Dirichlet Sampling for Persistent Diagrams

Zijian Chen, Moo K. Chung

{zijian.chen,mkchung}@wisc.edu

Abstract. In this short tutorial paper, we explain how to sample points from a persistent diagram that is triangle shaped domain widely used in topological data analysis (TDA). Since there is no ground truth or common database in TDA, simulation based method is often used toe establish the ground truth in TDA.

1 Sampling from a Dirichlet distribution

The persistent diagrams are scatter points in the unbounded domain

$$\mathcal{T}_\infty = \{(x_1, x_2) : x_2 \geq x_1\} \subset \mathbb{R}^2.$$

But this is not very convenient so we often constrain the domain as the bounded upper triangle

$$\mathcal{T} = \{(x_1, x_2) : x_2 \geq x_1, 0 \leq x_1, x_2 \leq 1\} \subset \mathbb{R}^2.$$

by scaling or thresholding data. We are interested in sampling points in $\mathcal{T}$ using the Dirichlet distribution somehow. The Dirichlet-like distribution defined on this domain is given by

$$f_\alpha^{\mathcal{T}}(x_1, x_2) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1}(1 - x_2)^{\alpha_2 - 1}(x_2 - x_1)^{\alpha_3 - 1}, \tag{1}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ are positive parameters [2]. To generate samples from $f_\alpha^{\mathcal{T}}$, we first need to sample from Dirichlet distribution defined on

$$\mathcal{T}_1 = \{(x_1, x_2, x_3) : x_1 + x_2 + x_3 = 1,\ x_1, x_2, x_3 \geq 0\} :$$

$$f_\alpha^{\mathcal{T}_1}(x_1, x_2, x_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} x_3^{\alpha_3 - 1}.$$

This can be achieved using the Gamma distribution:

Theorem 1. *[1,4] Let $Y_i \sim \mathrm{Gamma}(\alpha_i, 1), i = 1, 2, 3$ be independent Gamma random variables with $\alpha_i > 0$, and let*

$$X_i = \frac{Y_i}{\sum_{j=1}^{3} Y_j}, \quad i = 1, \cdots, 3.$$

Then (X_1, X_2, X_3) follows Dirichlet distribution on $\mathcal{T}_1$ with parameters $(\alpha_1, \alpha_2, \alpha_3)$.

Theorem 1 gives the Dirichlet distribution on $\mathcal{T}_1$. Then through the change of coordinates, we transform the distribution from $\mathcal{T}_1$ to $\mathcal{T}$.

Theorem 2. *Let* $Y_i \sim \mathrm{Gamma}(\alpha_i, 1), i = 1, 2, 3$ *be independent Gamma random variables with* $\alpha_i > 0$*, and let*

$$X_i = \frac{Y_i}{\sum_{j=1}^{3} Y_j}, \quad i = 1, 2, 3.$$

Then $(X_1, 1 - X_2)$ *follows the Dirichlet distribution on* $\mathcal{T}$ *with parameters* $(\alpha_1, \alpha_2, \alpha_3)$.

Proof. From Theorem 1, the random vector (X_1, X_2, X_3) follows the Dirichlet distribution on $\mathcal{T}_1$ with parameters $(\alpha_1, \alpha_2, \alpha_3)$. We then project this random vector onto $\mathcal{T}_2$ given by.

$$\mathcal{T}_2 = \{(x_1, x_2) : x_1, x_2 \geq 0, x_1 + x_2 \leq 1\}.$$

For any point $(x_1, x_2, x_3) \in \mathcal{T}_1$, x_3 is uniquely determined as $x_3 = 1 - x_1 - x_2$. Thus, the Dirichlet distribution on $\mathcal{T}_2$ is simply given by

$$f_\alpha^{\mathcal{T}_2}(x_1, x_2) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} (1 - x_1 - x_2)^{\alpha_3 - 1},$$

for $(x_1, x_2) \in \mathcal{T}_2$. We than transform $\mathcal{T}_2$ to $\mathcal{T}$ by rotation $(x_1, x_2) \longmapsto (x_1, 1 - x_2)$. Since the Jacobian of the rotation is 1, the density function $\mathcal{T}$ is given by

$$f_\alpha^{\mathcal{T}}(x_1, x_2) = f_\alpha^{\mathcal{T}_2}(x_1, 1 - x_2) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} x_1^{\alpha_1 - 1} (1 - x_2)^{\alpha_2 - 1} (x_2 - x_1)^{\alpha_3 - 1},$$

□

The codes are packaged into the function `Dirichlet_sample.m`, which inputs `alpha`, a 3×1 column vector and `n`, the number of samples. The output is a $n \times 2$ matrix. Each row of the output corresponds to one sample. To generate 500 scatter points from $f_\alpha^{\mathcal{T}}$ with $\alpha = (3, 2, 3)$, we run

```
alpha = [3,2,3]';
samples = Dirichlet_sample(alpha,1,500);
```

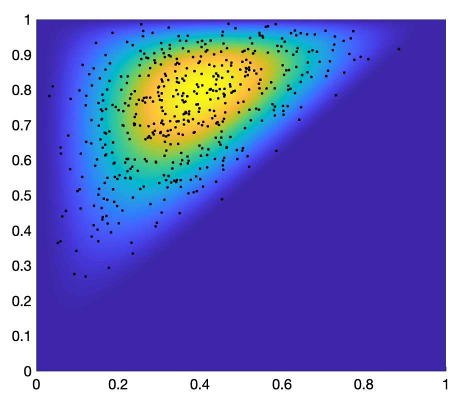

Fig. 1. 500 samples from $f_\alpha^{\mathcal{T}}$ with $\alpha = (3, 2, 3)$. The sampled points are overlaid on the contour plot of the theoretical densities.

2 Sampling from a mixture of Dirichlet distributions

Sampling from a single Dirichlet distribution provides a concentric pattern with unimodal peak somewhere inside the triangle. This may not be realistic for various data. We propose sample from a mixture of Dirichlet distributions. Suppose $f^{\mathcal{T}}_{\alpha_i}$ are Dirichlet distributions with parameters α_i. A more realistic model is to sample from a mixture $\sum_{i=1}^{k} w_i f^{\mathcal{T}}_{\alpha_i}$ with $\sum_{i=1}^{k} w_i = 1$. This is a proper distribution in $\mathcal{T}$. This will provide multiple concentrated regions. To implement the sampling procedure, we first need to generate a uniform sample $U \sim \mathrm{Unif}(0,1)$. If $U \in (\sum_{i=1}^{l} w_i, \sum_{i=1}^{l+1} w_i)$, then we generate a sample from the distribution of the $l-$th component (i.e., $f^{\mathcal{T}}_{\alpha_l}$). This process is repeated until the desired number of samples from the mixture distribution is obtained [3]. To generate 500 scatter points from the mixture

$$f^{\mathcal{T}} = 0.25 f^{\mathcal{T}}_{\alpha_1} + 0.25 f^{\mathcal{T}}_{\alpha_2} + 0.25 f^{\mathcal{T}}_{\alpha_3} + 0.25 f^{\mathcal{T}}_{\alpha_4},$$

with parameters $\alpha_1 = (3,8,2)$, $\alpha_2 = (8,3,2)$, $\alpha_3 = (7,3,8)$, $\alpha_4 = (3,7,8)$. The sampling is done through by extending the functionality of the previous MATLAB function `Dirichlet_sample.m`

```
alpha = [3,8,2;8,3,2;7,3,8;3,7,8]';
weight = [0.25,0.25,0.25,0.25];
samples = Dirichlet_sample(alpha,weight,500);
```

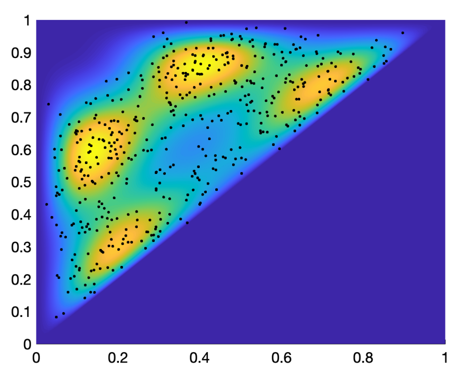

Fig. 2. 500 samples from $f^{\mathcal{T}}$ given above. The sampled points are overlaid on the contour plot of the theoretical densities.

References

1. Devroye, L.: Non-Uniform Random Variate Generation. Springer New York (2013)
2. Kotz, S., Balakrishnan, N., Johnson, N.: Continuous Multivariate Distributions, Volume 1: Models and Applications. Continuous Multivariate Distributions, Wiley (2004)
3. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics, Wiley (2004)
4. Ng, K., Tian, G., Tang, M.: Dirichlet and Related Distributions: Theory, Methods and Applications. Wiley-Blackwell, United States (2011)