Reflection - Dirichlet reflection - Frigyik et al.
Nicholas Hoernle **March 21, 2018**

1. The GEM distribution with concentration parameter α, is closely related to the Dirichlet stick-breaking process. For a K dimensional Dirichlet distribution, with parameters $a_{1:K}$, then a draw (ρ) from the Dirichlet distribution also can be written as a stick breaking process with $K-1$ breaks (and K pieces): $\rho_1 \sim Beta(\alpha_1, \sum_{k=1}^{K}(a_k - a_1)), \frac{\rho_2}{1-\rho_1} \sim Beta(\alpha_2, \sum_{k=2}^{K}(a_k - a_2)), \ldots, \rho_K = 1 - \sum_{k=1}^{K-1} \rho_k$. If we did not place a limit on K and rather let $K \to \infty$, and if all $a_k = \alpha \forall K$ then we have described the GEM distribution. So the GEM distribution is like an infinite dimensional Dirichlet distribution.

2. Two elements drawn from the GEM ρ_i and ρ_j for $i < j$ are not independent. Consider the stick breaking convention for the $GEM(\alpha)$ distribution where $\rho_k = \left[\prod_{j=1}^{k-1}(1 - V_j)\right] V_k = \left\{\left[\prod_{j=1}^{k-2}(1 - V_j)\right] - \rho_{k-1}\right\} V_k$ (where V_k is the proportion that is broken off at each step). Here we can see that ρ_k depends on the value of ρ_{k-1} (and previous values of ρ_i) and thus for two samples ρ_i and ρ_j for $i < j$, ρ_j is dependent on ρ_i ($P(\rho_j \mid \rho_i) \neq P(\rho_j)$).

3. As $N \to \infty$, the number of unique values $\{z_n\}$ also tends to ∞. This is exactly the paradigm that we desire as we'd like the number of assignments to grow appropriately with more observations of data (e.g. as the number of documents in a corpus tends to ∞, we might need new and unseen topics to help describe the data).