3 Inference with Gibbs Sampling

A segmentation of an observed image is found by posterior inference. The problem is to compute the conditional distribution of the latent variables—the customer assignments $c_{1:N}$—conditioned on the observed image features $x_{1:N}$, the scaling parameter α, the distances between pixels D, the window size a, and the base distribution hyperparameter λ:

$$p(c_{1:N} \mid x_{1:N}, \alpha, d, a, \lambda) = \frac{\left(\prod_{i=1}^{N} p(c_i \mid D, a, \alpha)\right) p(x_{1:N} \mid z(c_{1:N}), \lambda)}{\sum_{c_{1:N}} \left(\prod_{i=1}^{N} p(c_i \mid D, a, \alpha)\right) p(x_{1:N} \mid z(c_{1:N}), \lambda)} \tag{2}$$

where $z(c_{1:N})$ is the cluster representation that is derived from the customer representation $c_{1:N}$. Notice again that the prior term uses the customer representation to take into account distances between data points; the likelihood term uses the cluster representation.

The posterior in Equation (2) is not tractable to directly evaluate, due to the combinatorial sum in the denominator. We instead use Gibbs sampling [3], a simple form of Monte Carlo Markov chain (MCMC) inference [18]. We define the Markov chain by iteratively sampling each latent variable c_i conditioned on the others and the observations,

$$p(c_i \mid c_{-i}, x_{1:N}, D, \alpha, \lambda) \propto p(c_i \mid D, \alpha) p(x_{1:N} \mid z(c_{1:N}), \lambda). \tag{3}$$

The prior term is given in Equation (1). We can decompose the likelihood term as follows:

$$p(x_{1:N} \mid z(c_{1:N}), \lambda) = \prod_{k=1}^{K(c_{1:N})} p(x_{z(c_{1:N})=k} \mid z(c_{1:N}), \lambda). \tag{4}$$

We have introduced notation to more easily move from the customer representation—the primary latent variables of our model—and the cluster representation. Let $K(c_{1:N})$ denote the number of unique clusters in the customer assignments, $z(c_{1:N})$ the cluster assignments derived from the customer assignments, and $x_{z(c_{1:N})=k}$ the collection of observations assigned to cluster k. We assume that the cluster parameters ϕ_k have been analytically marginalized. This is possible when the base distribution G_0 is conjugate to the data generating distribution, e.g. Dirichlet to multinomial.

Sampling from Equation (3) happens in two stages. First, we remove the customer link c_i from the current configuration. Then, we consider the prior probability of each possible value of c_i and how it changes the likelihood term, by moving from $p(x_{1:N} \mid z(c_{-i}), \lambda)$ to $p(x_{1:N} \mid z(c_{1:N}), \lambda)$.

In the first stage, removing c_i either leaves the cluster structure intact, i.e., $z(c_{1:N}^{\text{old}}) = z(c_{-i})$, or splits the cluster assigned to data point i into two clusters. In the second stage, randomly reassigning c_i either leaves the cluster structure intact, i.e., $z(c_{-i}) = z(c_{1:N})$, or joins the cluster assigned to data point i to another. See Figure 1 for an illustration of these cases. Via these moves, the sampler explores the space of possible segmentations.

Let ℓ and m be the indices of the tables that are joined to index k. We first remove c_i, possibly splitting a cluster. Then we sample from

$$p(c_i \mid c_{-i}, x_{1:N}, D, \alpha, \lambda) \propto \begin{cases} p(c_i \mid D, \alpha)\Gamma(x, z, \lambda) & \text{if } c_i \text{ joins } \ell \text{ and } m; \\ p(c_i \mid D, \alpha) & \text{otherwise,} \end{cases} \tag{5}$$

where

$$\Gamma(x, z, \lambda) = \frac{p(x_{z(c_{1:N})=k} \mid \lambda)}{p(x_{z(c_{1:N})=\ell} \mid \lambda) p(x_{z(c_{1:N})=m} \mid \lambda)}. \tag{6}$$

This defines a Markov chain whose stationary distribution is the posterior of the spatial ddCRP defined in Section 2. Though our presentation is slightly different, this algorithm is equivalent to the one developed for ddCRP mixtures in [1].

In the rddCRP, the algorithm for sampling the customer indicators is nearly the same, but with two differences. First, when c_i is removed, it may spawn a new cluster. In that case, the region identity of the new table must be sampled from the region level CRP. Second, the likelihood term in Equation (4) depends only on the superpixels in the image assigned to the segment in question. In the rddCRP, it