We further develop a principled framework for escaping local optima in the online setting, by integrating birth and merge moves within our algorithm's coordinate ascent steps. Most existing mean-field algorithms impose a restrictive fixed truncation in the number of components, which is hard to set *a priori* on big datasets: either it is too small and inexpressive, or too large and computationally inefficient. Our birth and merge moves, together with a *nested* variational approximation to the posterior, enable adaptive creation and pruning of clusters on-the-fly. Because these moves are validated by an exactly tracked global variational objective, we avoid potential instabilities of stochastic online split-merge proposals [4]. The structure of our moves is very different from split-merge MCMC methods [5, 6]; applications of these algorithms have been limited to hundreds of data points, while our experiments show scaling of memoized split-merge proposals to millions of examples.

We review the Dirichlet process mixture model and variational inference in Sec. 2, outline our novel memoized algorithm in Sec. 3, and evaluate on clustering and denoising applications in Sec. 4.

2 Variational inference for Dirichlet process mixture models

The *Dirichlet process* (DP) provides a nonparametric prior for partitioning exchangeable datasets into discrete clusters [7]. An instantiation G of a DP is an infinite collection of atoms, each of which represents one mixture component. Component k has mixture weight w_k sampled as follows:

$$G \sim \text{DP}(\alpha_0 H), \quad G \triangleq \sum_{k=1}^{\infty} w_k \delta_{\phi_k}, \quad v_k \sim \text{Beta}(1, \alpha_0), \quad w_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell). \tag{1}$$

This *stick-breaking* process provides mixture weights and parameters. Each data item n chooses an assignment $z_n \sim \text{Cat}(w)$, and then draws observations $x_n \sim F(\phi_{z_n})$. The data-generating parameter ϕ_k is drawn from a base measure H with natural parameters λ_0. We assume both H and F belong to exponential families with log-normalizers a and sufficient statistics t:

$$p(\phi_k \mid \lambda_0) = \exp\left\{\lambda_0^T t_0(\phi_k) - a_0(\lambda_0)\right\}, \qquad p(x_n \mid \phi_k) = \exp\left\{\phi_k^T t(x_n) - a(\phi_k)\right\}. \tag{2}$$

For simplicity, we assume unit reference measures. The goal of inference is to recover stick-breaking proportions v_k and data-generating parameters ϕ_k for each global mixture component k, as well as discrete cluster assignments $z = \{z_n\}_{n=1}^N$ for each observation. The joint distribution is

$$p(\mathbf{x}, \mathbf{z}, \phi, v) = \prod_{n=1}^{N} F(x_n \mid \phi_{z_n}) \text{Cat}(z_n \mid w(v)) \prod_{k=1}^{\infty} \text{Beta}(v_k \mid 1, \alpha_0) H(\phi_k \mid \lambda_0) \tag{3}$$

While our algorithms are directly applicable to any DP mixture of exponential families, our experiments focus on D-dimensional real-valued data x_n, for which we take F to be Gaussian. For some data, we consider full-mean, full-covariance analysis (where H is normal-Wishart), while other applications consider zero-mean, full-covariance analysis (where H is Wishart).

2.1 Mean-field variational inference for DP mixture models

To approximate the full (but intractable) posterior over variables z, v, ϕ, we consider a fully-factorized variational distribution q, with individual factors from appropriate exponential families:[1]

$$q(\mathbf{z}, v, \phi) = \prod_{n=1}^{N} q(z_n|\hat{r}_n) \prod_{k=1}^{K} q(v_k|\hat{\alpha}_1, \hat{\alpha}_0) q(\phi_k|\hat{\lambda}_k), \tag{4}$$

$$q(z_n) = \text{Cat}(z_n \mid \hat{r}_{n1}, \ldots \hat{r}_{nK}), \quad q(v_k) = \text{Beta}(v_k \mid \hat{\alpha}_{k1}, \hat{\alpha}_{k0}), \quad q(\phi_k) = H(\phi_k \mid \hat{\lambda}_k). \tag{5}$$

To tractably handle the infinite set of components available under the DP prior, we truncate the discrete assignment factor to enforce $q(z_n = k) = 0$ for $k > K$. This forces all data to be explained by only the first K components, inducing *conditional independence* between observed data and any global parameters v_k, ϕ_k with index $k > K$. Inference may thus focus exclusively on a finite set of K components, while reasonably approximating the true infinite posterior for large K.

[1] To ease notation, we mark variables with hats to distinguish parameters $\hat{\theta}$ of variational factors q from parameters θ of the generative model p. In this way, θ_k and $\hat{\theta}_k$ always have equal dimension.