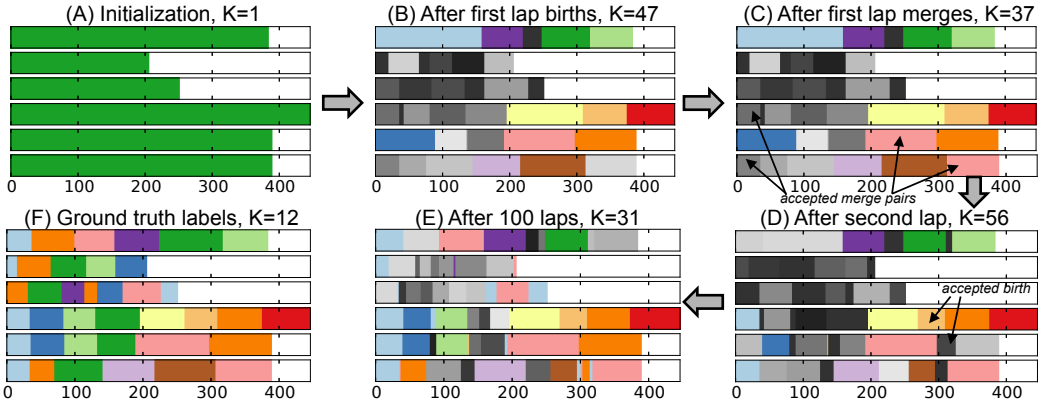

Figure 1: Illustration of our new birth/merge/delete variational algorithm as it learns to segment motion capture sequences into common exercise types (Sec. 5). Each panel shows segmentations of the same 6 sequences, with time on the horizontal axis. Starting from just one state (A), birth moves at the first sequence create useful states. Local updates to each sequence in turn can use existing states or birth new ones (B). After all sequences are updated once, we perform merge moves to clean up and *lap* is complete (C). After another complete lap of birth updates at each sequence followed by merges and deletes, the segmentation is further refined (D). After many laps, our final segmentation (E) aligns well to labels from a human annotator (F), with some true states aligning to multiple learned states that capture subject-specific variability in exercises.

2 Hierarchical Dirichlet Process Hidden Markov Models

We wish to jointly model N sequences, where sequence n has data $x_n = [x_{n1}, x_{n2}, \ldots, x_{nT_n}]$ and observation x_{nt} is a vector representing interval or timestep t. For example, $x_{nt} \in \mathbb{R}^D$ could be the spectrogram for an instant of audio, or human limb positions during a 100ms interval.

The HDP-HMM explains this data by assigning each observation x_{nt} to a single hidden state z_{nt}. The chosen state comes from a countably infinite set of options $k \in \{1, 2, \ldots\}$, generated via Markovian dynamics with initial state distributions π_0 and transition distributions $\{\pi_k\}_{k=1}^{\infty}$:

$$p(z_{n1} = k) = \pi_{0k}, \quad p(z_{nt} = \ell \mid z_{n,t-1} = k) = \pi_{k\ell}. \tag{1}$$

We draw data x_{nt} given assigned state $z_{nt} = k$ from an exponential family likelihood F:

$$F : \log p(x_{nt} \mid \phi_k) = s_F(x_{dn})^T \phi_k + c_F(\phi_k), \qquad H : \log p(\phi_k \mid \bar{\tau}) = \phi_k^T \bar{\tau} + c_H(\bar{\tau}). \tag{2}$$

The natural parameter ϕ_k for each state has conjugate prior H. Cumulant functions c_F, c_H ensure these distributions are normalized. The chosen exponential family is defined by its sufficient statistics s_F. Our experiments consider Bernoulli, Gaussian, and auto-regressive Gaussian likelihoods.

Hierarchies of Dirichlet processes. Under the HDP-HMM prior and posterior, the number of states is unbounded; it is possible that every observation comes from a unique state. The *hierarchical Dirichlet process* (HDP) [5] encourages sharing states over time via a latent root probability vector β over the infinite set of states (see Fig. 2). The *stick-breaking representation* of the prior on β first draws independent variables $u_k \sim \text{Beta}(1, \gamma)$ for each state k, and then sets $\beta_k = u_k \prod_{\ell=1}^{k-1}(1 - u_\ell)$. We interpret u_k as the conditional probability of choosing state k among states $\{k, k+1, k+2, \ldots\}$.

In expectation, the K most common states are first in stick-breaking order. We represent their probabilities via the vector $[\beta_1 \ \beta_2 \ \ldots \ \beta_K \ \beta_{>K}]$, where $\beta_{>K} = \sum_{k=K+1}^{\infty} \beta_k$. Given this $(K+1)$-dimensional probability vector β, the HDP-HMM generates transition distributions π_k for each state k from a Dirichlet with mean equal to β and variance governed by concentration parameter $\alpha > 0$:

$$[\pi_{k1} \ \ldots \ \pi_{kK} \ \pi_{k>K}] \sim \text{Dir}(\alpha\beta_1, \alpha\beta_2, \ldots, \alpha\beta_{>K}). \tag{3}$$

We draw starting probability vector π_0 from a similar prior with much smaller variance, $\pi_0 \sim \text{Dir}(\alpha_0\beta)$ with $\alpha_0 \gg \alpha$, because few starting states are observed.

Sticky self-transition bias. In many applications, we expect each segment to persist for many timesteps. The "sticky" parameterization of [4, 6] favors self-transition by placing extra prior mass on the transition probability π_{kk}. In particular, $[\pi_{k1} \ \ldots \ \pi_{k>K}] \sim \text{Dir}(\alpha\beta_1, \ldots \alpha\beta_k + \kappa, \ldots \alpha\beta_{>K})$ where $\kappa > 0$ controls the degree of self-transition bias. Choosing $\kappa \approx 100$ leads to long segment lengths, while avoiding the computational cost of semi-Markov alternatives [7].