a concrete example of the kind of problem we wish to solve, imagine an intrepid field biologist seeking to determine the impact of deforestation on bird diversity in the Amazon. Year after year, she ventures into the jungle to observe birds and sort her observations into species, including potentially previously undiscovered ones. She cannot bag all the birds she observes, and instead has to make on-the-fly determinations of species and counts within each species.

In this paper, we construct an efficient filter that this biologist could use. We place a CRP prior on the sequence of latent states and propose a novel inference algorithm, the Recursive Chinese Restaurant Process (R-CRP), which filters a posterior over latent states. Assuming the posterior behaves asymptotically like the prior, our algorithm has average case time complexity $O(t \log t)$ and average case space complexity $O(\log t)$, where t is the number of observations. R-CRP is a Bayesian recursion that constructs the prior for the current latent state (henceforth referred to as the latent prior) as a running sum of the posteriors for previous latent states. We then show that our online algorithm is competitive with commonly used offline and online inference algorithms. Our code is publicly available at `github.com/RylanSchaeffer/FieteLab-RCRP`.

Related Work: Many BNP latent variable models for time series exist, including infinite Hidden Markov Models (Beal et al. [2002]) and dependent Dirichlet Processes (MacEachern, Steven [1999], Lin et al. [2010]). Inference algorithms

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

fall into several categories, including (sequential) Monte Carlo (Neal [2000], Fearnhead [2004], Ulker et al. [2010]), variational inference (Blei and Jordan [2006], Zhang and Paisley [2016]), maximum a-posteriori estimates (Anderson [1991], Broderick et al. [2013]) and low variance asymptotic approximations (Kulis and Jordan [2012], Campbell et al. [2013]). Many of these inference algorithms make multiple passes through the entire sequence of observations (e.g. Zhang and Paisley [2016]), while others re-infer posteriors over the entire latent sequence with each new observation (e.g. Bartunov and Vetrov [2014]) or collapse probabilistic quantities into point estimates (e.g. Kulis and Jordan [2012]).

A smaller body of work specifically addresses online/streaming inference in nonparametric mixture models. Fearnhead [2004] introduced a sequential Monte Carlo algorithm with a heuristic to keep the number of particles manageable. Wang and Dunson [2011] proposed a maximum a-posteriori algorithm SUGS that greedily chooses the best possible cluster assignment for each observation, which was extended by Zhang et al. [2014] to the variational Bayes setting. Lin [2013] similarly introduces an online variational inference algorithm. Although not discussed in the original paper, Kulis and Jordan [2012]'s DP-Means can readily be extended to the streaming setting in the same manner that K-Means can be extended to the streaming setting. Most relevant to our work is Liu et al. [2014]'s Online Chinese Restaurant Process. Our work differs from theirs in three key ways. First, our algorithm does not require the sequence of latent variables to be exchangeable, allowing us to lay the groundwork for subsequent filters that are performant even if latent variables are non-exchangeable. Second, our algorithm does not rely on the correct values of previous latent variables being made available at later times, to allow retroactive corrections in a supervised fashion. Third, our algorithm is a Bayesian recursion that maintains full distributions at all times, whereas their algorithm relies on sampling. Our recursion R-CRP is also similar to Newton [1999, 2002]'s Predictive Recursion, but differs in that their algorithm is not designed for the online setting and their algorithm averages over permutations of the observations.

2 BACKGROUND

2.1 GENERATIVE PROCESS

We consider a latent-variable time series model with discrete latent variables $z_{1:T}$ and observable variables $o_{1:T}$, where $\cdot_{1:T}$ denotes the sequence $(\cdot_1, \cdot_2, ..., \cdot_T)$. Our generative process assumes a Chinese Restaurant Process (CRP) prior over the sequence of latent states:

$$\begin{aligned} z_{1:T} &\sim CRP(\alpha) \\ o_t | z_t &\sim p(o|z) \end{aligned} \tag{1}$$

2.2 CHINESE RESTAURANT PROCESS

The Chinese Restaurant Process (CRP; Blackwell and MacQueen [1973], Aldous [1985]) is a one-parameter (concentration parameter $\alpha > 0$) stochastic process that defines a discrete distribution over the partitions of a set. The CRP defines a conditional distribution for the tth categorical variable z_t given the preceding variables:

$$p(z_t = k | z_{<t}, \alpha) = \begin{cases} \frac{\sum_{t'=1}^{t-1} \mathbb{1}(z_{t'}=k)}{\alpha+t-1} & \text{if } 1 \leq k \leq K_{t-1} \\ \frac{\alpha}{\alpha+t-1} & \text{if } k = K_{t-1}+1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where K_t is the integer number of categories (clusters) with at least one variable in $\{z_{t'}\}_{t'=1}^{t'=1}$ belonging to that category (cluster). The term CRP arises from an analogy of seating a sequence of customers at a Chinese restaurant that has an infinite number of tables, each with an infinite number of chairs. Each customer is randomly placed either at a populated table with probability proportional to the number of previous customers at that table, or at a new, un-populated table with probability proportional to α. K_t denotes the number of non-empty tables after the t-th customer is seated. The CRP can be equivalently defined using indicator variables, a fact we later exploit:

$$\begin{aligned} p(z_t = k | z_{<t}, \alpha) = \frac{1}{\alpha + t - 1} \sum_{t' < t} \mathbb{1}(z_{t'} = k \leq K_{t-1}) \\ + \frac{\alpha}{\alpha + t - 1} \mathbb{1}(k = K_{t-1} + 1) \end{aligned} \tag{3}$$

2.3 CHINESE RESTAURANT TABLE DISTRIBUTION

The (random) number of non-empty tables after t customers have been seated K_t is described by the Chinese Restaurant Table (CRT) distribution:

$$p(K_t = k) = \frac{\Gamma(\alpha)}{\Gamma(t+\alpha)} |s(t,k)| \alpha^k \mathbb{1}(k \leq t) \tag{4}$$

where $|s(t,k)|$ are unsigned Stirling numbers of the first kind. The CRT can equivalently be defined as a sum of independent but non-identically distributed Bernoulli random variables indicating the tth customer was placed at a new table.

$$K_t = \sum_{t'=1}^{t} b_{t'} \quad \text{where} \quad b_{t'} \sim Bernoulli\Big(\frac{\alpha}{\alpha + t' - 1}\Big)$$

Per Le Cam's Theorem (Le Cam [1960]), K_t is well approximated by a Poisson distribution with rate $\lambda = \alpha \log(1 + t/\alpha)$, showing that the average number of tables grows logarithmically with t. This detail becomes important in our complexity analysis.

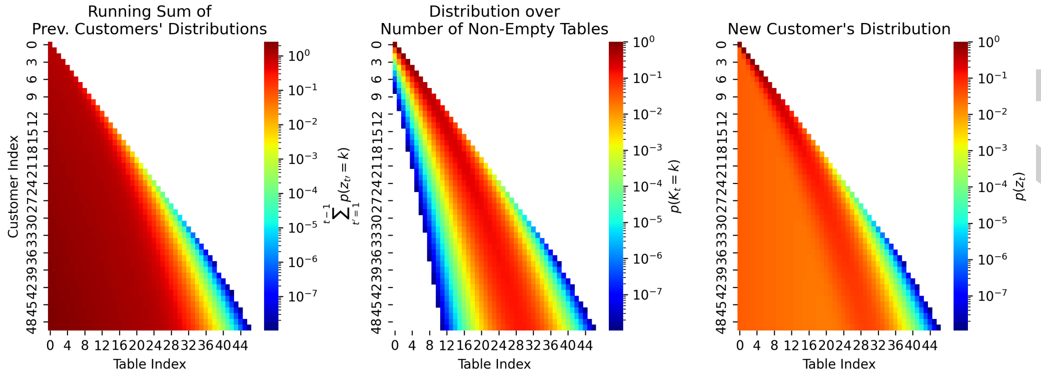


Figure 1: **Visualization of Recursion.** To make streaming inference possible, we break the CRP's conditional distribution's $p(z_t|z_{<t}, \alpha)$ dependence on the entire history $z_{<t}$ by replacing it with a marginal distribution $p(z_t|\alpha)$. The running sum of the previous marginal distributions $\sum_{t'<t} p(z_{t'} = k)$ (left) and the Chinese restaurant table distribution $p(K_{t-1} = k)$ (middle) together determine the next marginal distribution $p(z_t = k)$ (right). Note the logarithmic scaling. Here, $\alpha = 30.91$.

3 RECURSION FOR ONLINE FILTERING

3.1 OBJECTIVE

Our goal is to infer a posterior over the discrete latent state $p(z_t|o_{\leq t})$, subject to two constraints:

1. Inference must be performed online, meaning the filter cannot make use of the (possibly) infinite past nor can the filter be used to revise the past.
2. Inference must be efficient in the large t limit

Our biologist seeking to cluster birds into species shows why both constraints are important. She can't capture or carry every bird she observes, nor can she remember or revise every latent posterior over each bird's most likely species, but she must nonetheless be able to form a well-founded belief as to the most recently seen bird's species for as long as her research continues. This problem, of inferring a posterior over the current latent variable given past and current observations, is often referred to as filtering (e.g., Kalman filter, particle filter).

3.2 BAYESIAN RECURSION

We derive a streaming inference algorithm, R-CRP, which recursively computes the desired posterior on the most recent latent variables z_t. The CRP's conditional distribution complicates inference because one latent depends on all preceding latents; our approach is to break that dependence by converting the CRP's conditional distribution into a marginal distribution, which can be expressed as a running sum of previously computed quantities. For brevity, we refer to the prior on the current latent variable $p(z_t|o_{<t})$ as the "latent prior" and the posterior on the current variable $p(z_t|o_{\leq t})$ as the "latent posterior". Bayes' rule relates the latent prior to the latent posterior:

$$\underbrace{p(z_t = k|o_{\leq t})}_{\text{Latent Posterior}} = \frac{p(o_t|z_t = k)}{p(o_t|o_{<t})} \underbrace{p(z_t = k|o_{<t})}_{\text{Latent Prior}}. \qquad (5)$$

The latent prior $p(z_t = k|o_{<t})$ can be rewritten as the expectation of an indicator random variable that we can expand using the Law of Total expectation.

$$\begin{aligned}
\underbrace{p(z_t = k|o_{<t})}_{\text{Latent Prior}} &= \mathbb{E}_{p(z_t|o_{<t})}[\mathbb{1}(z_t = k)] \\
&= \mathbb{E}_{p(z_{<t}, K_{t-1}|o_{<t})}\Big[\mathbb{E}_{p(z_t|z_{<t}, K_{t-1}, o_{<t})}[\mathbb{1}(z_t = k)]\Big] \\
&= \mathbb{E}_{p(z_{<t}, K_{t-1}|o_{<t})}\Big[p(z_t = k|z_{<t}, K_{t-1}, o_{<t})\Big]
\end{aligned}$$

where K_t denotes the random number of non-empty tables after the t-th customer has been seated. The distribution inside the expectation is the conditional distribution specified by the CRP. Substituting Eqn. (3) and taking the expectation yields

$$\begin{aligned}
p(z_t = k|o_{<t}) &= \frac{1}{\alpha + t - 1} \sum_{t'<t} p(z_{t'} = k \leq K_{t-1}|o_{<t}) \\
&\quad + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1|o_{<t}) \\
&= \frac{1}{\alpha + t - 1} \sum_{t'<t} p(z_{t'} = k|o_{<t}) \\
&\quad + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1|o_{<t})
\end{aligned}$$

where the second equality is possible because $k \leq K_{t-1}$ is unnecessary; a previous customer cannot be assigned to a table that does not later exist. Specifically:

$$p(z_{t'} = k|o_{<t}) = p(z_{t'} = k \leq K_{t-1}|o_{<t}) + \underbrace{p(z_{t'} = k > K_{t-1}|o_{<t})}_{=0}$$

For a pure CRP, this recursion is exact. However, in the online setting, we must introduce one approximation: previous posteriors (on $z_{t'<t}$) cannot be retroactively revised based on data that arrives later $o_{t>t'}$, since this would require require remembering all previous latents. With this approximation, we reach our final recursion for the latent prior:

$$\begin{aligned}\underbrace{p(z_t = k|o_{<t})}_{\text{Latent Prior}} \approx & \frac{1}{\alpha+t-1}\sum_{t'<t} p(z_{t'} = k|o_{\leq t'}) \\ & + \frac{\alpha}{\alpha+t-1} p(K_{t-1} = k-1|o_{<t})\end{aligned} \tag{6}$$

Intuitively, Eqn. (6) says that the prior probability that the t-th latent variable belongs to the kth cluster is the sum of all preceding latents' posteriors' masses of belonging to the kth cluster, plus a new term. The new term depends on K_{t-1}, the number of non-empty tables, and the concentration parameter α; it pushes the current latent toward a new cluster. Fig. 1 visually displays how the accumulating mass of from previous customers and the pressure to create a new cluster compete to determine where the new customer will sit.

Substituting Eqn. (6) into Eqn. (5), our recursion is:

$$\begin{aligned}\underbrace{p(z_t = k|o_{\leq t})}_{\text{Latent Posterior}} \approx \frac{p(o_t|z_t = k)}{p(o_t|o_{<t})} \Bigg[& \frac{1}{\alpha+t-1}\sum_{t'<t} \underbrace{p(z_{t'} = k|o_{\leq t'})}_{\text{Previous Posteriors}} \\ & + \frac{\alpha}{\alpha+t-1} p(K_{t-1} = k-1|o_{<t})\Bigg]\end{aligned} \tag{7}$$

The sum over posteriors $p(z_{t'} = k|o_{\leq t'})$ can be computed by a simple iteration, $R_t(k) = R_{t-1}(k) + p(z_t = k|o_{\leq t})$. The recursive application of Eqn. (7) permits online inference of the latent states in that we can compute the posterior without re-accessing (and thus without storing) past observations.

We can also recursively compute the posterior distribution over the number of non-empty tables (which does not have a closed-form analytical expression):

$$\begin{aligned} p(K_t|o_{\leq t}) &= \sum_{K_{t-1}, z_t} p(K_t, K_{t-1}, z_t|o_{\leq t}) \\ &= \sum_{K_{t-1}, z_t} p(K_t|K_{t-1}, z_t) p(K_{t-1}|o_{\leq t}) p(z_t|o_{\leq t}) \\ &\approx \sum_{K_{t-1}, z_t} p(K_t|K_{t-1}, z_t) p(K_{t-1}|o_{\leq t-1}) p(z_t|o_{\leq t}) \end{aligned} \tag{8}$$

where the conditional $p(K_t|K_{t-1}, z_t)$ is a transition-like function such that K_{t-1} is incremented if $z_t > K_{t-1} + 1$ and kept constant otherwise:

$$p(K_t = K_{t-1} + 1|K_{t-1}, z_t) = \begin{cases} 1 & \text{if } K_{t-1} < z_t \\ 0 & \text{otherwise} \end{cases}$$

$$p(K_t = K_{t-1}|K_{t-1}, z_t) = \begin{cases} 1 & \text{if } K_{t-1} \geq z_t \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

This recursion requires storing only a running sum of latent states' filtered posteriors and a posterior over the number of tables. It does not require exchangability of the sequence. It does not collapse probability distributions via sampling (Liu et al. [2014]) or point estimators (Wang and Dunson [2011], Gomes et al. [2008]), nor require storing every previous latent variable's posterior (Zhang et al. [2014], Lin [2013]).

3.3 COMPLEXITY ANALYSIS

The total time complexity of the R-CRP recursion is determined by the number of latent states K_t, which is upper bounded by t. The recursion for the posterior over the number of non-empty tables has worst-case time complexity $O(K_t)$ per step, and the recursion for the posterior over the current latent state has worst-case time complexity $O(K_t)$ per step. The worst-case space complexity is $O(K_t)$. Consequently, the total worst-case complexity is $O(t^2)$ time and $O(t)$ space. However, if we assume that the posterior behaves asymptotically like the prior in which K_t grows logarithmically with t, then the average-case complexity is quasilinear time $O(t \log t)$ and logarithmic space $O(\log t)$.

4 EXPERIMENTS

4.1 CORRECTNESS FOR CRP PRIOR

According to the derivation, the recursion should hold exactly for the CRP prior in the absence of observations. We test that the recursion correctly computes the marginal distributions by comparing the recursion's analytical expression to 5000 Monte Carlo samples, each of 50 customers, drawn from $CRP(\alpha)$ for $\alpha \in \{1.1, 10.78, 15.37, 30.91\}$. First, we visually compared the analytical expressions versus the Monte Carlo estimates (one example $\alpha = 10.78$ shown in Fig 2) and found excellent agreement.

Second, we computed the mean squared error between the

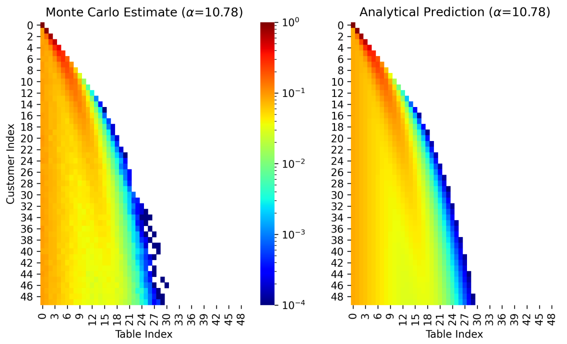


Figure 2: Comparison of $CRP(\alpha = 10.78)$ marginal probabilities $p(z_t|\alpha)$ for Monte Carlo estimates (left) and R-CRP analytical expression (right). Note the logarithmic scaling.

analytical R-CRP expression and Monte Carlo estimation as a function of the number of Monte Carlo samples. For all four α values, the squared error falls approximately as a power law with the number of samples (Fig 3), attesting to the accuracy of R-CRP.

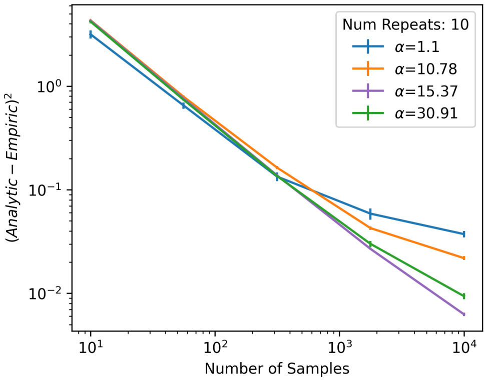


Figure 3: The mean-squared error between the recursively computed marginal distributions $p(z_t = k)$ (R-CRP) and Monte Carlo estimates falls as an approximate power law with the number of Monte Carlo samples.

Third, we tested the recursion by how well it matched the expected table occupancies. The expected number of customers at the kth table after t customers have been seated can be written as the sum of the customer seating marginal probabilities by linearity of expectation: $\mathbb{E}[N_{T,k}] = \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(z_t = k)] = \sum_{t=1}^{T} p(z_t = k)$. For $\alpha \in \{1.1, 10.01, 15.51, 30.03\}$, we found excellent matches for all values of α (Fig 4).

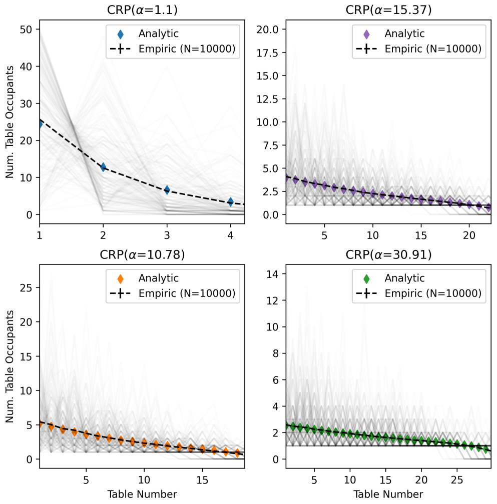


Figure 4: Monte Carlo estimates of table sizes for the CRP prior (black) closely match analytical expressions of table sizes (colored) for varying α values.

4.2 GAUSSIAN MIXTURE MODEL

We returned to performing inference and tested R-CRP on the synthetic mixture of Gaussians from Kulis and Jordan [2012]: Data are generated by drawing a sequence of cluster assignments $z_1, ..., z_T \sim CRP(\alpha)$, and then drawing a sequence of observations from the corresponding Gaussian: $x_t | z_t, \{\mu_k\}_{k=1}^{3} \sim \mathcal{N}(\mu_{z_t}, \Sigma_{z_t})$. The cluster means are drawn as $\mu_k \sim \mathcal{N}(0, \rho I)$ and the clusters have identical isotropic covariances $\Sigma_k = I$, (Fig. 5d). We consider two families of baseline algorithms: offline and online. Offline baselines have access to the entire dataset and can make multiple passes through it; we considered three such baselines:

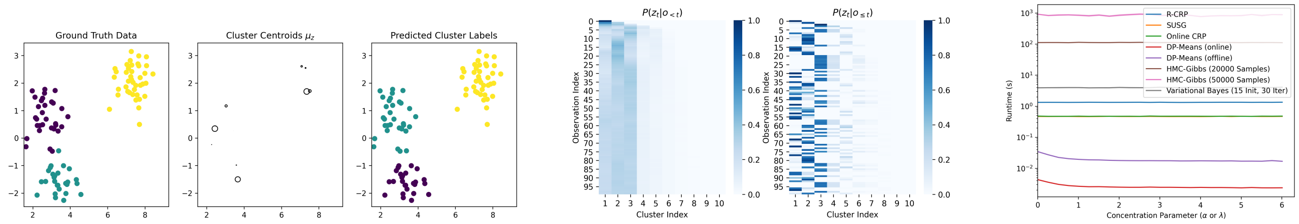
1. Variational Bayes from Blei and Jordan [2006], implemented in Scikit-Learn (Pedregosa et al. [2011]).
2. Hamiltonian-Gibbs Monte-Carlo Sampling, implemented in Pyro (Bingham et al. [2019]).
3. DP-Means, a low-variance asymptotic approximation from Kulis and Jordan [2012]. We call this DP-Means (offline).

The online baselines are constrained identically to our R-CRP, in that inference must be performed online. We considered three online baselines:

1. DP-Means, but limited to a single forward pass through the data, identical to how K-Means is performed in the streaming setting. We call this DP-Means (online).
2. Sequential Updating and Greedy Search (SUGS) from Wang and Dunson [2011], which uses a "local MAP approximation" i.e. $\hat{z}_t = \arg\max_k p(z_t = k | \hat{z}_{<t}, \alpha)$.
3. Online CRP from Liu et al. [2014], which uses sampling i.e. $\hat{z}_t \sim p(z_t = k | \hat{z}_{<t}, \alpha)$.

R-CRP performs online inference and parameter estimation by first inferring a latent posterior over which cluster the current observation belongs to i.e. $p(z_t | o_{\leq t})$, and then updates each cluster's parameters modulated by the posterior probability the observation belongs to that cluster. Table 1 provides pseudocode. As stated in the pseudocode, we entertain the notion that each observation could add a new cluster, which is a maximally conservative step. Others have explored fixed thresholds for creating new clusters (e.g. Lin [2013]), but we intentionally exclude such heuristics to maintain focus on our proposed recursion. Empirically, the probability mass placed on new clusters after the first few clusters have been established becomes vanishingly small, which suggests to us that thresholds are probably fine.

Despite being a filtering inference algorithm, R-CRP recovers highly plausible cluster centroids and makes accurate cluster predictions (Fig 5a). We also visualize the time evolution of cluster assignment priors, $p(z_t | o_{<t})$ and cluster assignment posteriors $p(z_t | o_{\leq t})$ (Fig 5b). The number of



(a)

(b)

(c)

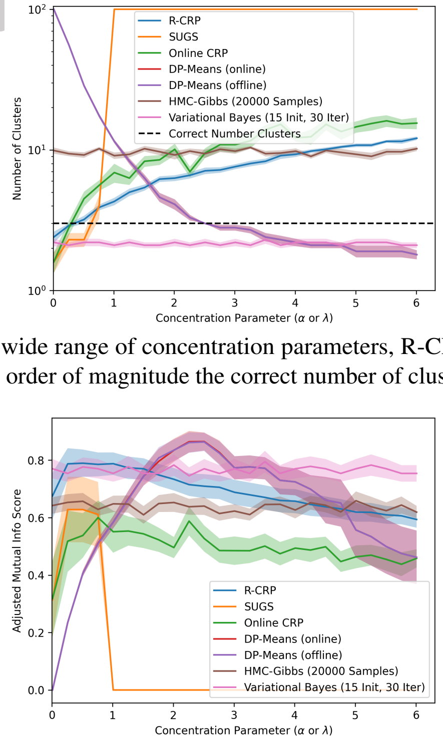
Figure 5: (a) One random sample from a mixture of 3 Gaussians (left). Centroids learned by R-CRP (center). Predicted class labels by R-CRP (up to an arbitrary aliasing) (right). (b) Evolution of cluster assignment priors $p(z_t|o_{<t})$ (left) and cluster assignment posteriors $p(z_t|o_{\leq t})$ (right) as more data are observed from the Mixture of Gaussians. (c) R-CRP is slower than most online baselines but signficantly faster than offline baselines.

non-empty tables rapidly increases, but after several observations, the model learns that some previously created clusters (i.e., clusters 4, 5, 6) were likely not genuine clusters and starts allocating more mass to earlier clusters. Although past mistakes are not corrected, their proportional influence decreases as more data arrives. Importantly, when we sweep the range of possible concentration parameters (λ for DP-Means, α for the rest), we find that over a wide range, R-CRP finds a more reasonable number of clusters (Fig. 6a) and has higher adjusted mutual information (Vinh et al. [2010]) between its predictions and the true class labels than the baselines (Fig 6b). R-CRP offers both high performance while being significantly faster than most offline baselines (Fig 5c). Despite the model's capacity to use as many clusters as observations, the likelihoods constraint the model to a reasonable number of clusters.

```
1. Initialize $p(K_0 = 0) = 1$, and $\forall k \in \mathbb{Z}$, set running sum of posteriors $R_0(k) = 0$
2. For each $t = 1, ..., T$:
    • Create new cluster mean $\mu_{k=t}^{(t)} = o_t$
    • Compute prior for new observation:
        $$p(z_t = k|o_{<t}) \approx \frac{1}{\alpha + t - 1} R_{t-1}(k) + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1|o_{<t})$$
    • Compute posterior for new observation:
        $$p(z_t = k|o_{\leq t}) \propto \mathcal{N}(o_t; \mu_k, \Sigma) p(z_t = k|o_{<t})$$
    • Update running sum of posteriors:
        $$R_t(k) = R_{t-1}(k) + p(z_t = k|o_{\leq t})$$
    • Update existing cluster means:
        $$\mu_k^{(t)} \leftarrow \mu_k^{(t-1)} + \frac{p(z_t = k|o_{\leq t})}{R_t(k)} (o_t - \mu_k^{(t-1)})$$
    • Compute new posterior $p(K_t|o_{\leq t})$ on number of tables using Eqn. (8)
```

Table 1: Pseudo-code for performing inference with R-CRP(α) in a Gaussian Mixture Model.

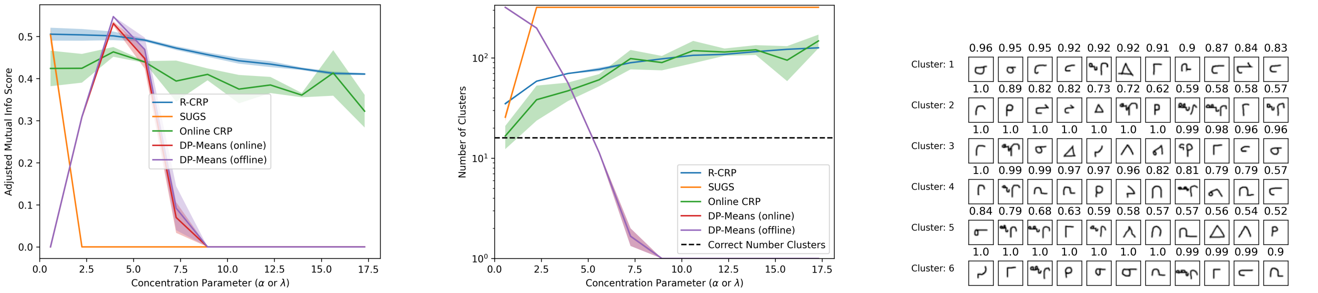
Figure 6



(a) For a wide range of concentration parameters, R-CRP learns within an order of magnitude the correct number of clusters.

(b) R-CRP displays high adjusted mutual information (Vinh et al. [2010]) with the correct class labels, comparable to offline baselines and better than online baselines[a].

[a] SUGS is not incorrectly implemented; its pathological performance has been noted before e.g. in Lin [2013]



(a) (b) (c)

Figure 7: Omniglot (Lake et al. [2015]) Handwritten Character Recognition. (a) R-CRP displays higher adjusted mutual information than other online baselines. (b) For a wide range of concentration parameters, R-CRP learns within an order of magnitude the correct number of clusters. (c) R-CRP recovers plausible clusters of Omniglot characters.

4.3 HANDWRITTEN CHARACTER RECOGNITION

Moving beyond simulated data, we turned to the problem of clustering handwritten characters. Humans are adept at determining which character a never-before-seen set of strokes represents, or whether the set of strokes represents a new character altogether. We used the Omniglot dataset (Lake et al. [2015]), which consists of 1623 different handwritten characters from 50 different alphabets, each drawn 20 times by different humans. An Omniglot datum is a greyscale 105 pixel by 105 pixel image. To be able to model the images with a Gaussian likelihood, we trained a variational autoencoder (Rezende et al. [2014], Kingma and Welling [2014]) using an open source VAE library[1] and took the Gaussian means at the bottleneck as representations of each image. We then tested how well R-CRP clusters this dataset compared against the previous Gaussian-likelihood online baselines (SUGS, Online CRP, DP-Means (online)).

We found R-CRP outperforms all online baselines per adjusted mutual information for all concentration parameters (except SUGS at $\alpha = 0.01$) (Fig 7a), and recovers within an order of magnitude the correct number of clusters for all concentration parameters (Fig 7b). R-CRP recovers plausible clusters of handwritten characters (Fig 7c), which admittedly aren't excellent, but are significantly better than the clusters produced by other baselines.

4.4 CATEGORICAL MIXTURE MODEL

To show that R-CRP works for non-Gaussian likelihoods, we next studied the performance of R-CRP on a Categorical mixture model, sometimes called a mixture-of-unigrams (Nigam et al. [2000]). A mixture-of-unigrams can be used to describe a corpus of documents, in which each document belongs to exactly one of a discrete number of topics. For instance, each Wikipedia page might concern history, science, pop culture or something else; as users create new Wikipedia pages, we would like to cluster these pages into the appropriate topics without needing to refit to the entire Wikipedia corpus.

Data are generated by drawing a sequence of cluster assignments $z_1, ..., z_T$, and then by drawing a probability vector $p_k \sim Dir(\beta \mathbf{1}) \in \Delta^{V-1}$ where V is the vocabulary size and Δ^{V-1} is the $V-1$ simplex. Each observation conditioned on the cluster assignment follows a Multinomial distribution $o_t \sim Multinomial(M, p_{z_t})$ where M is the document length; each document can be thought of as word counts for each word in the vocabulary.

To perform inference and parameter estimation, R-CRP uses a Dirichlet-Multinomial likelihood. Specifically, each cluster has a pseudocounts parameter $n_k \in \mathbb{R}_+^V$ that parameterizes a Dirichlet distribution over the cluster's probability vector $\hat{p}_k \sim Dir(n_k)$. The likelihood is then:

$$p(o_t | z_t = k; n_k^{(t-1)}) = \int_{\Delta^{V-1}} p(o_t | \hat{p}_k) p(\hat{p}_k | n_k^{(t-1)}) d\hat{p}_k$$

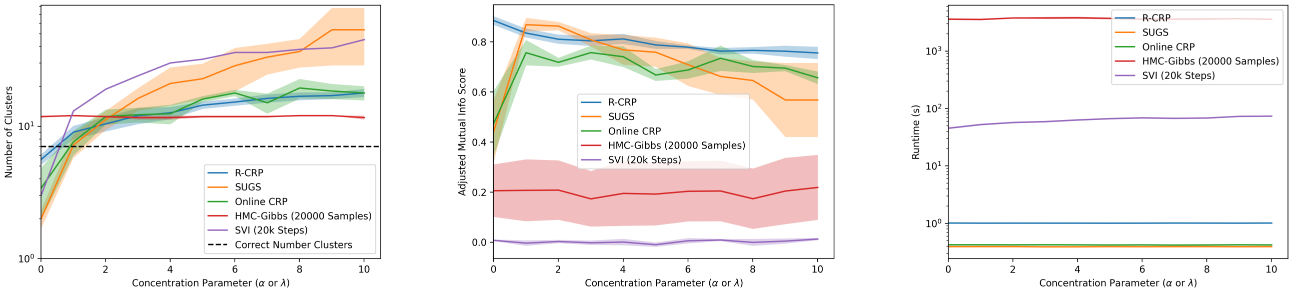
$$p(o_t | z_t = k; n_k^{(t-1)}) = \int_{\Delta^{V-1}} \prod_{w=1}^{W_t} p(o_{t,w} | \hat{p}_k) p(\hat{p}_k | n_k^{(t-1)}) d\hat{p}_k$$

To update each cluster's pseudocounts, we increment the pseudocounts by the newest observation weighed by the posterior probability the observation belongs to the cluster:

$$n_k^{(t)} \leftarrow n_k^{(t-1)} + p(z_t = k | o_{\leq t}) o_t$$

Pseudocode is provided in Table 2. We again compare R-CRP against offline and online baselines, with two changes: (1) we drop DP-Means because the algorithm is defined for a Gaussian likelihood, (2) we replace Variational Bayes (VB) with Stochastic Variational Inference (SVI; Hoffman et al. [2013]) because Scikit-Learn does not have an implementation of VB and Pyro offers better support for SVI.

[1] https://github.com/jmtomczak/vae_vampprior



(a)

(b)

(c)

Figure 8: For a wide range of concentration parameters, R-CRP (a) recovers close to the correct number of topics and (b) has outstanding adjusted mutual information with the true topic labels. (c) R-CRP is slightly slower than online baselines but significantly faster than offline baselines.

We find that over a wide range of concentration parameters, R-CRP recovers close to the correct number of topics (8a). We also find that R-CRP has higher or comparable adjusted mutual information (Vinh et al. [2010]) over a wide range of concentration parameters than many of the baselines, including HMC-Gibbs and SVI [2].

5 DISCUSSION

In this paper, we presented an efficient streaming inference algorithm, the Recursive Chinese Restaurant Process (R-CRP), for nonparametric mixture models. The time and space complexities are controlled by the number of unique values the discrete latent variables can take, which is a property of the task; assuming the posterior behaves asymptotically like the prior, R-CRP has quasilinear expected time complexity $O(t \log t)$ and logarithhmic expected space complexity $O(\log(t))$. In addition to functioning online and being efficient in the large t limit, R-CRP maintains full posteriors over the latent variables without relying on exchangeability of the latent variables. Experimentally, we showed that if the latent variables are directly observable, R-CRP exactly describes the marginal distribution of the most recent variable $p(z_t|\alpha)$. We then showed that when the latent variables are indeed latent, under one approximation, R-CRP can recover the latent structure as well as or better than commonly used inference algorithms that require simultaneous access to the entire dataset and make multiple passes through the dataset.

Future Directions: This paper lays the groundwork for a variety of extensions in different directions. One possible direction is encouraging others to pursue practical approximate inference algorithms to make Bayesian nonparametrics, a powerful tool, more usable in practice. Another possible direction is studying criteria for when to create new clusters or merge existing clusters.

Author Contributions

RS conceived the idea, sharpened it through discussions with other coauthors, implemented all experiments and co-wrote the paper. BB worked on the mixture of unigrams

```
1. Initialize $p(K_0 = 0) = 1$, and $\forall k \in \mathbb{Z}$, set running sum of posteriors $R_0(k) = 0$
2. For each $t = 1, ..., T$:
    • Create new cluster pseudocounts $n_{k=t}^{(t)} = o_t + \varepsilon$, where $\varepsilon$ is a small fixed positive quantity
    • Compute prior for new observation:
        $$p(z_t = k|o_{<t}) \approx \frac{1}{\alpha+t-1} R_{t-1}(k) + \frac{\alpha}{\alpha+t-1} p(K_{t-1} = k-1|o_{<t})$$
    • Compute posterior for new observation:
        $$p(z_t = k|o_{\leq t}) \propto Multi(o_t; p_k) p(z_t = k|o_{<t})$$
    • Update running sum of posteriors:
        $$R_t(k) = R_{t-1}(k) + p(z_t = k|o_{\leq t})$$
    • Update existing cluster pseudocounts:
        $$n_k^{(t)} \leftarrow n_k^{(t-1)} + p(z_t = k|o_{\leq t}) o_t$$
    • Compute new posterior $p(K_t|o_{\leq t})$ on number of tables using Eqn. (8)
```

Table 2: Pseudo-code for performing inference with R-CRP(α) in a Categorical Mixture Model.

[2] Their poor performance suggests an implementation error, but the Pyro developers told us that inferring a large number of discrete latent variables with both algorithms performs poorly this Pyro post and this Pyro post.

and topic modeling experiments, and co-wrote the corresponding sections. MK worked on the mixture of unigrams and helped with figures. WP helped RS in early discussions and edited the paper. IRF provided funding and compute, helped select experiments, advised the team and co-wrote the paper.

REFERENCES

David J. Aldous. Exchangeability and related topics. In David J. Aldous, Illdar A. Ibragimov, Jean Jacod, and P. L. Hennequin, editors, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, Lecture Notes in Mathematics, pages 1–198, Berlin, Heidelberg, 1985. Springer. ISBN 978-3-540-39316-0. doi: 10.1007/BFb0099421.

John Anderson. The adaptive nature of human categorization. *Psychological Review*, 1991. URL http://act-r.psy.cmu.edu/wordpress/wp-content/uploads/2012/12/89AdaptiveNature.pdf.

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.

Sergey Bartunov and Dmitry P Vetrov. Variational Inference for Sequential Distance Dependent Chinese Restaurant Process. *International Conference on Machine Learning*, page 9, 2014.

Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Annals of Mathematical Statistics*, 37(6):1554–1563, December 1966. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177699147. URL https://projecteuclid.org/euclid.aoms/1177699147. Publisher: Institute of Mathematical Statistics.

Matthew J Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. The Infinite Hidden Markov Model. *Advances in Neural Information Processing Systems*, page 8, 2002.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 20:6, 2019.

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

Trevor Campbell, Miao Liu, Brian Kulis, Jonathan P. How, and Lawrence Carin. Dynamic Clustering via Asymptotics of the Dependent Dirichlet Process Mixture. *arXiv:1305.6659 [cs, stat]*, November 2013. URL http://arxiv.org/abs/1305.6659. arXiv: 1305.6659.

Paul Fearnhead. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14(1):11–21, January 2004. ISSN 1573-1375. doi: 10.1023/B:STCO.0000009418.04621.cd. URL https://doi.org/10.1023/B:STCO.0000009418.04621.cd.

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.

Ryan Gomes, Max Welling, and Pietro Perona. Incremental learning of nonparametric Bayesian mixture models. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, AK, USA, June 2008. IEEE. ISBN 978-1-4244-2242-5. doi: 10.1109/CVPR.2008.4587370. URL http://ieeexplore.ieee.org/document/4587370/.

Hjort, N., C Holmes, P Mueller, and Walker, S. *Bayesian Non-parametrics: Principles and Practice*. Cambridge University Press, 2010.

Matthew D Hoffman, David Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, page 45, 2013.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *International Conference on Learning Representations*, 2014. URL https://dare.uva.nl/search?identifier=cf65ba0f-d88f-4a49-8ebd-3a7fce86edd7. Publisher: Ithaca, NYarXiv.org.

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

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, December 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aab3050. URL https://science.sciencemag.org/content/350/6266/1332. Publisher: American Association for the Advancement of Science Section: Research Article.

Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960. ISSN 0030-8730. URL https://projecteuclid.org/euclid.pjm/1103038058. Publisher: Pacific Journal of Mathematics.

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

Dahua Lin, Eric Grimson, and John W Fisher. Construction of Dependent Dirichlet Processes based on Poisson Processes. *Advances in Neural Information Processing Systems*, page 9, 2010.

Chien-Liang Liu, Tsung-Hsun Tsai, and Chia-Hoang Lee. Online chinese restaurant process. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 591–600, New York New York USA, August 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623636. URL https://dl.acm.org/doi/10.1145/2623330.2623636.

MacEachern, Steven. Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999.

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

M. Newton. A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26, March 1999. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/86.1.15. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/86.1.15.

Michael A. Newton. On a Nonparametric Recursive Estimator of the Mixing Distribution. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 64(2):306–322, 2002. ISSN 0581-572X. URL https://www.jstor.org/stable/25051398. Publisher: Springer.

Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103–134, May 2000. ISSN 1573-0565. doi: 10.1023/A:1007692713085. URL https://doi.org/10.1023/A:1007692713085.

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.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, June 2014. URL http://proceedings.mlr.press/v32/rezende14.html. ISSN: 1938-7228.

Yener Ulker, Bilge Gunsel, and Taylan Cemgil. Sequential Monte Carlo Samplers for Dirichlet Process Mixtures. International Conference on Artificial Intelligence and Statistics:8, 2010.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, page 18, 2010.

Lianming Wang and David B. Dunson. Fast Bayesian Inference in Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, January 2011. ISSN 1061-8600, 1537-2715. doi: 10.1198/jcgs.2010.07081. URL http://www.tandfonline.com/doi/abs/10.1198/jcgs.2010.07081.

Aonan Zhang and San Gultekin John Paisley. Stochastic Variational Inference for the HDP-HMM. *International Conference on Machine Learning*, page 9, 2016.

Xiaole Zhang, David J. Nott, Christopher Yau, and Ajay Jasra. A Sequential Algorithm for Fast Fitting of Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 23(4):1143–1162, October 2014. ISSN 1061-8600. doi: 10.1080/10618600.2013.870906. URL https://doi.org/10.1080/10618600.2013.870906. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2013.870906.