Supplementary Materials for "An Alternative Prior Process for Nonparametric Bayesian Clustering"

Hanna M. Wallach
Department of Computer Science
University of Massachusetts Amherst

Shane T. Jensen
Department of Statistics
The Wharton School, University of Pennsylvania

Lee Dicker
Department of Biostatistics
Harvard School of Public Health

Katherine A. Heller
Engineering Department
University of Cambridge

1 Proof of Law for $\mathbb{E}\left(K_N \,|\, \mathrm{UN}\right)$

We start by defining $T_k = \inf\{m > T_{k-1};\; X_m \notin \{X_1, \ldots, X_{m-1}\}\}$. T_k is the "waiting time" (number of observations needed) until the k^{th} new cluster is generated by the uniform process. Under the uniform process, $T_k = \sum_{i=1}^{k} \tau_i$ where $\tau_i \sim \text{Geometric}\left(\theta \,/\, (\theta + i - 1\right)$ and the τ_i variables are independent, so

$$\mathbb{E}\left(T_k\right) = \sum_{i=1}^{k} \frac{\theta + i - 1}{\theta} = \frac{k^2}{2\theta} + k\left(1 - \frac{1}{2\theta}\right)$$

and

$$\begin{aligned} \mathrm{Var}\left(T_k\right) &= \sum_{i=1}^{k} \frac{\left(\theta + i - 1\right)\left(i - 1\right)}{\theta^2} &(1)\\ &= \frac{k^3}{3\theta^2} + k^2 \frac{1}{2\theta}\left(1 - \frac{1}{\theta}\right) + k \frac{1}{2\theta}\left(\frac{1}{3\theta} - 1\right). \end{aligned}$$

In terms of T_k, $K_N = \max\{k;\; T_k \leq N\} = \sum_{k=1}^{N} \mathrm{I}\left(T_k \leq N\right)$. We first prove a strong law for the convergence of T_k. Let $\epsilon > 0$. From Chebychev's inequality and (1), we have the following:

$$\mathrm{P}\left(\left|T_k - \mathbb{E}\left(T_k\right)\right| > \epsilon k^2\right) \leq \frac{\mathrm{Var}\left(T_k\right)}{\epsilon^2 k^4} \leq \frac{C(\theta, \epsilon)}{k}. \quad (2)$$

From (2),

$$\mathrm{P}\left(\left|T_{k^2} - \mathbb{E}\left(T_{k^2}\right)\right| > \epsilon k^4\right) \leq \frac{C(\theta, \epsilon)}{k^2},$$

and so by the Borel-Cantelli lemma, we have $\mathrm{P}\left(\left|T_{k^2} - \mathbb{E}\left(T_{k^2}\right)\right| > \epsilon k^4\right) = 0$. Since $\epsilon > 0$ was chosen arbitrarily, it follows that $\frac{T_{k^2} - \mathbb{E}\left(T_{k^2}\right)}{k^4} \to 0$ almost surely and hence $\frac{T_{k^2}}{k^4} \to \frac{1}{2\theta}$ almost surely. Now, let $m = \lfloor \sqrt{k} \rfloor$. Since T_k is increasing, we have:

$$\frac{T_{m^2}}{(m+1)^4} \leq \frac{T_k}{k^2} \leq \frac{T_{(m+1)^2}}{m^4}. \quad (3)$$

Since $\frac{m+1}{m} \to 1$, both sides of the inequality (3) converge to $(2\theta)^{-1}$ almost surely, and so

$$\frac{T_k}{k^2} \to \frac{1}{2\theta} \text{ almost surely.} \quad (4)$$

The strong law (4) implies a strong law for K_N as follows. $T_{K_N} \leq N < T_{K_N + 1}$ and, consequently,

$$\frac{T_{K_N}}{K_N^2} \leq \frac{N}{K_N^2} < \frac{T_{K_N+1}}{K_N^2}.$$

Since $K_N \to \infty$ almost surely and $T_k \,/\, k^2 \to 1 \,/\, (2\theta)$ almost surely, it follows that the left and right hand side above both converge to $1 \,/\, (2\theta)$ almost surely. Thus, $K_N^2 \,/\, N \to 2\theta$ almost surely and so

$$\frac{K_N}{\sqrt{N}} \to \sqrt{2\theta} \text{ almost surely.} \quad (5)$$

From the strong law (5) and the dominated convergence theorem, we have the following:

$$\frac{\mathbb{E}\left(K_N\right)}{N} \to 0. \quad (6)$$

Combining (6) with following result from section 2,

$$\mathbb{E}\left(K_N^2\right) = \mathbb{E}\left(K_N\right) + 2\theta\left(N - \mathbb{E}\left(K_N\right)\right). \quad (7)$$

gives us

$$\frac{\mathbb{E}\left(K_N^2\right)}{N} \to 2\theta. \quad (8)$$

Finally, using (8) together with Fatou's lemma and Jensen's inequality, gives us the following:

$$\begin{aligned} \sqrt{2\theta} \leq \liminf_{N \to \infty} \frac{\mathbb{E}\left(K_N\right)}{\sqrt{N}} &\leq \limsup_{N \to \infty} \frac{\mathbb{E}\left(K_N\right)}{\sqrt{N}} \\ &\leq \limsup_{N \to \infty} \sqrt{\frac{\mathbb{E}\left(K_N^2\right)}{N}} = \sqrt{2\theta}. \end{aligned}$$

This then proves the result

$$\frac{\mathbb{E}\left(K_N\right)}{\sqrt{N}} \rightarrow \sqrt{2\theta}$$

under the uniform process.

2 Result relating $\mathbb{E}\left(K_N\right)$ to $\mathbb{E}\left(K_N^2\right)$

Recall the definition of T_k from above and now define $M_N = K_N + 1$. Consider the "waiting time" T_{M_N} until the observation that creates the $(K_N+1)^{\text{th}}$ unique cluster. We relate $\mathbb{E}\left(K_N\right)$ to $\mathbb{E}\left(K_N^2\right)$ by calculating $\mathbb{E}\left(T_{M_N}\right)$ in two different ways. First, observe that

$$\begin{aligned}
\mathbb{E}\left(T_{M_N}\right) &= \mathbb{E}\left(\sum_{k=1}^{\infty} \tau_k \cdot \mathrm{I}\left(k \leq M_N\right)\right) \\
&= \frac{\theta-1}{\theta} \sum_{k=1}^{\infty} \mathrm{P}\left(k \leq M_N\right) \\
&\quad + \frac{1}{\theta} \sum_{k=1}^{\infty} k \cdot \mathrm{P}\left(k \leq M_N\right) \\
&= \frac{\theta-1}{\theta}\, \mathbb{E}\left(M_N\right) + \frac{1}{2\theta}\, \mathbb{E}\left(M_N(M_N+1)\right),
\end{aligned}$$

which, since $M_N = K_N + 1$, simplifies to

$$\mathbb{E}\left(T_{M_N}\right) = 1 + \mathbb{E}\left(K_N\right)\left(1 + \frac{1}{2\theta}\right) + \mathbb{E}\left(K_N^2\right)\frac{1}{2\theta}. \quad (9)$$

Now $T_{M_N} = N + \sum_j \mathrm{I}\left(M_{N+j} = M_N\right)$ and so $\mathbb{E}\left(T_{M_N}\right) = N + \sum_j \mathrm{P}\left(M_{N+j} = M_N\right)$ where

$$\begin{aligned}
\mathrm{P}\left(M_{N+j} = M_N\right) &= \sum_k \mathrm{P}\left(T_k \leq N,\ N + j < T_{k+1}\right) \\
&= \sum_k \mathrm{P}\left(M_N = k+1\right) \mathrm{P}\left(j < \tau_{k+1}\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\left(T_{M_N}\right) &= n + \sum_j \sum_k \mathrm{P}\left(M_N = k+1\right) \mathrm{P}\left(j < \tau_{k+1}\right) \\
&= N + \sum_k \mathrm{P}\left(M_N = k+1\right) \mathbb{E}\left(\tau_{k+1}\right) \\
&= N + \sum_k \mathrm{P}\left(K_N = k\right) \frac{k+\theta}{\theta},
\end{aligned}$$

which can be simplified to

$$\mathbb{E}\left(T_{M_N}\right) = N + 1 + \mathbb{E}\left(K_N\right)\frac{1}{\theta}. \quad (10)$$

Combining (9) and (10) gives (7):

$$\mathbb{E}\left(K_N^2\right) = \mathbb{E}\left(K_N\right) + 2\theta\left(N - \mathbb{E}\left(K_N\right)\right).$$

3 Evaluation Algorithm

The evaluation algorithm used to approximate $\log P(\mathcal{W}^{\text{test}} \mid \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$ is based on the "left-to-right" evaluation algorithm introduced by Wallach *et al.* (2009), adapted to marginalize out test cluster assignments. Pseudocode is given in algorithm 1.

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by CIA, NSA and NSF under NSF grant #IIS-0326249, and in part by subcontract #B582467 from Lawrence Livermore National Security, LLC, prime contractor to DOE/NNSA contract #DE-AC52-07NA27344. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *26th International Conference on Machine Learning.*

initialize $l := 0$
for each document d in $\mathcal{W}^{\text{test}}$ **do**
  initialize $p_d := 0$
  for each particle $r = 1$ to R **do**
    for $d' < d$ **do**
      $c_{d'}^{(r)} \sim P(c_{d'}^{(r)} \,|\, \mathcal{W}_{<d}^{\text{test}}, \{\boldsymbol{c}_{<d}^{(r)}\}_{\setminus d'}, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$
    end for
    $p_d := p_d + \sum_c P(\boldsymbol{w}_d^{\text{test}}, c_d^{(r)} \!=\! c \,|\, \mathcal{W}_{<d}^{\text{test}}, \boldsymbol{c}_{<d}^{(r)}, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$
    $c_d^{(r)} \sim P(c_d^{(r)} \,|\, \boldsymbol{w}_d^{\text{test}}, \mathcal{W}_{<d}^{\text{test}}, \boldsymbol{c}_{<d}^{(r)}, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$
  end for
  $p_n := p_n \,/\, R$
  $l := l + \log p_n$
end for
$\log P(\mathcal{W}^{\text{test}} \,|\, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta}) \simeq l$

Algorithm 1: "Left-to-right" evaluation algorithm for computing $\log P(\mathcal{W}^{\text{test}} \,|\, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$.