Efficient Online Inference for Nonparametric Mixture Models

Rylan Schaeffer[1,2], Blake Bordelon[2], Mikail Khona[1], Weiwei Pan[2] and Ila Fiete[1]

[1]Department of Brain and Cognitive Sciences, MIT [2]SEAS, Harvard University contact: fiete@mit.edu

INTRODUCTION

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 $(\cdot_1, \cdot_2, ..., \cdot_T)$.

We draw cluster assignments from a Chinese restaurant prior and assume each cluster has a emission distribution:

$$z_{1:T} \sim CRP(\alpha)$$

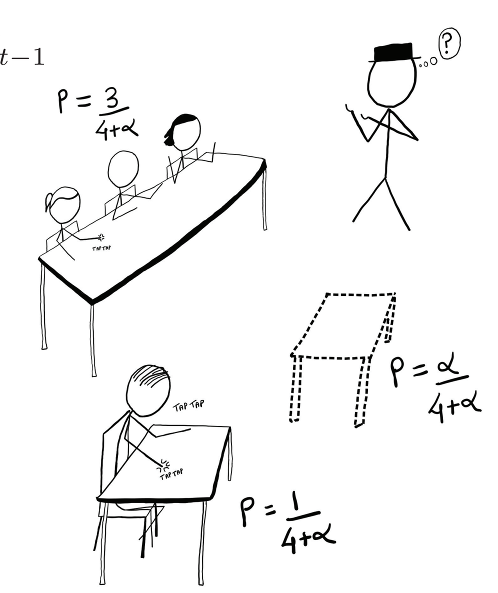
$$o_t | z_t \sim p(o|z)$$

Goal: infer a posterior $p(z_t|o_{\leq t})$ over the discrete latent state subject to:

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.

THE CHINESE RESTAURANT PROCESS

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



The Chinese Restaurant Table Distribution

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

K_t is the (random) integer number of clusters after the t-th observation.

RECURSION FOR ONLINE FILTERING

Using Bayes' Rule we can write the exact posterior as:

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

The latent prior is obtained by marginalizing over cluster assignments:

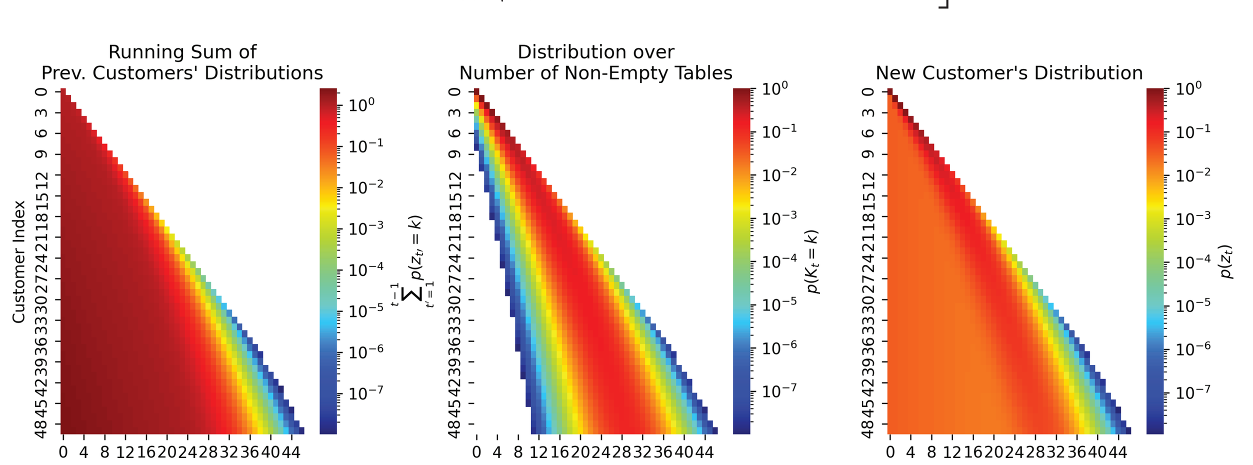
$$\underbrace{p(z_t = k | o_{<t})}_{\text{Latent Prior}} = \mathbb{E}_{p(z_{<t}, K_{t-1} | o_{<t})} \Big[p(z_t = k | z_{<t}, K_{t-1}, o_{<t}) \Big]$$

Approximate recursion for the prior:

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

Putting it all together, we have a Bayesian recursion for the posterior:

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



Recursion for the table numbers:

$$p(K_t | o_{\leq t}) \approx \sum_{K_{t-1}, z_t} p(K_t | K_{t-1}, z_t) p(K_{t-1} | o_{\leq t-1}) p(z_t | o_{\leq t})$$

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

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

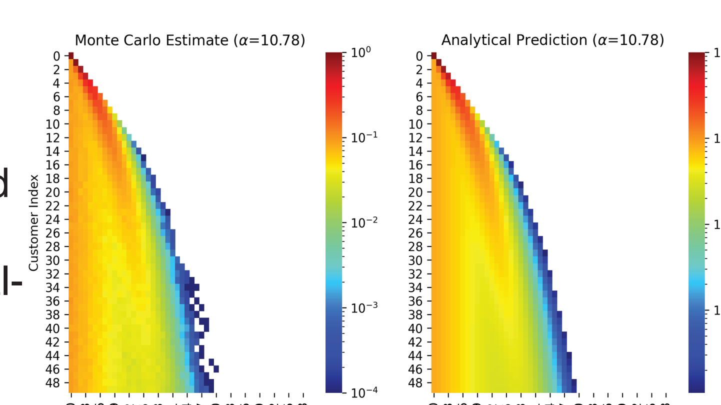
R-CRP performs online inference and parameter estimation by:

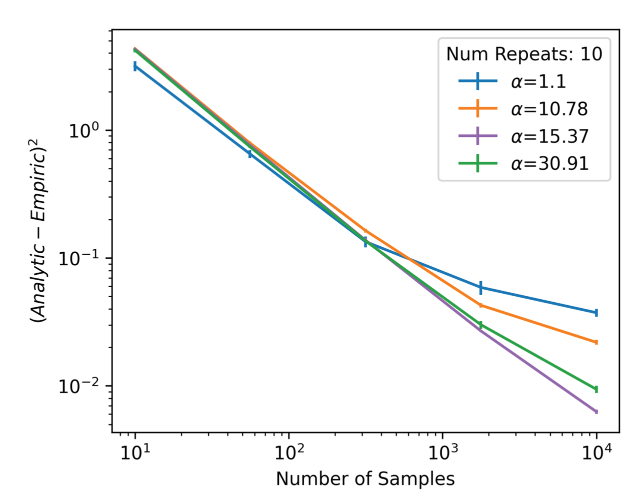
1) Inferring a latent posterior over which cluster the current observation belongs to i.e. $p(z_t|o_{\leq t})$,

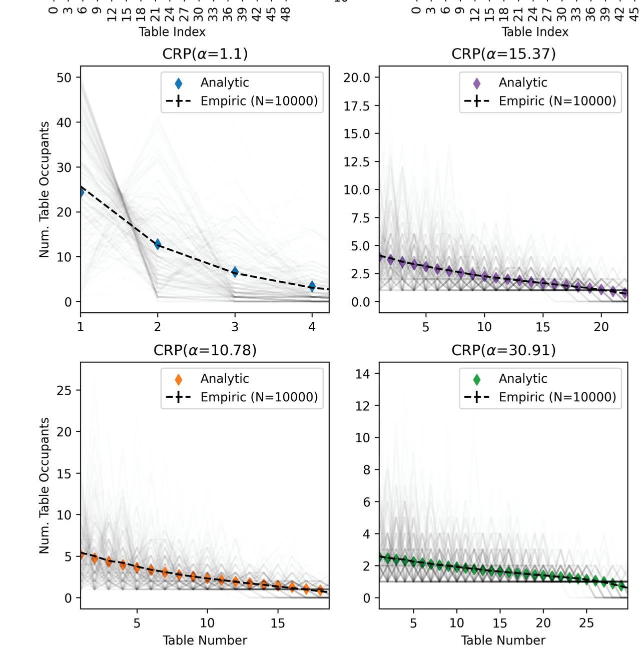
2) Updates each cluster's parameters modulated by the posterior probability the observation belongs to that cluster.

CORRECTNESS FOR CRP PRIOR

According to the derivation, the recursion should hold exactly for the CRP prior in the absence of observations.The recursion should hold exactly for the CRP prior. We confirm this by comparing our analytical expression to 5k Monte Carlo samples, each of 50 customers, drawn from CRP(α), for α ={1.1,10.78,15.37,30.91}.







EXPERIMENTS

We compare R-CRP with several baselines:

Offline Baselines:

1) Variational Bayes implemented in Scikit-Learn.
2) Hamiltonian-Gibbs Monte-Carlo Sampling, implemented in Pyro.
3) DP-Means, a low-variance asymptotic approximation. We call this DP-Means (offline).

Online Baselines:

1) DP-Means, but limited to a single forward pass through the data, We call this DP-Means (online).
2) Sequential Updating and Greedy Search (SUGS) which uses a "local MAP approximation" i.e. $\hat{z}_t = \mathrm{argmax}_k\, p(z_t = k | \hat{z}_{<t}, \alpha)$
3) Online CRP, which uses sampling i.e. $\hat{z}_t \sim p(z_t = k | \hat{z}_{<t}, \alpha)$

.

GAUSSIAN MIXTURE MODEL

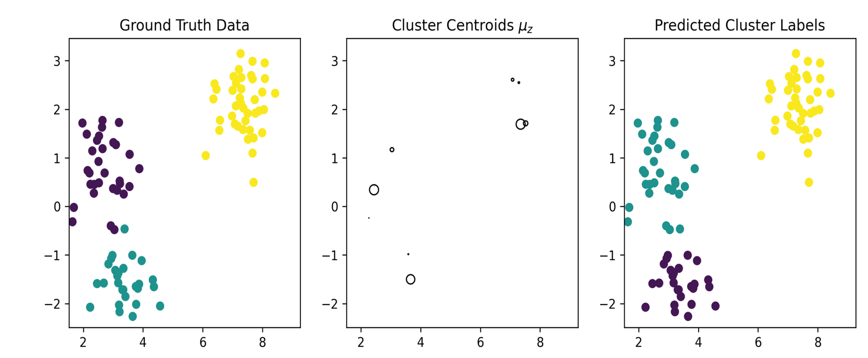
Data are generated by drawing a sequence of cluster assignments $z_1, ..., z_T \sim CRP(\alpha)$ and then drawing a sequence of observations from the corresponding Gaussian:

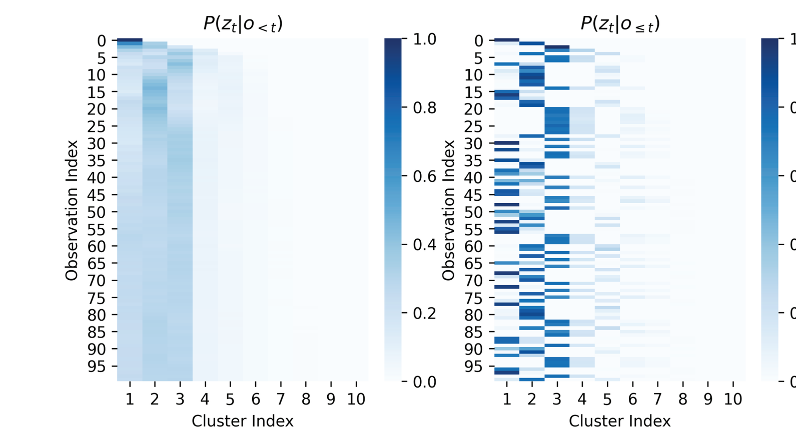
$$x_t | z_t, \{\mu_k\}_{k=1}^3 \sim \mathcal{N}(\mu_{z_t}, \Sigma_{z_t}).$$

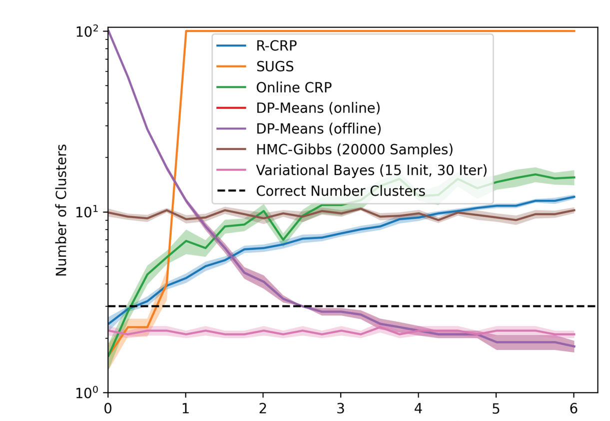
The cluster means are drawn as $\mu_k \sim \mathcal{N}(0, \rho I)$ and the clusters have identical isotropic covariances: $\Sigma_k = I$.

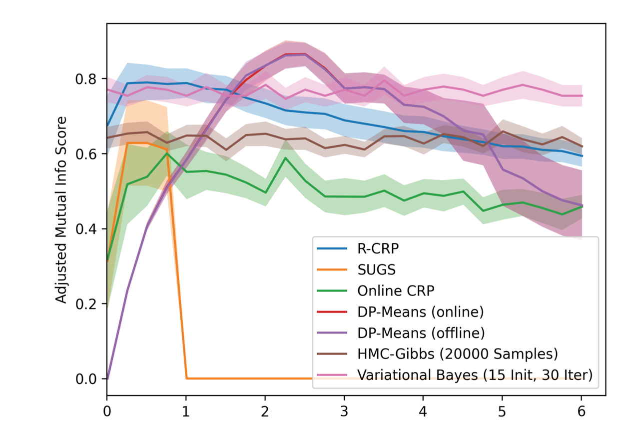
1. Initialize $p(K_0 = 0) = 1$, and $\forall k \in \mathbb{Z}$, set running sum of posteriors $R_0(k) = 0$
2. For each $t = 1, ..., T$:
   - Create new cluster mean $\mu_{k=t}^{(t)} = o_t$
   - Compute prior for new observation:
     $$p(z_t = k | o_{<t}) \approx \frac{1}{\alpha + t - 1} R_{t-1}(k) + \frac{\alpha}{\alpha + t - 1} p(K_{t-1} = k - 1 | o_{<t})$$
   - 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 online inference with R-CRP(α) in a Gaussian Mixture Model.



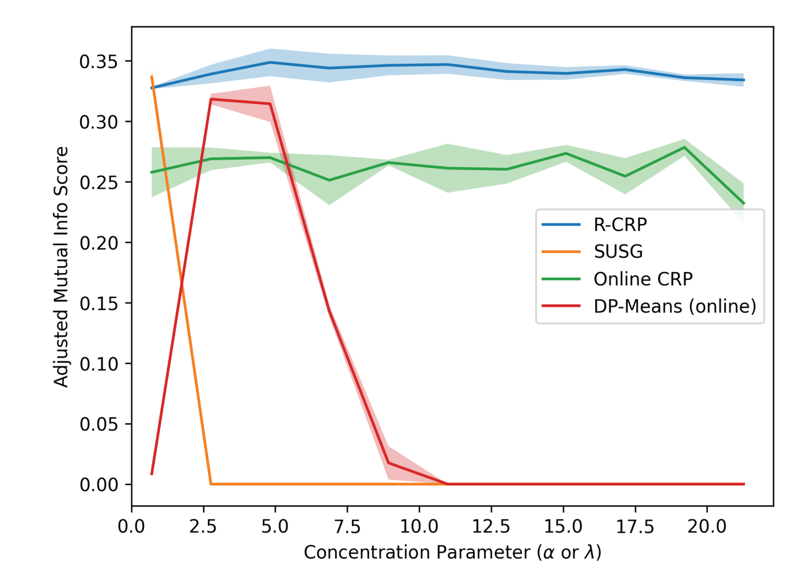


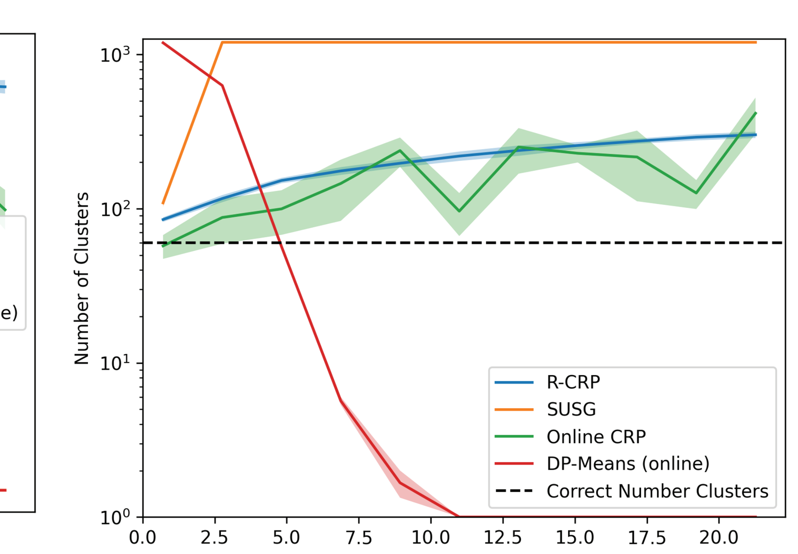




HANDWRITTEN CHARACTER RECOGNITION (OMNIGLOT)

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 strokesrepresents, or whether the set of strokes represents a new character altogether.

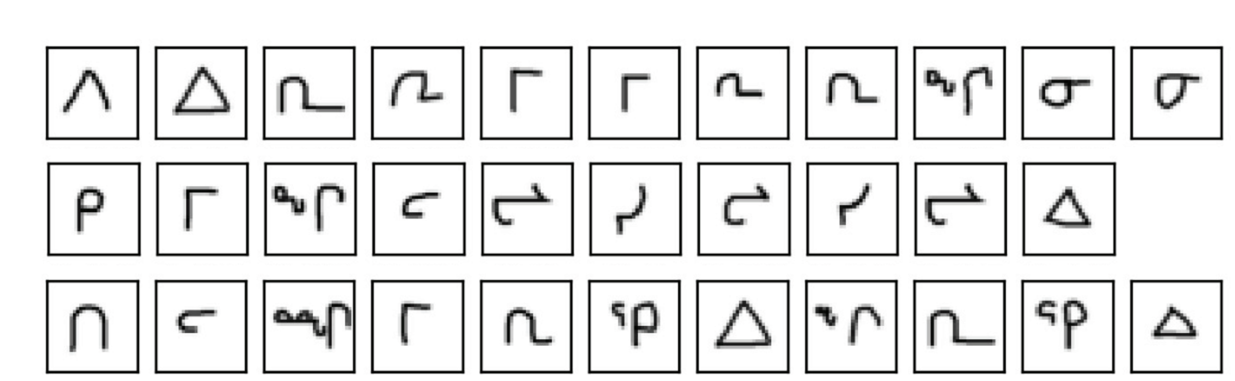




Cluster: 1

Cluster: 2

Cluster: 3



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. 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.

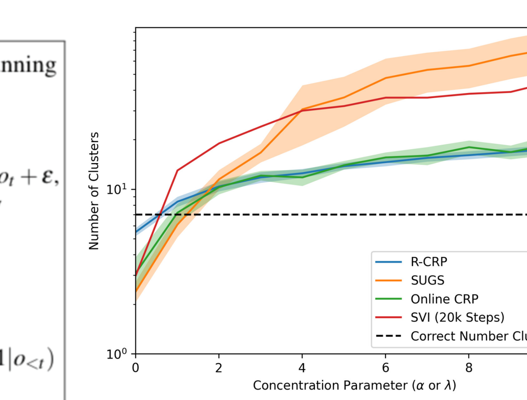
Data are generated by drawing cluster assignments $z_1, ..., z_T$ And then drawing a probability vector $p_k \sim Dir(\beta \mathbf{1}) \in \Delta^{V-1}$ Each observation then follows $o_t \sim \text{Multinomial}(M, p_{z_t})$. Here, M is the document length and V is the vocabulary length

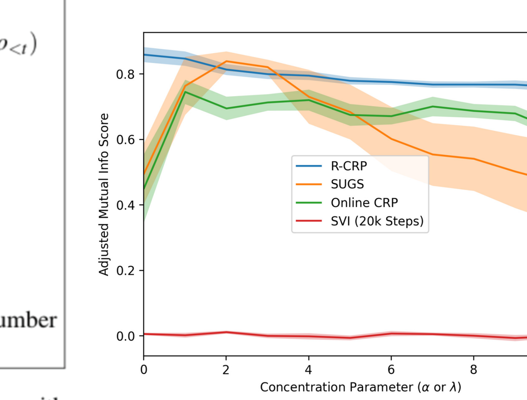
To update each cluster's pseudocounts,

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

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 ε 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 online inference with R-CRP(α) in a Categorical Mixture Model.





COMPLEXITY ANALYSIS

K_t is well approximated by a Poisson with rate:

$$\lambda = \alpha \log(1 + t/\alpha)$$

If inference follows the prior's asymptotic behavior, our expected time complexity is $O(t \log(t))$ and expected space complexity is $O(\log(t))$.

CONCLUSIONS

- We presented an efficient online inference algorithm for nonparametric discrete latent variable time series data. The R-CRP is a novel Bayesian recursion that performs inference online with quasilinear average case time complexity $O(t \log(t))$ and logarithmic average case space complexity $O(\log(t))$.
- When the latent variables are indeed hidden, under one small approximation, R-CRP recovers the latent structure as well as or better than commonly used inference algorithms that make multiple passes through the data and require simultaneous access to the entire dataset.

ACKNOWLEDGEMENTS

IRF is supported by the Simons Foundation and HHMI. MK acknowledges funding from the MIT Department of Physics and McGovern Institute.