Reliable and Scalable Variational Inference for the Hierarchical Dirichlet Process

Michael C. Hughes, Dae Il Kim, and Erik B. Sudderth
mhughes@cs.brown.edu daeil@cs.brown.edu sudderth@cs.brown.edu
Dept. of Computer Science, Brown University, Providence, RI, USA.

Abstract

We introduce a new variational inference objective for hierarchical Dirichlet process admixture models. Our approach provides novel and scalable algorithms for learning nonparametric topic models of text documents and Gaussian admixture models of image patches. Improving on the point estimates of topic probabilities used in previous work, we define full variational posteriors for all latent variables and optimize parameters via a novel surrogate likelihood bound. We show that this approach has crucial advantages for data-driven learning of the number of topics. Via merge and delete moves that remove redundant or irrelevant topics, we learn compact and interpretable models with less computation. Scaling to millions of documents is possible using stochastic or memoized variational updates.

1 INTRODUCTION

Bayesian nonparametric models are increasingly applied to data with rich hierarchical structure, such as words within documents (Teh et al., 2006) or patches within images (Sudderth et al., 2008). *Hierarchical Dirichlet process* (HDP) admixture models provide a natural way to discover shared clusters, or *topics*, in grouped data. The HDP prior expects the number of topics to smoothly grow as more examples appear, making it attractive for analyzing big datasets.

While there are numerous existing inference algorithms for the HDP, all suffer from some combination of inability to scale to large datasets, vulnerability to poor local optima, or the need for external specification of the target model complexity. Simple Markov chain Monte Carlo (MCMC) samplers (Teh et al., 2006) can dynamically add or remove topics, but are computationally demanding with more than a few thousand documents and may take a long time to mix from a poor initialization. Collapsed variational methods (Teh et al., 2008) are based on a sophisticated family of marginal likelihood bounds, but lead to challenging optimization problems and sensitivity to initialization. Stochastic variational methods (Wang et al., 2011) and streaming methods (Broderick et al., 2013) are by design more scalable, but are easily trapped at a fixed point near a poor initialization. More recent variational algorithms have dynamically inserted or removed topics to escape local optima, but either lack guarantees for improving whole-data model quality (Bryant and Sudderth, 2012) or rely on slow-to-mix Gibbs sampler steps (Wang and Blei, 2012).

We develop a scalable HDP learning algorithm that enables reliable selection of the number of active topics. After reviewing HDP admixtures in Sec. 2, we develop a novel variational bound (Sec. 3) that captures posterior uncertainty in topic appearance probabilities, and leads to sensible model selection behavior (see Fig. 2). Sec. 4 then develops novel stochastic (Hoffman et al., 2013) and memoized (Hughes and Sudderth, 2013) variational inference algorithms for the HDP. The memoized approach supports merge and delete moves (Sec. 5) that remove redundant or irrelevant topics, leading to compact and interpretable models. Sec. 6 demonstrates faster and more accurate learning of HDP models for documents and images.

2 HDP ADMIXTURE MODELS

Consider data partitioned into D exchangeable groups $x = \{x_1 \dots x_D\}$, for example documents or images. Each group d contains N_d tokens $x_d = \{x_{d1}, \dots x_{dN_d}\}$, for example words or small pixel patches. For large datasets we divide groups into B predefined *batches*, where $\mathcal{D}_b$ is the set of documents in batch b.

Appearing in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

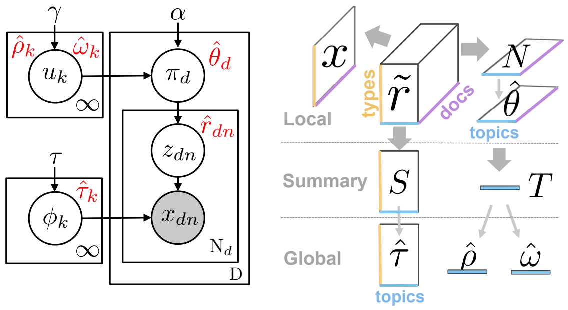

Figure 1: *Left:* Directed graphical model for the HDP admixture (Sec. 2). Free parameters for mean-field variational inference (Sec. 3) shown in red. *Right:* Flow chart for our inference algorithm, specialized for bag-of-words data, where we can use sparse type-based assignments $\tilde{r}$ instead of per-token variables $\hat{r}$. We define $\tilde{r}_{dwk}$ to be the total mass of all tokens in document d of type w assigned to k: $\tilde{r}_{dwk} = \sum_{n=1}^{N_d} \hat{r}_{dnk} \delta_{x_{dn},w}$. Updates flow from $\tilde{r}$ to global topic-type parameters $\hat{\tau}$ and (separately) to global topic weight parameters $\hat{\rho}, \hat{\omega}$. Each variable's shape gives its dimensionality. Thick arrows indicate summary statistics; thin arrows show free parameter updates.

To discover themes or topics common to all groups, while capturing group-specific variability in topic usage, we use the HDP admixture model (Teh et al., 2006) of Fig. 1. The HDP uses group-specific frequencies to cluster tokens into an *a priori* unbounded set of topics. To generate each token, a global topic (indexed by integer k) is first drawn, and an observation is then sampled from the likelihood distribution for topic k.

Topic-specific data generation. HDP admixtures are applicable to any real or discrete data for which an appropriate exponential family likelihood is available. Data assigned to topic k is generated from a distribution F with parameters ϕ_k, and conjugate prior H:

$$\begin{aligned} F: \quad & \log p(x_{dn}|\phi_k) = s_F(x_{dn})^T \phi_k + c_F(\phi_k), \\ H: \quad & \log p(\phi_k|\bar{\tau}) = \phi_k^T \bar{\tau} + c_H(\bar{\tau}). \end{aligned}$$

Here c_H and c_F are cumulant functions, and $s_F(x_{dn})$ is a sufficient statistic vector. For discrete data x, F is multinomial and H is Dirichlet. For real-valued x, we take F to be Gaussian and H Normal-Wishart.

Allocating topics to tokens. Each topic k is defined by two global variables: the data-generating exponential family parameters ϕ_k, and a frequency weight u_k. Each scalar $0 < u_k < 1$ defines the conditional probability of sampling topic k, given that the first $k-1$ topics were *not* sampled:

$$u_k \sim \text{Beta}(1, \gamma), \qquad \beta_k \triangleq u_k \prod_{\ell=1}^{k-1} (1 - u_\ell). \tag{1}$$

This *stick-breaking process* (Sethuraman, 1994; Blei and Jordan, 2006) transforms $\{u_\ell\}_{\ell=1}^k$ to define the marginal probability β_k of selecting topic k.

Each group or document has unique topic frequencies $\pi_d = [\pi_{d1}, \ldots, \pi_{dk}, \ldots]$, where the HDP prior induces a finite Dirichlet distribution on the first K probabilities:

$$[\pi_{d1} \ldots \ \pi_{dK} \ \pi_{d>K}] \sim \text{Dir}(\alpha\beta_1, \ldots \alpha\beta_K, \alpha\beta_{>K}). \tag{2}$$

This implies that π_d has mean β and variance determined by the concentration parameter α. The subscript $_{>K}$ denotes the aggregate mass of all topics with indices larger than K, so that $\beta_{>K} \triangleq \sum_{\ell=K+1}^{\infty} \beta_\ell$.

To generate token n in document d, we first draw a topic assignment $z_{dn} \sim \text{Cat}(\pi_d)$, where integer $z_{dn} \in \{1, 2, \ldots\}$ indicates the chosen topic k. Second, we draw the observed token x_{dn} from density F, using the parameter ϕ_k indicated by z_{dn}.

3 VARIATIONAL INFERENCE

Given observed data x, we wish to learn global topic parameters u, ϕ and local document structure π_d, z_d. Taking an optimization approach (Wainwright and Jordan, 2008), we seek an approximate distribution q over these variables that is as close as possible to the true, intractable posterior in KL divergence but belongs to a simpler, fully factorized family $q(\cdot) = q(u)q(\phi)q(\pi)q(z)$ of exponential family densities.

Previous variational methods for HDP topic models (Wang et al., 2011) have employed a *Chinese restaurant franchise* (CRF) model representation (Teh et al., 2006). Here each document has its own local topics, a stick-breaking prior on their frequencies, and latent categorical variables linking each local topic to some global cluster. With this expanded set of highly-coupled latent variables, the factorizations inherent in mean field variational methods induce many local optima. We thus develop an alternative bound based on the *direct assignment* HDP representation in Fig. 1.

3.1 Direct Assignment Variational Posteriors

Deferring discussion of the global topic weight posterior $q(u)$ until Sec. 3.2, we define other variational posteriors below, marking free parameters with hats to make clear which quantities are optimized:

$$\begin{aligned} q(z|\hat{r}) &= \textstyle\prod_{d=1}^{D} \prod_{n=1}^{N_d} \text{Cat}(z_{dn} \mid \hat{r}_{dn1}, \hat{r}_{dn2}, \ldots \hat{r}_{dnK}), \\ q(\pi) &= \textstyle\prod_{d=1}^{D} \text{Dir}(\pi_d|\hat{\theta}_{d1}, \ldots \hat{\theta}_{dK+1}), \\ q(\phi|\hat{\tau}) &= \textstyle\prod_{k=1}^{\infty} H(\phi_k|\hat{\tau}_k). \end{aligned} \tag{3}$$

This posterior models data using K *active* topics. Crucially, as in Teh et al. (2008) and Bryant and Sudderth (2012), the chosen truncation level K defines only the form of local factors $q(z)$ and $q(\pi)$. Global factors do not require an explicit truncation, as those with indices greater than K are conditionally independent of the data. This approach allows optimization of K and avoids artifacts that arise with non-nested truncations of stick-breaking processes (Blei and Jordan, 2006).

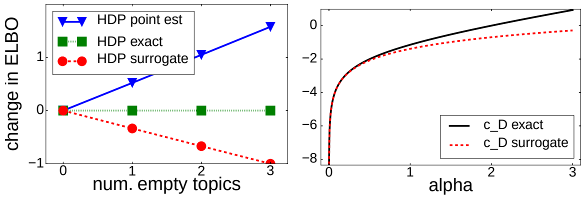

Figure 2: *Left:* Comparison of variational objectives resulting from different choices for $q(u)$ on the model selection task of Sec. 3.2. Our new surrogate bound sensibly prefers models without empty topics, while using point estimation does not. *Right:* Illustration of Eq. (12)'s tight lower bound on $c_D(\alpha\beta)$, shown for $K = 1, \beta = [0.5, 0.5]$. This bound makes our surrogate objective tractable.

Factor $q(z)$. Given truncation level K, token indicator z_{dn} must be assigned to one of the K active topics. The categorical distribution $q(z_{dn})$ is parameterized by a positive vector $\hat{r}_{dn}$ of size K that sums to one.

Factor $q(\pi)$. π_d can be represented by a positive vector of size $K+1$ encoding the K active topic probabilities in document d and (at the last index) the aggregate mass $\pi_{d>K}$ of all inactive topics. Thus, $q(\pi_d)$ is a Dirichlet distribution with parameters $\hat{\theta}_d \in \mathbb{R}^{K+1}$.

Factor $q(\phi)$. Data-generating factors $q(\phi_k)$ for each topic k come from the conjugate family H with free parameter $\hat{\tau}_k$. For discrete data H is Dirichlet and $\hat{\tau}_k$ is a vector the length of the vocabulary W.

Objective function. Mean field methods optimize an evidence lower bound $\log p(x|\gamma, \alpha, \tau) \geq \mathcal{L}(\cdot)$, where

$$\mathcal{L}(\cdot) \triangleq \mathcal{L}_{data}(\cdot) + H_z(\cdot) + \mathcal{L}_{HDP}(\cdot) + \mathcal{L}_u(\cdot). \quad (4)$$

The final term $\mathcal{L}_u(\cdot)$, which depends only on $q(u)$, is discussed in the next section. The first three terms account for data generation, the assignment entropy, and the document-topic allocations. These are defined below, with expectations taken with respect to Eq. (3):

$$\mathcal{L}_{data}(\cdot) \triangleq \mathbb{E}_q[\log p(x|z,\phi) + \log \tfrac{p(\phi|\bar{\tau})}{q(\phi|\hat{\tau})}], \quad (5)$$

$$H_z(\cdot) \triangleq -\textstyle\sum_{k=1}^{K} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \hat{r}_{dnk} \log \hat{r}_{dnk},$$

$$\mathcal{L}_{HDP}(\cdot) \triangleq \mathbb{E}_q\Big[\log \tfrac{p(z|\pi)p(\pi|\alpha,u)}{q(\pi|\hat{\theta})}\Big].$$

The forms of $\mathcal{L}_{data}$ and H_z are unchanged from the simpler case of mean-field for DP mixtures. Closed-form expressions are in the Supplement.

3.2 Topic Weights and Model Selection

Previous work on the direct assignment HDP suggested a point estimate approximation for topic appearance parameters β (Liang et al., 2007; Bryant and Sudderth, 2012), or equivalently $q(u_k) = \delta_{\hat{u}_k}(u_k)$. While efficient, this approach creates problems with model selection. The resulting objective lower bounds a joint evidence that *includes* the point estimate u: $\log p(x, u|\alpha, \gamma, \tau)$. Consequently, the point estimate for u is a MAP estimate, with prior defined by $\mathcal{L}_u$:

$$\mathcal{L}_u^{PE} = \textstyle\sum_{k=1}^{K} \log \mathrm{Beta}(\hat{u}_k|1, \gamma). \quad (6)$$

Consider instead a different $q(u)$ that places a proper Beta distribution over each parameter u_k:

$$q(u|\hat{\rho}, \hat{\omega}) = \textstyle\prod_{k=1}^{\infty} \mathrm{Beta}(u_k \mid \hat{\rho}_k\hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k). \quad (7)$$

Here, free parameter $0 < \hat{\rho}_k < 1$ defines the mean: $\mathbb{E}[u_k] = \hat{\rho}_k$, while $\hat{\omega}_k > 0$ controls the variance of u_k. Under this proper Beta family, we can integrate the variable u away to obtain a proper marginal evidence $\log p(x|\alpha, \gamma, \tau)$. Consequently, $\mathcal{L}_u$ term has the form

$$\mathcal{L}_u^{Beta}(\cdot) = \textstyle\sum_{k=1}^{K} \mathbb{E}_q[\log \tfrac{p(u_k)}{q(u_k)}] \quad (8)$$

Model selection. Given our chosen family for $q(z, \pi, \phi)$ in Eq. (3) and a proper $q(u)$ in Eq. (7), the objective $\mathcal{L}$ can be used to compare two alternative sets of free parameters, even if they have different numbers of active topics K. Our recommended setting of $q(u)$ enjoys the benefits of marginalization, while MAP point estimation can yield pathological behavior when comparing $\mathcal{L}$ at different truncation levels.

To illustrate, consider two candidate models, A and E. Candidate A has K topics and token parameters $\hat{r}^A$. Candidate E_J has the *same* token parameters as well as J additional topics with zero mass. For each token n, we set vector $\hat{r}_n^E$ so the first K topics are equal to $\hat{r}_n^A$, and the extra J topics are set to zero. We desire an objective that prefers A by penalizing the "empty" topics in E, or at least one that does not favor E.

The behavior of different objectives is shown in Fig. 2, where we plot $\mathcal{L}(E_J) - \mathcal{L}(A)$ for $J = \{0, 1, 2, 3\}$ empty topics. When using the Beta form for $q(u)$, we find that exact numerical evaluation of the HDP objective is invariant to empty topics, while our scalable surrogate objective from Sec. 3.3 penalizes empty topics slightly. However, point-estimation of $q(u)$ always *favors* adding empty topics. Thus, we focus on the Beta form of $q(u)$ to learn compact, interpretable models.

3.3 Surrogate bound for tractable inference.

Motivated by Fig. 2, we wish to employ the proper Beta form for $q(u)$. However, this leads to a non-conjugate relationship between $q(u)$ and $q(\pi)$, complicating inference. Some terms of the resulting objective have no closed-form. To gain tractability, we develop a surrogate bound on the ideal objective.

Consider the ELBO term $\mathcal{L}_{HDP}$ under $q(u)$ in Eq. (7).

$$\mathbb{E}_q[\log \tfrac{p(z)p(\pi)}{q(\pi)}] = \textstyle\sum_{d=1}^{D} \mathbb{E}_q[c_D(\alpha\beta)] - c_D(\hat{\theta}_d) \quad (9)$$
$$+ \textstyle\sum_{k=1}^{K+1} \Big(N_{dk} + \alpha\mathbb{E}_q[\beta_k] - \hat{\theta}_{dk}\Big)\mathbb{E}_q[\log \pi_{dk}]$$

Here, sufficient statistic N_{dk} counts the usage of topic k in document d: $N_{dk} \triangleq \sum_{n=1}^{N_d} \hat{r}_{dnk}$. Furthermore,

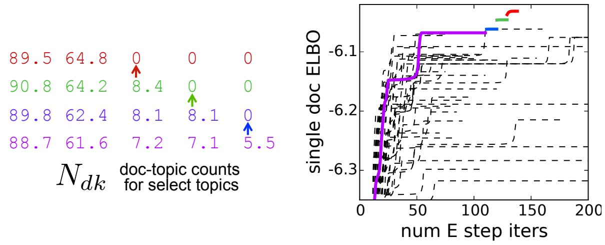

Figure 3: Sparsity-promoting restarts for local steps on the Science corpus with $K = 100$. *Left:* Example fixed points of the topic usage statistic N_{dk} for one document. *Right:* Trace of single-document ELBO objective during E-step inference for 50 random initializations (dashed lines), plus one sparsity-promoting run (solid) which climbs through the color-coded fixed points in the adjacent plot.

two required expectations have closed-form expressions. $\mathbb{E}[\beta_k]$ comes from Eq. (1), and

$$\mathbb{E}[\log \pi_{dk}] = \psi(\hat{\theta}_{dk}) - \psi(\textstyle\sum_{\ell=1}^{K+1} \hat{\theta}_{d\ell}). \tag{10}$$

However, c_D is the cumulant function of the Dirichlet,

$$c_D(a_1, \dots a_W) = \log \frac{\Gamma(\sum_{w=1}^{W} a_w)}{\prod_{w=1}^{W} \Gamma(a_w)}, \tag{11}$$

and $\mathbb{E}_q[c_D(\alpha\beta)]$ has no closed form. To avoid this problematic expectation of log Gamma functions, we introduce a novel bound on $c_D(\cdot)$:

$$\begin{aligned} c_D(\alpha\beta) \geq K \log \alpha &+ \textstyle\sum_{k=1}^{K} \log u_k \\ &+ \textstyle\sum_{k=1}^{K} (K+1-k) \log 1-u_k. \end{aligned} \tag{12}$$

Fig. 2 shows this bound is valid for all $\alpha > 0$. For proof, see the Supplement. We can tractably compute the expectation of Eq. (12), because expectations of logs of Beta random variables have a closed form.

Substituting Eq. (12) into our original objective $\mathcal{L}$ yields a surrogate objective $\mathcal{L}_{sur}$ which can be used for model selection because it remains a valid lower bound on the log evidence $\log p(x|\alpha, \gamma, \bar{\tau})$. Our surrogate objective induces a small penalty for empty components in Fig. 2, which is superior to the reward for empty components induced by point estimates.

4 INFERENCE ALGORITHM

We now describe an algorithm for optimizing the free parameters of our chosen approximation family q. We first give concrete updates to local and global factors. Later, we introduce memoized and stochastic methods for scalable online learning.

4.1 Local updates.

In the local step, we visit each document d and update token indicators r_{dn} via Eq. (13) and document-topic parameters $\hat{\theta}_d$ via Eq. (14). These steps are interdependent: updating $\hat{r}_{dn}$ requires an expectation computed from $\hat{\theta}_d$, and vice versa. Thus, at each document we need to initialize $\hat{\theta}_d$ and then alternate these updates until convergence. We discuss initialization and convergence strategies in the Supplement.

Update of $q(z)$. We update the free parameter $\hat{r}_{dn}$ for each token n in document d according to

$$\hat{r}_{dnk} \propto \exp\Big(\mathbb{E}_q[\log \pi_{dk}] + \mathbb{E}_q[\log p(x_{dn}|\phi_k)]\Big), \tag{13}$$

which uses known expectations. The vector $\hat{r}_{dn}$ is normalized over all topics k so its sum is one.

Update of $q(\pi_d)$. We update free parameter $\hat{\theta}_d$ given N_{dk}, which summarizes usage of topic k across all tokens in document d. The update is

$$\hat{\theta}_{dk} = \alpha \mathbb{E}_q[\beta_k] + N_{dk}, \tag{14}$$

where the expectation $\mathbb{E}_q[\beta_k]$ follows from Eq. (1). This update applies to all $K+1$ entries of $\hat{\theta}_d$. The last index aggregates all inactive topics, and is simply set to $\alpha \mathbb{E}[\beta_{>K}]$, since $N_{d>K}$ is zero by truncation.

Sparse Restarts. When visiting document d, the joint inference of $\hat{\theta}$ and $\hat{r}$ can be challenging. Many local optima exist even for this single-document task, as shown Fig. 3. A common failure mode occurs when a few tokens are assigned to a rare "junk" topic. Reassignment of these tokens may not happen under Eq. (13) updates due to a valley in the objective between keeping the current junk assignments and setting the junk topic to zero.

To more adequately escape local optima, we develop sparsity-promoting restart moves which take a final document-topic count vector $[N_{d1} \dots N_{dK}]$ produced by coordinate ascent, propose an alternative which has one entry set to zero, and accept if this improves the ELBO after further ascent steps. In practice, the acceptance rate varies from 30-50% when trying the 5 smallest non-zero topics. We observe huge gains in the whole-dataset objective due to these restarts.

4.2 Global updates.

Fig. 1 shows global parameter updates to $\hat{\tau}, \hat{\rho}$, and $\hat{\omega}$ require compact *sufficient statistics* of local parameters. The updates below focus on these summaries.

Update for $q(\phi)$. We update free parameter $\hat{\tau}$ to

$$\hat{\tau}_k = S_k + \bar{\tau}, \qquad S_k \triangleq \textstyle\sum_{d=1}^{D} \sum_n s_F(x_{dnk}) \hat{r}_{dnk}, \tag{15}$$

where S_k is the statistic summarizing data assigned to topic k across all tokens. For topic models, S_k is a vector of counts for each vocabulary type.

Update for $q(u)$. Finally, we consider the free parameters $\hat{\rho}, \hat{\omega}$ for all K active topics. No closed-form update exists due to non-conjugacy. Instead, we numerically optimize our surrogate objective, finding the best vectors $\hat{\rho}, \hat{\omega}$ simultaneously. The constrained optimization problem is:

$$\begin{aligned}\hat{\rho}, \hat{\omega} = \text{argmax}_{\rho,\omega} \mathcal{L}_{HDP}(\rho, \omega, T, \alpha) + \mathcal{L}_u(\rho, \omega, \gamma) \quad (16)\\ \text{s.t.} \quad 0 < \rho_k < 1, \omega_k > 0 \text{ for } k \in \{1, 2, \ldots K\}\end{aligned}$$

where sufficient statistic $T = [T_1 \ldots T_K \; T_{K+1}]$ sums the expectation of Eq. (10) across documents:

$$T_k(\hat{\theta}) \triangleq \textstyle\sum_{d=1}^{D} \mathbb{E}[\log \pi_{dk}]. \qquad (17)$$

The Supplement provides implementation details, including the exact function and gradients we provide to a modern L-BFGS optimization algorithm.

4.3 Memoized algorithm.

We now provide a memoized coordinate ascent update algorithm. The update cycle comes from Hughes and Sudderth (2013), which was inspired by the incremental EM approach of Neal and Hinton (1998). Data is visited one batch at a time, where the batches are predefined. We call each complete pass through all batches a *lap*. At each batch, we perform a local step update to $q(z_d), q(\pi_d)$ for each document d in the batch, and then a global-step update to $q(u), q(\phi)$.

Affordable batch-by-batch processing is possible by tracking sufficient statistics and exploiting their additivity. For each statistic, we track a batch-specific quantity (denoted N^b) for each batch and an aggregated whole-dataset quantity (N). By definition, $N_k = \sum_{b=1}^{B} N_k^b$. After visiting each batch b, we perform an incremental update to make the aggregate summaries reflect the new batch summaries and remove any previous contribution from batch b.

This algorithm requires storing per-batch summaries N^b, S^b, T^b for every batch during inference. This requirement is modest, remaining size $\mathcal{O}(BK)$ no matter how many tokens or documents occur in each batch.

ELBO computation. Computing the objective $\mathcal{L}$ is possible after each batch visit, so long as we track sufficient statistics as well as a few ELBO-specific quantities. First, we store the entropy H_z from Eq. (5) at each batch, as in Hughes and Sudderth (2013).

Second, consider the computation of $\mathcal{L}_{HDP}$ in Eq. (9). Naively, this computation requires sums over all documents. However, by tracking the following terms we can perform rapid evaluation:

$$G_k^b \triangleq \textstyle\sum_{d \in \mathcal{D}_b} (N_{dk} - \hat{\theta}_{dk}) \mathbb{E}[\log \pi_{dk}], \qquad (18)$$

$$Q_0^b = \textstyle\sum_{d \in \mathcal{D}_b} \log \Gamma(\sum_{k=1}^{K+1} \hat{\theta}_{dk}), Q_k^b = \sum_{d \in \mathcal{D}_b} \log \Gamma(\hat{\theta}_{dk}).$$

After aggregating these tracked statistics across all batches, such as $Q_k = \sum_{b=1}^{B} Q_k^b$, Eq. (9) becomes

$$\begin{aligned}\mathcal{L}_{HDP}(\cdot) = D\mathbb{E}_q[c_D(\alpha\beta)] - Q_0 \qquad (19)\\ + \textstyle\sum_{k=1}^{K+1} Q_k + G_k + \alpha \mathbb{E}_q[\beta_k] T_k\end{aligned}$$

which given tracked statistics can be evaluated with cost independent of the number of documents D.

4.4 Stochastic algorithm.

Our objective $\mathcal{L}$ can also be optimized with stochastic variational inference (Hoffman et al., 2013). The stochastic global step at iteration t updates the natural parameters of $q(u)$ and $q(\phi)$ with learning rate ξ_t. For example, the new $\hat{\tau}_t$ interpolates between the previous value $\hat{\tau}_{t-1}$ and an amplified estimate from the current batch $\hat{\tau}^b$. When ξ_t decays appropriately, this method guarantees convergence to a local optimum.

4.5 Computational complexity

Our direct assignment representation is more efficient than the CRF approach of Wang et al. (2011). The dominant cost of both algorithms is the local step for each token. We require $\mathcal{O}(N_d K)$ computations to update the free parameters $\hat{r}$ for a single document via Eq. (13). The CRF method requires $\mathcal{O}(N_d K J)$ operations, where $J < K$ is the number of global topics allowed in each document (for more details, see Eq. 18 of Wang et al. (2011)). For any reasonable value of $J > 1$, the CRF approach is more expensive. When $J = \mathcal{O}(K)$, the CRF local step is quadratic in the number of topics, while our approach is always linear.

5 MERGE AND DELETE MOVES.

Here, we develop two moves, merge and delete, which help discover a compact set of interpretable topics. As illustrated in Fig. 4, merges combine redundant topics, while deletes remove unnecessary "junk" topics or empty topics. Both moves enable faster subsequent iterations by making the active set of topics smaller.

5.1 Merge moves.

Each merge move transforms a current variational posterior q of size K into a candidate q' of size $K - 1$ by combining two topics in a single *merged* topic. During each pass we consider several candidate pairs. For each pair $\ell < m$, we imagine simply pooling together all tokens assigned to either topic ℓ or m in the original model to create topic ℓ in q'. All other parameters are copied over unchanged. Formally,

$$\hat{r}'_{dn\ell} = \hat{r}_{dn\ell} + \hat{r}_{dnm}, \forall d, n, \; \hat{\theta}'_{d\ell} = \hat{\theta}_{d\ell} + \hat{\theta}_{dm}, \forall d. \quad (20)$$

A global update to create $\hat{\tau}', \hat{\rho}', \hat{\omega}'$ completes the candidate, and we keep it if the objective $\mathcal{L}$ improves.

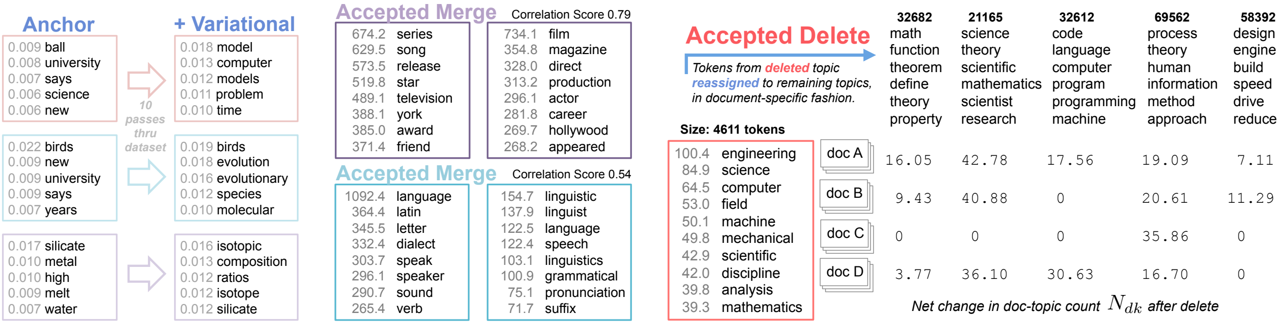

Figure 4: *Left:* Anchor topics (Arora et al., 2013) can be improved significantly by variational updates. *Center:* Topic pairs accepted by merge moves during run on Wikipedia. Combining each pair into one topic improves our objective $\mathcal{L}$, saves space, and removes redundancy. *Right:* Accepted delete move during run on Wikipedia. Red topic is rarely used and lacks semantic focus. Removing it and reassigning its mass to remaining topics improves $\mathcal{L}$ and interpretability.

For large datasets, explicitly retaining both $\hat{r}$ and $\hat{r}'$ via Eq. (20) is prohibitive. Instead, we can exploit additive statistics to rapidly evaluate a proposed merge. Eq. (20) implies that $S'_\ell = S_\ell + S_m$ and $N'_\ell = N_\ell + N_m$. This allows constructing candidate $\hat{\tau}'$ values and evaluating $\mathcal{L}_{data}$ without visiting any batches.

Not all statistics can be computed in this way, so some modest tracking must occur. For each candidate merge, we must compute T'^b_ℓ from Eq. (17) as well as the ELBO statistics G'^b_ℓ, Q'^b_ℓ from Eq. (18) at each batch. Finally, we track the entropy H_z for each candidate, as did Hughes and Sudderth (2013).

The first step of a merge is to select candidate pairs using a correlation score (Bryant and Sudderth, 2012):

$$\text{score}(\ell, m) = \text{Corr}(N_{:\ell}, N_{:m}), \quad -1 < \text{score} < 1. \quad (21)$$

Large scores identify topic pairs frequently used in the same documents. Before each lap we select at most 50 pairs to track with score above 0.05.

Next, we visit each batch in order, tracking relevant merge summaries during standard memoized updates. Finally, we evaluate each candidate using both tracked summaries and additive summaries, accepting or rejecting as needed. Many merges can be accepted after each lap, so long as no two share a topic in common.

5.2 Delete moves

Delete moves provide a more powerful alternative to merges for removing rarely used "junk" topics. For an illustration of an accepted delete on Wikipedia data, see Fig. 4. After identifying a candidate topic with small mass to delete, we reassign *all* its tokens to the remaining topics and then accept if the objective $\mathcal{L}$ improves. This move can succeed when a merge would fail because each document's tokens can be reassigned in a customized way, as shown in Fig. 4.

To make this move scalable for our memoized algorithm, we identify a candidate delete topic j in advance and collect a *target dataset* x' of all documents which use selected topic j significantly: $\{d : N_{dj} > 0.01\}$. Given the target set, we initialize candidate sufficient statistics by simply removing entries associated with topic j. From this initialization, we run several local-global updates on the target and then accept the move if the target's variational objective $\mathcal{L}(\cdot)$ improves. Further details can be found in the Supplement. To be sure of deleting a topic, the target set x' must contain *all* documents which pass our threshold test. Thus, deletes are only applicable to topics of below some critical size to remain affordable. We set a maximum budget of 500 documents for the target dataset size in our topic modeling experiments.

Acceptance rates in practice. Here, we summarize acceptance rates for merges and deletes during a typical run on the Wikipedia dataset with $K = 200$ initial topics. During the first 4 passes, we accept 73 of 79 proposed deletes (92%), and 12 of 194 merges (6%). These moves crucially remove bad topics from the random initialization. After the first few laps, no further merges are accepted and only 10% of deletes are accepted (at most 1 or 2 attempts per lap).

6 EXPERIMENTS

Our experiments compare inference methods for fitting HDP topic models. For our new HDP objective, we study stochastic with fixed K (SOfix), memoized with fixed K (MOfix), and memoized with deletes and merges (MOdm). For baselines, we consider the collapsed sampler (Gibbs) of Teh et al. (2006), the stochastic CRF method (crfSOfix) of Wang et al. (2011), and the stochastic split-merge method (SOsm) of Bryant and Sudderth (2012). For each method, we perform several runs from various initial K values.

For each run, we measure its predictive power via a heldout document completion task, as in Bryant and

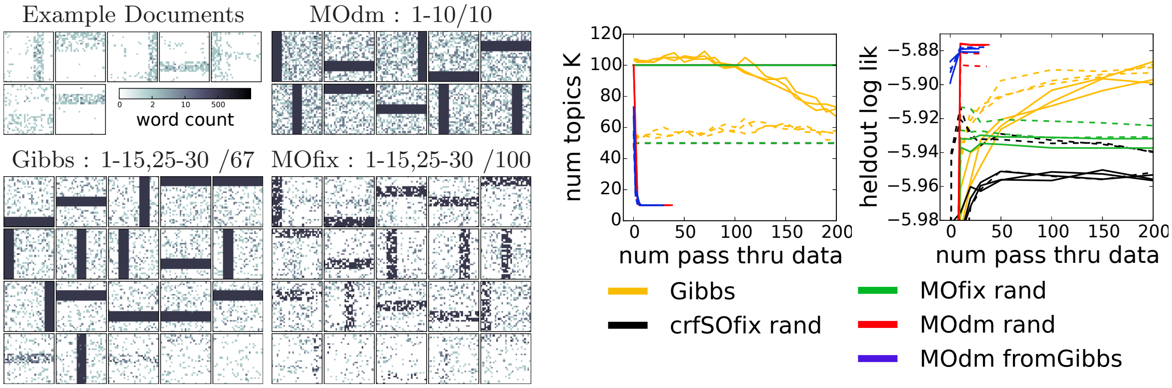

Figure 5: Comparison of inference methods on toy bars dataset from Sec. 6.1. *Top Left:* Word count images for 7 example documents and the final 10 estimated topics from MOdm. Each image shows all 900 vocabulary types arranged in square grid. *Bottom left:* Final estimated topics from Gibbs and MOfix. We rank topics from most to least probable, and show ranks 1-15 and 25-30. *Right:* Trace plots of the number of topics K and heldout likelihood during training. Line style indicates number of initial topics: dashed is $K = 50$, solid is $K = 100$.

Sudderth (2012). Each model is summarized by a point-estimate of the topic-word probabilities ϕ. For each heldout document d we randomly split its word tokens into two halves: x'_d, x''_d. We use the first half to infer a point-estimate of π_d, then estimate log-likelihood of each token in the second half x''_d.

$$\text{heldout-lik}(x|\phi) = \frac{\sum_{d \in \mathcal{D}_{test}} \log p(x''_d | \pi_d, \phi)}{\sum_{d \in \mathcal{D}_{test}} |x''_d|} \qquad (22)$$

Hyperparameters. In all runs, we set $\gamma = 10$, $\alpha = 0.5$ and topic-word pseudocount $\bar{\tau} = 0.1$. Stochastic runs use the learning rate decay recommended in Bryant and Sudderth (2012): $\kappa = 0.5, \delta = 1$.

6.1 Toy bars dataset.

We study a variant of the toy bars dataset of Griffiths and Steyvers (2004), shown in Fig. 5. There are 10 ideal bar topics, 5 horizontal and 5 vertical. The bars are noisier than the original and cover a larger vocabulary (900 words). We generate 1000 documents for training and 100 more for heldout test. Each one has 200 tokens drawn from 1-3 topics.

Fig. 5 shows many runs of all algorithms on this benchmark. Variational methods initialized with 50 or 100 topics get stuck rapidly, while the Gibbs sampler finds a redundant set of the ideal topics and is unable to effectively merge down to the ideal 10.

In contrast, our MOdm method uses merges and deletes to rapidly recover the 10 ideal bars after only a few laps. Without these moves, MOfix runs remain stuck at suboptimal fragments of bars. Furthermore, our MOdm method initialized with the sampler's final topics (fromGibbs) easily recovers the ideal bars.

6.2 Academic and news articles.

Next, we apply all methods to papers from the NIPS conference, articles from Wikipedia, and articles from the journal Science (Paisley et al., 2011), with 80%-20% train-test splits. Online methods process each training set in 20 batches. Trace plots in Fig. 6 compare predictive power and model complexity as more data is processed. We summarize conclusions below.

Anchor topics are good; variational is better. Using the anchor word method (Arora et al., 2013) for initial topic-word parameters yields better predictions than random initialization (`rand`). However, our methods can still make big, useful changes from this starting point. See Fig. 4 for some examples.

Deletes and merges make big, useful changes. Across all 3 datasets in Fig. 6, merges and deletes remove many topics. On Wikipedia, we reduce 200 topics to under 100 while improving predictions. Similar gains occur from the final result of the Gibbs sampler.

Competitors get stuck or improve slowly. The Gibbs sampler needs many laps to make quality predictions. The CRF method gets stuck quickly, while our methods (using the direct assignment representation) do better from similar initializations. The stochastic split-merge method (SOsm) grows to a prescribed maximum number of topics but fails to make better predictions. This indicates problems with heuristic acceptance rules, and motivates our moves governed by exact evaluation of a whole-dataset objective.

Next, we analyze the New York Times Annotated Corpus: 1.8 million articles from 1987 to 2007. We withhold 800 documents and divide the remainder into 200 batches (9084 documents per batch). Fig. 6 shows the predictive performance of the more-scalable methods.

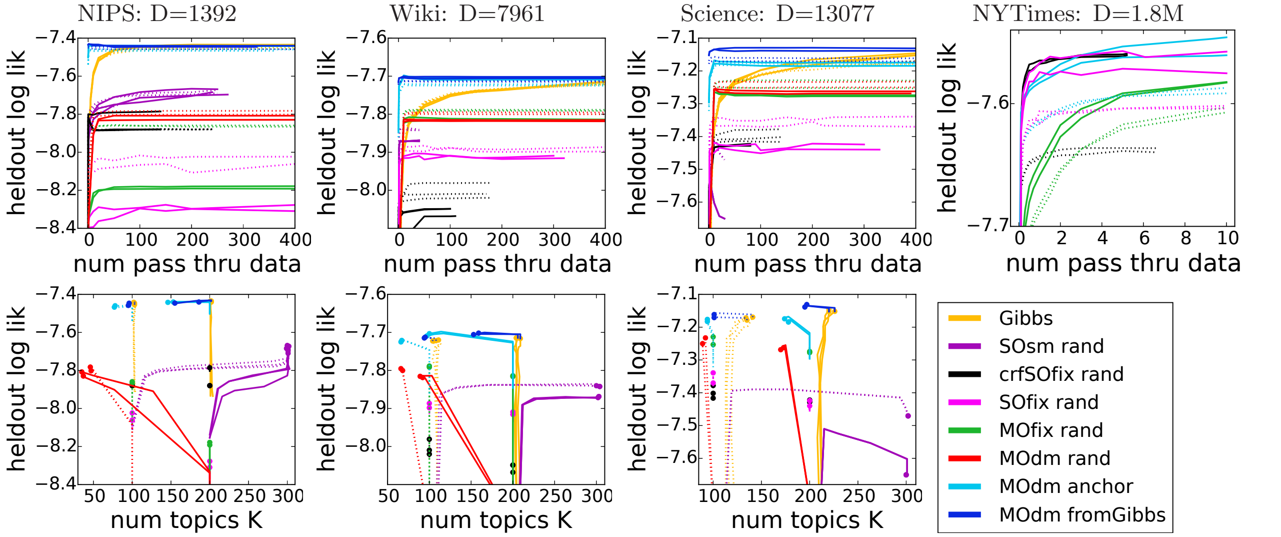

Figure 6: Comparison of inference methods on academic and news article datasets (Sec. 6.2). Line style indicates initial number of topics K: 100 is dots, 200 is solid. *Top row:* Heldout likelihood (larger is better) as more training data is seen. *Bottom row:* Trace plots of heldout likelihood and number of topics. Each solid dot marks the final result of a single run, with the trailing line its trajectory from initialization. Ideal runs move toward the upper left corner.

For this large-scale task, our direct assignment representation is more efficient than the CRF code released by Wang et al. (2011). With $K = 200$ topics, our memoized algorithm with merge and delete moves (MOdm) completes 8 laps through the 1.8 million documents in the amount of time the CRF code completes a single lap. No deletes or merges are accepted from any MOdm run, likely because 1.8M documents require more than a few hundred topics. However, the acceptance rate of sparsity-promoting restarts is 75%. With a more efficient, parallelized implementation, we believe our variational approach will enable reliable large-scale learning of topic models with larger K.

6.3 Image patch modeling.

Finally, we study 8×8 patches from grayscale natural images as in Zoran and Weiss (2012). We train on 3.5 million patches from 400 images, comparing HDP admixtures to Dirichlet process (DP) mixtures using a zero-mean Gaussian likelihood. The HDP model captures within-image patch similarity via image-specific mixture component frequencies. Both methods are evaluated on 50 heldout images scored via Eq. (22).

Fig. 7 shows merges and deletes removing junk topics while improving predictions, justifying the generality of these moves. Further, the HDP earns better prediction scores than the DP mixture. We illustrate this success by plotting sample patches from the top 4 topics (ranked by topic weight π) for several heldout images. The HDP adapts topic weights to each image, favoring smooth patches for some images (d) and textured patches for others (e-f). The less-flexible DP must use the same weights for all images (c).

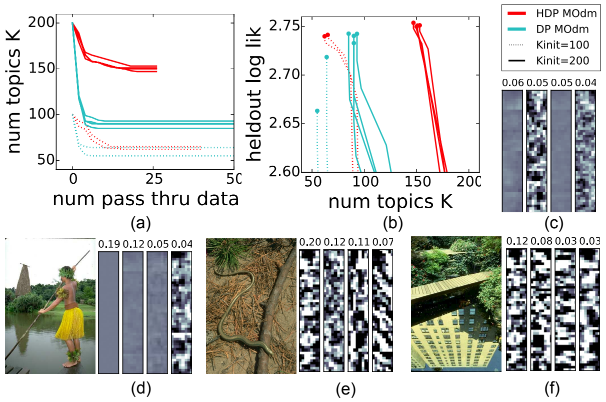

Figure 7: Comparison of DP mixtures and HDP admixtures on 3.5M image patches (Sec. 6.3). (a-b) Trace plots of number of topics and heldout likelihood, as in Fig 6. (c) Patches from the top 4 estimated DP clusters. Each column shows 6 stacked 8×8 patches sampled from one cluster. (d-f) Patches from 4 top-ranked HDP clusters for select test images from BSDS500 (Arbelaez et al., 2011).

7 CONCLUSION

We have developed a scalable variational algorithm for learning compact, interpretable HDP models from millions of examples. Our novel objective applies to any exponential family likelihood and could prove useful for sequential or relational models based on the HDP.

Acknowledgments This research supported in part by NSF CAREER Award No. IIS-1349774. M. Hughes supported in part by an NSF Graduate Research Fellowship under Grant No. DGE0228243.

References

P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, 2013.

D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *Neural Information Processing Systems*, 2013.

M. Bryant and E. B. Sudderth. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Neural Information Processing Systems*, 2012.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 2004.

M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1), 2013.

M. C. Hughes and E. B. Sudderth. Memoized online variational inference for Dirichlet process mixture models. In *Neural Information Processing Systems*, 2013.

P. Liang, S. Petrov, M. I. Jordan, and D. Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Empirical Methods in Natural Language Processing*, 2007.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

J. Paisley, C. Wang, and D. Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*, 2011.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77:291–330, 2008.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for HDP. In *Neural Information Processing Systems*, 2008.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

C. Wang and D. Blei. Truncation-free online variational inference for Bayesian nonparametric models. In *Neural Information Processing Systems*, 2012.

C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2011.

D. Zoran and Y. Weiss. Natural images, Gaussian mixtures and dead leaves. In *Neural Information Processing Systems*, 2012.