Cruising The Simplex: Hamiltonian Monte Carlo and the Dirichlet Distribution

Michael Betancourt
Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

Due to its constrained support, the Dirichlet distribution is uniquely suited to many applications. The constraints that make it powerful, however, can also hinder practical implementations, particularly those utilizing Markov Chain Monte Carlo (MCMC) techniques such as Hamiltonian Monte Carlo. I demonstrate a series of transformations that reshape the canonical Dirichlet distribution into a form much more amenable to MCMC algorithms.

The Dirichlet Distribution

Given $m$ random variables, $x$, with the constrained support

$$0 \leq x_i \leq 1 \quad \sum_{i=1}^{m} x_i = 1,$$

the Dirichlet distribution \cite{1} \cite{2} is defined by the parameterized probability density

$$\text{Dir} \left( x | \alpha \right) = \frac{\Gamma \left( \sum_{i=1}^{m} \alpha_i \right)}{\prod_{i=1}^{m} \Gamma \left( \alpha_i \right)} \cdot \prod_{i=1}^{m} x_i^{\alpha_i - 1},$$

where the $\alpha$ are positive-valued reals and $\Gamma$ is the usual gamma function.

Because of its distinctive support, the Dirichlet distribution is particularly well-suited for modeling the allocation of conserved quantities such as probability. The distribution becomes invaluable when studying categorical problems: inference with histograms a pervasive example.

\footnote{betan@mit.edu}
Generating Dirichlet Samples

Sampling from the Dirichlet distribution is made feasible due to a convenient property of the Gamma distribution \[3\]. An ensemble of Gamma variates,

\[ u_i \sim \text{Ga} \left( u_i | \alpha_i, \beta_i \right) \]

follows a Dirichlet distribution upon normalization,

\[ \left\{ \frac{u_i}{\sum_j u_j} \right\} = \{ x_i \} \sim \text{Dir} \left( x | \alpha \right) \].

Given the efficiency of modern Gamma generators \[4\], the generation of independent Dirichlet variates is particularly undemanding.

The same property admits a parallel Markov Chain Monte Carlo approach. Here separate chains generate the independent Gamma variates, which are then normalized to produce the desired Dirichlet sample. Because the chains interact only in the normalization, however, their evolutions are uncorrelated and consequently useless to algorithms that rely on the local correlations of the distribution such as nested sampling \[5\]. Moreover, the need to interrupt the evolution to normalize prohibits the generalization to chains sampling from a posterior distribution predicated on a Dirichlet prior.

Creating a Markov chain that samples directly from the Dirichlet distribution is complicated by the constraints, as the Dirichlet variates manifest not in \( m \) dimensions but rather in a \( m - 1 \) dimensional submanifold known as a simplex. Unless the Markov transitions accommodate the constraints directly, proposed samples will fall outside of the simplex (with probability 1) and the chain will be unable to progress beyond an initial seed.

By parameterizing the simplex directly, the constraints are implicitly taken into account and the chain will have no problem exploring the full support of the Dirichlet distribution. Constructing a systematic map between the original \( m \) dimensional manifold and the simplex is straightforward, if a bit ungainly, but real difficulties begin to arise when considering the boundary of the simplex.

Because the simplex is the inclusive volume of a \( m - 1 \) dimensional polytope, the support is bounded by a surface of piecewise faces, the number of which grows exponentially with the dimension of the original distribution. Modeling this complex boundary in terms of simplex coordinates is awkward at best, especially when appealing to more sophisticated algorithms like constrained Hamiltonian Monte Carlo \([6, 7]\) that require a careful description of the boundary geometry.
Smoothing Out The Simplex

A simple change of variables dramatically simplifies the structure of the simplex. Taking

\[ y_i^2 = x_i, \]

the original Dirichlet distribution becomes

\[
\text{Dir}(\mathbf{y}|\mathbf{\alpha}) = 2^m \frac{\Gamma \left( \sum_{i=1}^{m} \alpha_i \right)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \cdot \prod_{i=1}^{m} y_i^{2\alpha_i - 1}
\]

with the support

\[
0 \leq y_i \leq 1,
\]

\[
\sum_{i=1}^{m} y_i^2 = 1.
\]

The quadratic constraint defines a substantially simpler submanifold: the surface of an \( m \) dimensional hypersphere within the positive orthant.

Transforming to hyperspherical coordinates [8],

\[
y_i = r \left( \prod_{k=1}^{i-1} \sin \theta_k \right) \cdot \left\{ \begin{array}{l}
\cos \theta_i, \quad i < m \\
1, \quad i = m
\end{array} \right.,
\]

the constrained support becomes

\[
0 \leq \theta_i \leq \frac{\pi}{2},
\]

\[
r^2 = 1
\]

with the distribution

\[
\text{Dir}(r, \mathbf{\theta}|\mathbf{\alpha}) = 2^m \frac{\Gamma \left( \sum_{i=1}^{m} \alpha_i \right)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \cdot r^{2\tilde{\alpha}_0 - 1} \prod_{i=1}^{m-1} (\cos \theta_i)^{2\alpha_i - 1} (\sin \theta_i)^{2\tilde{\alpha}_i - 1},
\]

where

\[
\tilde{\alpha}_i = \sum_{k=i+1}^{m} \alpha_k.
\]

3
Marginalization over the radial coordinate is immediate due to the constraint, giving a distribution for the hyperspherical angles alone

$$\text{Dir} (\theta | \alpha) = 2^{m-1} \frac{\Gamma \left( \sum_{i=1}^{m} \alpha_i \right)}{\prod_{i=1}^{m} \Gamma (\alpha_i)} \cdot \prod_{i=1}^{m-1} (\cos \theta_i)^{2\alpha_i-1} (\sin \theta_i)^{2\tilde{\alpha}_i-1},$$

with the remaining constraints

$$0 \leq \theta_i \leq \frac{\pi}{2}.$$

By distorting the original Dirichlet distribution, the simplex has become the constrained surface of a hypersphere readily parameterized with hyperspherical coordinates. The complexity of the hyperspherical boundary scales linearly with dimension and the entire submanifold becomes particularly accommodating to constrained Hamiltonian Monte Carlo.

**Differentiating on the Hypersphere**

The ability to run a Markov chain that samples directly from the Dirichlet distribution accommodates algorithms such as nested sampling, or extending the chain to sample from a posterior based on a Dirichlet prior. These applications, however, require the gradient of a function, in these cases the log likelihood, with respect to the hyperspherical coordinates.

Appealing to the chain rule,

$$\frac{\partial f (x)}{\partial \theta_i} = \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{\partial f (x)}{\partial x_j} \frac{\partial x_j}{\partial y_k} \frac{\partial y_k}{\partial \theta_i}$$

$$= \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{\partial f (x)}{\partial x_j} 2y_k \delta_{jk} \frac{\partial y_k}{\partial \theta_i}$$

$$= 2 \sum_{j=1}^{m} \frac{\partial f (x)}{\partial x_j} y_j \frac{\partial y_j}{\partial \theta_i}$$

Substituting the hyperspherical derivative,

$$\frac{\partial y_j}{\partial \theta_i} = \begin{cases} 0, & i > j \\ -y_j \cdot \tan \theta_i, & i = j \\ y_j / \tan \theta_i, & i < j \end{cases}$$
gives
\[
\frac{\partial f (x)}{\partial \theta_i} = 2 \left[ -\frac{\partial f (x)}{\partial x_i} y_i^2 \tan \theta_i + \sum_{j=i+1}^{m} \frac{\partial f (x)}{\partial x_j} y_j^2 \frac{1}{\tan \theta_i} \right],
\]
or, in terms of the original variables,
\[
\frac{\partial f (x)}{\partial \theta_i} = 2 \left[ -\frac{\partial f (x)}{\partial x_i} x_i \tan \theta_i + \sum_{j=i+1}^{m} \frac{\partial f (x)}{\partial x_j} x_j \frac{1}{\tan \theta_i} \right].
\]

Once the original gradient, \( \partial f/\partial x_i \), has been computed, the hyperspherical derivative requires only a few additional tangent evaluations. This additional burden can be eliminated entirely by storing the sine and cosine evaluations when projecting the hyperspherical coordinates \( \theta \) to the original coordinates \( x \) beforehand.

## Acknowledgements

I thank Kat Deck, John Rutherford, Leo Stein, and Phil Zukin for helpful discussion and comments.

## References

[1] MacKay, D. J. C. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, New York

[2] Bishop, C.M. (2007) *Pattern Classification and Machine Learning*. Springer, New York

[3] Cheng, R. C. H. (1998) Random Variate Generation. In *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice* (ed. Banks). Wiley & Sons, New York.

[4] Marsaglia, G. and Tsang, W. W. *ACM TOMS* 26:3 365-372 (2000)

[5] Skilling, J. (2004) Nested Sampling. In *Maximum Entropy and Bayesian methods in science and engineering* (ed. G. Erickson, J. T. Rychert, C. R. Smith). AIP Conf. Proc., 735: 395-405.

[6] Neal, R. M. *MCMC using Hamiltonian dynamics*, [http://www.cs.utoronto.ca/~radford/ham-mcmc.abstract.html](http://www.cs.utoronto.ca/~radford/ham-mcmc.abstract.html), March 5, 2010.

[7] Betancourt, M. J. *arXiv:1005.0157*

[8] Hassani, S. (2002) *Mathematical Physics: A Modern Introduction to Its Foundations*. Springer, New York