Efficient Unconstraining Parameter Transforms for Hamiltonian Monte Carlo

Meenal Jhajharia
Flatiron Institute

Seth Axen
University of Tübingen

Adam Haber
Weizmann Institute

Sean Pinkney
Omnicom Media Group

Bob Carpenter
Flatiron Institute

DRAFT: July 27, 2022

Abstract

This paper evaluates the statistical and computational efficiency of unconstraining parameter transforms for Hamiltonian Monte Carlo sampling.

1 Introduction

In statistical computing, we often need to compute high-dimensional integrals over densities $\pi(x)$ (e.g., Bayesian estimation or prediction, p-value calculations, etc.). The only black-box techniques that work for general high-dimensional integrals are Markov chain Monte Carlo (MCMC) methods. The most effective MCMC method in high dimensions is Hamiltonian Monte Carlo (HMC). HMC works by simulating the Hamiltonian dynamics of a fictitious particle representing the value being sampled coupled with a momentum term.

Although it is possible to write HMC samplers that work for simply constrained values such as upper- and/or lower-bounds [4] or unit vectors [2], it is much more challenging to do the same for complex constraints such as simplexes or positive definite matrices or for densities involving multiple constrained values. Instead, it is far more common to map the constrained values to unconstrained values before sampling [3]. For example, the Probabilistic Programming language Stan, defines distributions with an unconstrained support, in favour of making sampling an easier process. The variables with constraints (or with constrained support in this case) are transformed to an unconstrained space. Inverse transforms of these variables are used for log density adjustments. These are usually in the form of a Jacobian matrix comprised of the gradient $\frac{\partial x}{\partial y}$, where x and y are the constrained and unconstrained parameters respectively.

These gradients are computationally expensive, certain transforms are entirely inefficient, and some work well on certain distributions or parametrizations and vice versa. This presents the issue of selecting which transform to use among an infinite set of options, which is the topic of this paper. We begin by defining the general theory of transforms, followed by transforms from a unit simplex Δ^N to the R^M. Finally, we discuss notions of "better" transforms defined on the simplex, using statistical measures like Effective Sample Size, Leapfrog steps taken by the sampler etc.

2 Unconstraining Transforms

To draw samples from a parameter defined on a constrained space $\mathcal{X}$ that can be uniquely parameterized by n real degrees of freedom, we instead perform sampling in an unconstrained space $\mathcal{Y} \in \mathbb{R}^m$ for $m \geq n$. Let $p_Y(y)$ be an improper density function defined over $\mathcal{Y} \to (0, \infty)$ with support over all elements of $\mathcal{Y}$ and a surjective, smooth and continuous map $g : \mathcal{Y} \to \mathcal{X}$. Then, for any proper density π_X over $\mathcal{X}$, the constraining transform can be defined as (p_Y, g): $\mathcal{Y} \to \mathcal{X}$. In this case, the density $p_Y(y) = \pi_Y(y) \cdot \pi_X(g(y))$ is proper, and $Y \sim \pi_Y \implies g(Y) \sim \pi_X$. The required change of variables for this unconstraining transform is:

$$p_Y(y) = p_X(f^{-1}(y)) \left| J_{f^{-1}}(y) \right|,$$

where the Jacobian of the inverse transform is defined by

$$J_{f^{-1}}(y) = \frac{\partial}{\partial y} f^{-1}(y)$$

In this paper we consider transforms where $m = n$, for which the constraining transform is defined as (π_Y, f): $\mathcal{Y} \to \mathcal{X}$, given that $\pi_Y(y) = \left| J_{f^{-1}}(y) \right|$.

3 Unit simplex

A unit N-simplex is an $N + 1$-dimensional vector of non-negative values that sums to one. Simplexes are useful for representations of multinomial probabilities (e.g., probabilities of categories in a classification problem). The set of unit N-simplexes is conventionally denoted

$$\Delta^N = \left\{ x \in \mathbb{R}^{N+1}_{\geq 0} \;\middle|\; \sum_{i=1}^{N} x_i = 1 \right\}$$

Geometrically, an N-simplex is the convex closure of $N + 1$ points that are 1 in one coordinate and 0 elsewhere. For example, the 3-simplex is the complex closure of $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, and $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$. As such, there are only N degrees of freedom, because if x is an N-simplex, then

$$x_N = 1 - (x_1 + x_2 + \cdots + x_{N-1}).$$

We will use Δ^N_- to denote $N - 1$ elements of a simplex, this is sufficient to uniquely determine Δ^N.

3.1 Stick-Breaking Transform

The Stick-Breaking transform carries forward the intuition in the stick-breaking construction for Dirichlet [5]. It is a process of recursively breaking a piece x_i from a stick of unit length, where the leftover stick in the i^{th} iteration is $1 - \sum_1^i x$. Let $y = f(x)$, then we define the stick-breaking mapping f: $\Delta^{N-1} \to R^{N-1}$, for $1 \leq i \leq N$ as:

$$y_i = \text{logit}(z_i) - \log\left(\frac{1}{N-i}\right)$$

for break proportion

$$z_i = \frac{x_i}{1 - \sum_{i'=1}^{i-1} x_{i'}}.$$

The inverse transform $f^{-1}\colon R^{N-1} \to \Delta^{N-1}$ is defined as:

$$x_i = \left(1 - \sum_{i'=1}^{i-1} x_{i'}\right)$$

for break proportion

$$z_i = \text{logit}^{-1}\left(y_i + \log\left(\frac{1}{N-i}\right)\right)$$

The determinant of the Jacobian

$$|\mathbf{J}| = \prod_{i=1}^{N-1} z_i\,(1-z_i)\left(1 - \sum_{i'=1}^{i-1} x_{i'}\right)$$

3.2 Additive log ratio transform

The unconstraining transform for the identified softmax is known as the additive log ratio (ALR) transform [1], which is a bijection $\text{alr} : \Delta^{N-1} \to \mathbb{R}^{N-1}$ defined for $x \in \Delta^{N-1}$ by

$$\text{alr}(x) = \left[\log\frac{x_1}{x_N} \cdots \log\frac{x_{N-1}}{x_N}\right]$$

The inverse additive log ratio transform maps values in $\mathbb{R}^{N-1}$ to Δ^{N-1} defined for $y \in \mathbb{R}^{N-1}$ by

$$\text{alr}^{-1}(y) = \text{softmax}(\begin{bmatrix} y & 0 \end{bmatrix}),$$

where for $u \in \mathbb{R}^N$,

$$\text{softmax}(u) = \frac{exp(u)}{\sum exp(u)}$$

$$\text{Here, } |J| = \prod \exp(y)\ \left(\frac{1}{1+\sum(\exp(y))}\right)^N$$

3.3 Augmented-Softmax Transform

We define the transformation $\phi : \mathbb{R}^n \to \Delta^{n-1} \times \mathbb{R}_{>0} : y \mapsto (x_-, r)$, where Δ^N_- and x_- denote $N-1$ elements of the simplex and x, respectively. Here $r = \sum_{i=1}^n exp(y_i)$, $x_i = \frac{1}{r} exp(y_i)$ for $i \in [1, N-1]$ and δ_{ij} is the Kronecker delta function. If $\text{diag}(x)_{i,j} = \delta_{ij} x_i$ and $\mathbf{1}_n$ is the n-vector of ones.

$$J = (I_{n-1} - x_- \mathbf{1}_{n-1}^\top)\,\text{diag}(x_-),$$

Using Sylvester's determinant theorem, $|I_{n-1} - x_- \mathbf{1}_{n-1}^\top| = 1 - \mathbf{1}_{n-1}^\top x_- = 1 - \sum_{i=1}^{n-1} x_i = x_n$, so

$$|J| = x_n \prod_{i=1}^{n-1} x_i = \prod_{i=1}^{n} x_i = \exp\left(\sum_{i=1}^{n-1} y_i\right)\left(1 + \sum_{i=1}^{n-1} e^{y_i}\right)^{-n}$$

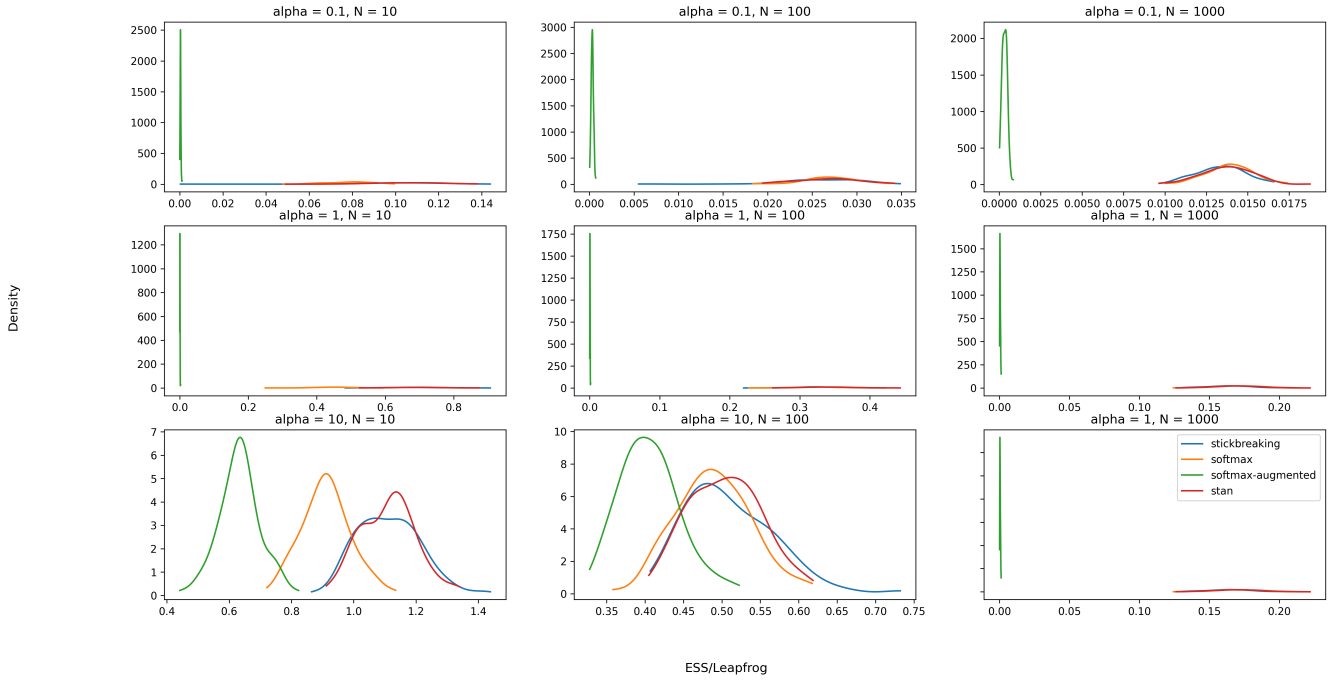

Figure 1: Effective Sample Size/Total LeapFrog Steps

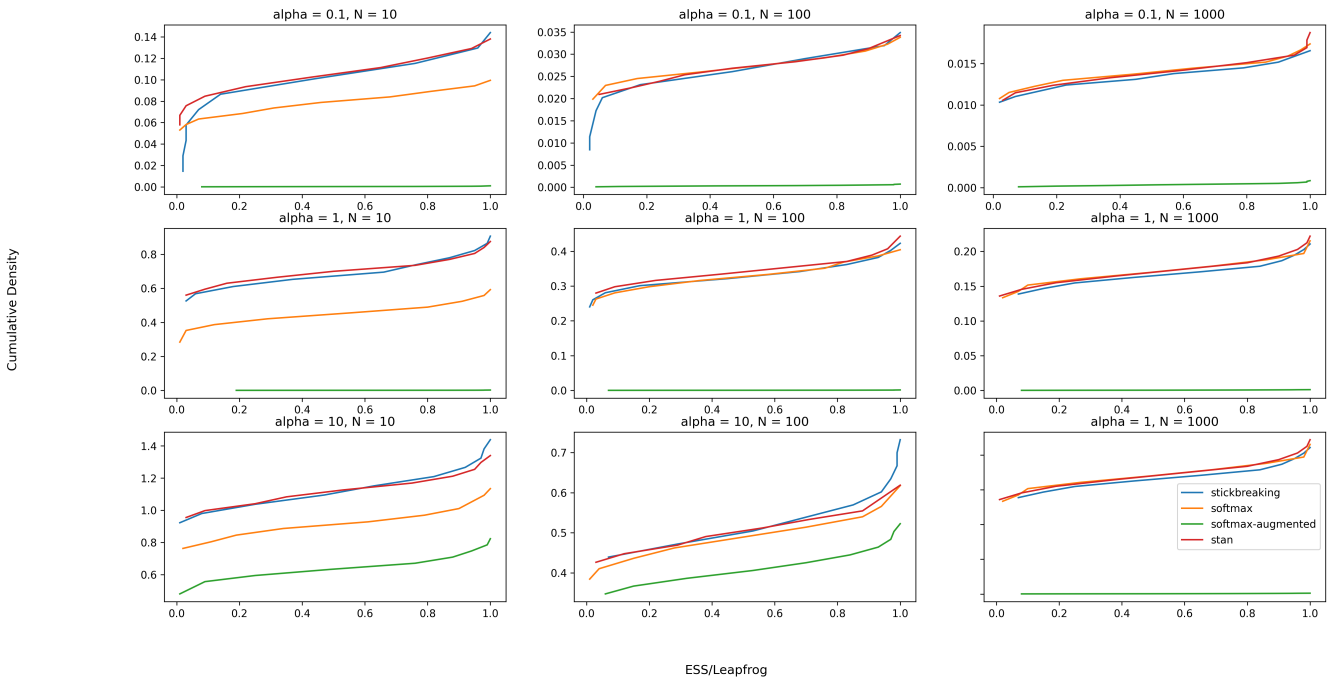

Figure 2: Effective Sample Size/Total LeapFrog Steps

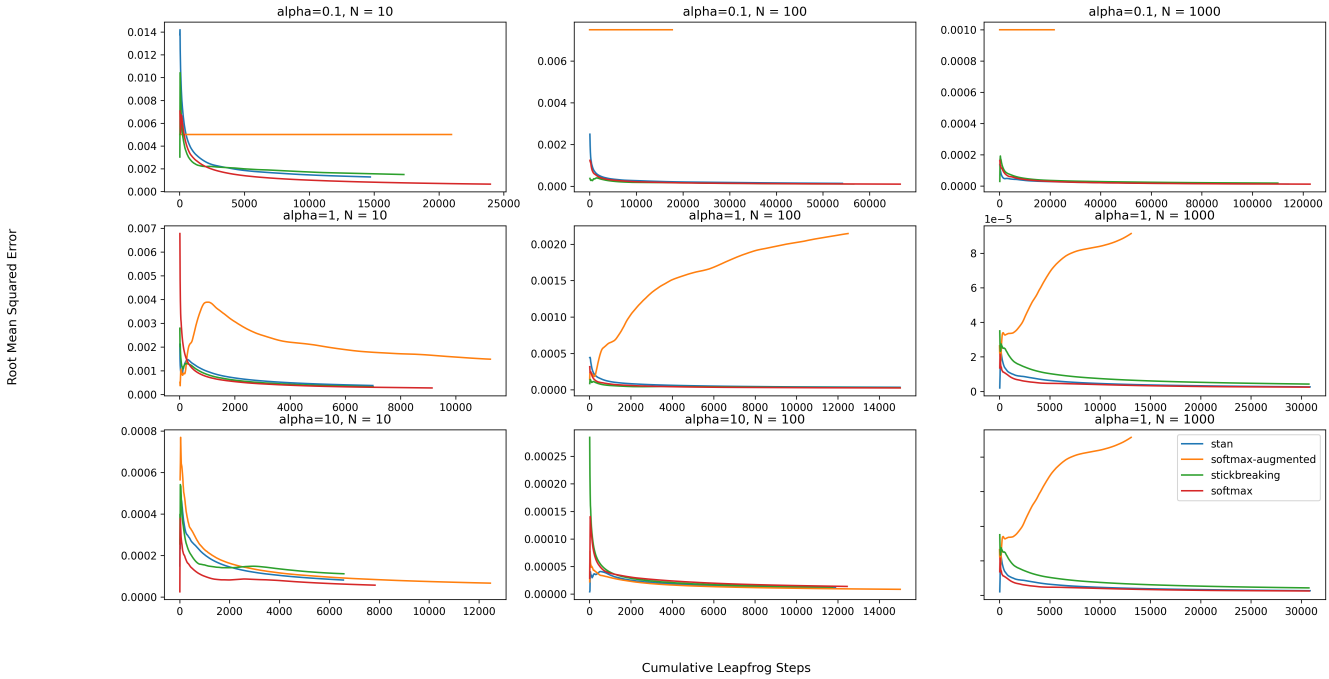

Figure 3: Root Mean Squared Error vs Cumulative Leapfrog Steps

4 Results

Acknowledgements

We would like to thank `matrixcalculus.org` for providing an easy-to-use symbolic matrix derivative calculator.

References

[1] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[2] Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40(4):825–845, 2013.

[3] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1âĂŞ32, 2017.

[4] Radford Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, , and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman and Hall/CRC, 2011.

[5] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

A Jacobian Determinants

A.1 Stickbreaking Transform

The Jacobian matrix for f^{-1} is a lower-triangular diagonal matrix, so for the change of variables we evaluate $\mathbf{J}_{i,i}$ where $i \in 1 : N-1$.

$$\begin{aligned}
\mathbf{J}_{i,i} &= \frac{\partial x_i}{\partial y_i} = \frac{\partial x_i}{\partial z_i}\frac{\partial z_i}{\partial y_i} \\
\mathbf{J}_{i,i} &= \left(1 - \sum_{k'=1}^{k-1} x_{k'}\right) z_k(1 - z_k), \\
|\mathbf{J}| &= \prod_{i=1}^{N-1} \mathbf{J}_{i,i}
\end{aligned}$$

The change of variables adjustment $p_Y(y) = p_X(f^{-1}(y)) \prod_{i=1}^{N-1} z_i \left(1 - z_i\right) \left(1 - \sum_{i'=1}^{i-1} x_{i'}\right)$.

A.2 Additive Log ratio transform

To calculate the determinant of the Jacobian of the inverse transform, we start by noting that $s = \exp \circ \text{norm}$, where exp is the elementwise exponential function and norm is defined by

$$\text{norm}(z) = \frac{z}{\text{sum}(z) + 1}.$$

As such, the resulting Jacobian determinant is the product of the Jacobian determinants of the component functions,

$$|J_s(y)| = \left|J_{\exp}(y)\right| \left|J_{\text{norm}}(z)\right|,$$

where $z = \exp(y)$. The Jacobian for the exponential function is diagonal, so the determinant is the product of the diagonal of the Jacobian, which for $y \in \mathbb{R}^{N-1}$ is

$$\left|J_{\exp}(y)\right| = \text{prod}(\exp(y)).$$

As above, let $z = \exp(y) \in (0, \text{inf})^{N-1}$. We can differentiate norm to derive the Jacobian,

$$J_{\text{norm}} = \frac{1}{1 + \text{sum}(z)} \mathbb{I}_{N-1} - \left(\frac{1}{(1 + \text{sum}(z))^2} \beta\right) \text{vector}_{N-1}(1)^\top,$$

where $\mathbb{I}_{N-1}$ is the $(N-1) \times (N-1)$ unit matrix and $\text{vector}_{N-1}(1)$ is the $N-1$-vector with values 1. Using the matrix determinant lemma,[1] we have

$$\begin{aligned}
\text{absdet}(J_{\text{norm}}(z)) &= \left(1 + \text{vector}_{N-1}(1)^\top \left(\frac{1}{1+\text{sum}(z)}\mathbb{I}\right)^{-1} \frac{-z}{(1+\text{sum}(z))^2}\right) \det\left(\frac{1}{1+\text{sum}(z)}\mathbb{I}\right) \\
&= \left(1 + \text{sum}\left(\frac{-(1+\text{sum}(z))z}{(1+\text{sum}(z))^2}\right)\right) \left(\frac{1}{1+\text{sum}(z)}\right)^{N-1} \\
&= \left(1 + \text{sum}\left(\frac{-z}{1+\text{sum}(z)}\right)\right) \left(\frac{1}{1+\text{sum}(z)}\right)^{N-1} \\
&= (1 - \text{sum}(\text{norm}(z))) \left(\frac{1}{1+\text{sum}(z)}\right)^{N-1} \\
&= \left(\frac{1}{1+\text{sum}(z)}\right)^{N}.
\end{aligned}$$

Thus the entire absolute determinant of the Jacobian is defined by the product,

$$|J_s(y)| \;=\; \text{prod}(\exp(y)) \left(\frac{1}{1+\text{sum}(\exp(y))}\right)^{N}.$$

and our final expression for densities for unconstrained $y \in \mathbb{R}^{N-1}$ is

$$p_Y(y) = p_X(\text{alr}^{-1}(y)) \,\text{prod}(\exp(y)) \left(\frac{1}{1+\text{sum}(\exp(y))}\right)^{N}$$

[1] The matrix determinant lemma is

$$\det(A + uv^\top) = (1 + v^\top A^{-1} u)\det(A).$$