Information theory of Exchangeable Estimators

N. P. Santhanam∗, M. M. Madiman†, A. D. Sarwate‡

December 5, 2012

Abstract

Exchangeable random partition processes are the basis for Bayesian approaches to statistical inference in large alphabet settings. On the other hand, the notion of the pattern of a sequence provides a framework for data compression in large alphabet scenarios. Because data compression and parameter estimation are intimately related, we study the redundancy of Bayes estimators coming from Poisson-Dirichlet priors (or “Chinese restaurant processes”) and the Pitman-Yor prior. This provides an understanding of these estimators in the setting of unknown discrete alphabets from the perspective of universal compression. In particular, we identify relations between alphabet sizes and sample sizes where the redundancy is small— and hence, characterize useful regimes for these estimators.

1 Introduction

A number of statistical inference problems of significant contemporary interest, such as text classification, language modeling, and DNA microarray analysis, requires computing inferences based on observed sequences of symbols in which the sequence length or sample size is comparable or even smaller than the set of symbols, the alphabet. For instance, language models for speech recognition estimate distributions over English words using text examples much smaller than the vocabulary.

Inference in this setting has received a lot of attention, from Laplace [1, 2, 3] in the 18th century, to Good [4] in the mid-20th century, to an explosion of work in the statistics [5, 6, 7, 8, 9, 10, 11, 12, 13], information theory [14, 15, 16, 17, 18, 19] and machine learning [20, 21, 22, 23] communities in the last few decades.

A major strand in the information theory literature on the subject has been based on the notion of patterns that captures information in sequences that can be described well (see [24] for formal characterizations of the above idea), a major strand in the statistical literature has been based on the notion of exchangeability that can be thought of as a generalization of independence.

Let \( \mathcal{P} \) be any collection of measures over infinite sequences \( X_1, X_2, \ldots \), where \( X_i \) come from a countable (infinite) set (the alphabet). We will assume that \( X_i \) are independent and identically distributed (i.i.d.), and in a slight abuse of notion, refer to \( \mathcal{P} \) as a collection of i.i.d.

∗Department of Electrical Engineering, University of Hawaii at Manoa, 2540 Dole Street, Honolulu, HI 96822, USA, nsanthan@hawaii.edu
†Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA, mokshay.madiman@yale.edu
‡Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave., Chicago, IL 60637, USA, asarwate@ttic.edu
measures. Let \( q \) be a measure over the infinite sequences as well. Suppose we consider the distributions induced on length \( n \) sequences by measures in \( \mathcal{P} \)—call this set of distributions \( \mathcal{P}^n \), and we ask how well does \( q \) approximate an arbitrary unknown distribution \( p \in \mathcal{P}^n \)? Clearly no matter what \( p \) is, for any number \( R_n > 0 \)

\[
p(X^n_1 : q(X^n_1) > R_n p(X^n_1)) < 1/R_n.
\]

Now if \( R_n \) is any function that grows sufficiently fast with \( n \), \( q \) does not asymptotically overestimate probabilities of length-\( n \) sequences by more than \( R_n \) with probability 1, no matter what measure in \( \mathcal{P} \) generated the sequences.

However, underestimation is not so simple. Unless we choose \( q \) carefully, we could underestimate probabilities quite badly. The redundancy of an estimator (which we define formally in Section ??) for length \( n \) sequences is an measure of how closely the probabilities induced on length-\( n \) sequences by \( q \) match the highest probability assigned to that sequence by any distribution in \( \mathcal{P}^n \). Redundancy captures how well \( q \) performs against the collection \( \mathcal{P} \), but the connections run deeper. Ideally, the redundancy will be sublinear in the sequence length \( n \), so the per-sample redundancy vanishes as \( n \to \infty \). If an estimator has \( o(n) \) redundancy for any measure \( p \in \mathcal{P} \), we call the estimator universal for \( \mathcal{P} \).

Suppose we have a measure (prior) on \( \mathcal{P}^n \). Different priors induce different distributions on the data \( X^n_1 \). We think of the prior as choosing a distribution in \( \mathcal{P}^n \) and then the distribution generates the data we see. How much information about the model can we obtain from the data (assuming we know the prior)? Indeed, a well known resul \[ \] proves that the redundancy of the best possible distribution for \( \mathcal{P}^n \) equals the maximum possible information (in bits) that is present about \( \mathcal{P}^n \) in the length \( n \) sequences for any choice of prior.

When the symbols come from a countably infinite alphabet, the usual notions of redundancy seem to break down. However, a series of papers starting from [16] have formalized a similar framework for countably infinite alphabets—one based on the notion of patterns of sequences that simply abstract the identities of symbols, and indicate only the relative order of appearance. For example, the pattern of FEDERER is 1232424, while that of PATTERN is 1233456. The crux of the idea here is that instead of considering the set of measures \( \mathcal{P} \) over infinite sequences, we consider the set of measures induced over patterns of the sequences. It then follows that now our estimate \( q_\Phi \) is a measure over patterns.

The conditional distributions are particularly interesting—given a sequence of symbols, our new estimate tells us the probability that the next symbol is “new” (has not appeared in the given sequence), and the probability of symbols that have appeared in the sequence thus far. It is not an accident that similar estimators come up in the Bayesian non-parametric literature that focuses on exchangeability.

In probability and statistics literature, Kingman [25] advocated the use of exchangeable random partitions to accommodate the analysis of data from an alphabet that is not bounded or known in advance. A more detailed discussion of the history and philosophy of this problem can be found in the works of Zabell [11, 26] collected in [27]. One of the most popular exchangeable random partition processes is the “Chinese restaurant process,” [10], which is a special case of the Poisson-Dirichlet or Pitman-Yor process [28, 29]. These processes can be viewed as prior distributions on the set of all discrete distributions that can be used as the basis for Bayesian estimation of probabilities.

Our objective is to evaluate the performance of the sequential estimators corresponding to these exchangeable partition processes. In each case above—Chinese restaurant process, the Pitman Yor
process, or the Ewen’s sampling formula, given \( n \) symbols from a distribution, these sequential estimators give the chance that the \((n + 1)\)-th symbol from the distribution has the same value as a previously observed symbol and the chance that it has a new previously unobserved value.

Let \( \mathcal{I} \) be the collection of all distributions over countable (potentially infinite) alphabets, and let \( \mathcal{I}^\infty \) be the collection of all \( i.i.d. \) measures with single letter marginals in \( \mathcal{I} \). Let \( \mathcal{I}_\Phi \) be the collection of all measures over patterns induced by measures in \( \mathcal{I}^\infty \). We evaluate the redundancy of the above estimators against \( \mathcal{I}_\Phi \).

In the context of sequential estimation, universal estimators do exist \([16]\) for the collection \( \mathcal{I}_\Phi \) of measures over patterns, and the normalized redundancy can scale as \( O(n^{1/2}) \) (for a computationally intensive method) or \( O(n^{2/3}) \) (for a linear-time estimator). However, both of these estimators are quite complicated; it would be desirable to have simple estimators that can also have vanishing per-sample redundancy.

For the case of the estimators studied in nonparametric Bayesian statistics, our results show that they are in general neither weakly nor strongly universal. Because the redundancy is measured in the worst case over \( p \), the alphabet size may be arbitrarily large with respect to the sample size, so the estimators do not have \( o(n) \) redundancy. However, we can show that a mixture of estimators corresponding to Chinese restaurant process estimators is weakly universal. This mixture is made by optimizing individual CRP estimators that assume a bound on the support of \( p \). If such a bound is known in advance, we can derive much tighter bounds on the redundancy. In this setting the two-parameter Poisson-Dirichlet (or Pitman-Yor) estimator is superior to the estimator derived from the Chinese restaurant process.

In order to describe our results, we require a variety of definitions from different research communities. In Section 2, we describe this preliminary material and place it in context, and in Section 3 we describe our main results.

## 2 Preliminaries

Let \( \mathcal{I}_k \) denote the set of all probability distributions on alphabets of size \( k \), \( \mathcal{I}^\infty \) be all probability distributions on countably infinite alphabets, and let

\[
\mathcal{I} = \mathcal{I}^\infty \cup \bigcup_{k \geq 1} \mathcal{I}_k
\]

be the set of all discrete distributions irrespective of support and support size.

For a fixed \( p \), let \( x^n_1 = (x_1, x_2, \ldots, x_n) \) be a sequence drawn \( i.i.d. \) according to \( p \). We denote the pattern of \( x^n_1 \) by \( \psi^n_1 \). The pattern is formed by taking \( \psi_1 = 1 \) and

\[
\psi_1 = \begin{cases} 
\psi_j & x_i = x_j, \ j < i \\
1 + \max_{j < i} \psi_j & x_i \neq x_j, \ \forall j < i
\end{cases}
\]

For example, the pattern of \( x^n_1 = \text{FEDERER} \) is \( \psi^n_1 = 1232424 \). Let \( \psi^n \) be the set of all patterns of length \( n \). We write \( p(\psi^n) \) for the probability that a length-\( n \) sequence generated by \( p \) has pattern \( \psi^n \). For a pattern \( \psi^n \) we write \( \phi_\mu \) for the number of symbols that appear \( \mu \) times in \( \psi^n_1 \) and \( m = \sum \phi_\mu \) is the number of distinct symbols in \( \psi^n_1 \). We call \( \phi_\mu \) the prevalence of \( \mu \). Thus for \( \text{FEDERER} \), we have \( \phi_1 = 2 \), \( \phi_2 = 1 \), and \( \phi_3 = 1 \), and \( m = 4 \).
2.1 Exchangeable partition processes

An exchangeable random partition refers to a sequence \((C_n : n \in \mathbb{N})\), where \(C_n\) is a random partition of the set \([n] = \{1, 2, \ldots, n\}\), satisfying the following conditions: (i) the probability that \(C_n\) is a particular partition depends only on the vector \((s_1, s_2, \ldots, s_n)\), where \(s_k\) is the number of parts in the partition of size \(k\), and (ii) the realizations of the sequence are consistent in that all the parts of \(C_n\) are also parts of the partition \(C_{n+1}\), except that the new element \(n + 1\) may either be in a new part of \(C_{n+1}\) by itself or has joined one of the existing parts of \(C_n\).

For a sequence \(X_1, \ldots, X_n\) from a discrete alphabet, one can partition the set \([n]\) into component sets \(\{A_x\}\) where \(A_x = \{i : X_i = x\}\) is the indices corresponding to the positions in which \(x\) has appeared. When such partitions are generated from i.i.d. data \(X_i\), the corresponding sequence of random partitions is called a paintbox process.

The remarkable Kingman representation theorem \([8]\) states that the probability measure induced by any exchangeable random partition is a mixture of paintbox processes, where the mixture is taken using a probability measure ("prior" in Bayesian terminology) on the class of paintbox processes. Since each paintbox process corresponds to a discrete probability measure (the one such that i.i.d. \(X_1\) drawn from it produced the paintbox process), the prior may be viewed as living on the set of probability measures on a countable alphabet.

2.2 Dirichlet priors and Chinese restaurant processes

Not surprisingly, special classes of priors give rise to special classes of exchangeable random partitions. One particularly nice class of priors on the set of probability measures on a countable alphabet is that of the Poisson-Dirichlet priors \([30, 5, 31]\) (sometimes called Dirichlet processes since they live on the infinite-dimensional space of probability measures and generalize the usual finite-dimensional Dirichlet distribution).

The Chinese restaurant process (or CRP) is related to the so-called Griffiths-Engen-McCloskey (GEM) distribution with parameter \(\theta\), denoted by \(\text{GEM}(\theta)\). Consider \(W_1, W_2, \ldots\) drawn i.i.d. according to a \(\text{Beta}(1, \theta)\) distribution, and set

\[
\begin{align*}
p_1 &= W_1 \\
p_i &= W_i \prod_{j<i} (1 - W_i) \quad \forall i > 1
\end{align*}
\]

This can be interpreted as follows: take a stick of unit length and break it into pieces of size \(W_1\) and \(1 - W_1\). Now take the piece of size \(1 - W_1\) and break off a \(W_2\) fraction of that. Continue in this way. The resulting lengths of the sticks create a distribution on a countably infinite set. The distribution of the sequence \(p = (p_1, p_2, \ldots)\) is the \(\text{GEM}(\theta)\) distribution.

**Remark 1.** Let \(\pi\) denote the elements of \(p\) sorted in decreasing order so that \(\pi_1 \geq \pi_2 \geq \cdots\). Then the distribution of \(\pi\) is the Poisson-Dirichlet distribution \(\text{PD}(\theta)\) as defined by Kingman.

Another popular class of distributions on probability vectors is the Pitman-Yor family of distributions \([29]\), also known as the two-parameter Poisson-Dirichlet family of distributions \(\text{PD}(\alpha, \theta)\). The two parameters here are a discount parameter \(\alpha \in [0, 1]\), and a strength parameter \(\theta > -\alpha\). The distribution \(\text{PD}(\alpha, \theta)\) can be generated in a similar way as the Poisson-Dirichlet distribution.
\( \text{PD}(\theta) = \text{PD}(0, \theta) \) described earlier. Let \( W_1, W_2, \ldots \) be drawn i.i.d. according to a Beta\((1 - \alpha, \theta + n\alpha)\) distribution, and again set

\[
\hat{p}_1 = W_1 \\ \\
\hat{p}_i = W_i \prod_{j<i} (1 - W_j) \quad \forall i > 1
\]

A similar “stick-breaking” interpretation holds here as well. Now let \( \tilde{p} \) be equal to the sequence \( \hat{p} \) sorted in descending order. The distribution of \( \tilde{p} \) is \( \text{PD}(\alpha, \theta) \).

### 2.3 Pattern probability estimators

Given a sample \( x_1^n \) with pattern \( \psi^n \) we would like to produce an pattern probability estimator. This is a function of the form \( q(\psi_{n+1} | \psi^n) \) that assigns a probability of seeing a symbol previously seen in \( \psi^n \) as well as a probability of seeing a new symbol. In this paper we will investigate two different pattern probability estimators based on Bayesian models.

The *Ewens sampling formula* \([32]\), which has its origins in theoretical population genetics, is a formula for the probability mass function of a marginal of a CRP corresponding to a fixed population size. In other words, it specifies the probability of an exchangeable random partition of \([n]\) that is obtained when one uses the Poisson-Dirichlet \( \text{PD}(\theta) \) prior to mix paintbox processes. Because of the equivalence between patterns and exchangeable random partitions, it estimates the probability of a pattern \( \psi^n \) via the following formula:

\[
q_{\theta}^{\text{CRP}}(\psi_1, \ldots, \psi_n) = \frac{\theta^m}{\theta(\theta + 1) \cdots (\theta + n - 1) \prod_{\mu=1}^n (\mu - 1)!} \phi_\mu. 
\]

Recall that \( \phi_\mu \) is the number of symbols that appear \( \mu \) times in \( \psi^n \). In particular, the predictive distribution associated to the Ewens sampling formula or Chinese restaurant process is

\[
q_{\theta}^{\text{CRP}}(\psi | \psi_1, \ldots, \psi_n) = \begin{cases} 
\frac{\mu}{n+\theta} & \psi \text{ appeared } \mu \text{ times in } \psi_1, \ldots, \psi_n; \\
\frac{\theta}{n+\theta} & \psi \text{ corresponds to new.}
\end{cases}
\]

More generally, one can define the Pitman-Yor predictor as

\[
q_{\alpha, \theta}^{\text{PY}}(\psi | \psi_1, \ldots, \psi_n) = \begin{cases} 
\frac{\mu-\alpha}{n+\theta} & \psi \text{ appeared } \mu \text{ times in } \psi_1, \ldots, \psi_n; \\
\frac{\theta + m\alpha}{n+\theta} & \psi \text{ corresponds to new.}
\end{cases}
\]

where \( m \) is the number of distinct symbols in \( \psi^n \). The probability assigned by the Pitman-Yor predictor to a pattern \( \psi^n \) is

\[
q_{\alpha, \theta}^{\text{PY}}(\psi_1, \ldots, \psi_n) = \frac{\theta(\theta + \alpha) \cdots (\theta + (m - 1)\alpha)}{\theta(\theta + 1) \cdots (\theta + n - 1)} \prod_{\mu=1}^n \left( \frac{\Gamma(\mu - \alpha)}{\Gamma(1 - \alpha)} \right)^{\phi_\mu}. 
\]

Note that \( \Gamma(\mu - \alpha)/\Gamma(1 - \alpha) = (\mu - \alpha - 1)(\mu - \alpha - 2) \cdots (1 - \alpha) \).
2.4 Worst-case and average redundancy

How should we measure the quality of a pattern probability predictor \( q \)? We investigate two criteria here: the worst-case and the average-case redundancy. The redundancy of \( q \) on a given pattern \( \psi^n \) is

\[
R(q) \overset{\text{def}}{=} \sup_{p \in \mathcal{I}} \log \frac{p(\psi^n)}{q(\psi^n)},
\]

(11)

The worst-case redundancy of \( q \) is defined to be

\[
\hat{R}(q) \overset{\text{def}}{=} \sup_{p \in \mathcal{I}} \max_{\psi^n \in \Psi^n} \log \frac{p(\psi^n)}{q(\psi^n)},
\]

(12)

Recall that \( p(\psi^n) \) just denotes the probability that a length-\( n \) sequence generated by \( p \) has pattern \( \psi^n \) — it is unnecessary to specify the support here. This is called worst-case because the internal supremum is taken over all patterns.

The average-case redundancy replaces the max over patterns with an expectation over \( p \):

\[
\bar{R}(q) \overset{\text{def}}{=} \sup_{p \in \mathcal{I}} \mathbb{E}_p \left[ \log \frac{p(\psi^n)}{q(\psi^n)} \right],
\]

(13)

\[
= \sup_{p \in \mathcal{I}} D(p \parallel q),
\]

(14)

where \( D(\cdot \parallel \cdot) \) is the Kullback-Leibler divergence or relative entropy. That is, the average-case redundancy is nothing but the worst-case Kullback-Leibler divergence between the distribution \( p \) and the predictor \( q \).

A pattern probability estimator is considered “good” if the worst-case or average-case redundancies are sublinear in \( n \). Succinctly put, redundancy that is sublinear in \( n \) implies that the underlying probability of a sequence can be estimated accurately almost surely.

3 Redundancy results

We now describe our main results on the redundancy of estimators derived from the prior distributions on \( \mathcal{I} \).

3.1 Chinese restaurant process predictors

Previously [33] it was shown by some of the authors that the worst-case and average-case redundancies for the CRP estimator are both \( \Omega(n \log n) \), which means it is neither strongly nor weakly universal. However, this estimator is not quite as bad as this result might suggest, and a simple twist yields an estimator that is weakly universal.

Our first new result is for the CRP estimator when we have a bound on the number \( m \) of distinct elements in the pattern \( \psi^n \).

**Theorem 1** (Redundancy for CRP estimators). Consider the estimator \( q^{\text{CRP}}_{\theta}(\psi^n_1) \) in (7) and (8). Then for sufficiently large \( n \) and for patterns \( \psi^n_1 \) whose number of distinct symbols \( m \) satisfies

\[
m \leq C \cdot \frac{n}{\log n} (\log \log n)^2,
\]

(15)
the redundancy of the predictor $q_\theta^{\text{CRP}}(\psi^n_1)$ with $|\theta| = m/\log n$ satisfies:

\[
\log \frac{p(\psi^n_1)}{q_\theta^{\text{CRP}}(\psi^n_1)} \leq 3C \cdot \frac{n(\log \log n)^3}{\log n} = o(n). \tag{16}
\]

**Proof.** The number of patterns with prevalences $\{\phi_\mu\}$ is

\[
\frac{n!}{\prod_{\mu=1}^{n} \mu!^{\phi_\mu} \phi_\mu!},
\]

and therefore

\[
p(\psi^n_1) \leq \frac{\prod_{\mu=1}^{n} [\mu!]^{\phi_\mu} \phi_\mu!}{n!}, \tag{17}
\]

since patterns with prevalences $\{\phi_\mu\}$ all have the same probability.

Using the upper bound on $p(\psi^n_1)$ and (7) yields

\[
\log \frac{p(\psi^n_1)}{q_\theta^{\text{CRP}}(\psi^n_1)} \leq \log \prod_{\mu=1}^{n} \left[ \frac{[\mu!]^{\phi_\mu} \phi_\mu!}{[(\mu - 1)!]^{\phi_\mu}} \right] + \log \frac{\theta(\theta + 1) \cdots (\theta + n - 1)}{\theta^n n!}
\]

\[
= \log \left( \prod_{\mu=1}^{n} \mu^{\phi_\mu} \right) + \log \left( \prod_{\mu=1}^{n} \phi_\mu! \right) + \log \frac{\theta(\theta + 1) \cdots (\theta + n - 1)}{\theta^n n!} \tag{18}
\]

Let $\bar{\theta} = [\theta]$. The following bound follows from Stirling’s approximation:

\[
\frac{\theta(\theta + 1) \cdots (\theta + n - 1)}{n!} \leq \frac{(\bar{\theta} + n)!}{\theta^n n!} \leq \frac{(\bar{\theta} + n)!}{\theta^{\bar{\theta} n} n!} \leq \left( \frac{\bar{\theta} + n}{\bar{\theta}} \right)^{\bar{\theta}} \left( \frac{\bar{\theta} + n}{n} \right)^n. \tag{19}
\]

The first term of (18) can be upper bounded by $\log(n/m)^m$ since the argument of the $\log(\cdot)$ is maximized over $\mu \in [1, n]$ when $\mu = n/m$. The second term is also maximized when all symbols appear the same number of times, corresponding to $\phi_\mu = m$ for one $\mu$. Therefore

\[
\log \frac{p(\psi^n_1)}{q_\theta^{\text{CRP}}(\psi^n_1)} \leq \log \left( \frac{n}{m} \right)^m + \log \frac{m!}{\theta^m} + \log \left( \frac{\bar{\theta} + n}{\bar{\theta}} \right)^{\bar{\theta}} \left( 1 + \frac{\bar{\theta}}{\bar{\theta}} \right)^n \tag{20}
\]

Now the last term can be upper bounded by $e^{\bar{\theta}}$ for sufficiently large $n$:

\[
\log \frac{p(\psi^n_1)}{q_\theta^{\text{CRP}}(\psi^n_1)} \leq \log \left( \frac{n}{m} \right)^m + \log \frac{m!}{\theta^m} + \log \left( \frac{(\bar{\theta} + n)e^{\bar{\theta}}}{\theta} \right)^{\bar{\theta}} \tag{21}
\]

Choose $\bar{\theta} = m/\log n$. This gives the bound:

\[
\log \frac{p(\psi^n_1)}{q_\theta^{\text{CRP}}(\psi^n_1)} \leq m \log \left( \frac{n}{m} \right) + \log \frac{m!}{m^m} \left( \frac{\bar{\theta}}{\theta} \right)^m + m \log n + \log \left( \frac{(\bar{\theta} + n)e^{\bar{\theta}}}{\theta} \right)^{\bar{\theta}}
\]
the second term is negative for sufficiently large \( m \). Therefore

\[
\log \frac{p(\psi^n)}{q_{\theta}^{\text{CRP}}(\psi^n)} \leq m \log \left( \frac{n}{m} \right) + m \log \log n + \frac{m}{\log n} \log \left( 2 + \frac{n \log n}{m} \right)
\]  

(22)

Now choosing

\[
m = C \frac{n}{\log n} (\log \log n)^2,
\]

the bound becomes

\[
\log \frac{p(\psi^n)}{q_{\theta}^{\text{CRP}}(\psi^n)}
\leq Cn \frac{(\log \log n)^2}{\log n} \log \left( \frac{\log n}{(\log \log n)^2} \right) + Cn \frac{(\log \log n)^3}{\log n} + Cn \left( \frac{\log \log n}{\log n} \right)^2 \log \left( 2 + \left( \frac{\log n}{\log \log n} \right)^2 \right)
\]

\[
\leq 3Cn \frac{(\log \log n)^3}{\log n}
\]

\[
= o(n).
\]

This theorem is slightly dissatisfying, since it requires us to have a bound on \( m \). It turns out that by taking mixtures of CRP estimators we can arrive at an estimator that is weakly universal. That is, let \( \tilde{q}_{m,n}^{\text{CRP}}(\cdot) \) be the CRP estimator with \( \theta = m/\log n \). Then define

\[
q^*(\cdot) = \sum_m c_{n,m} \tilde{q}_{m,n}^{\text{CRP}}(\cdot)
\]

(23)

for a set of positive coefficients \( c_{n,m} \) that sum to 1:

\[
\sum_m c_{n,m} = 1.
\]

(24)

We will furthermore assume that there exists a \( \beta \geq 0 \) such that for any \( n \),

\[
\min_m c_{n,m} \geq n^{-\beta}.
\]

(25)

We can, for example, choose \( c_{m,n} = \frac{36}{m^2 n^2 \pi^2} \). It is clear that \( q^* \) is a pattern probability estimator.

**Lemma 1.** For all discrete i.i.d. processes \( P \) with entropy rate \( H \), let \( M_n \) be the random variable counting the number of distinct symbols in a sample of length \( n \) drawn from \( P \). The following bound holds

\[
\mathbb{E}[M_n] \leq \frac{nH}{\log n} + 1.
\]

(26)
Proof. Let $P(i) = p_i$. Writing out the expectation of $M_n$, we see:

$$
E[M_n] = \sum_{i=1}^{\infty} (1 - (1 - p_i)^n)
$$

$$
= \sum_{i=1}^{\infty} \frac{p_i}{1 - (1 - p_i)} \frac{1 - (1 - p_i)^n}{1 - (1 - p_i)}
$$

$$
= \sum_{i=1}^{\infty} p_i \sum_{j=0}^{n-1} (1 - p_i)^j
$$

$$
\leq 1 + \sum_{i=1}^{\infty} p_i \sum_{j=1}^{n-1} (1 - p_i)^j.
$$

Now, using the fact that $\sum_{k=1}^{n-1} \frac{1}{k} \geq \log n$,

$$
E[M_n] \log n \leq \log n + \sum_{i=1}^{\infty} p_i \left( \sum_{j=1}^{n-1} (1 - p_i)^j \right) \left( \sum_{k=1}^{n-1} \frac{1}{k} \right)
$$

$$
\leq \log n + \sum_{i=1}^{\infty} p_i n \sum_{j=1}^{n-1} \frac{(1 - p_i)^j}{j}
$$

$$
\leq \log n + n \sum_{i=1}^{\infty} p_i \sum_{j=1}^{\infty} \frac{(1 - p_i)^j}{j}
$$

$$
= \log n + n \sum_{i=1}^{\infty} -p_i \log(1 - (1 - p_i))
$$

$$
= \log n + nH.
$$


\[\square\]

**Theorem 2** (Weak universality for CRP mixtures). For all discrete i.i.d. processes $p \in \mathcal{I}$ with finite entropy rate,

$$
D(p \parallel q^*) = o(n).
$$

That is, $q^*$ is weakly universal.

Proof. We first apply Markov's inequality to $M_n$ from the previous lemma:

$$
\mathbb{P} \left( M_n > \frac{n(\log \log n)^2}{\log n} \right) \leq \frac{\log n}{n(\log \log n)^2} \left( \frac{nH}{\log n} + 1 \right)
$$

$$
\leq \frac{H}{(\log \log n)^2} + \frac{\log n}{n(\log \log n)^2}.
$$

(28)

Therefore for all finite entropy processes, this probability goes to 0 as $n \to \infty$. 

9
From (22) and (25) for sufficiently large $n$ we have
\[
\log \frac{p(\psi^n_1)}{c_{m,n}q_{m,n}(\psi^n_1)} \leq \log \frac{1}{\min_m c_{m,n}} + m \log \frac{n}{m} + m \log \log n + 2m \\
\leq \beta \log n + m \log \frac{n}{m} + m \log \log n + 2m \\
\leq 3n \log \log n.
\]

Then we can upper bound the divergence using the preceding:
\[
\mathbb{E}_p \left[ \log \frac{p(\psi^n_1)}{\sum_m c_{m,n}q_{m,n}(\psi^n_1)} \right] \leq \mathbb{P} \left( M_n > \frac{n(\log \log n)^2}{\log n} \right) 3n \log \log n \\
+ \mathbb{P} \left( M_n \leq \frac{n(\log \log n)^2}{\log n} \right) \mathbb{E}_p \left[ \log \frac{p(\psi^n_1)}{c_{m^*,n}q_{m^*,n}(\psi^n_1)} \right]
\]
where $m^* = \frac{n(\log \log n)^2}{\log n}$. Then using (28) on the first term and Theorem 1 on the second,
\[
\mathbb{E}_p \left[ \log \frac{p(\psi^n_1)}{\sum_m c_{m,n}q_{m,n}(\psi^n_1)} \right] \leq \frac{3H n}{(\log \log n)} + \frac{3 \log n}{\log \log n} + \frac{3C n(\log \log n)^3}{\log n} + \beta \log n.
\]
This upper bound is $o(n)$.

What the preceding theorem shows is that the mixture of CRP estimators $q^*$ is weakly universal. However, note that $q^*$ is not itself a CRP estimator.

### 3.2 Pitman-Yor predictors

We now turn to the more general class of Pitman-Yor predictors. We can obtain a similar result as for the CRP estimator, but we can handle patterns with $m = o(n)$.

**Theorem 3** (Worst-case redundancy). Consider the estimator $q_{\alpha,\theta}^{PY}(\psi^n_1)$. Then for sufficiently large $n$ and for patterns $\psi^n_1$ whose number of distinct symbols $m$ satisfies $m = o(n)$, the worst-case redundancy of the predictor $q_{\alpha,\theta}^{PY}(\psi^n_1)$ with $\theta = m/\log n$ satisfies:
\[
\log \frac{p(\psi^n_1)}{q_{\alpha,\theta}^{PY}(\psi^n_1)} = o(n).
\]

**Proof.** For a pattern $\psi^n_1$, from the definition of $q_{\alpha,\theta}^{PY}(\psi^n_1)$ in (10) and (17),
\[
\log \frac{p(\psi^n_1)}{q_{\alpha,\theta}^{PY}(\psi^n_1)} \leq \log \left( \frac{\prod_{\mu=1}^{n}[\mu!]\phi^n_\mu!}{n!} \cdot \frac{(\theta + 1) \cdots (\theta + n - 1)}{(\theta + \alpha)(\theta + 2\alpha) \cdots (\theta + m\alpha)} \prod_{\mu=1}^{n} \left( \frac{\Gamma(1 - \alpha)}{\Gamma(\mu - \alpha - 1)} \phi^n_{\mu} \right) \right).
\]

We can bound the components separately. First, as before we have:
\[
\prod_{\mu=1}^{n} \phi^n_\mu! \leq m!
\]
Since \( \theta > -\alpha \), we have \( \theta + \alpha > 0 \) and

\[
(\theta + \alpha)(\theta + 2\alpha) \cdots (\theta + (m-1)\alpha) \geq (\theta + \alpha)\alpha(2\alpha) \cdots ((m-2)\alpha)
= (\theta + \alpha)(m-2)!\alpha^{m-2}.
\]

Again, letting \( \tilde{\theta} = [\theta] \), from the same arguments as in (19)-(21),

\[
\log \frac{\theta(\theta + 1) \cdots (\theta + n - 1)}{n!} \leq \tilde{\theta} \log \frac{(\tilde{\theta} + n)e}{\tilde{\theta}}.
\]

Finally, note that \((1 - \alpha)(2 - \alpha) \cdots (\mu - 1 - \alpha) \geq (1 - \alpha)(\mu - 2)!\), so

\[
\prod_{\mu=1}^{n} \mu!^{\phi_{\mu}} (1 - \alpha)(\mu - 2)! = \left( \prod_{\mu=1}^{n} \frac{\mu!}{(1 - \alpha)(\mu - 2)!} \right)^{\phi_{\mu}} \leq \left( \prod_{\mu=1}^{n} \frac{\mu!}{(1 - \alpha)(\mu - 2)!} \right)^{2\phi_{\mu}} \leq \frac{2^{m}}{(1 - \alpha)^{m}}.
\]

Putting this together:

\[
\log \frac{p(\psi_{1}^{n})}{q_{\alpha,\tilde{\theta}}(\psi_{1}^{n})} \leq \log \frac{m!}{(\theta + \alpha)(m-2)!\alpha^{m-2}} + \tilde{\theta} \log \frac{(\tilde{\theta} + n)e}{\tilde{\theta}} + \log \frac{(n/m)^{2m}}{(1 - \alpha)^{m}} + 2m \log \frac{n}{m} + (m - 2) \log \frac{1}{(1 - \alpha)\alpha} + \tilde{\theta} \log \frac{(\tilde{\theta} + n)e}{\tilde{\theta}} + \log \frac{m^{2}}{(\theta + \alpha)} + \log \frac{1}{(1 - \alpha)^{2}}.
\]

If \( m = o(n) \) then this is less than \( o(n) \), as desired.

It is well known that the Pitman-Yor process can produce patterns whose relative frequency is 0, e.g. the pattern \( 1^k 2^{n-k} \cdots (n-k) \). Therefore, it is not surprising that the worst-case redundancy and average case redundancies can be bad. However, as the next theorem shows, the actual redundancy of the Pitman-Yor estimator is \( \Theta(n) \), which is significantly better than the lower bound of \( \Omega(n \log n) \) proved in [33] for Chinese restaurant processes.

**Theorem 4 (Redundancies).** Consider the estimator \( q_{\alpha,\tilde{\theta}}^{PY}(\psi_{1}^{n}) \). Then for sufficiently large \( n \), the worst-case redundancy and average case redundancy satisfy:

\[
\hat{R}(q_{\alpha,\tilde{\theta}}^{PY}(\cdot)) = \Theta(n) \quad (32)
\]
\[
\tilde{R}(q_{\alpha,\tilde{\theta}}^{PY}(\cdot)) = \Theta(n). \quad (33)
\]

That is, \( q_{\alpha,\tilde{\theta}}^{PY}(\cdot) \) is neither strongly nor weakly universal.

**Proof.** For the upper bound, we start with (31) and note that in the worst case, \( m = O(n) \) so \( \hat{R}(q_{\alpha,\tilde{\theta}}^{PY}(\cdot)) = O(n) \) and a fortiori \( \tilde{R}(q_{\alpha,\tilde{\theta}}^{PY}(\cdot)) = O(n) \).
For the lower bound, consider the patterns $1 \cdots 1$ and $12 \cdots n$. For the Pitman-Yor estimator,

$$q_{\alpha, \theta}^{\text{PY}}(1 \cdots 1)q_{\alpha, \theta}^{\text{PY}}(12 \cdots n) = \frac{\theta(1 - \alpha) \cdots (n - 1 - \alpha) \theta(\theta + \alpha) \cdots (\theta + (n - 1)\alpha)}{\theta(\theta + 1) \cdots (\theta + n - 1) \theta(\theta + 1) \cdots (\theta + n - 1)}$$

$$= \frac{(1 - \alpha)(\theta + \alpha)}{(\theta + 1)^2} \cdot \frac{(2 - \alpha)(\theta + 2\alpha)}{(\theta + 2)^2} \cdots \frac{(n - 1 - \alpha)(\theta + (n - 1)\alpha)}{(\theta + n - 1)^2}$$

Now for $j \geq 2$:

$$\frac{(j - \theta)(\theta + j\alpha)}{(\theta + j)^2} \leq \frac{j\theta + j^2\alpha}{j^2 + 2j\theta}$$

$$\leq \frac{\theta + j\alpha}{2\theta + j}$$

$$\leq \max\left\{\frac{1}{2}, \alpha\right\}.$$ 

So each such term is less than 1. Therefore for $\alpha < 1$, there exists a constant $0 < c < 1$ such that

$$q_{\alpha, \theta}^{\text{PY}}(1 \cdots 1)q_{\alpha, \theta}^{\text{PY}}(12 \cdots n) \leq c^n.$$

Thus

$$\log \frac{1}{q_{\alpha, \theta}^{\text{PY}}(1 \cdots 1)} + \log \frac{1}{q_{\alpha, \theta}^{\text{PY}}(12 \cdots n)} \geq n \log \frac{1}{c}.$$ 

Let the distribution $p_1$ be a singleton, so $p_1(1 \cdots 1) = 1$. For any small $\delta > 0$ we can find a distribution $p_n$ such that $p_n(12 \cdots n) = 1 - \delta$ by choosing $p_n$ to be uniform on a sufficiently large set. Thus

$$\hat{R}(q_{\alpha, \theta}^{\text{PY}}(\cdot)) \geq \max \left\{ \log \frac{1 - \delta}{q_{\alpha, \theta}^{\text{PY}}(1 \cdots 1)}, \log \frac{1 - \delta}{q_{\alpha, \theta}^{\text{PY}}(12 \cdots n)} \right\}$$

$$\geq \frac{1}{2} \left( \log \frac{1}{q_{\alpha, \theta}^{\text{PY}}(1 \cdots 1)} + \log \frac{1}{q_{\alpha, \theta}^{\text{PY}}(12 \cdots n)} \right) + \log(1 - \delta)$$

$$\geq \frac{n}{2} \log \frac{1}{c} + \log(1 - \delta)$$

This shows that $\hat{R}(q_{\alpha, \theta}^{\text{PY}}(\cdot)) = \Omega(n)$. Furthermore,

$$\hat{R}(q_{\alpha, \theta}^{\text{PY}}(\cdot)) \geq \max \left\{ (1 - \delta) \frac{1 - \delta}{q_{\alpha, \theta}^{\text{PY}}(1 \cdots 1)}, (1 - \delta) \frac{1 - \delta}{q_{\alpha, \theta}^{\text{PY}}(12 \cdots n)} \right\}$$

$$\geq (1 - \delta) \left( \frac{n}{2} \log \frac{1}{c} + \log(1 - \delta) \right),$$ 

so $\hat{R}(q_{\alpha, \theta}^{\text{PY}}(\cdot)) = \Omega(n)$.
4 Conclusions and future work

In this note we investigated the worst-case and average-case redundancies of pattern probability estimators derived from priors on $I$ that are popular in Bayesian statistics. Both the CRP and Pitman-Yor estimators give a vanishing redundancy per symbol for patterns whose number of distinct symbols $m$ is sufficiently small. The Pitman-Yor estimator requires only that $m = o(n)$, which is an improvement on the CRP. However, when $m$ can be arbitrarily large (or the alphabet size is arbitrarily large) the worst-case and average-case redundancies do not scale like $o(n)$. Here again, the Pitman-Yor estimator is superior, in that the redundancies scale like $\Theta(n)$ as opposed to the $\Omega(n \log n)$ for the CRP estimator. While these results show that these estimators are not strongly universal, we constructed a mixture of CRP process (which is not itself a CRP estimator) that is weakly universal.

On the other hand, one of the estimators derived in [16] is exchangeable and has near optimal worst case redundancy, growing as $O(\sqrt{n})$. From Kingman’s results, this estimator can be obtained using a prior on $I$—however, this prior is yet unknown. Finding this prior may potentially reveal new interesting classes of priors other than the Poisson-Dirichlet priors.

Acknowledgments

The authors thank Persi Diaconis, Miroslav Dudik, Fan Chung Graham, Ron Graham, Susan Holmes, Olgica Milenkovic, Alon Orlitsky, Ofer Shayevitz, Krishna Viswanathan, Aaron Wagner and Junan Zhang for helpful conversations related to the paper, as well as the American Institute of Mathematics and NSF for sponsoring a workshop on probability estimation topics. In particular, the authors also thank Alon Orlitsky and Krishna Viswanathan for being co-organizers of the above workshop along with one of the authors.

N. Santhanam was supported by a startup grant from the University of Hawaii and NSF Grant CCF-1018984. M. Madiman was supported by a Junior Faculty Fellowship from Yale University. A.D. Sarwate was supported by the California Institute for Telecommunications and Information Technology (CALIT2) at the University of California, San Diego.

References


