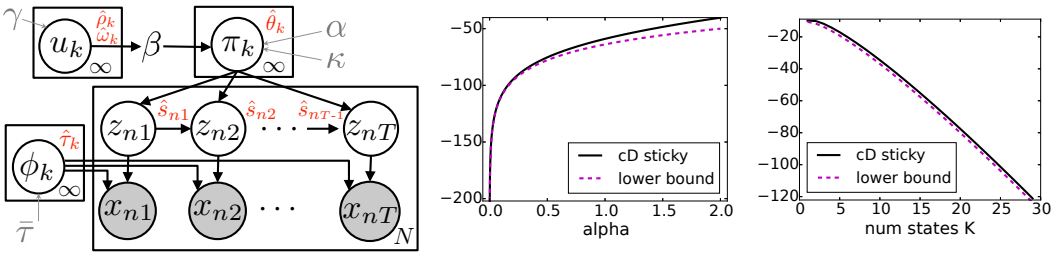

Figure 2: *Left:* Graphical representation of the HDP hidden Markov model. Variational parameters are shown in red. *Center:* Our surrogate bound for the sticky Dirichlet cumulant function c_D (Eq. 9) as a function of α, computed with $\kappa = 100$ and uniform β with $K = 20$ active states. *Right:* Surrogate bound vs. K, with fixed $\kappa = 100, \alpha = 0.5$. This bound remains tight when our state adaptation moves insert or remove states.

3 Memoized and Stochastic Variational Inference

After observing data x, our inferential goal is posterior knowledge of top-level conditional probabilities u, HMM parameters π, ϕ, and assignments z. We refer to u, π, ϕ as *global* parameters because they generalize to new data sequences. In contrast, the states z_n are *local* to a specific sequence x_n.

3.1 A Factorized Variational Lower Bound

We seek a distribution q over the unobserved variables that is close to the true posterior, but lies in the simpler factorized family $q(\cdot) \triangleq q(u)q(\phi)q(\pi)q(z)$. Each factor has exponential family form with free parameters denoted by hats, and our inference algorithms update these parameters to minimize the Kullback-Leibler (KL) divergence $\mathrm{KL}(q \,||\, p)$. Our chosen factorization for q is similar to [7], but includes a substantially more accurate approximation to $q(u)$ as detailed in Sec. 3.2.

Factor $q(z)$. For each sequence n, we use an independent factor $q(z_n)$ with Markovian structure:

$$q(z_n) \triangleq \left[\prod_{k=1}^{K} \hat{r}_{n1k}^{\delta_k(z_{n1})}\right] \prod_{t=1}^{T-1} \prod_{k=1}^{K} \prod_{\ell=1}^{K} \left[\frac{\hat{s}_{ntk\ell}}{\hat{r}_{ntk}}\right]^{\delta_k(z_{nt})\delta_\ell(z_{n,t+1})} \tag{4}$$

Free parameter vector $\hat{s}_{nt}$ defines the joint assignment probabilities $\hat{s}_{ntk\ell} \triangleq q(z_{n,t+1} = \ell, z_{nt} = k)$, so the K^2 non-negative entries of $\hat{s}_{nt}$ sum to one. The parameter $\hat{r}_{nt}$ defines the marginal probability $\hat{r}_{ntk} = q(z_{nt} = k)$, and equals $\hat{r}_{ntk} = \sum_{\ell=1}^{K} \hat{s}_{ntk\ell}$. We can find the expected count of transitions from state k to ℓ across all sequences via the sufficient statistic $M_{k\ell}(\hat{s}) \triangleq \sum_{n=1}^{N} \sum_{t=1}^{T_n-1} \hat{s}_{ntk\ell}$.

The truncation level K limits the total number of states to which data is assigned. Under our approximate posterior, only $q(z_n)$ is constrained by this choice; no global factors are truncated. Indeed, if data is only assigned to the first K states, the conditional independence properties of the HDP-HMM imply that $\{\phi_k, u_k \mid k > K\}$ are independent of the data. Their optimal variational posteriors thus match the prior, and need not be explicitly computed or stored [15, 16]. Simple variational algorithms treat K as a fixed constant [7], but Sec. 4 develops novel algorithms that fit K to data.

Factor $q(\pi)$. For the starting state ($k = 0$) and each state $k \in 1, 2, \ldots$, we define $q(\pi_k)$ as a Dirichlet distribution: $q(\pi_k) \triangleq \mathrm{Dir}(\hat{\theta}_{k1}, \ldots, \hat{\theta}_{kK}, \hat{\theta}_{k>K})$. Free parameter $\hat{\theta}_k$ is a vector of $K + 1$ positive numbers, with one entry for each of the K active states and a final entry for the aggregate mass of all other states. The expected log transition probability between states k and ℓ, $P_{k\ell}(\hat{\theta}) \triangleq \mathbb{E}_q[\log \pi_{k\ell}] = \psi(\hat{\theta}_{k\ell}) - \psi(\sum_{m=1}^{K+1} \hat{\theta}_{km})$, is a key sufficient statistic.

Factor $q(\phi)$. Emission parameter ϕ_k for state k has factor $q(\phi_k) \triangleq H(\hat{\tau}_k)$ conjugate to the likelihood F. The supplement provides details for Bernoulli, Gaussian, and auto-regressive F.

We score the approximation q via an objective function $\mathcal{L}$ that assigns a scalar value (higher is better) to each possible input of free parameters, data x, and hyperparameters $\gamma, \alpha, \kappa, \bar{\tau}$:

$$\mathcal{L}(\cdot) \triangleq \mathbb{E}_q\left[\log p(x, z, \pi, u, \phi) - \log q(z, \pi, u, \phi)\right] = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{entropy}} + \mathcal{L}_{\text{hdp-local}} + \mathcal{L}_{\text{hdp-global}}. \tag{5}$$