This function provides a lower bound on the marginal evidence: $\log p(x|\gamma, \alpha, \kappa, \bar{\tau}) \geq \mathcal{L}$. Improving this bound is equivalent to minimizing KL($q \,||\, p$). Its four component terms are defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{data}}(x, \hat{r}, \hat{\tau}) &\triangleq \mathbb{E}_q \left[\log p(x \mid z, \phi) + \log \tfrac{p(\phi)}{q(\phi)} \right], & \mathcal{L}_{\text{entropy}}(\hat{s}) &\triangleq -\mathbb{E}_q \left[\log q(z) \right], \\ \mathcal{L}_{\text{hdp-local}}(\hat{s}, \hat{\theta}, \hat{\rho}, \hat{\omega}) &\triangleq \mathbb{E}_q \left[\log p(z \mid \pi) + \log \tfrac{p(\pi)}{q(\pi)} \right], & \mathcal{L}_{\text{hdp-global}}(\hat{\rho}, \hat{\omega}) &\triangleq \mathbb{E}_q \left[\log \tfrac{p(u)}{q(u)} \right]. \end{aligned} \tag{6}$$

Detailed analytic expansions for each term are available in the supplement.

3.2 Tractable Posterior Inference for Global State Probabilities

Previous variational methods for the HDP-HMM [7], and for HDP topic models [16] and HDP grammars [17], used a zero-variance point estimate for the top-level state probabilities β. While this approximation simplifies inference, the variational objective no longer bounds the marginal evidence. Such pseudo-bounds are unsuitable for model selection and can favor models with redundant states that do not explain any data, but nevertheless increase computational and storage costs [14].

Because we seek to learn compact and interpretable models, and automatically adapt the truncation level K to each dataset, we instead place a proper beta distribution on u_k, $k \in 1, 2, \ldots K$:

$$q(u_k) \triangleq \text{Beta}(\hat{\rho}_k \hat{\omega}_k, (1-\hat{\rho}_k)\hat{\omega}_k), \text{ where } \hat{\rho}_k \in (0, 1), \hat{\omega}_k > 0. \tag{7}$$

Here $\hat{\rho}_k = \mathbb{E}_{q(u)}[u_k]$, $\mathbb{E}_{q(u)}[\beta_k] = \hat{\rho}_k \mathbb{E}[\beta_{>k-1}]$, and $\mathbb{E}_{q(u)}[\beta_{>k}] = \prod_{\ell=1}^{k}(1-\hat{\rho}_\ell)$. The scalar $\hat{\omega}_k$ controls the variance, where the zero-variance point estimate is recovered as $\hat{\omega}_k \to \infty$.

The beta factorization in Eq. (7) complicates evaluation of the marginal likelihood bound in Eq. (6):

$$\begin{aligned} \mathcal{L}_{\text{hdp-local}}(\hat{s}, \hat{\theta}, \hat{\rho}, \hat{\omega}) = \mathbb{E}_{q(u)}[c_D(\alpha_0 \beta)] + \textstyle\sum_{k=1}^{K} \mathbb{E}_{q(u)}[c_D(\alpha\beta + \kappa\delta_k)] \\ - \textstyle\sum_{k=0}^{K} c_D(\hat{\theta}_k) + \sum_{k=0}^{K} \sum_{\ell=1}^{K+1} (M_{k\ell}(\hat{s}) + \alpha_k \mathbb{E}_{q(u)}[\beta_\ell] + \kappa\delta_k(\ell) - \hat{\theta}_{k\ell}) P_{k\ell}(\hat{\theta}). \end{aligned} \tag{8}$$

The Dirichlet cumulant function c_D maps $K+1$ positive parameters to a log-normalization constant. For a non-sticky HDP-HMM where $\kappa = 0$, previous work [14] established the following bound:

$$c_D(\alpha\beta) \triangleq \log \Gamma(\alpha) - \textstyle\sum_{k=1}^{K+1} \log \Gamma(\alpha\beta_k) \geq K \log \alpha + \sum_{\ell=1}^{K+1} \log \beta_\ell. \tag{9}$$

Direct evaluation of $\mathbb{E}_{q(u)}[c_D(\alpha\beta)]$ is problematic because the expectations of log-gamma functions have no closed form, but the lower bound has a simple expectation given beta distributed $q(u_k)$.

Developing a similar bound for sticky models with $\kappa > 0$ requires a novel contribution. To begin, in the supplement we establish the following bound for any $\kappa > 0, \alpha > 0$:

$$c_D(\alpha\beta + \kappa\delta_k) \geq K \log \alpha - \log(\alpha + \kappa) + \log(\alpha\beta_k + \kappa) + \textstyle\sum_{\ell=1\, \ell \neq k}^{K+1} \log(\beta_\ell). \tag{10}$$

To handle the intractable term $\mathbb{E}_{q(u)}[\log(\alpha\beta_k + \kappa)]$, we leverage the concavity of the logarithm:

$$\log(\alpha\beta_k + \kappa) \geq \beta_k \log(\alpha + \kappa) + (1 - \beta_k) \log \kappa. \tag{11}$$

Combining Eqs. (10) and (11) and taking expectations, we can evaluate a lower bound on Eq. (8) in closed form, and thereby efficiently optimize its parameters. As illustrated in Fig. 2, this rigorous lower bound on the marginal evidence $\log p(x)$ is quite accurate for practical hyperparameters.

3.3 Batch and Stochastic Variational Inference

Most variational inference algorithms maximize $\mathcal{L}$ via coordinate ascent optimization, where the best value of each parameter is found given fixed values for other variational factors. For the HDP-HMM this leads to the following updates, which when iterated converge to some local maximum.

Local update to $q(z_n)$. The assignments for each sequence z_n can be updated independently via dynamic programming [18]. The forward-backward algorithm takes as input a $T_n \times K$ matrix of log-likelihoods $\mathbb{E}_q[\log p(x_n \mid \phi_k)]$ given the current $\hat{\tau}$, and log transition probabilities P_{jk} given the current $\hat{\theta}$. It outputs the optimal marginal state probabilities $\hat{s}_n, \hat{r}_n$ under objective $\mathcal{L}$. This step has cost $\mathcal{O}(T_n K^2)$ for sequence n, and we can process multiple sequences in parallel for efficiency.

Global update to $q(\phi)$. Conjugate priors lead to simple closed-form updates $\hat{\tau}_k = \bar{\tau} + S_k$, where sufficient statistic S_k summarizes the data assigned to state k: $S_k \triangleq \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{r}_{ntk} s_F(x_{nt})$.

Global update to $q(\pi)$. For each state $k \in \{0, 1, 2 \ldots K\}$, the positive vector $\hat{\theta}_k$ defining the optimal Dirichlet posterior on transition probabilities from state k is $\hat{\theta}_{k\ell} = M_{k\ell}(\hat{s}) + \alpha\beta_\ell + \kappa\delta_k(\ell)$. Statistic $M_{k\ell}(\hat{s})$ counts the expected number of transitions from state k to ℓ across all sequences.