For the SCNT model, there is a related but simpler update (see supplemental material).

As in collapsed sampling algorithms for LDA [7], we can analytically marginalize the word distribution Ω_k for each topic. Let $M_{kw}^{\backslash dn}$ denote the number of instances of word w assigned to topic k, excluding token n in document d, and $M_{k\cdot}^{\backslash dn}$ the number of total tokens assigned to topic k. For a vocabulary with W unique word types, the posterior distribution of topic indicator z_{dn} is then

$$p(z_{dn} = k \mid \pi_{:d}, z_{\backslash dn}) \propto \pi_{kd} \left(\frac{M_{kw}^{\backslash dn} + \beta}{M_{k\cdot}^{\backslash dn} + W\beta} \right). \tag{10}$$

Recall that the topic probabilities $\pi_{:d}$ are determined from $v_{:d}$ via Equation (2).

3.2 Metropolis Independence Sampler Updates for Topic Activations

The posterior distribution of $v_{:d}$ does not have a closed analytical form due to the logistic nonlinearity underlying our stick-breaking construction. We instead employ a Metropolis-Hastings independence sampler, where proposals $q(v_{:d}^* \mid v_{:d}, A, u_{:d}, \lambda_v) = N(v_{:d}^* \mid Au_{:d}, \lambda_v^{-1} I_{\bar{K}})$ are drawn from the prior. Combining this with the likelihood of the N_d word tokens, the proposal is accepted with probability $\min(\mathbb{A}(v_{:d}^*, v_{:d}), 1)$, where

$$\begin{aligned}
\mathbb{A}(v_{:d}^*, v_{:d}) &= \frac{p(v_{:d}^* \mid A, u_{:d}, \lambda_v) \prod_{n=1}^{N_d} p(z_{dn} \mid v_{:d}^*) q(v_{:d} \mid v_{:d}^*, A, u_{:d}, \lambda_v)}{p(v_{:d} \mid A, u_{:d}, \lambda_v) \prod_{n=1}^{N_d} p(z_{dn} \mid v_{:d}) q(v_{:d}^* \mid v_{:d}, A, u_{:d}, \lambda_v)} \\
&= \prod_{n=1}^{N_d} \frac{p(z_{dn} \mid v_{:d}^*)}{p(z_{dn} \mid v_{:d})} = \prod_{k=1}^{K} \left(\frac{\pi_{kd}^*}{\pi_{kd}} \right)^{\sum_{n=1}^{N_d} \delta(z_{dn}, k)}
\end{aligned} \tag{11}$$

Because the proposal cancels with the prior distribution in the acceptance ratio $\mathbb{A}(v_{:d}^*, v_{:d})$, the final probability depends only on a ratio of likelihood functions, which can be easily evaluated from counts of the number of words assigned to each topic by z_d.

4 Experimental Results

4.1 Toy Bars Dataset

Following related validations of the LDA model [7], we ran experiments on a toy corpus of "images" designed to validate the features of the DCNT. The dataset consisted of 1,500 images (documents), each containing a vocabulary of 25 pixels (word types) arranged in a 5x5 grid. Documents can be visualized by displaying pixels with intensity proportional to the number of corresponding words (see Figure 2). Each training document contained 300 word tokens.

Ten topics were defined, corresponding to all possible horizontal and vertical 5-pixel "bars". We consider two toy datasets. In the first, a random number of topics is chosen for each document, and then a corresponding subset of the bars is picked uniformly at random. In the second, we induce topic correlations by generating documents that contain a combination of either only horizontal (topics 1-5) or only vertical (topics 6-10) bars. For these datasets, there was no associated metadata, so the input features were simply set as $\phi_d = 1$.

Using these toy datasets, we compared the LDA model to several versions of the DCNT. For LDA, we set the number of topics to the true value of $K = 10$. Similar to previous toy experiments [7], we set the parameters of its Dirichlet prior over topic distributions to $\alpha = 50/K$, and the topic smoothing parameter to $\beta = 0.01$. For the DCNT model, we set $\gamma_\mu = 10^6$, and all gamma prior hyperparameters as $a = b = 0.01$, corresponding to a mean of 1 and a variance of 100. To initialize the sampler, we set the precision parameters to their prior mean of 1, and sample all other variables from their prior. We compared three variants of the DCNT model: the singly correlated SCNT (A constrained to be diagonal) with $K = 10$, the DCNT with $K = 10$, and the DCNT with $K = 20$. The final case explores whether our stick-breaking prior can successfully infer the number of topics.

For the toy dataset with correlated topics, the results of running all sampling algorithms for 10,000 iterations are illustrated in Figure 2. On this relatively clean data, all models limited to $K = 10$