Efficient Streaming Inference for Infinite Latent Feature Models

Anonymous Authors[1]

Abstract

Bayesian nonparametrics are a classic Bayesian approach to continual learning, in which the nonparametric model grows in representational capacity as more data are observed. One Bayesian nonparametric model, the Indian Buffet Process (IBP), is used when the stream of data is assumed to be generated by an unknown number of binary latent features. However, use of the IBP has been limited by the lack of inference algorithms for the online setting. Here we propose a Bayesian recursion called the Recursive IBP (R-IBP) that efficiently filters a posterior distribution over latent features from a sequence of observations in a streaming manner. R-IBP can be combined with a variety of other models whenever the number of components is unknown, including independent component analysis (ICA), non-negative matrix factorization (NMF), and factor analysis (FA). Our recursive filter has quasilinear average time complexity and logarithmic average space complexity in the number of observations.

1. Introduction

A central challenge of continual learning is defining a model capable of growing in representational capacity as more data are observed. One approach is to define a model with an infinite number of parameters, under the constraint that only a finite number of parameters may be used for a finite amount of data; then, as more data are observed, more parameters become available for use and the model can more flesibly model the data. This approach, called Bayesian nonparametrics (BNPs), offers a principled and powerful solution to continual learning.

However, despite the seemingly clear relevance of BNPs, these models are not widely used (although see (Mehta et al., 2021; Lee et al., 2020; Kessler et al., 2019)). One reason may be that most, if not all, inference algorithms (both sampling and variational) for BNP models are not applicable in the streaming setting. Recent work sought to rectify this shortcoming by providing an efficient streaming inference algorithm for an "infinite" clustering model called the Chinese Restaurant Process (Schaeffer et al., 2021). In this paper, we propose an efficient streaming inference algorithm for an "infinite" feature model called the Indian Buffet Process (IBP), which has been used in the continual learning literature (Kessler et al., 2019; Mehta et al., 2021). We start by introducing our precise inference problem, then define the IBP and its utility, and finally detail our proposed inference algorithm.

2. Background

2.1. Generative Model

We consider a latent variable time series model with K-dimensional binary latent variables $z_{1:T}$ (i.e. $z_t \in \{0,1\}^K$) and observable variables $o_{1:T}$, where K is unknown and $\cdot_{1:T}$ denotes the sequence $(\cdot_1, \cdot_2, ..., \cdot_T)$. Our generative process assumes an Indian Buffet Process (IBP) prior over the sequence of latent states:

$$z_{1:T} \sim IBP(\alpha, \beta)$$
$$o_t | z_t \sim p(o|z)$$

2.2. Indian Buffet Process

The Indian Buffet Process (IBP; (Griffiths & Ghahramani, 2011)) is a two-parameter[1] ($\alpha > 0, \beta > 0$) stochastic process that defines a discrete distribution over binary matrices with finite rows and unbounded number of columns. The term IBP arises from an analogy of a sequence of customers (rows) arriving at an Indian buffet and choosing an unbounded number of dishes (columns). The tth customer samples an integer number of new dishes $\lambda_t \sim Poisson(\alpha\beta/(\beta+t-1))$ and then samples previous dishes with probability proportional to the number of previous customers who sampled said dishes. Denoting the total number of dishes after the first t customers as $\Lambda_t = \sum_{t'=1}^{t} \lambda_{t'}$, the IBP defines a conditional distribution for the tth row and

[1]The original IBP paper (Griffiths & Ghahramani, 2005) defined a single parameter model that (Ghahramani et al., 2007) extended to two parameters and that (Teh & Görür, 2009) extended to three. This paper applies equally to all, but since our focus is on efficient streaming inference and not particular properties of a specific model, we chose the 2 parameter IBP to balance expositional simplicity and model complexity.

kth column's binary variable z_{tk}, given the preceding rows:

$$P(z_{t,k} = 1 | z_{<t,k}, \Lambda_{t-1}, \lambda_t, \alpha, \beta) = \begin{cases} \frac{1}{\beta+t-1} \sum_{t'<t} \mathbb{I}(z_{t',k} = 1) & \text{if } k \leq \Lambda_{t-1} \\ p(\lambda_t + \Lambda_{t-1} \geq k) & \text{if } \Lambda_{t-1} < k \end{cases} \quad (1)$$

The IBP can be equivalently expressed with random indicator variables, a fact we later exploit:

$$p(z_{t,k} = 1 | z_{<t,k}, \Lambda_{t-1}, \lambda_t, \alpha, \beta) = \frac{1}{\beta+t-1} \sum_{t'<t} \mathbb{I}(z_{t',k} = 1)\mathbb{I}(k \leq \Lambda_{t-1}) + \mathbb{I}(\Lambda_{t-1} < k)\mathbb{I}(k \leq \Lambda_{t-1} + \lambda_t) \quad (2)$$

Because each of the λ_t are independent Poissons with rate $\alpha\beta/(\beta + t - 1)$ and because the sum of independent Poisson random variables is itself Poisson, we know that $\Lambda_t \sim Poisson(\sum_{t'=1}^{t} \alpha\beta/(\beta + t' - 1))$. This implies the expected number of dishes grows logarithmically with t because $\mathbb{E}[\Lambda_t] = \sum_{t'=1}^{t} \alpha\beta/(\beta + t' - 1)) \approx \alpha\beta \int_{t'=1}^{t} dt'/(\beta + t' - 1) \approx \alpha\beta \log(1 + t/\beta)$. This detail becomes important in our later complexity analysis.

2.3. Settings of the IBP

By offering a distribution over an unbounded set of binary variables, the IBP provides a useful probabilistic tool for three reasons. First, the IBP can add additional columns as necessary, allowing the model to grow in complexity as more data are observed. By associating each column with a feature (alternatively called factors), data can be expressed as an "infinite" combination of said factors. This makes the IBP a useful addition to methods that would otherwise require specifying a fixed number of components; for instance, the IBP has been utilized with independent component analysis (ICA) (Knowles & Ghahramani, 2007), Non-Negative Matrix Factorization (NMF) (Gupta et al., 2012; Zhou et al., 2012), and Factor Analysis (FA) (Paisley & Carin, 2009).

3. Recursion for Online Filtering (R-IBP)

3.1. Objective

Our goal is to infer a posterior over the current observation's binary latent variables $z_t \stackrel{\text{def}}{=} \{z_{tk}\}_{k=1}^{k=\infty}$ given the entire history of observations i.e. filter $p(z_t|o_{\leq t})$, subject to two constraints:

1. Inference must be performed online
2. Inference must be efficient in the large observation (i.e. time) limit

In a time series context, inferring $p(z_t|o_{\leq t})$ is often called filtering (e.g. Kalman filter, particle filter).

3.2. Bayesian Recursion

The challenge with using the IBP in a streaming setting is that its conditional form renders the current latent variables z_t dependent on all previous latent variables $z_{<t}$. Our approach is to break the dependence by converting the conditional into a marginal, which can be expressed as a running sum. For brevity, we refer to the prior on the current latent variables $p(z_t|o_{<t})$ as the "latent prior" and the posterior on the current variables $p(z_t|o_{\leq t})$ as the "latent posterior". Bayes' rule relates the latent prior to the latent posterior:

$$\underbrace{p(z_{tk} = 1|o_{\leq t})}_{\text{Latent Posterior}} = \frac{p(o_t|z_{tk} = 1)}{p(o_t|o_{<t})} \underbrace{p(z_{tk} = 1|o_{<t})}_{\text{Latent Prior}} \quad (3)$$

The latent prior $p(z_{tk} = 1|o_{<t})$ can be rewritten as the expectation of an indicator random variable that we can expand using the Law of Total expectation. Suppressing α, β for brevity, the IBP's marginal distribution is the expectation of the IBP's now-random conditional distribution:

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

Substituting Eqn. 2, taking the expectations and simplifying yields

$$p(z_{tk} = 1|o_{<t}) = \frac{1}{\beta+t-1} \sum_{t'<t} p(z_{t'k} = 1|o_{<t}) + p(\Lambda_{t-1} \leq k - 1|o_{<t}) - p(\Lambda_{t-1} + \lambda_t \leq k - 1|o_{<t}) \quad (4)$$

where $k \leq \Lambda_{t-1}$ disappears because the probability of a customer eating a non-existent dish is 0 (i.e. $p(z_{t'k} = 1, k > \Lambda_{t-1}|o_{<t}) = 0$ and $p(z_{t'k} = 1|o_{<t}) = p(z_{t'k} = 1, k \leq \Lambda_{t-1}|o_{<t}) + p(z_{t'k} = 1, k > \Lambda_{t-1}|o_{<t})$) and where the other two terms are a difference of two cumulative distribution functions (CDFs)

$$\begin{aligned} &p(\Lambda_{t-1} < k \leq \Lambda_{t-1} + \lambda_t|o_{<t}) \\ &= 1 - p(\Lambda_{t-1} < k \leq \Lambda_{t-1} + \lambda_t|o_{<t})^C \\ &= 1 - \Big(p(\Lambda_{t-1} \geq k|o_{<t}) + p(k > \Lambda_{t-1} + \lambda_t|o_{<t})\Big) \\ &= p(\Lambda_{t-1} \leq k - 1|o_{<t}) - p(\Lambda_{t-1} + \lambda_t \leq k - 1|o_{<t}) \end{aligned}$$

For the IBP prior, Eqn. 4 is an exact recursion. However, once data is observed, we must introduce one approximation. The reason why is $p(z_{t'k}|o_{<t})$ requires revising every previous latent posteriors using all observations, which is

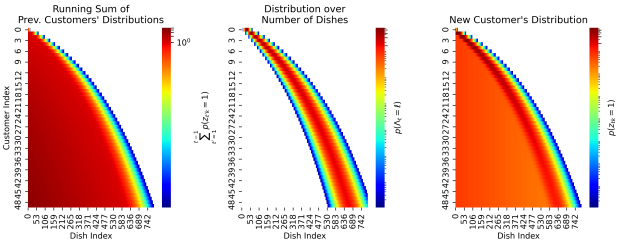

Figure 1. Visualization of R-IBP. The running sum of previous latent features' posterior masses (left) competes with the pressure to create new latent features (center) to generate a prior for the next observations latent features (right).

not possible in the streaming context. We therefore assume latent posteriors depend only on data seen up to that point i.e. i.e. $p(z_{t'k}|o_{\leq t}) = p(z_{t'k}|o_{\leq t'})$ for $t' \leq t$, precluding retroactive revision. This yields our key expression for the latent prior:

$$p(z_{tk} = 1|o_{<t}) \approx \frac{1}{\beta + t - 1} \sum_{t'<t} p(z_{t'k} = 1|o_{<t'}) + p(\Lambda_{t-1} \leq k - 1|o_{<t}) - p(\Lambda_{t-1} + \lambda_t \leq k - 1|o_{<t}) \tag{5}$$

We call this streaming inference algorithm the **Recursive Indian Buffet Process (R-IBP)**. Intuitively, Eqn. 5 tells us that the prior probability the tth observation was generated by the kth feature is approximately given by the running sum of the previous customers' posterior probabilities of also being generated by the kth feature, plus one other term. This running sum of probability masses of the latent binary variables is a "soft" version of the IBP's natural behavior, which would ordinarily add the realizations of the binary variables if the variables were observed. The other term can be shown to be a difference of two Poisson CDFs, which we discuss below; its role is to incentivize the creation of new features. As new observations arrive, these two terms compete to explain the observations: the first term advises the observation can be explained using previous features, proportional to how common those features are, while the second term advises the new observation might be better explained by creating new, additional features. This recursion is visually displayed in Fig 1 for $\alpha = 30.91, \beta = 13.82$.

We now return to the difference of CDFs to show why the posterior over the total number of dishes remains Poisson and to give its exact rate. As before, let Λ_t denote the total number of dishes after the tth customer:

$$\Lambda_t \overset{\text{def}}{=} \sum_{k=1}^{k=\infty} \min\left(1, \sum_{t'=1}^{t'=t} z_{t',k}\right)$$

Each term in the sum represents whether the kth feature was present in the first t customers. Consider one term in the sum, $M_{tk} \overset{\text{def}}{=} \min(1, \sum_{t' \leq t} z_{t',k})$. We can use the following proposition to determine the distribution of M_{tk}:

Proposition 1 *Let X be a random variable with CDF $F_X(x) = p(X \leq x)$ and let $c \in \mathbb{R}$ be a constant. Then the random variable $Y \overset{def}{=} \min(c, X)$ has a CDF $F_Y(y) = p(Y \leq y)$ given by*

$$F_Y(y) = \begin{cases} F_X(y) & \text{if } y < c \\ 1 & \text{if } y \geq c \end{cases}$$

Substituting M_{tk} for Y, $\sum_{t' \leq t} z_{t'k}$ for X and 1 for c, it follows that

$$F_{M_{tk}|o_{\leq t}}(0) = F_{\sum z_{t'k}|o_{\leq t}}(0)$$

and

$$F_{M_{tk}|o_{\leq t}}(1) = 1.$$

We can now determine the probability mass function (PMF) of M_{tk}:

$$\begin{aligned} p(M_{tk} = 0|o_{\leq t}) &= p(M_{tk} \leq 0|o_{\leq t}) \\ &= F_{M_{tk}|o_{\leq t}}(0) \\ &= F_{\sum z_{t'k}|o_{\leq t}}(0) \\ &= p\Big(\sum_{t' \leq t} z_{t'k} \leq 0 \Big| o_{\leq t}\Big) \\ &= p\Big(\sum_{t' \leq t} z_{t'k} = 0 \Big| o_{\leq t}\Big) \end{aligned}$$

where the first and last steps follow because M_{tk} and $\sum_{t' \leq t} z_{t',k}$ can only take values in $\{0, 1, 2, ..., t\}$. Each $z_{t'k}$ is a Bernoulli random variable with distribution given by $p(z_{t'k}|o_{\leq t'})$. The sum can only be 0 if all $z_{t'k} = 0$, which occurs with probability $\prod_{t' \leq t} p(z_{t'k} = 0|o_{\leq t'})$. The PMF of M_{tk} is therefore

$$\begin{aligned} p(M_{tk} = 0|o_{\leq t}) &= \prod_{t' \leq t} p(z_{t'k} = 0|o_{\leq t}) \\ p(M_{tk} = 1|o_{\leq t}) &= F_{M_{tk}|o_{\leq t}}(1) - F_{M_{tk}|o_{\leq t}}(0) \\ &= 1 - \prod_{t' \leq t} p(z_{t'k} = 0|o_{\leq t'}) \end{aligned}$$

and $p(M_k = n) = 0$ for $n = 2, 3, ..., t$. This tells us that $M_k \sim Bernoulli(1 - \prod_{t' \leq t} p(z_{t'k} = 0|o_{\leq t}))$, which is sensible as M_k describes the presence of the kth feature. Then, because Λ_t is the sum of independent but non-identically distributed Bernoullis, Le Cam's Theorem

(Le Cam, 1960) tells us that Λ_t closely follows a Poisson distribution:

$$p(\Lambda_t|o_{\leq t}) = Poisson\Big(\sum_{k=1}^{k=\infty}\Big(1-\prod_{t'=1}^{t'=t} p(z_{t'k}=0|o_{\leq t'})\Big)\Big) \tag{6}$$

As a sanity check, we know that for infinite data, the IBP should fill the entire feature space; here, as $t \to \infty$, the product approaches 0 and the probability that the kth feature exists goes to 1, meaning the entire feature space is filled.

Eqn. 5 also requires a prior on the next number of dishes, which can be straightforwardly constructed. Because the number of new dishes added by the tth customer λ_t does not depend on the preceding total number of dishes or previous observations, the prior on the next number of dishes $p(\Lambda_{t+1}|o_{\leq t}) = p(\Lambda_t + \lambda_{t+1}|o_{\leq t})$ is Poisson with rate

$$\sum_{k=1}^{k=\infty}\Big(1-\prod_{t'=1}^{t'=t-1} p(z_{t'k}=0|o_{\leq t'})\Big) + \alpha\beta/(\beta+t-1)\Big)$$

3.3. Complexity Analysis

The worst-case time and space complexity of R-IBP is determined by the number of latent features Λ_t, which is unbounded. This is a property of the IBP itself, as the IBP permits adding an arbitrarily large number of features for even a single observation (albeit with exponentially vanishing probability). Consequently, we instead consider the average-case time and space complexity as dictated by the IBP prior.

Computing the posterior over the number of dishes has time complexity $O(\Lambda_t)$ per step, for t steps, and computing the posterior over the present features has time complexity $O(\Lambda_t)$ per step, for t steps. Storing the running sum of features' probability masses requires $O(\Lambda_t)$ space and storing the Poisson rate that describes the posterior over the number of features requires $O(1)$ space. Recalling that Λ_t grows logarithmically with t, the average-case complexity is quasilinear $O(t\Lambda_t) \approx O(t \log t)$ with time and logarithmic $O(\Lambda_t) \approx O(\log t)$ with space.

4. Experimental Results

4.1. IBP Prior

According to the derivation, the recursion should hold exactly for the IBP prior in the absence of observations. We test this by checking whether the recursion correctly captures the sequence of marginal distributions by comparing the recursion's analytical expressions to 5000 Monte Carlo samples, each of 50 customers, drawn using the IBP's conditional distribution, for $\alpha \in \{2.64, 30.91\} \times \beta \in \{3.11, 13.82\}$. First, we visually compared the analytical expressions versus the Monte Carlo estimates and found excellent agreement (Fig. 2).

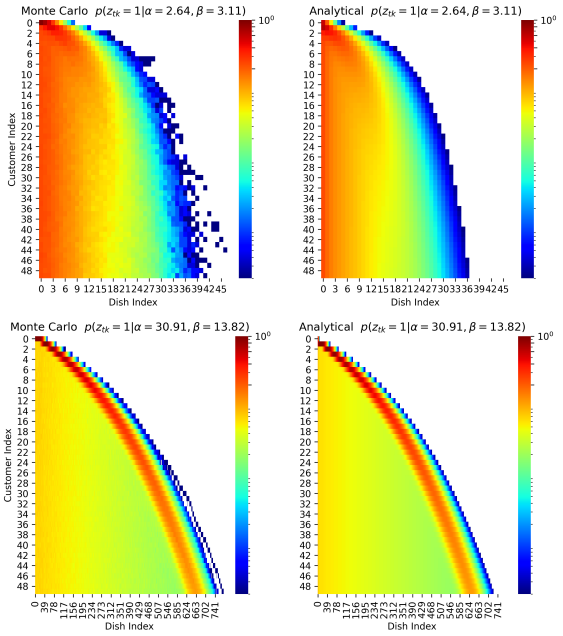

Figure 2. Two examples (top: $\alpha = 2.64, \beta = 3.11$; bottom: $\alpha = 30.91, \beta = 13.82$) comparing Monte Carlo estimates of $p(z_{tk} = 1)$ against our analytical expression.

Second, we computed the mean squared error between the analytical expression and Monte Carlo estimates as a function of the number of Monte Carlo samples. For all (α, β) values, the squared error falls as a power law with the number of samples (Fig 3), supporting our claim that the recursion is exact for the IBP prior.

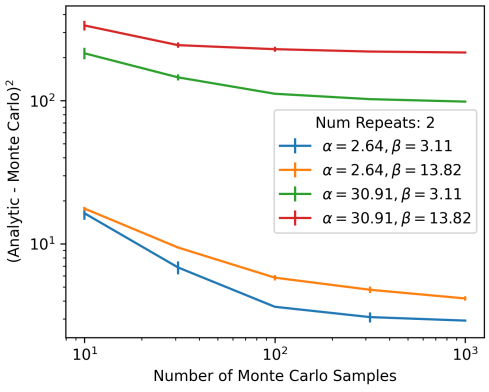

Figure 3. For all settings of (α, β), the mean-squared error of the Monte Carlo estimate and our analytical expression falls as a power law with the number of samples.

References

Ghahramani, Z., Griffiths, T. L., and Sollich, P. Bayesian Nonparametric Latent Feature Models. *Bayesian Statistics*, 8:25, 2007.

Griffiths, T. L. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. *Neural Information Processing Systems*, pp. 8, 2005.

Griffiths, T. L. and Ghahramani, Z. The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, 12(32):1185–1224, 2011. ISSN 1533-7928. URL http://jmlr.org/papers/v12/griffiths11a.html.

Gupta, S. K., Phung, D., and Venkatesh, S. A nonparametric Bayesian Poisson gamma model for count data. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp. 1815–1818, November 2012. ISSN: 1051-4651.

Kessler, S., Nguyen, V., Zohren, S., and Roberts, S. Indian Buffet Neural Networks for Continual Learning. *NeurIPS Workshop on Bayesian Deep Learning*, pp. 14, 2019.

Knowles, D. A. and Ghahramani, Z. Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. *undefined*, 2007. URL /paper/Infinite-Sparse-Factor-Analysis-and-Infinite-Knowles-Ghahramani/76e171e8de3fe77d4532ed235e0a0669e420b782.

Le Cam, L. 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.

Lee, S., Ha, J., Zhang, D., and Kim, G. 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.

Mehta, N., Liang, K. J., Verma, V. K., and Carin, L. Continual Learning using a Bayesian Nonparametric Dictionary of Weight Factors. *AISTATS*, pp. 11, 2021.

Paisley, J. and Carin, L. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pp. 1–8, Montreal, Quebec, Canada, 2009. ACM Press. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553474. URL http://portal.acm.org/citation.cfm?doid=1553374.1553474.

Schaeffer, R., Bordelon, B., Khona, M., and Fiete, I. R. Efficient Online Inference for Nonparametric Latent Variable Time Series. *Uncertainty in Artificial Intelligence*, 2021.

Teh, Y. W. and Görür, D. Indian buffet processes with power-law behavior. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pp. 1838–1846, Red Hook, NY, USA, December 2009. Curran Associates Inc. ISBN 978-1-61567-911-9.

Zhou, M., Hannah, L., Dunson, D., and Carin, L. Beta-Negative Binomial Process and Poisson Factor Analysis. In *Artificial Intelligence and Statistics*, pp. 1462–1471. PMLR, March 2012. URL http://proceedings.mlr.press/v22/zhou12c.html. ISSN: 1938-7228.