Supplement to: Scaling the Indian Buffet Process via Submodular Maximization

Colorado Reed CR478@CAM.AC.UK
Zoubin Ghahramani ZOUBIN@ENG.CAM.AC.UK
Engineering Department, Cambridge University, Cambridge UK

S.1. Truncated Gaussian Properties

In the main text we examined a truncated Gaussian of the form:

$$T\mathcal{N}(\tilde{\mu}_{kd}, \tilde{\sigma}^2_{kd}) = \frac{2}{\operatorname{erfc}\left(-\frac{\tilde{\mu}_{kd}}{\tilde{\sigma}_{kd}\sqrt{2}}\right)}\mathcal{N}(\tilde{\mu}_{kd}, \tilde{\sigma}^2_{kd}) \tag{1}$$

with $\mathcal{N}$ representing a Gaussian distribution. The first two moments of $T\mathcal{N}(\tilde{\mu}_{kd}, \tilde{\sigma}^2_{kd})$ are:

$$\mathbb{E}\left[a_{kd}\right] = \tilde{\mu}_{kd} + \tilde{\sigma}_{kd}\frac{\sqrt{2/\pi}}{\operatorname{erfcx}\left(\wp_{kd}\right)} \tag{2}$$

$$\mathbb{E}\left[a^2_{kd}\right] = \tilde{\mu}^2_{kd} + \tilde{\sigma}^2_{kd} + \tilde{\sigma}_{kd}\tilde{\mu}_{kd}\frac{\sqrt{2/\pi}}{\operatorname{erfcx}\left(\wp_{kd}\right)} \tag{3}$$

with $\wp_{kd} = -\frac{\tilde{\mu}_{kd}}{\tilde{\sigma}_{kd}\sqrt{2}}$ and $\operatorname{erfcx}(y) = e^{y^2}(1 - \operatorname{erf}(y))$ representing the scaled complementary error function. The entropy is

$$H(q(a_{kd})) = \frac{1}{2}\ln\frac{\pi e\tilde{\sigma}^2_{kd}}{2} + \ln\operatorname{erfc}\left(-\frac{\tilde{\mu}_{kd}}{\tilde{\sigma}_{kd}\sqrt{2}}\right) \tag{4}$$

$$+ \frac{\tilde{\mu}_{kd}}{\tilde{\sigma}_{kd}}\sqrt{\frac{1}{2\pi}}\left(\operatorname{erfcx}\left(-\frac{\tilde{\mu}_{kd}}{\tilde{\sigma}_{kd}\sqrt{2}}\right)\right)^{-1}. \tag{5}$$

S.2. Shifted Equivalence Classes

Here we discuss the "shifted" equivalence class of binary matrices first proposed by Ding et al. (2010). For a given $N \times K$ binary matrix $\boldsymbol{Z}$, the equivalence class for this binary matrix $[\boldsymbol{Z}]$ is obtained by shifting all-zero columns to the right of the non-zero columns while maintaining the non-zero column orderings, see Figure 1. Placing independent Beta$(\frac{\alpha}{K}, 1)$ priors on the Bernoulli entries of $\boldsymbol{Z}$ and integrating over these priors yields the following probability for $\boldsymbol{Z}$, see Eq. 27 in Griffiths & Ghahramani (2005):

$$P(\boldsymbol{Z}) = \prod_{k=1}^{K}\frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})} \tag{6}$$

where $m_k = \sum_{n=1}^{N} z_{nk}$. Letting $K \to \infty$ yields $P(\boldsymbol{Z}) = 0$ for all $\boldsymbol{Z}$. However, the probability of certain equivalence classes of binary matrices, $P([\boldsymbol{Z}])$, can remain non-zero as $K \to \infty$. Specifically, Griffiths & Ghahramani (2005) show $P([\boldsymbol{Z}])$ remains non-zero for the "left-ordered form" equivalence class of binary matrices, whereby the columns of $\boldsymbol{Z}$ are ordered such that the binary values of the columns are non-increasing, where the first row is the most significant bit. Here we outline a similar result for the shifted equivalence class.[1]

We obtain the probability of the shifted equivalence class by multiplying the multiplicity of the equivalence class by the probability of a matrix within the class. For a given matrix with K columns and K_+ non-zero columns, each shifted equivalence class has $\binom{K}{K_+}$ matrices that map to it, yielding:

$$P\left([\boldsymbol{Z}]\right) = \binom{K}{K_+}\prod_{k=1}^{K}\frac{\frac{\alpha}{K}\Gamma(m_k + \frac{\alpha}{K})\Gamma(N - m_k + 1)}{\Gamma(N + 1 + \frac{\alpha}{K})}. \tag{7}$$

Following a similar algebraic rearrangement as Griffiths & Ghahramani (2005) Eqs. 30-33, except replacing the $\frac{K!}{\prod_{h=0}^{2^N-1} K_h!}$ term with $\binom{K}{K_+}$—which occurs because of the different equivalence class multiplicities—results in:

$$P([\boldsymbol{Z}]) = \frac{\alpha^{K_+}}{K_+!}\cdot\frac{K!}{(K-K_+)!K^{K_+}}\cdot\left(\frac{N!}{\prod_{j=1}^{2^N-1}(j + \frac{\alpha}{K})}\right)^{K}$$

$$\cdot\prod_{k=1}^{K_+}\frac{(N - m_k)!\prod_{j=1}^{m_k-1}(j + \frac{\alpha}{K})}{N!}. \tag{8}$$

We then take the limit $K \to \infty$ for each of the four terms. The first term has no K dependence and does

[1] Ding et al. (2010) proposed this equivalence class but did not explicitly show that it remains well defined as $K \to \infty$. Furthermore, they did not discuss the collapsed case where we first marginalize over the beta priors on $\boldsymbol{Z}$.

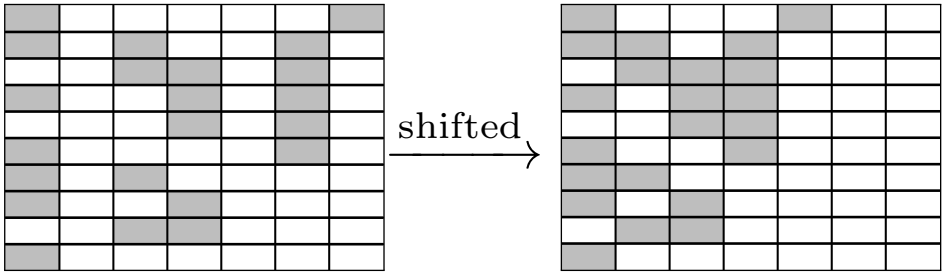

Figure 1. Example of a binary matrix (left) and its shifted equivalence matrix (dark squares are 1, white squares are 0)—placing the two all-zero columns anywhere in the matrix will yield the same equivalence matrix.

not change in the infinite limit. For the second term we let $K_0 = K - K+$ and have $\frac{K!}{K_0! K^{K_+}}$. Equations 60-62 in Griffiths & Ghahramani (2005) show that this term becomes 1 as $K \to \infty$. The infinite limit of the third and fourth terms are determined in the Appendix of Griffiths & Ghahramani (2005). Combining all four terms together yields:

$$P([\boldsymbol{Z}]) = \frac{\alpha^{K_+}}{K_+!} \mathrm{e}^{-\alpha H_N} \prod_{k=1}^{K^+} \frac{(N-m_k)!(m_k-1)!}{N!} \tag{9}$$

where H_N is the N^{th} harmonic number.

The probability of the shifted equivalence class is nearly identical to the probability of the left-ordered-form equivalence class:

$$P([\boldsymbol{Z}]_{\text{lof}}) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^N-1} K_h!} \mathrm{e}^{-\alpha H_N} \prod_{k=1}^{K^+} \frac{(N-m_k)!(m_k-1)!}{N!}, \tag{10}$$

where K_h is the number of columns of $\boldsymbol{Z}$ with binary value $h \in \{1, \ldots, 2^{N-1}\}$ when the first row is taken to be the most significant bit. The only difference between Eq. 9 and Eq. 10 is the denominator of the first fraction. For the left-ordered-form, this term penalizes $\boldsymbol{Z}$ matrices with identical columns. In the feature assignment view, this term penalizes features that are assigned to the exact same set of observations. The $K_+!$ term in the shifted equivalence class prior does not distinguish between identical and distinct columns of $\boldsymbol{Z}$, and in turn, does not penalize repeated feature assignments. These two equivalence class probabilities are proportional in the limit of large N as the probability of two columns being identical approaches 0.

S.3. Hyperparameter Inference

In the main text we assumed the hyperparameters $\boldsymbol{\theta} = \{\sigma_{\boldsymbol{X}}, \sigma_{\boldsymbol{A}}, \alpha\}$ were known (i.e. estimated from the data). Placing conjugate gamma hyperpriors on these parameters allows for a straightforward extension in which we infer their values. Formally, let

$$p(\tau_X) = \text{Gamma}(\tau_X; a_X, b_X) \tag{11}$$
$$p(\tau_A) = \text{Gamma}(\tau_A; a_A, b_A) \tag{12}$$
$$p(\alpha) = \text{Gamma}(\alpha; a_\alpha, b_\alpha) \tag{13}$$

where τ represents the precision, equivalent to the inverse variance $\frac{1}{\sigma^2}$, for the variance parameter indicated in the subscript. Update equations for the variational distributions follow from standard update equations for variational inference in exponential families, cf. Attias (2000), and yield:

$$q(\tau_X) = \text{Gamma}(\widetilde{a}_X, \widetilde{b}_X) \tag{14}$$
$$q(\tau_A) = \text{Gamma}(\widetilde{a}_A, \widetilde{b}_A) \tag{15}$$
$$q(\alpha) = \text{Gamma}(\widetilde{a}_\alpha, \widetilde{b}_\alpha) \tag{16}$$

with variance updates

$$\widetilde{a}_A = a_A + \frac{KD}{2} \tag{17}$$
$$\widetilde{b}_A = b_A + \frac{1}{2} \sum_{k=1}^{K_+} \sum_{d=1}^{D} \mathbb{E}\left[a_{kd}^2\right] \tag{18}$$

and

$$\widetilde{a}_X = a_X + \frac{ND}{2} \tag{19}$$
$$\widetilde{b}_X = b_X + \frac{1}{2} \sum_{n=1}^{N} \sum_{d=1}^{D} \Big[x_{nd}^2 + \sum_{k=1}^{K_+} \Big[\mathbb{E}\left[a_{kd}^2\right] z_{nk} \tag{20}$$
$$- 2\mathbb{E}[a_{kd}] z_{nk} x_{nd} + 2 \sum_{k'=k+1}^{K_+} z_{nk} z_{nk'} a_{kd} a_{k'd} \Big] \Big] \tag{21}$$

and $q(\alpha)$ updates

$$\widetilde{a}_\alpha = a_\alpha + K_+ \tag{22}$$
$$\widetilde{b}_\alpha = b_\alpha + H_N. \tag{23}$$

MEIBP inference is carried out exactly as discussed in the main text except all instances of $\sigma_{\boldsymbol{X}}, \sigma_{\boldsymbol{A}}$, and α are replaced with the expectation from their respective variational distribution. Furthermore the variational lower bound also has three additional entropy terms for gamma distributions, one for each hyperparameter.

S.4. Evidence as a function of $\mathbf{Z}_{n\cdot}$

As shown in the main text, we obtain a submodular objective function for each $\boldsymbol{Z}_{n\cdot}$, $n \in \{1, \ldots, N\}$ by examining the evidence as a function of $\boldsymbol{Z}_{n\cdot}$ while holding constant all $n' \in \{1, \ldots, N\} \setminus n$. The evidence is

$$\frac{1}{\sigma_{\boldsymbol{X}}^2} \sum_{n=1}^{N} \left[-\frac{1}{2} \boldsymbol{Z}_{n\cdot} \boldsymbol{\Phi} \boldsymbol{\Phi}^T \boldsymbol{Z}_{n\cdot}^T + \boldsymbol{Z}_{n\cdot} \boldsymbol{\xi}_{n\cdot}^T \right] - \ln K_+! + \sum_{k=1}^{K^+} \left[\ln \frac{(N-m_k)!(m_k-1)!}{N!} + \eta_k \right] + \text{const} \tag{24}$$

$$\xi_{nk} = \boldsymbol{\Phi}_{k\cdot} \boldsymbol{X}_{n\cdot}^T + \frac{1}{2} \sum_{d=1}^{D} \left[\mathbb{E}[a_{kd}]^2 - \mathbb{E}[a_{kd}^2] \right] \tag{25}$$

$$\eta_k = \sum_{d=1}^{D} \left[-\frac{\ln \frac{\pi \sigma_{\boldsymbol{A}}^2}{2\alpha^{2/D}}}{2} - \frac{\mathbb{E}[a_{kd}^2]}{2\sigma_{\boldsymbol{A}}^2} + H(q(a_{kd})) \right], \tag{26}$$

which nearly factorizes over the $\boldsymbol{Z}_{n\cdot}$ because the likelihood component and parts of the prior components naturally fit into a quadratic function of $\boldsymbol{Z}_{n\cdot}$. The $\ln K_+!$ and η_k only couple the rows of $\boldsymbol{Z}$ when K_+ changes, while the log-factorial term couples the rows of $\boldsymbol{Z}$ through the sums of the columns. Both of these terms only depend on statistics of $\boldsymbol{Z}$ (the m_k values and K_+), not the $\boldsymbol{Z}$ matrix itself, e.g. permuting the rows of $\boldsymbol{Z}$ would not affect these terms. Furthermore, $\ln K_+$ and η_k have no N dependence and become insignificant as N increases. These observations, in conjunction with the MEIBP performance in the experimental section of the main text, indicate that optimizing Eq. 24 for $\boldsymbol{Z}_{n\cdot}$ is a reasonable surrogate for optimizing $\boldsymbol{Z}$.

Here we explicitly decompose Eq. 24 to show its $\boldsymbol{Z}_{n\cdot}$ dependency. Decomposing $\ln \frac{(N-m_k)!(m_k-1)!}{N!}$ is straightforward if we first define the function:

$$\nu(z_{nk}) = \begin{cases} \ln (N - m_{k\setminus n} - z_{nk})!(m_{k\setminus n} + z_{nk} - 1)!/N! \\ 0, \quad \text{if } m_{k\setminus n} = 0 \text{ and } z_{nk} = 0. \end{cases} \tag{27}$$

where the "$\setminus n$" subscript indicates the variable with the n^{th} row removed from $\boldsymbol{Z}$. For a given n we have:

$$\begin{aligned} \sum_{k=1}^{K_+} \nu(z_{nk}) &= \sum_{k=1}^{K_+} \ln (N-m_k)!(m_k-1)!/N! \\ &= \sum_{k=1}^{K_+} z_{nk} \left(\nu(z_{nk}=1) - \nu(z_{nk}=0) \right) \\ &\quad + \nu(z_{nk}=0), \end{aligned} \tag{28}$$

which makes the $\boldsymbol{Z}_{n\cdot}$ dependency explicit and lets us add $\nu(z_{nk}=1) - \nu(z_{nk}=0)$ into the inner-product term, $\boldsymbol{\xi}_{n\cdot}$, and place $\nu(z_{nk}=0)$ into a constant term. We can incorporate η_k into the inner-product term in a similar manner for a given $n \in \{1, \ldots, N\}$:

$$\sum_{k=1}^{K_+} \eta_k = \sum_{k:m_{k\setminus n}>0} \eta_k + \sum_{k=1}^{K_+} \mathbf{1}_{\{m_{k\setminus n}>0\}} z_{nk} \eta_k, \tag{29}$$

where the first term does not depend on $\boldsymbol{Z}_{n\cdot}$ and is added to the constant term, while the second term is added to the inner-product term. Finally, for a given $n \in \{1, \ldots, N\}$ the $\ln K!$ term becomes

$$\ln K_+! = \ln \left(K_{+\setminus n} + \sum_{k=1}^{K_+} \left[\mathbf{1}_{\{m_{k\setminus n}=0\}} z_{nk} \right] \right)!, \tag{30}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. As stated in the main text, combining the above terms yields the following submodular objective function for $n = 1, \ldots, N$:

$$\mathcal{F}(\boldsymbol{Z}_{n\cdot}) = -\frac{1}{2\sigma_{\boldsymbol{X}}^2} \boldsymbol{Z}_{n\cdot} \boldsymbol{\Phi} \boldsymbol{\Phi}^T \boldsymbol{Z}_{n\cdot}^T + \boldsymbol{Z}_{n\cdot} \boldsymbol{\omega}_{n\cdot}^T + const - \ln \left(K_{+\setminus n} + \sum_{k=1}^{K_+} \left[\mathbf{1}_{\{m_{k\setminus n}=0\}} z_{nk} \right] \right)! \tag{31}$$

$$\boldsymbol{\Phi}_{k\cdot} = \left(\mathbb{E}\left[a_{k1}\right], \ldots, \mathbb{E}\left[a_{kD}\right] \right) \tag{32}$$

$$\begin{aligned} \omega_{nk} =& \frac{1}{\sigma_{\boldsymbol{X}}^2} \left(\boldsymbol{\Phi}_{k\cdot} \boldsymbol{X}_{n\cdot}^T + \frac{1}{2} \sum_{d=1}^{D} \left[\mathbb{E}[a_{kd}]^2 - \mathbb{E}[a_{kd}^2] \right] \right) \\ &+ \nu(z_{nk}=1) - \nu(z_{nk}=0) + \mathbf{1}_{\{m_{k\setminus n}>0\}} \eta_k, \end{aligned} \tag{33}$$

$\mathbf{1}\{\cdot\}$ is the indicator function, and the subscript "$\setminus n$" is the value of the given variable after removing the n^{th} row from $\boldsymbol{Z}$.

S.5. Additional MEIBP Characterization

In this section, we will maintain a growing list of additional MEIBP characterization experiments. See http://arxiv.org/abs/1304.3285 for the current version.

S.5.1. Learning K_+

An ostensible advantage of using Bayesian nonparametric priors is that a user does not need to specify the multiplicity of the prior parameters. Clever sampling techniques such as slice sampling and retrospective sampling allow samples to be drawn from these nonparametric priors, c.f. Teh et al. (2007) and Papaspiliopoulos & Roberts (2008). However variational methods are not directly amenable to Bayesian nonparametric priors as the variational optimization cannot be performed over an unbounded prior space. Instead, variational methods must specify a maximum model complexity (parameter multiplicity). Several heuristics have been proposed to address this limitation: Wang & Blei (2012) sampled from the variational distribution for the local parameters—which included sampling from the unbounded prior— and used the empirical distributions of the local samples to update the global parameters, while Ding et al. (2010) simply started with $K_+ = 1$ and greedily added features. We did not address these techniques in this work as the MEIBP performed competitively with the unbounded sampling techniques without employing these types of heuristics. Furthermore, here we demonstrate that the MEIBP can robustly infer the true number of latent features when the K_+ bound is greater than the true number of latent features.

For this experiment we generated the binary images dataset used in Griffiths & Ghahramani (2005), where the dataset, $\boldsymbol{X}$, consisted of 2000 6×6 images. Each row of $\boldsymbol{X}$ was a 36 dimensional vector of pixel intensity values that was generated by using $\boldsymbol{Z}$ to linearly combine a subset of the four binary factors shown in Figure 2. Gaussian white noise, $\mathcal{N}(0, \sigma_X)$, was then added to each image, yielding $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{A} + \mathbf{E}$. The feature vectors, $\boldsymbol{Z}_{n\cdot}$ were sampled from a distribution in which each factor was present with probability 0.5. Figure 3 shows four of these images with different σ_X values.

Figure 2. The four binary latent factors used in the sensitivity analysis in this section. The white squares are ones and the dark squares are zeros.

We initialized the MEIBP with $K = 20$, σ_X=1.0, $\sigma_A = 1.0$, $\alpha = 2$, $\tilde{\mu}_{kd} \sim |\mathcal{N}(0, 0.05)|$ (variational factor means), $\tilde{\sigma}_{kd} \sim |\mathcal{N}(0, 0.1)|$ (variational factor standard deviations), $z_{nk} \sim \text{Bernoulli}(\frac{1}{3})$. With this initialization, we tested the MEIBP robustness by performing MEIBP inference on $\boldsymbol{X}$ for $\sigma_X = 0.1, \ldots, 1.0$ in 100 evenly spaced increments with all hyperparameters and algorithm options unchanged during the experiment. MEIBP convergence was determined in the same way as the main experimental section. Figure 4 (top) shows a histogram of the final number of MEIBP features ($K_{\text{true}} = 4$) and Figure 4 (bottom) shows the final number of MEIBP features as a function of σ_X.

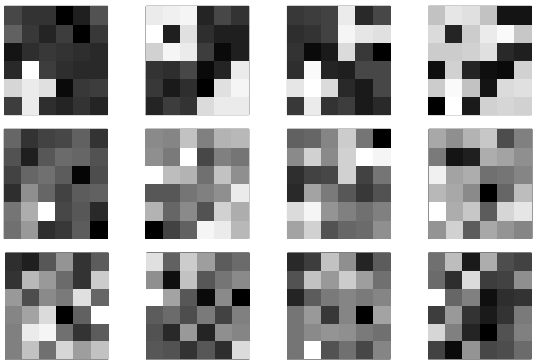

Figure 3. Example data used in the sensitivity analysis discussed in §S.5.1. Each column contains the same combination of latent factors, where the top row has a data noise term of $\sigma_X = 0.1$, the middle row has $\sigma_X = 0.5$, and the bottom row has $\sigma_X = 1.0$. Top: histogram of final K_+ value. Bottom: final K_+ value as a function of σ_X.

These results indicate that the regularizing nature of the IBP prior tends to lead to the correct number of latent features even when the K_+ bound is much larger than the true K_+. Furthermore this experiment indicates that MEIBP inference is robust to model noise, at least, for the simple data used in this experiment. At a medium level of data noise, the inference occasionally finished with $K_+ = 3$, which resulted from two true latent factors collapsing to the same inferred latent feature. Once this occurred, MEIBP did not have a mechanism for splitting the features. For σ_X comparable to the latent factors, $\sigma_X \geq 0.9$, MEIBP often inferred "noise features," which were essentially whitenoise and were typically active for less than 4% of the data instances. In future experiments we will attempt to flesh out the practical differences between unbounded priors and priors that operate in a large bounded latent space.

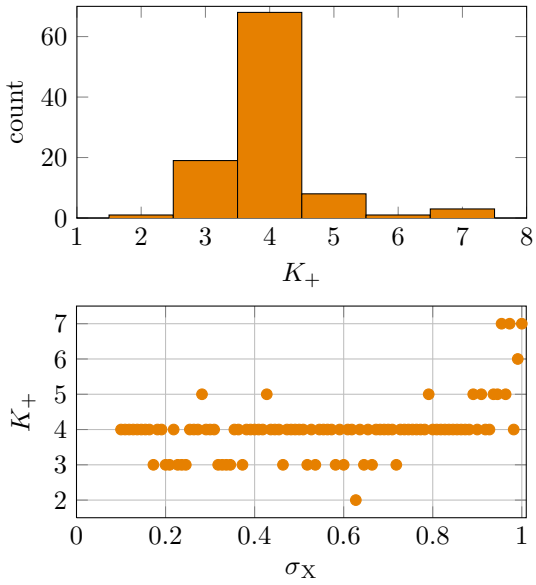

Figure 4. Final feature count (K_+ value) for MEIBP inference where $K_{\text{true}} = 4$ for the binary image data with K_+ initialized to 20 for $\sigma_X = 0.1, \ldots, 1.0$ in 100 evenly spaced increments with all hyperparameters and algorithm options fixed during the experiment.

References

Attias, H. A variational bayesian framework for graphical models. *Advances in Neural Information Processing Systems*, 12:209–215, 2000.

Ding, N., Qi, Y.A., Xiang, R., Molloy, I., and Li, N. Nonparametric Bayesian matrix factorization by Power-EP. In *14th Int'l Conf. on AISTATS*, volume 9, pp. 169–176, 2010.

Griffiths, T. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. Technical report, Gatsby Unit, UCL, London, UK, 2005.

Papaspiliopoulos, O. and Roberts, G. O. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1): 169–186, 2008.

Teh, Y.W., Gorur, D., and Ghahramani, Z. Stick-breaking construction for the indian buffet process. In *Int'l Conference on AISTATS*, volume 11, 2007.

Wang, C. and Blei, D. Truncation-free online variational inference for bayesian nonparametric models. In *Advances in Neural Information Processing Systems*, volume 25, 2012.