Supplementary Material: Reliable and Scalable Variational Inference for the Hierarchical Dirichlet Process

Michael C. Hughes MHUGHES@CS.BROWN.EDU
Dae Il Kim DAEIL@CS.BROWN.EDU
Erik B. Sudderth SUDDERTH@CS.BROWN.EDU
Department of Computer Science
Brown University
Providence, RI 02912-1910, USA

Abstract

This document contains supplementary details for the AISTATS 2015 paper "Reliable and scalable variational inference for the Hierarchical Dirichlet process" (Hughes et al., 2015). First, we show more detailed traceplots from the topic model experiments. Next, we provide expanded closed-form expressions for various ELBO terms in Sec. B, and a detailed derivation of our surrogate bound in Sec. C. Next, we describe how to optimize the free parameters $\hat{\rho}, \hat{\omega}$ of our variational objective in Sec. D. We discuss the local step algorithm in Sec E. Finally, Sec. F provides more formal details on the delete move.

A. Topic Model Experiments

Fig. 1 shows an expanded version of the main paper's Fig. 6 here, including another row of trace plots of active topic counts over passing through training data for different methods.

All methods are allowed initializations from $K = 100$ and $K = 200$ for NIPS, Wikipedia, and Science datasets. For New York Times articles, we show those same initializations for a subset of algorithms scalable enough to run on 1.8M documents.

We do not show trace plots of active topics K for NYTimes, because we find these do not change much at all. With 1.8M documents, it is easy for a few hundred topics to be used without redundancy or junk to remove.

B. Data-generation ELBO computation.

In the main paper, we define the variational objective $\mathcal{L}(\cdot)$ as

$$\mathcal{L}(\cdot) \triangleq \mathcal{L}_{data}(\cdot) + H_z(\cdot) + \mathcal{L}_{HDP}(\cdot) + \mathcal{L}_u(\cdot). \tag{1}$$

In this section we provide complete expressions for calculating $\mathcal{L}_{data}$. First, we give required equations for the Dirichlet-Multinomial case used in topic models. Second, we give a general closed-form for any choice of exponential family data-generation.

Later in Sec. (D) we define $\mathcal{L}_u(\cdot)$ and $\mathcal{L}_{HDP}(\cdot)$.

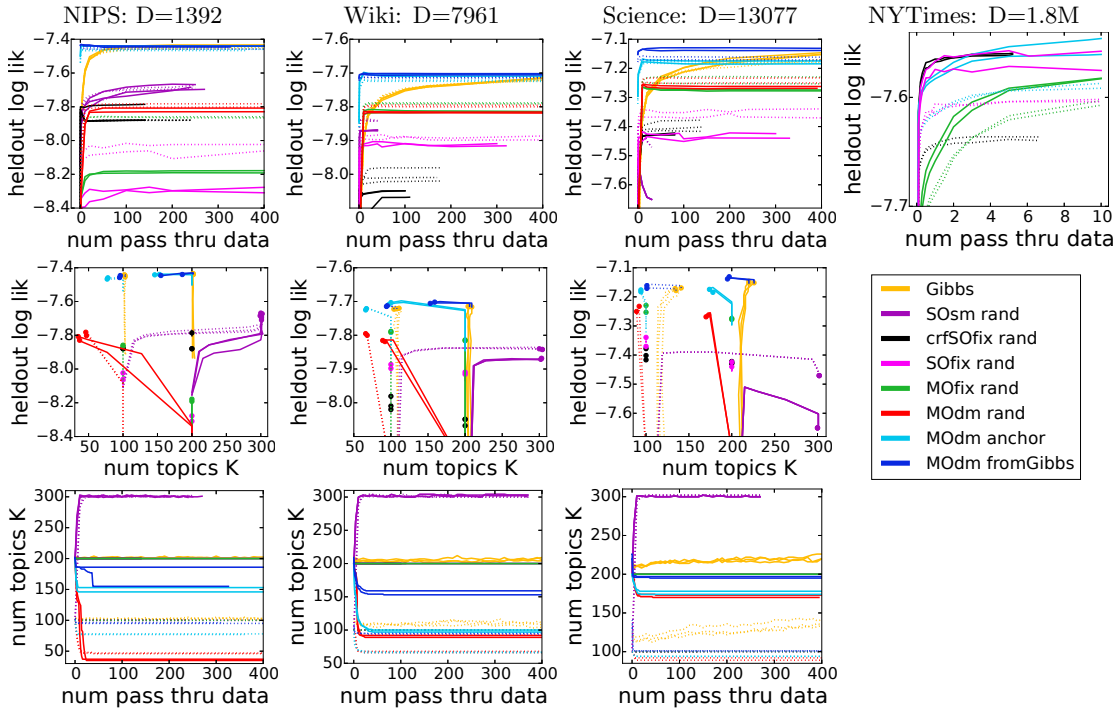

Figure 1: Topic modeling algorithm comparisons from main paper on NIPS, Wikipedia, Science, and NYTimes datasets (one dataset per column). Color indicates the algorithm used. Each algorithm given several runs at different values of the initial number of topics. $K = 100$ runs shown with dotted lines, $K = 200$ runs shown with solid lines. *Top Row:* Trace plots of heldout likelihood as more training data is seen. *Middle Row:* Traces of predictive power and number of active topics K during training. Solid dot indicates final result of each algorithm, trailing line indicates algorithm's trajectory from initialization. *Bottom Row:* Trace plots of the number of active topics K as more training data is seen.

B.1 Data generation term : Dirichlet-Multinomial

When H is Dirichlet over W words with parameter vector $\bar{\tau}$ of size W, we have

$$\begin{aligned}\mathcal{L}_{data}(\cdot) &\triangleq \mathbb{E}_q[\log p(x|z,\phi) + \log \tfrac{p(\phi|\bar{\tau})}{q(\phi|\hat{\tau})} \\ &= \sum_{k=1}^{K} c_D(\bar{\tau}) - c_D(\hat{\tau}_k) \\ &\quad + \sum_{k=1}^{K}\sum_{w=1}^{W} (S_{kw} + \bar{\tau}_w - \hat{\tau}_{kw})\mathbb{E}[\log \phi_{kw}]\end{aligned} \tag{2}$$

where $c_D(\cdot)$ is the log cumulant function of the Dirichlet defined in the main paper and S_{kw} counts the total number of words of type w assigned to topic k.

B.2 Data generation term : General EF Form.

Here, we provide a more formal treatment than the main paper of a general-purpose data-generation model, which subsumes Dirichlet-Multinomial, Beta-Bernoulli, Wishart-Normal, and many other well-known conjugate data generation models.

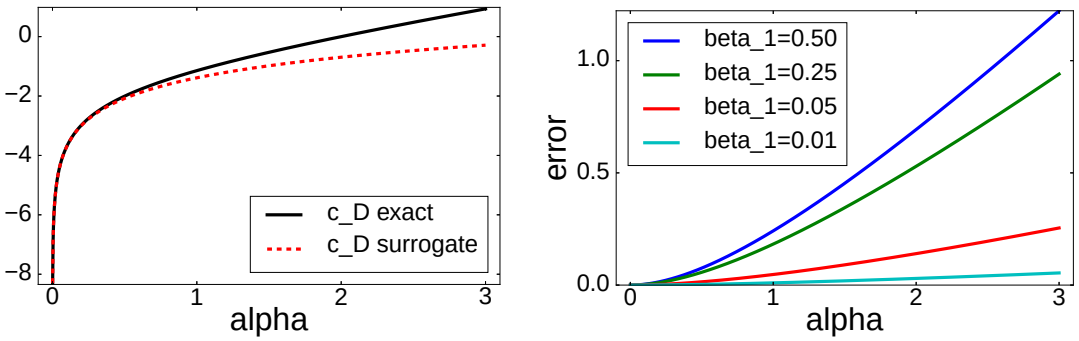

Figure 2: *Left:* Comparison of the exact value of $c_D(\alpha\beta)$ (Eq. (7), solid black) alongside our tight surrogate bound (Eq. (13), dashed red), across a range of possible $\alpha > 0$. We fix $K = 1$ and set $\beta = [\beta_1, 1 - \beta_1] = [0.5, 0.5]$. *Right:* The error in our surrogate bound across various values of α and β_1 for the $K = 1$ case. The function $c_D(\alpha\beta_1, \alpha(1 - \beta_1))$ is symmetric for β_1 around 0.5, so we need only consider the range $\beta_1 \in [0, 0.5]$ instead of $[0, 1]$.

First, each data item x_{dn} comes from an exponential family (EF) density family F with natural parameter ϕ_k and log cumulant $c_F(\cdot)$.

$$F : \log p(x_n|\phi_k) = s(x_n)^T \phi_k + c_F(\phi_k) \tag{3}$$

In turn, ϕ_k comes from exponential family H that is conjugate to F with natural parameters $\bar{\tau}, \bar{C}$, and cumulant function $c_H(\cdot)$.

$$H : \log p(\phi_k|\bar{\tau}, \bar{C}) = \bar{\tau}^T \phi_k + \bar{C}^T c_F(\phi_k) + c_H(\bar{\tau}, \bar{C}) \tag{4}$$

Here, $\bar{C}$ is interpreted as a scalar pseudo count, while $\bar{\tau}$ is a vector that acts like a pseudo-sufficient statistic.

Next, We assume the variational factor $q(\phi_k)$ also comes from H, with free parameters $\hat{C}_k, \hat{\tau}_k$. Under this assumption, we can evaluate the expectations that define $\mathcal{L}_{data}$ in closed form. First, define compact sufficient statistics S_k, N_k as follows

$$\begin{aligned} N_k(\hat{r}) &= \textstyle\sum_{d=1}^{D} \sum_{n=1}^{N_d} \hat{r}_{dnk} \\ S_k(\hat{r}) &= \textstyle\sum_{d=1}^{D} \sum_{n=1}^{N_d} x_{dn} \hat{r}_{dnk} \end{aligned} \tag{5}$$

Then, we have

$$\begin{aligned} \mathcal{L}_{data}(\cdot) &\triangleq \mathbb{E}_q[\log p(x|z, \phi) + \log \tfrac{p(\phi|\bar{\tau})}{q(\phi|\hat{\tau})} \\ &= \sum_{k=1}^{K} c_H(\bar{\tau}, \bar{C}) - c_H(\hat{\tau}_k, \hat{C}_k) \\ &\quad + (S_k + \bar{\tau} - \hat{\tau}_k)^T \mathbb{E}_{q(\phi_k)}[\phi_k] \\ &\quad + (N_k + \bar{C} - \hat{C}_k) \mathbb{E}_{q(\phi_k)}[c_F(\phi_k)] \end{aligned} \tag{6}$$

C. Surrogate Bound

Here, we provide formal details for a surrogate bound on the intractable expectation of the cumulant of the Dirichlet function. Plots describing our bound and its associated error can be found in Fig. 2.

C.1 Bound for the Dirichlet cumulant function

As in the main paper, we define the cumulant function c_D of the Dirichlet distribution as

$$c_D(\alpha\beta) = c_D(\alpha\beta_1, \alpha\beta_2, \ldots \alpha\beta_K, \alpha\beta_{K+1}) \triangleq \log\Gamma(\alpha) - \sum_{k=1}^{K+1} \log\Gamma(\alpha\beta_k) \tag{7}$$

where $\alpha > 0$ is a positive scalar, and $\beta = \{\beta_k\}_{k=1}^{K+1}$ is a vector of positive numbers that sum-to-one.

The log-Gamma function $\log\Gamma(\cdot)$ has the following definition[1] for scalar input $x > 0$:

$$-\log\Gamma(x) = \log x + \gamma x + \sum_{n=1}^{\infty} \left(\log\left(1 + \frac{x}{n}\right) - \frac{x}{n} \right) \tag{8}$$

where $\gamma \approx .57721$ is the Euler-Mascheroni constant.

Substituting this expansion for every $\log\Gamma(\cdot)$ in the definition of c_D, we find

$$\begin{aligned} c_D(\alpha\beta) = &-\log\alpha - \gamma\alpha - \sum_{n=1}^{\infty} \left(\log\left(1 + \frac{\alpha}{n}\right) - \frac{\alpha}{n} \right) \\ &+ \sum_{k=1}^{K+1} \left[\log\alpha\beta_k + \gamma\alpha\beta_k + \sum_{n=1}^{\infty} \left(\log\left(1 + \frac{\alpha\beta_k}{n}\right) - \frac{\alpha\beta_k}{n} \right) \right] \end{aligned} \tag{9}$$

Here, all the infinite sums are convergent. This allows some regrouping, and we find that several terms cancel to zero. Our expression for $c_D(\alpha\beta$ now simplifies to:

$$\begin{aligned} c_D(\alpha\beta) = &-\log\alpha + \sum_{k=1}^{K+1} \log\alpha\beta_k \\ &+ \sum_{n=1}^{\infty} \left(\log\left(\prod_{k=1}^{K+1} \left(1 + \frac{\alpha\beta_k}{n}\right) \right) - \log\left(1 + \frac{\alpha}{n}\right) \right) \end{aligned} \tag{10}$$

Finally, via the binomial product expansion below, we realize that the infinite sum must be larger than zero.

$$\prod_{k=1}^{K+1} \left(1 + \frac{\alpha\beta_k}{n}\right) = 1 + \sum_{k=1}^{K+1} \frac{\alpha\beta_k}{n} + \text{pos. const.} \quad \rightarrow \quad \prod_{k=1}^{K+1} \left(1 + \frac{\alpha\beta_k}{n}\right) \geq \left(1 + \frac{\alpha}{n}\right) \tag{11}$$

Thus, by simply leaving off the infinite sum from Eq. (11) we have a valid lower bound on $c_D(\cdot)$:

$$c_D(\alpha\beta) \geq -\log\alpha + \sum_{k=1}^{K+1} \log\alpha\beta_k \tag{12}$$

Expanding $\log\alpha\beta_k = \log\alpha + \log\beta_k$, we can further simplify to

$$c_D(\alpha\beta) \geq K\log\alpha + \sum_{k=1}^{K+1} \log\beta_k \tag{13}$$

1. `http://mathworld.wolfram.com/LogGammaFunction.html`

C.2 Bound for the Expected value of the Dirichlet cumulant function

We now wish to compute the expected value of the bound in Eq. (13), under $q(u|\hat{\rho}, \hat{\omega})$.

First, we can recall how to write $\log \beta_k$ in terms of our stick length variables u:

$$\log \beta_k \triangleq \begin{cases} \log \Big(u_k \prod_{\ell=1}^{k-1} (1-u_\ell) \Big) & \text{if } k \in \{1, 2, \dots K\} \\ \log \Big(\prod_{\ell=1}^{K} (1-u_\ell) \Big) & \text{if } k = K+1 \end{cases} \tag{14}$$

Using this definition and carefully expanding the log product into a sum of logs, we can write Eq. (13) in terms of u as follows

$$c_D(\alpha\beta) \geq K \log \alpha + \sum_{k=1}^{K} \Big(\log u_k + (K+1-k) \log (1-u_k) \Big) \tag{15}$$

Finally, applying the expectation operator and using linearity of expectations we have

$$\mathbb{E}_q[c_D(\alpha\beta)] \geq K \log \alpha + \sum_{k=1}^{K} \Big(\mathbb{E}_q[\log u_k] + (K+1-k) \mathbb{E}_q[\log 1-u_k] \Big) \tag{16}$$

D. Optimization updates for ρ, ω

In this section, we describe how we determine optimal values for free parameters ρ, ω given fixed values of other variational free parameters $\hat{\theta}, \hat{\rho}, \hat{\tau}$. We proceed in three steps: First showing mathematically how the ELBO terms $\mathcal{L}_{HDP}$ and $\mathcal{L}_u$ can be expanded and manipulated to give a closed form objective depending only on ρ, ω. Second, we frame a constrained optimization problem for ρ, ω given this objective. Finally, we describe a transformation to an unconstrained optimization problem that yields optimal ρ, ω values and can be implemented using modern gradient descent methods like L-BFGS.

D.1 Derivation of Optimization Objective

The free parameters ρ, ω appear in both the surrogate bound on $\mathcal{L}_{HDP}$, and the global term $\mathcal{L}_u$. We give complete forms of each below, then bring them together to form the function to optimize to find the best ρ, ω values.

First, $\mathcal{L}_u$ is defined as:

$$\begin{aligned} \mathcal{L}_u(\hat{\rho}, \hat{\omega}) &= \mathbb{E}_q \Big[\log \frac{p(u|\gamma)}{q(u|\hat{\rho}, \hat{\omega})} \Big] \\ &= \sum_{k=1}^{K} c_B(1, \gamma) - c_B(\hat{\rho}_k \hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k) \\ &\quad + \Big(1 - \hat{\rho}_k \hat{\omega}_k\Big) \mathbb{E}[\log u_k] \\ &\quad + \Big(\gamma - (1-\hat{\rho}_k)\hat{\omega}_k\Big) \mathbb{E}[\log 1-u_k] \end{aligned} \tag{17}$$

where $c_B(\cdot)$ is the log cumulant function of the Beta distribution (a simplified two argument case of $c_D(\cdot)$). Furthermore, all expectations here have closed-form

$$\begin{aligned} \mathbb{E}[\log u_k] &= \psi(\hat{\rho}_k \hat{\omega}_k) - \psi(\hat{\omega}_k) \\ \mathbb{E}[\log 1 - u_k] &= \psi((1-\hat{\rho}_k)\hat{\omega}_k) - \psi(\hat{\omega}_k) \end{aligned} \tag{18}$$

The other term $\mathcal{L}_{HDP}$ is given by

$$\begin{aligned}\mathcal{L}_{HDP}(\cdot) &= \mathbb{E}_q[\log\Big[p(z)\frac{p(\pi)}{q(\pi)}\Big]] \\ &= \textstyle\sum_{d=1}^{D}\mathbb{E}_q[c_D(\alpha\beta)] - c_D(\hat{\theta}_d) \\ &+ \textstyle\sum_{k=1}^{K+1}\Big(N_{dk}(\hat{r}) + \alpha\mathbb{E}_q[\beta_k] - \hat{\theta}_{dk}\Big)T_k(\hat{\theta})\end{aligned} \tag{19}$$

where $T_k(\theta) = \sum_{d=1}^{D}\mathbb{E}_q[\log\pi_{dk}]$, and $N_{dk}(\hat{r}) = \sum_{n=1}^{N_d}\hat{r}_{dnk}$.
Applying our surrogate bound to deal with intractable expectation $\mathbb{E}[c_D(\alpha\beta)]$, we have

$$\begin{aligned}\mathcal{L}_{HDP}(\cdot) \geq DK\log\alpha &- \sum_{d=1}^{D}c_D(\hat{\theta}_d) \\ &+ D\sum_{k=1}^{K}\Big(\mathbb{E}_q[\log u_k] + (K{+}1{-}k)\mathbb{E}_q[\log 1{-}u_k]\Big) \\ &+ \sum_{d=1}^{D}\sum_{k=1}^{K+1}\Big(N_{dk}(\hat{r}) + \alpha\mathbb{E}_q[\beta_k] - \hat{\theta}_{dk}\Big)T_k(\hat{\theta})\end{aligned} \tag{20}$$

Optimization objective: Combining Eq. (20) and Eq. (19) and keeping only terms that depend on ρ, ω, we can define a new objective function $\mathcal{L}_G$:

$$\begin{aligned}\mathcal{L}_G(\rho,\omega) = \sum_{k=1}^{K} &-c_B(\hat{\rho}_k\hat{\omega}_k, (1{-}\hat{\rho}_k)\hat{\omega}_k) \\ &+ \Big(D+1-\hat{\rho}_k\hat{\omega}_k\Big)\Big[\psi(\hat{\rho}_k\hat{\omega}_k) - \psi(\hat{\omega}_k)\Big] \\ &+ \Big(D(K{+}1{-}k) + \gamma - (1{-}\hat{\rho}_k)\hat{\omega}_k\Big)\Big[\psi((1{-}\hat{\rho}_k)\hat{\omega}_k) - \psi(\hat{\omega}_k)\Big] \\ &+ \alpha\mathbb{E}[\beta_k]T_k\end{aligned} \tag{21}$$

D.2 Constrained optimization problem.

Eq. 21 is the objective function of the constrained optimization problem we solve to determine optimal free parameters ρ, ω.

$$\hat{\rho},\hat{\omega} = \text{argmax}_{\rho,\omega}\ \ \mathcal{L}_G(\rho,\omega) \tag{22}$$

$$\text{subject to}\ \ 0<\rho_k<1,\ \ \ \omega_k>0\ \ \text{for}\ k=1,2,...K \tag{23}$$

Below, we supply closed-form analytical gradient expressions for both ρ and ω, which can be used with modern first-order constrained optimization solvers.

Gradient computation for ω: Taking the derivative of Eq. (21) with respect to entry m of vector ω, we have

$$\begin{aligned}\frac{\delta}{\delta\omega_m}[\mathcal{L}_G] = &\Big(D+1-\rho_m\omega_m\Big)\Big[\rho_m\psi'(\rho_m\omega_m) - \psi'(\omega_m)\Big] \\ &+ \Big(D(K{+}1{-}k) + \gamma - (1{-}\rho_m)\omega_m\Big)\Big[(1{-}\rho_m)\psi'((1{-}\rho_m)\omega_m) - \psi'(\omega_m)\Big]\end{aligned} \tag{24}$$

Gradient computation for ρ: First, define Δ as a $K \times K+1$ matrix of partial derivatives of $\mathbb{E}_q[\beta_k]$ with respect to ρ

$$\Delta_{mk} \triangleq \frac{\delta}{\delta \rho_m}\mathbb{E}[\beta_k] = \begin{cases} -\frac{1}{1-\rho_m}\mathbb{E}[\beta_k] & \text{if } m < k \\ \frac{1}{\rho_m}\mathbb{E}[\beta_k] & \text{if } m = k \\ 0 & \text{if } m > k \end{cases} \tag{25}$$

Now, the derivative of Eq. (21) with respect to entry m of vector ρ is

$$\begin{aligned} \frac{\delta}{\delta \rho_m}[\mathcal{L}_G] &= \omega_m(D+1-\rho_m\omega_m)\psi'(\rho_m\omega_m) \\ &\quad - \omega_m\Big(D(K+1-k)+\gamma-(1-\rho_m)\omega_m\Big)\psi'((1-\rho_m)\omega_m) \\ &\quad + \alpha\sum_{k=1}^{K}\Delta_{mk}T_k \end{aligned} \tag{26}$$

D.3 Transformation to unconstrained optimization problem.

Both target variables ρ, ω have simple bound constraints on each of their K entries. Each entry of ρ lies in $[0,1]$, while each entry of ω must be larger than 0. We can define an invertible transform between constrained scalars ρ_k, ω_k and unconstrained real scalar variables c_k, d_k as follows:

$$\begin{aligned} c_k &\triangleq \text{sigmoid}^{-1}(\rho_k) & \rho_k &\triangleq \text{sigmoid}(c_k) = \frac{1}{1+e^{-c_k}} \\ d_k &\triangleq \log \omega_k & \omega_k &\triangleq e^{d_k} \end{aligned} \tag{27}$$

As shorthand, we write $\rho(c)$ to denote the vector ρ obtained by transforming the input vector c. Similarly, we write $\omega(d)$ to be the vector ω obtained by applying the transform to input d. We can then define an unconstrained optimization problem

$$c^*, d^* \leftarrow \text{argmax}_{c,d} \;\; \mathcal{L}_G\Big(\rho(c), \omega(d)\Big) \tag{28}$$

The optimal values c^*, d^* can be simply transformed to ρ^*, ω^*, which are by construction optimal solutions to our original problem defined in Eq. (22)

Our unconstrained objective can be solved via gradient descent, where the gradients can be easily computed by the chain rule using our original gradients with respect to ρ, ω as inputs.

The gradient at entry m of vector c is

$$\begin{aligned} \frac{\delta}{\delta c_m}[\mathcal{L}_G] &\triangleq \frac{\delta}{\delta c_m}[\rho_m] \cdot \frac{\delta}{\delta \rho_m}\mathcal{L}_G \\ &= \rho_m(1-\rho_m)\frac{\delta}{\delta \rho_m}\mathcal{L}_G, \quad \text{where } \rho_m \triangleq \frac{1}{1+e^{-c_m}} \end{aligned} \tag{29}$$

Similarly, the gradient at entry m of vector d is

$$\begin{aligned} \frac{\delta}{\delta d_m}[\mathcal{L}_G] &\triangleq \frac{\delta}{\delta d_m}[\omega_m] \cdot \frac{\delta}{\delta \omega_m}\mathcal{L}_G \\ &= \omega_m\frac{\delta}{\delta \omega_m}\mathcal{L}_G, \quad \text{where } \omega_m \triangleq e^{d_m} \end{aligned} \tag{30}$$

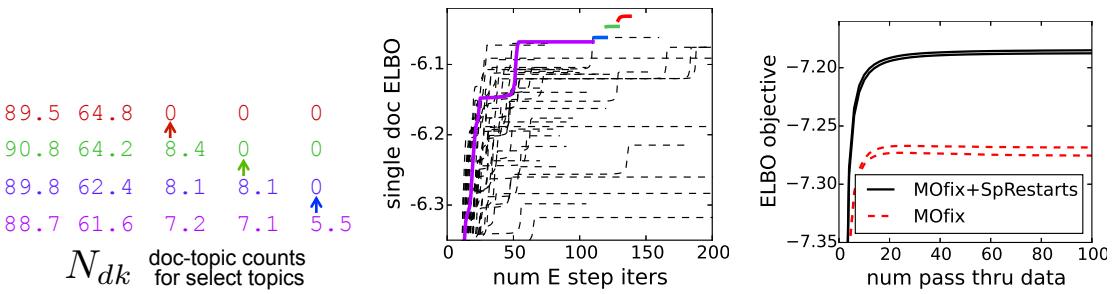

Figure 3: Sparsity-promoting restarts for local update steps on the Science corpus with $K = 100$. *Left:* Example fixed points of the document-topic count summary statistic N_{dk} for a single document in the Science corpus. We show only select topic indices out of all $K = 100$. *Center:* Trace of a single document's objective $\mathcal{L}$ during the local step inference for 50 random initializations (dashed lines). The solid lines show one run with sparsity-promoting moves enabled. This run climbs through the color coded fixed points in the left plot. *Right:* Trace plot of the whole-dataset objective $\mathcal{L}$ across many passes through the whole Science corpus. Using sparsity-promoting restarts yields noticeable improvements in model quality.

E. Local Step Algorithm

In this section, we describe the local step algorithm. When visiting a document d, we need to infer token soft assignments $\hat{r}_{dn}$ for the N_d tokens in this document, as well as the topic weight parameter vector $\hat{\theta}_d$ for this document (a vector of size $K + 1$). As described in the main paper, these two free parameters have inter-dependent updates. Thus, we need an initialization heuristic.

We suggest thinking of the initialization in terms of the initial value for $\tilde{\pi}_{dk} = \exp \mathbb{E}[\log \pi_{dk}]$, for each active topic $k \in 1, 2, \ldots K$. We interpret this quantity intuitively as the "probability" of topic k in in document d. We are free to set this initial vector to any valid vector on the simplex. In the main paper Fig. 3, reproduced and expanded here in Fig. 3, we show how 50 random initializations of $\tilde{\pi}_{dk}$ can lead to very diverse fixed points with different values for N_{dk} and consequently different objective score trajectories.

As a reasonable heuristic, we suggest setting $\tilde{\pi}_{dk}$ to $\mathbb{E}[\beta_k]$. This makes sense under the model: when visiting a new document, we sensibly guess that the probability of topic k in this document will be our estimate of the topic's probability across all documents.

Given an initial value of $\tilde{\pi}_d$, we obtain an initial value for $\hat{r}_d$ using the update in the main paper. Then, we iteratively alternate between updates to $\hat{\theta}_d$ and $\hat{r}_d$ until convergence occurs. To assess convergence, we monitor the sufficient statistic N_{dk} for each active topic, and halt when the absolute change from the previous iteration for all topics falls below a defined threshold (we use 0.0001), or until a prescribed budget of computation has been exhausted (we set this to 100 iterations).

If computational cost is not a concern, trying multiple restarts of this algorithm from different initializations and selecting the best one in terms of the resulting objective $\mathcal{L}(\cdot)$ would be good practice. However, each independent restart each requires many iterations to converge, which can be expensive. We find that our recommended initialization plus our sparsity-promoting restarts (where each restart runs for a very small budget of 2-5 iterations) provide successful performance while remaining relatively affordable. You can see in Fig. 3 that with both this initialization and sparse restarts enabled, trace plots of the whole-dataset objective as more data is seen show dramatic improvement over simply using the heuristic initialization alone (dashed red line).

F. Delete Move

Delete moves allow removal of unnecessary topics that is not possible via our pair-wise merge moves. As shown in our toy-bars experiment plots (main paper's Fig. 5), a common pitfall is that inference can get stuck with some extra "junk" topics which are assigned to few tokens in only a few documents. These topics are often not possible to eliminate via pair-wise merges, but rather require document-specific changes to local token parameters. To remove these "junk" topics, delete moves provide flexible document-specific reassignment at greater cost than pair-wise merge moves. Merges and deletes complement each other: merges remove redundant topics that appear in many documents, while deletes are for rare, junk topics. Together, these moves create compact, interpretable models that are not slowed down by useless computations.

Below, we describe first how a delete move would work if we could afford explicitly updating all documents in the dataset. Next, we describe how deletes work in our memoized framework.

F.1 Whole-dataset delete construction.

Delete moves remove some topic, indexed by j, from a current set of parameters and sufficient statistics of size $K+1$. For simplicity, this explanation assumes j is last in index order, but in fact j can be at any position. The move constructs new parameters and new sufficient statistics S', N' of size K, where any mass assigned to topic j has been reallocated among the other topics. Here, unlike the merge move, we have no one-step rule for constructing the candidate local parameters. Instead, we use a heuristic initialization followed by refinement coordinate ascent updates. We initialize sufficient statistics by simply removing any entries associated with topic j.

$$\begin{aligned} \text{Original: } N &= [N_1 \quad \dots \quad N_K \; N_j] \\ \text{Candidate Init: } N' &= [N_1 \quad \dots \quad N_K] \end{aligned} \tag{31}$$

Given these initial summaries, we take a global step to create K candidate global parameters. The main paper's Fig. 1 reviews the big picture for how summary statistics lead to global parameter updates. After creating the candidate global parameters, we realize that $\hat{\tau}'$ will have exactly the same first K topics as the original model. For $\hat{\rho}', \hat{\omega}'$, the resulting $\mathbb{E}[\beta]$ will be similar, too. Next, a local step reassigns *all* tokens (including those ignored) among the K remaining topics. After one more global step, we have a viable candidate model q' representing the whole dataset. This model can be compared to the original, and kept if the objective improves.

F.2 Memoized delete construction.

For large datasets, it is infeasible to perform several local/global update cycles for all documents just to evaluate one candidate move. A more scalable delete move is possible because we assume junk topic j has only a small subset of documents with appreciable mass, while most documents assign $N_{dj} \approx 0$. Thus, only the small set satisfying $N_{dj} > \epsilon$ need to be edited explicitly, where we set $\epsilon = 0.01$.

The memoized delete move happens in three steps. First, we gather all documents satisfying this threshold test into a target dataset during a standard pass of the dataset. Second, we construct the delete candidate model q' from the target set, performing the simple construction described above while holding the non-target sufficient statistics fixed. That is, for each additive sufficient statistic vector N, S, T in the previous model, we create candidates N', S', T' that satisfy the following relation:

$$N'_k = N_k - N_k^{before} + N_k^{after}, k \in \{1 \dots K\} \tag{32}$$

Here, N_k^{before} is the statistic for topic k on the target set before removing j, and N_k^{after} is the computed statistic on the target set after removing j and performing the several updates.

For the specific case of the token count statistic N' on the target set, we know that $N_\epsilon + \sum_{k=1}^K N'_k = N_j + \sum_{k=1}^K N_k$ where N_ϵ represents the small mass assigned to j from documents that did not pass the threshold test. If accepted, sufficient statistic vector N' will soon accurately reflect all data (including the small discarded mass) after a complete pass of local and global steps at all batches.

To determine acceptance, we evaluate the objective $\mathcal{L}(\cdot)$ using candidate global parameters $\hat{\rho}', \hat{\omega}', \hat{\tau}'$, which are obtained via direct updates from N', S', T'. For the local arguments to $\mathcal{L}(\cdot)$, we use the inferred parameters $\hat{r}_d, \hat{\theta}_d$ from documents in the target set.

If the candidate model improves this objective, we accept it. After accepting, we need to adjust all stored batch-specific summaries to reflect the new model. Otherwise, our new aggregate summaries will not be consistent with the sum of stored batch summaries, and subsequent incremental updates will be invalid. We thus edit the stored statistics for each batch to reflect the final state of the target-set documents from that batch.

Immediately after a delete move, we do not have the required ELBO summaries to exactly compute the bound after visiting the next batch. However, after completing a complete lap through all batches, the relevant summaries will be refreshed and the ELBO computable.

F.3 Selecting topics to delete.

Delete move costs scale with the number of documents in the target set. We specify a maximum cap for the total documents we can afford to process as a target set: $D_{target} = 500$. Any topic occuring in fewer than D_{target} documents is eligible for deletion. We select among this eligible set as many topics as possible until the total cap is reached, and build the target set as the union of all documents passing the threshold test for any selected topic. This allows potentially *multiple* topics to be deleted in one pass through the data, each one considered independently, while never exceeding the specified cap on target set size.

References

Michael C. Hughes, Dae Il Kim, and Erik B. Sudderth. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2015.