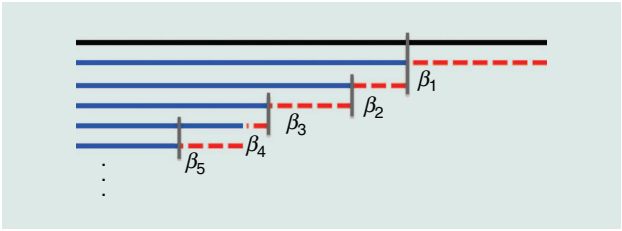

[FIG3] Pictorial representation of the stick-breaking construction of the DP.

STICKY HDP-HMM

In the Bayesian HMM of the previous section, we assumed that the number of HMM modes K is known. But what if this is not the case? For example, in the speaker diarization task considered later, determining the number of speakers involved in the meeting is one of the primary inference goals. Moreover, even when a model adequately describes previous observations, it can be desirable to allow new modes to be added as more data are observed. For example, what if more speakers enter the meeting? To avoid restrictions on the size of the mode space, such scenarios naturally lead to priors on probability measures G_j that have an unbounded collection of support points θ_k.

THE HMM IS THE MOST BASIC EXAMPLE OF A MARKOV SWITCHING PROCESS AND FORMS THE BUILDING BLOCK FOR MORE COMPLICATED PROCESSES EXAMINED LATER.

The DP, denoted by $\mathrm{DP}(\gamma, H)$, provides a distribution over countably infinite probability measures

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \quad \theta_k \sim H \tag{4}$$

on a parameter space Θ. The weights are sampled via a stick-breaking construction [11]

$$\beta_k = \nu_k \prod_{\ell=1}^{k-1} (1-\nu_\ell) \qquad \nu_k \sim \mathrm{Beta}(1, \gamma). \tag{5}$$

In effect, we have divided a unit-length stick into lengths given by the weights β_k: the kth weight is a random proportion v_k of the remaining stick after the first $(k-1)$ weights have been chosen. We denote this distribution by $\beta \sim \mathrm{GEM}(\gamma)$. See Figure 3 for a pictorial representation of this process.

The DP has proven useful in many applications due to its clustering properties, which are clearly seen by examining the predictive distribution of draws $\theta_i' \sim G_0$. Because probability measures drawn from a DP are discrete, there is a strictly positive probability of multiple observations θ_i' taking identical values within the set $\{\theta_k\}$, with θ_k defined as in (4). For each value θ_i', let z_i be an indicator random variable that picks out the unique value θ_k such that $\theta_i' = \theta_{z_i}$. Blackwell and MacQueen [12] derived a Pólya urn representation of the θ_i'

$$\theta_i' \mid \theta_1', \ldots, \theta_{i-1}' \sim \frac{\gamma}{\gamma + i - 1} H + \sum_{j=1}^{i-1} \frac{1}{\gamma + i - 1} \delta_{\theta_j'}$$
$$\sim \frac{\gamma}{\gamma + i - 1} H + \sum_{k=1}^{K} \frac{n_k}{\gamma + i - 1} \delta_{\theta_k}. \tag{6}$$

This implies the following predictive distribution on the indicator assignment variables

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \gamma) = \frac{\gamma}{N + \gamma} \delta(z, K+1) + \frac{1}{N + \gamma} \sum_{k=1}^{K} n_k \delta(z, k). \tag{7}$$

Here, $n_k = \sum_{i=1}^{N} \delta(z_i, k)$ is the number of indicator random variables taking the value k, and $K+1$ is a previously unseen value. The discrete Kronecker delta $\delta(z, k) = 1$ if $z = k$, and 0 otherwise. The distribution on partitions induced by the sequence of conditional distributions in (7) is commonly referred to as the Chinese restaurant process. Take i to be a customer entering a restaurant with infinitely many tables, each serving a unique dish θ_k. Each arriving customer chooses a table, indicated by z_i, in proportion to how many customers are currently sitting at that table. With some positive probability proportional to γ, the customer starts a new, previously unoccupied table $K+1$. From the Chinese restaurant process, we see that the DP has a reinforcement property that leads to a clustering at the values θ_k. This representation also provides a means of sampling observations from a DP without explicitly constructing the infinite probability measure $G_0 \sim \mathrm{DP}(\gamma, H)$.

One could imagine using the DP to define a prior on the set of HMM transition probability measures G_j. Taking each transition measure G_j as an independent draw from $\mathrm{DP}(\gamma, H)$ implies that these probability measures are of the form $\sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_{jk}}$, with weights $\pi_j \sim \mathrm{GEM}(\gamma)$ and atoms $\theta_{jk} \sim H$. Assuming H is absolutely continuous (e.g., a Gaussian distribution), this construction leads to transition measures with nonoverlapping support (i.e., $\theta_{jk} \neq \theta_{\ell k'}$ with probability one.) Based on such a construction, we would move from one infinite collection of HMM modes to an entirely new collection at each transition step, implying that previously visited modes would never be revisited. This is clearly not what we intended. Instead, consider the HDP [13], which defines a collection of probability measures $\{G_j\}$ on the same support points $\{\theta_1, \theta_2, \ldots\}$ by assuming that each discrete measure G_j is a variation on a global discrete measure G_0. Specifically, the Bayesian hierarchical specification takes $G_j \sim \mathrm{DP}(\alpha, G_0)$, with G_0 itself a draw from a $\mathrm{DP}(\gamma, H)$. Through this construction, one can show that the probability measures are described as

$$\begin{aligned} G_0 &= \textstyle\sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} & \beta \mid \gamma &\sim \mathrm{GEM}(\gamma) \\ G_j &= \textstyle\sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} & \pi_j \mid \alpha, \beta &\sim \mathrm{DP}(\alpha, \beta) \end{aligned} \qquad \theta_k \mid H \sim H. \tag{8}$$

Applying the HDP prior to the HMM, we obtain the HDP-HMM of Teh et al. [13].