

Sampling from the Pitman-Yor Diffusion Tree

David A. Knowles

Recall that at a branch point the probability of following one of the existing branches k is

$$\frac{b_k - \beta}{m + \alpha} \quad (1)$$

where b_k is the number of samples which previously took branch k and m is the total number of samples through this branch point so far. The probability of diverging at the branch point and creating a new branch is

$$\frac{\alpha + \beta K}{m + \alpha} \quad (2)$$

where K is the number of branches from this branch point.

It is straightforward to sample sequentially from the prior. This is most easily done by sampling the tree structure and divergence times first, followed by the divergence locations. We will need the inverse cumulative divergence function, $A^{-1}(y) = 1.0 - \exp(-y/c)$ for the divergence function $a(t) = \frac{c}{1-t}$.

Each point starts at the root of the tree. The cumulative distribution function for the divergence time of the i -th sample is

$$C(t) = 1 - \exp \left\{ -A(t) \frac{\Gamma(i - 1 - \beta)}{\Gamma(i + \alpha)} \right\}$$

We can sample from this distribution by drawing $U \sim \text{Uniform}[0, 1]$ and setting

$$t_d = C^{-1}(U) := A^{-1} \left(-\frac{\Gamma(i + \alpha)}{\Gamma(i - 1 - \beta)} \log(1 - U) \right)$$

If t_d is actually past the next branch point, we diverge at this branch point or choose one of the previous paths with the probabilities defined in Equations 2 and 1 respectively. If we choose one of the existing branches then we must again sample a divergence time. On an edge from node a to b previously traversed by $m(b)$ data points, the cumulative distribution function for a new divergence time is

$$C(t) = 1 - \exp \left\{ -[A(t) - A(t_a)] \frac{\Gamma(m(b) - \beta)}{\Gamma(m(b) + 1 + \alpha)} \right\}$$

which we can sample as follows

$$t_d := A^{-1} \left(A(t_a) - \frac{\Gamma(m(b) + 1 + \alpha)}{\Gamma(m(b) - \beta)} \log(1 - U) \right)$$

We do not actually need to be able to evaluate $A(t_a)$ since this will necessarily have been calculated when sampling t_a . If $t_d > t_b$ we again choose whether to follow an existing branch or diverge according to Equations 2 and 1.

Given the tree structure and divergence times sampling the locations simply involves a sweep down the tree sampling $x_b \sim N(x_a, \sigma^2(t_b - t_a)I)$ for each branch $[ab]$.