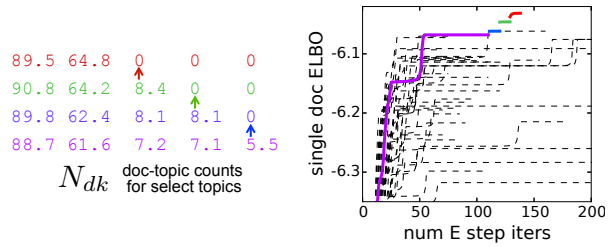


Figure 3: Sparsity-promoting restarts for local steps on the Science corpus with $K = 100$. *Left:* Example fixed points of the topic usage statistic $N_{dk}$ for one document. *Right:* Trace of single-document ELBO objective during E-step inference for 50 random initializations (dashed lines), plus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

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

However, $c_D$ is the cumulant function of the Dirichlet,

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

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 $c_D(\cdot)$:

$$c_D(\alpha\beta) \geq K \log \alpha + \textstyle\sum_{k=1}^{K} \log u_k \quad (12)$$
$$+ \textstyle\sum_{k=1}^{K} (K+1-k) \log 1-u_k.$$

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

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

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

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

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. Reasignment 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

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

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