Reparameterizing the Birkhoff Polytope for Variational Permutation Inference

Anonymous Authors
Anonymous Institutions

Abstract

Many matching, tracking, sorting, and ranking problems require probabilistic reasoning about possible permutations, a set that grows factorially with dimension. Combinatorial optimization algorithms may enable efficient point estimation, but fully Bayesian inference poses a severe challenge in this high-dimensional, discrete space. To surmount this, we start with the usual step of relaxing a discrete set (here, of permutation matrices) to its convex hull, which here is the Birkhoff polytope: the set of all doubly-stochastic matrices. We then introduce two novel transformations: first, an invertible and differentiable map from unconstrained space to the Birkhoff polytope, and second, a similar map to a ball around the polytope. Both transformations include a temperature parameter that, in the limit, concentrates the densities on permutation matrices. We then exploit these transformations and reparameterization gradients to introduce variational inference over permutation matrices, and we show via a series of simulated and real experiments the value of this approach.

1 Introduction

Permutation inference is central to many modern machine learning problems. Identity management [Guibas, 2008] and multiple-object tracking [Shin et al., 2005, Kondor et al., 2007] are fundamentally concerned with finding a permutation that maps an observed set of items to a set of canonical labels. Ranking problems, critical to search and recommender systems, require inference over the space of item orderings [Meilă et al., 2007, Lebanon and Mao, 2008, Adams and Zemel, 2011].

Preliminary work. Under review by AISTATS 2018. Do not distribute.

Furthermore, many probabilistic models, like preferential attachment network models [Bloem-Reddy and Orbanz, 2016] and repulsive point process models [Rao et al., 2016], incorporate a latent permutation into their generative processes; inference over model parameters requires integrating over the set of permutations that could have given rise to the observed data. In neuroscience, experimentalists now measure whole-brain recordings in *C. Elegans* [Kato et al., 2015, Nguyen et al., 2016], a model organism with a known synaptic network [White et al., 1986]; a current challenge is matching the observed neurons to corresponding nodes in the reference network. In Section 5, we address this problem from a Bayesian perspective in which permutation inference is a central component of a larger inference problem involving unknown model parameters and hierarchical structure.

The task of computing optimal point estimates of permutations under various loss functions has been well studied in the combinatorial optimization literature [Kuhn, 1955, Munkres, 1957, Lawler, 1963]. However, many probabilistic tasks, like the aforementioned neural identity inference problem, require reasoning about the posterior distribution over permutation matrices. A variety of Bayesian permutation inference algorithms have been proposed, leveraging sampling methods [Diaconis, 1988, Miller et al., 2013, Harrison and Miller, 2013], Fourier representations [Kondor et al., 2007, Huang et al., 2009], as well as convex [Lim and Wright, 2014] and continuous [Plis et al., 2011] relaxations for approximating the posterior distribution. Here, we address this problem from an alternative direction, leveraging stochastic variational inference [Hoffman et al., 2013] and reparameterization gradients [Rezende et al., 2014, Kingma and Welling, 2014] to derive a scalable and efficient permutation inference algorithm.

Section 2 lays the necessary groundwork, introducing definitions, prior work on permutation inference, variational inference, and continuous relaxations. Section 3 presents our primary contribution: a pair of transformations that enable variational inference over doubly-stochastic matrices, and, in the zero-temperature limit,

permutations, via stochastic variational inference. In the process, we show how these transformations connect to recent work on discrete variational inference [Maddison et al., 2016, Jang et al., 2016, Balog et al., 2017]. Sections 4 and 5 present a variety of experiments that illustrate the benefits of the proposed variational approach. Further details are in the supplement.

2 Background

2.1 Definitions and notation.

A permutation is a bijective mapping of a set onto itself. When this set is finite, the mapping is conveniently represented as a binary matrix $X \in \{0,1\}^{N\times N}$ where $X_{m,n} = 1$ implies that element m is mapped to element n. Since permutations are bijections, both the rows and columns of X must sum to one. From a geometric perspective, the Birkhoff-von Neumann theorem states that the convex hull of the set of permutation matrices is the set of doubly-stochastic matrices; i.e. non-negative square matrices whose rows and columns sum to one. The set of doubly-stochastic matrices is known as the *Birkhoff polytope*, and it is defined by,

$$\begin{aligned}\mathcal{B}_N = \Big\{ X : \quad & X_{m,n} \geq 0 \quad \forall\, m,n \in 1,\ldots,N; \\ & \sum_{n=1}^{N} X_{m,n} = 1 \quad \forall\, m \in 1,\ldots,N; \\ & \sum_{m=1}^{N} X_{m,n} = 1 \quad \forall\, n \in 1,\ldots,N \Big\}.\end{aligned}$$

These linear row- and column-normalization constraints restrict $\mathcal{B}_N$ to a $(N-1)^2$ dimensional subset of $\mathbb{R}^{N\times N}$. Despite these constraints, we have a number of efficient algorithms for working with these objects. The *Sinkhorn-Knopp algorithm* [Sinkhorn and Knopp, 1967] maps the positive orthant onto $\mathcal{B}_N$ by iteratively normalizing the rows and columns, and the *Hungarian algorithm* [Kuhn, 1955, Munkres, 1957] solves the minimum weight bipartite matching problem—optimizing a linear objective over the set of permutation matrices—in cubic time.

2.2 Related Work

A number of previous works have considered approximate methods of posterior inference over the space of permutations. When a point estimate will not suffice, sampling methods like Markov chain Monte Carlo (MCMC) algorithms may yield a reasonable approximate posterior for simple problems [Diaconis, 1988]. Harrison and Miller [2013] developed an importance sampling algorithm that fills in count matrices one row at a time, showing promising results for matrices with $O(100)$ rows and columns. Li et al. [2013] considered using the Hungarian algorithm within a Perturb-and-MAP algorithm for approximate sampling. Another line of work considers inference in the spectral domain, approximating distributions over permutations with the low frequency Fourier components [Kondor et al., 2007, Huang et al., 2009]. Perhaps most relevant to this work, Plis et al. [2011] propose a continuous relaxation from permutation matrices to points on a hypersphere, and then use the von Mises-Fisher (vMF) distribution to model distributions on the sphere's surface. We will relax permutations to points in the Birkhoff polytope and derive temperature-controlled densities such that as the temperature goes to zero, the distribution converges to an atomic density on permutation matrices. This will enable efficient variational inference with the reparameterization trick, which we describe next.

2.3 Variational inference and the reparameterization trick

Given an intractable model with data y, likelihood $p(y \,|\, x)$, and prior $p(x)$, variational Bayesian inference algorithms aim to approximate the posterior distribution $p(x \,|\, y)$ with a more tractable distribution $q(x;\theta)$, where "tractable" means that, at a minimum, we can sample q and evaluate it pointwise (including its normalization constant) [Blei et al., 2017]. We find this approximate distribution by searching for the parameters θ that minimize the Kullback-Leibler (KL) divergence between q and the true posterior, or equivalently, maximize the evidence lower bound (ELBO),

$$\mathcal{L}(\theta) \triangleq \mathbb{E}_q \left[\log p(x,y) - \log q(x;\theta)\right].$$

Perhaps the simplest method of optimizing the ELBO is stochastic gradient ascent. However, computing $\nabla_\theta \mathcal{L}(\theta)$ requires some care since the ELBO contains an expectation with respect to a distribution that depends on these parameters.

When x is a continuous random variable, we can sometimes leverage the *reparameterization trick* [Salimans and Knowles, 2013, Kingma and Welling, 2014]. Specifically, in some cases we can simulate from q via the following equivalence,

$$x \sim q(x;\theta) \qquad \Longleftrightarrow \qquad \xi \sim r(\xi), \quad x = g(\xi;\theta),$$

where r is a distribution on the "noise" ξ and where $g(\xi;\theta)$ is a deterministic and differentiable function. The reparameterization trick effectively "factors out" the randomness of q. With this transformation, we can bring the gradient inside the expectation as

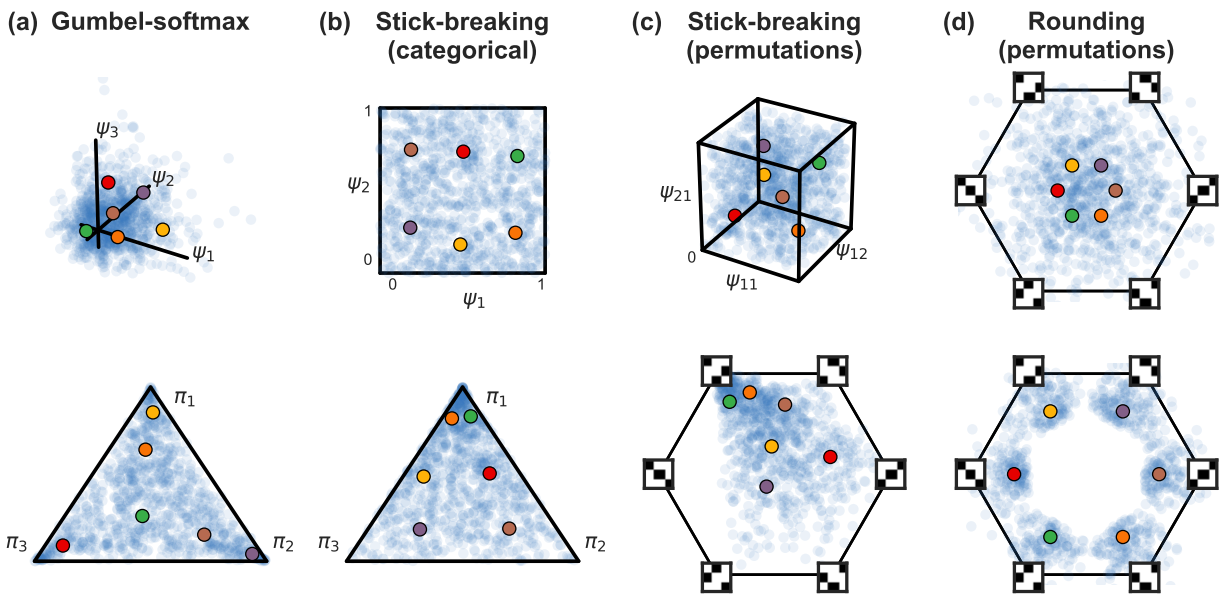

Figure 1: Reparameterizations of discrete polytopes. From left to right: (a) The Gumbel-softmax, or "Concrete" transformation maps Gumbel r.v.'s $\psi \in \mathbb{R}^N$ (blue dots) to points in the simplex $x \in \Delta_N$ by applying the softmax. Colored dots aid in visualizing the transformation. (b) Stick-breaking offers and alternative transformation for categorical inference, here from points $\psi \in [0,1]^{N-1}$ to Δ_N, but the ordering of the stick-breaking induces an asymmetry in the transformation. (c) We extend this stick-breaking transformation to reparameterize the Birkhoff polytope, i.e. the set of doubly-stochastic matrices. We show how $\mathcal{B}_3$ is reparameterized in terms of matrices $\Psi \in [0,1]^{2\times 2}$ (here, $\psi_{22} = 0$). These points are mapped to doubly-stochastic matrices, which we have projected onto $\mathbb{R}^2$ below. (d) Finally, we derive a "rounding" transformation that moves points in $\mathbb{R}^{N\times N}$ nearer to the closest permutation matrix, which is found with the Hungarian algorithm. This is more symmetric, but does not map strictly onto $\mathcal{B}_N$.

follows,

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{r(\xi)}\Big[\nabla_\theta \log p(g(\xi;\theta)\,|\,y) - \nabla_\theta \log q(g(\xi;\theta);\theta)\Big]. \quad (1)$$

This gradient can be estimated with Monte Carlo, and, in practice, this leads to lower variance estimates of the gradient than, for example, the score function estimator [Williams, 1992, Glynn, 1990].

Critically, the gradients in (1) can only be computed if x is continuous. Recently, Maddison et al. [2016] and Jang et al. [2016] proposed the "Gumbel-softmax" method for discrete variational inference. It is based on the following observation: one-hot vectors $x \in \{0,1\}^N$ can be viewed as vertices of the simplex Δ_N; likewise, discrete probability mass functions $q(x;\theta)$ can be seen as atomic densities on the vertices of the simplex. This motivates a natural relaxation: let $q(x;\theta)$ be a density on the interior of the simplex instead and anneal this density such that it converges to an atomic density on the vertices. Fig. 1a illustrates this idea. Gumbel random variates, are mapped through a temperature-controlled softmax function, $g(\psi;\tau) = \left[e^{\psi_1/\tau}/Z, \ldots, e^{\psi_N/\tau}/Z\right]$, where $Z = \sum_{n=1}^N e^{\psi_n/\tau}$, to obtain points in the simplex. As τ goes to zero, the density concentrates on one-hot vectors. We build on these ideas for variational permutation inference.

3 Variational permutation inference via reparameterization

The Gumbel-softmax method scales linearly with the support of the discrete distribution, rendering it prohibitively expensive for direct use on the set of $N!$ permutations. Instead, we develop two transformations to map $O(N^2)$-dimensional random variates to points in or near the Birkhoff polytope.[1] Like the Gumbel-softmax method, these transformations will be controlled by a temperature that concentrates the resulting density near permutation matrices. The first method is a novel "stick-breaking" construction; the second rounds points toward permutations with the Hungarian algorithm. We present these in turn and then discuss their relative merits. We provide further

[1] While Gumbel-softmax does not immediately extend to permutation inference, the methods presented herein easily extend to categorical inference. We explored this direction experimentally and show results in the supplement.

implementation details for both methods in the supplement.

3.1 Stick-breaking transformations to the Birkhoff polytope

Stick-breaking is well-known as a construction for the Dirichlet process [Sethuraman, 1994]; here we show how the same intuition can be extended to more complex discrete objects. Let Ψ be a matrix in $[0,1]^{(N-1)\times(N-1)}$; we will transform it into a doubly-stochastic matrix $X \in [0,1]^{N\times N}$ by filling in entry by entry, starting in the top left and raster scanning left to right then top to bottom. Denote the (m,n)-th entries of Ψ and X by ψ_{mn} and x_{mn}, respectively.

Each row and column has an associated unit-length "stick" that we allot to its entries. The first entry in the matrix is given by $x_{11} = \psi_{11}$. As we work left to right in the first row, the remaining stick length decreases as we add new entries. This reflects the row normalization constraints. The first row follows the standard stick-breaking construction,

$$x_{1n} = \psi_{1n}\left(1 - \sum_{k=1}^{n-1} x_{1k}\right) \quad \text{for } n = 2, \ldots, N-1$$
$$x_{1N} = 1 - \sum_{n=1}^{N-1} x_{1n}.$$

This is illustrated in Fig. 1b, where points in the unit square map to points in the simplex.

Subsequent rows are more interesting, requiring a novel advance on the typical uses of stick breaking. Here we need to conform to row and column sums (which introduce upper bounds), and a lower bound induced by stick remainders that must allow completion of subsequent sum constraints. Specifically, the remaining rows must now conform to both row- and column-constraints. That is,

$$x_{mn} \le 1 - \sum_{k=1}^{n-1} x_{mk} \qquad \text{(row sum)}$$
$$x_{mn} \le 1 - \sum_{k=1}^{m-1} x_{kn} \qquad \text{(column sum)}.$$

Moreover, there is also a lower bound on x_{mn}. This entry must claim enough of the stick such that what is leftover fits within the confines imposed by subsequent column sums. That is, each column sum places an upper bound on the amount that may be attributed to any subsequent entry. If the remaining stick exceeds the sum of these upper bounds, the matrix will not be doubly-stochastic. Thus,

$$\underbrace{1 - \sum_{k=1}^{n} x_{mk}}_{\text{remaining stick}} \le \underbrace{\sum_{j=n+1}^{N} \left(1 - \sum_{k=1}^{m-1} x_{kj}\right)}_{\text{remaining upper bounds}}.$$

Rearranging terms, we have,

$$x_{mn} \ge 1 - N + n - \sum_{k=1}^{n-1} x_{mk} + \sum_{k=1}^{m-1}\sum_{j=n+1}^{N} x_{kj}.$$

Of course, this bound is only relevant if the right hand side is greater than zero. Taken together, we have $\ell_{mn} \le x_{mn} \le u_{mn}$, where,

$$\ell_{mn} \triangleq \max\left\{0,\, 1 - N + n - \sum_{k=1}^{n-1} x_{mk} + \sum_{k=1}^{m-1}\sum_{j=n+1}^{N} x_{kj}\right\}$$
$$u_{mn} \triangleq \min\left\{1 - \sum_{k=1}^{n-1} x_{mk},\, 1 - \sum_{k=1}^{m-1} x_{kn}\right\}.$$

Accordingly, we define $x_{mn} = \ell_{mn} + \psi_{mn}(u_{mn} - \ell_{mn})$. The inverse transformation from X to Ψ is analogous. We start by computing ψ_{11} and then progressively compute upper and lower bounds and set $\psi_{mn} = (x_{mn} - \ell_{mn})/(u_{mn} - \ell_{mn})$.

To complete the reparameterization, we define a parametric, temperature-controlled density for Ψ. Let $\Xi \in \mathbb{R}^{(N-1)\times(N-1)}$ be a matrix of standard Gaussian random variables. We define,

$$\psi_{mn} = \sigma\left(\frac{\mu_{mn} + \eta_{mn}\xi_{mn}}{\tau}\right),$$

where $\theta = \{\mu_{mn}, \eta^2_{mn}\}_{m,n=1}^{N}$ are the mean and variance parameters of the mapping, $\sigma(u) = (1+e^{-u})^{-1}$ is the logistic function, and τ is a temperature parameter. As $\tau \to 0$, the values of ψ_{mn} are pushed to either zero or one, depending on whether the input to the logistic function is negative or positive, respectively. As a result, the doubly-stochastic output matrix X is pushed toward the extreme points of the Birkhoff polytope, the permutation matrices. This map is illustrated in Fig. 1c for permutations of $N = 3$ elements.

We compute gradients of this transformation with automatic differentiation. Since this transformation is "feed-forward," its Jacobian is lower triangular. The determinant of the Jacobian, necessary for evaluating the density $q(X;\theta)$, is a simple function of the upper and lower bounds and is given in the supplement. While this map is peculiar in its reliance on an ordering of the elements, as discussed in Section 3.3, it is a novel transformation to the Birkhoff polytope with the essential properties for gradient-based variational permutation inference.

3.2 Rounding toward permutation matrices

While relaxing permutations to the Birkhoff polytope is intuitively appealing, it is not strictly required. For example, consider the following procedure for sampling a point *near* the Birkhoff polytope:

(i) Input $\Xi \in \mathbb{R}^{N\times N}$, $M \in \mathbb{R}_+^{N\times N}$, and $H \in \mathbb{R}_+^{N\times N}$;

(ii) Map $M \to \mathsf{sink}(M)$, a point near the Birkhoff polytope, using the Sinkhorn-Knopp algorithm;

(iii) Set $\Psi = \mathsf{sink}(M) + H \odot \Xi$ where $\odot$ denotes elementwise multiplication;

(iv) Find $\mathsf{round}(\Psi)$, the nearest permutation matrix to Ψ, using the Hungarian algorithm;

(v) Output $X = \tau\Psi + (1-\tau)\mathsf{round}(\Psi)$.

This procedure is a mapping $X = g(\Xi; M, H)$, and when the elements of Ξ are independently sampled from a standard normal distribution, it implicitly defines a distribution over matrices X parameterized by $\theta = \{M, H\}$. Furthermore, as τ goes to zero, the density concentrates on permutation matrices. A simple example is shown in Fig. 1d, where $M = \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}$, $H = 0.4^2 I$, and $\tau = 0.5$. We use this procedure to define a variational distribution with density $q(X;\theta)$.

To compute the ELBO and its gradient (1), we need to evaluate $q(X;\theta)$. By construction, steps (i) and (ii) involve differentiable transformations of parameter M to set the mean close to the Birkhoff polytope, but since these do not influence the distribution of Ξ, the non-invertibility of the sink function poses no problems. Had we applied sink directly to Ξ, this would not be true. The challenge in computing the density stems from the rounding in steps (iv) and (v).

To compute $q(X;\theta)$, we need the inverse $g^{-1}(X; M, H)$ and its Jacobian. The inverse is straightforward: when $\tau \in [0, 1)$, $\mathsf{round}(\Psi)$ outputs a point strictly closer to the nearest permutation, implying $\mathsf{round}(\Psi) \equiv \mathsf{round}(X)$. Thus, the inverse is $\Psi = \frac{1}{\tau}X - \frac{1-\tau}{\tau}\mathsf{round}(X)$. A slight wrinkle arises from the fact that step (v) maps to a subset $\mathcal{X}_\tau \subset \mathbb{R}^{N\times N}$, but this inverse is valid for all $X \in \mathcal{X}_\tau$.[2]

The Jacobian is more challenging due to the non-differentiability of round. However, since the nearest permutation output only changes at points that are equidistant from two or more permutation matrices, round is a piecewise constant function with discontinuities only at a set of points with zero measure. In practice, we find that we can safely ignore these discontinuities.

With the inverse and its Jacobian, we have

$$q(X;\theta) = \frac{1}{\tau^{N^2}}\mathcal{N}\left(\frac{1}{\tau}X - \frac{1-\tau}{\tau}\mathsf{round}(X);\, \mathsf{sink}(M), H\right),$$

for $X \in \mathcal{X}_\tau$. In the zero-temperature limit we recover a discrete distribution on permutation matrices; otherwise the density concentrates near the vertices as $\tau \to 0$. This transformation leverages computationally efficient algorithms like Sinkhorn-Knopp and the Hungarian algorithm to define a temperature-controlled variational distribution near the Birkhoff polytope, and it enjoys many theoretical and practical benefits.

3.3 Theoretical considerations

The stick-breaking and rounding transformations introduced above each have their strengths and weaknesses. Here we list some of their conceptual differences. While these considerations aid in understanding the differences between the two transformations, the ultimate test is in their empirical performance, which we study in Section 4.

- Stick-breaking relaxes to $\mathcal{B}_N$ whereas rounding relaxes to $\mathbb{R}^{N\times N}$. The Birkhoff polytope is intuitively appealing, but as long as the likelihood, $p(y \mid X)$, accepts real-valued matrices, either may suffice.
- Rounding uses the $O(N^3)$ Hungarian algorithm in its sampling process, whereas stick-breaking has $O(N^2)$ complexity. In practice, the stick-breaking computations are slightly more efficient.
- Rounding can easily incorporate constraints. If certain mappings are invalid, i.e. $X_{mn} \equiv 0$, they are given an infinite cost in the Hungarian algorithm.[3] This is hard to do this with stick breaking as it would change the computation of the upper and lower bounds.
- Stick-breaking introduces a dependence on ordering. While the mapping is bijective, a desired distribution on the Birkhoff polytope may require a complex distribution for Ψ. Rounding, by contrast, is more "symmetric" in this regard.

In summary, stick-breaking offers an intuitive advantage—an exact relaxation to the Birkhoff polytope—but it suffers from its sensitivity to ordering

[2] Consider a simple example of rounding in the one-dimensional simplex, that is, the unit interval. If $\tau = 0.5$, the rounding operation maps $[0, 1]$ to $[0, 0.25) \cup [0.75, 1]$; the resulting density has zero measure in the interval $[0.25, 0.75)$. The same is true of rounding toward permutations: the inverse mapping is only defined for points within τ of a permutation.

[3] Constraints of the form $X_{m,n} \equiv 1$ simply reduce the dimension of the inference problem.

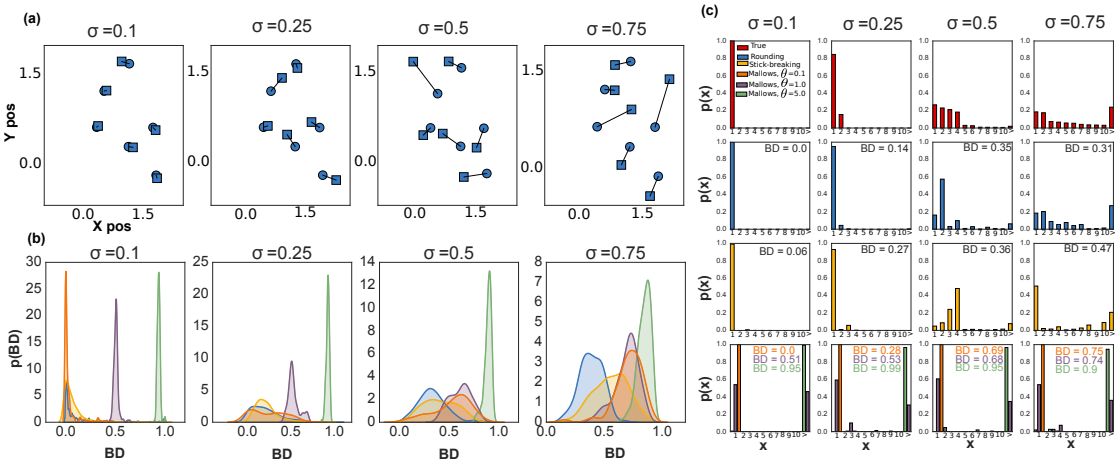

Figure 2: Synthetic matching experiment results. The goal is to infer the lines that match squares to circles. (a) Examples of center locations (circles) and noisy samples (squares), at different noise variances. (b) For illustration, we show the true and inferred probability mass functions for different method (rows) along with the Battacharya distance between them for a selected case of each σ (columns). Permutations (indices) are sorted from the highest to lowest actual posterior probability. Only the 10 most likely configurations are shown, and the 11st bar represents the mass of all remaining configurations. (c) KDE plots of Battacharya distances for each parameter configuration (based on 200 experiment repetitions) for each method and parameter configuration

and its inability to easily incorporate constraints. As we show next, these concerns ultimately lead us to favor the rounding based methods in practice.

4 Synthetic Experiments

We are interested in two principal questions: (i) how well can the stick-breaking and rounding reparameterizations of the Birkhoff polytope approximate the true posterior distribution over permutations in tractable, low-dimensional cases? and (ii) when do our proposed continuous relaxations offer advantages over alternative Bayesian permutation inference algorithms?

To assess the quality of our approximations for distributions over permutations, we considered a toy matching problem in which we are given the locations of N cluster centers and a corresponding set of N observations, one for each cluster, corrupted by Gaussian noise. Moreover, the observations are permuted so there is no correspondence between the order of observations and the order of the cluster centers. The goal is to recover the posterior distribution over permutations. For $N = 6$, we can explicitly enumerate the $N! = 720$ permutations and compute the posterior exactly.

As a baseline, we consider the Mallows distribution Mallows [1957] with density over a permutations ϕ given by $p_{\theta,\phi_0}(\phi) \propto \exp(-\theta d(\phi, \phi_0))$, where ϕ_0 is a central permutation, $d(\phi, \phi_0) = \sum_{i=1}^{N} |\phi(i) - \phi_0(i)|$ is a distance between permutations, and θ controls the spread around ϕ_0. This is the most popular exponential family model for permutations, but since it is necessarily unimodal, it can fail to capture complex permutation distributions.

Table 1: Mean BDs in the synthetic matching experiment for various methods and observation variances.

	Variance σ^2			
Method	$.1^2$	$.25^2$	$.5^2$	$.75^2$
Stick-breaking	.09	.23	.41	.55
Rounding	**.06**	**.21**	**.32**	**.38**
Mallows ($\theta = 0.1$)	.93	.92	.89	.85
Mallows ($\theta = 0.5$)	.51	.53	.61	.71
Mallows ($\theta = 2$)	.23	.33	.53	.69
Mallows ($\theta = 5$)	.08	.27	.54	.72
Mallows ($\theta = 10$)	.08	.27	.54	.72

We measured the discrepancy between true posterior and an empirical estimate of the inferred posteriors using using the Battacharya distance (BD). We fit $q(X; \theta)$ for both stick-breaking and rounding transformations, sampled the variational posterior, and rounded the samples to the nearest permutation matrix with the Hungarian algorithm. For the Mallows distribution, we set ϕ_0 to the MAP estimate, also found with the Hungarian algorithm, and sampled using MCMC.

We found our method outperforms the simple Mallows distribution and reasonably approximates non-trivial distributions over permutations. Fig 2 illustrates our findings, showing (a) sample experiment configurations; (b) examples of inferred, discrete, posteriors for stick breaking (top), rounding (middle), and Mallows (bot-

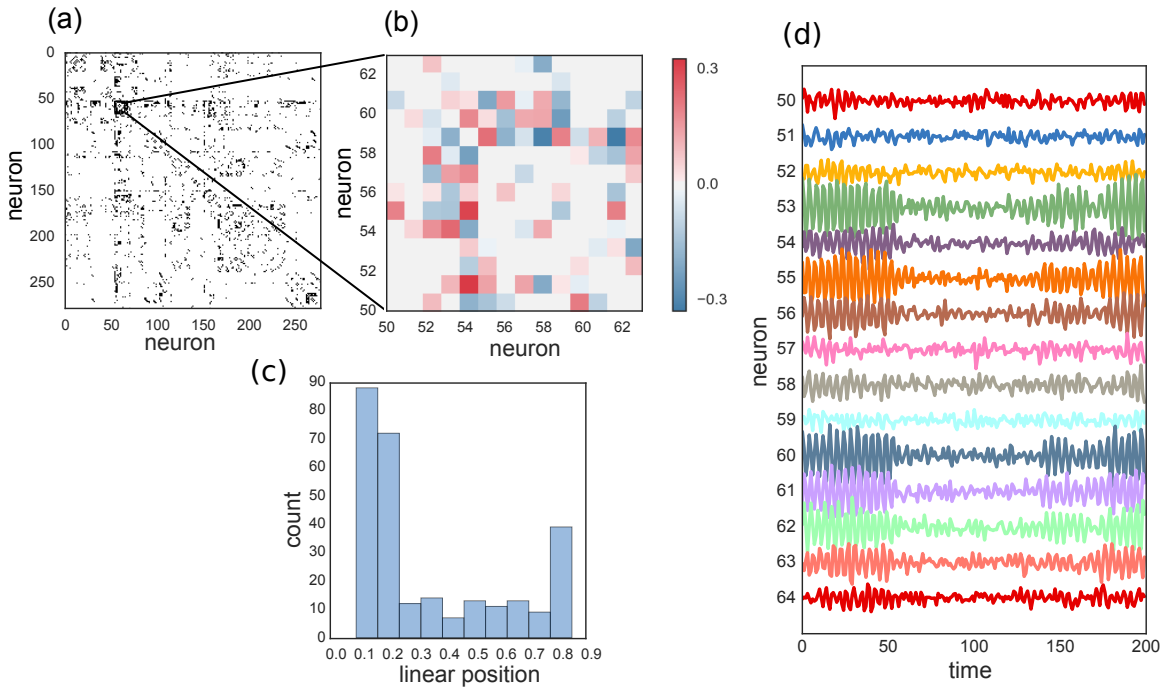

Figure 3: Problem setup. (a) Hermaphrodite C.elegans reference connectome (from Varshney et al. [2011], Lints et al. [2005]) consisting of 278 somatic neurons, merging two distinct types of synapses: chemical and electrical (gap junctions). (b) Example of matrix W consistent with the connectome information (only 14 neurons for visibility), (c) Distribution of neuron position in the body, zero means head and one means tail. From White et al. [1986], Lints et al. [2005] (d). Sampled linear dynamical system with matrix W.

tom); and (c) histogram of Battacharya distance. The latter are summarized in Table 1.

5 Inferring neuron identities in *C. elegans*

Finally, we consider an application motivated by the study of the neural dynamics in *C. elegans*. This worm is a model organism in neuroscience as its neural network is stereotyped from animal to animal and its complete neural wiring diagram is known [Varshney et al., 2011]. We represent this network, or connectome, as a binary adjacency matrix $A \in \{0,1\}^{N\times N}$, shown in Fig. 3a. The hermaphrodite has $N = 278$ somatic neurons, and (undirected) synaptic connections between neurons m and n are denoted by $A_{mn} = 1$.

Modern recording technology enables simultaneous measurements of hundreds of these neurons simultaneously [Kato et al., 2015, Nguyen et al., 2016]. However, matching the observed neurons to nodes in the reference connectome is still a manual task. Experimenters consider the location of the neuron along with its pattern of activity to perform this matching, but the process is laborious and the results prone to error. We prototype an alternative solution, leveraging the location of neurons and their activity in a probabilistic model. We resolve neural identity by integrating different sources of information from the connectome, some covariates (e.g. position) and neural dynamics. Moreover, we combine information from many individuals to facilitate identity resolution. The hierarchical nature of this problem and the plethora of prior constraints and observations motivates our Bayesian approach.

Probabilistic Model. Let J denote the number of worms and $Y^{(j)} \in \mathbb{R}^{T_j\times N}$ denote a recording of worm j with T_j time steps and N neurons. We model the neural activity with a linear dynamical system $Y_t^{(j)} = X^{(j)} W X^{(j)^\mathsf{T}} Y_{t-1}^{(j)} + \varepsilon_t^{(j)}$, where $\varepsilon_t^{(j)}$ is Gaussian noise. Here, $X^{(j)}$ is a latent permutation of neurons that must be inferred for each worm in order to align the observations with the shared dynamics matrix W. The hierarchical component of the model is that W is shared by all worms, and it encodes the influence of one neuron on another (the rows and columns of W are ordered in the same way as the known connectome A). The connectome specifies which entries of W may be non-zero: without a connection ($A_{mn} = 0$) the corresponding weight must be zero; if a connection exists ($A_{mn} = 1$), we must infer its weight. Fig. 3d shows simulated traces from a network that respects the connectivity of A and has random Gaussian weights. The linear model is a simple start; in future work we can incorporate nonlinear dynamics, more informed priors on W, etc.

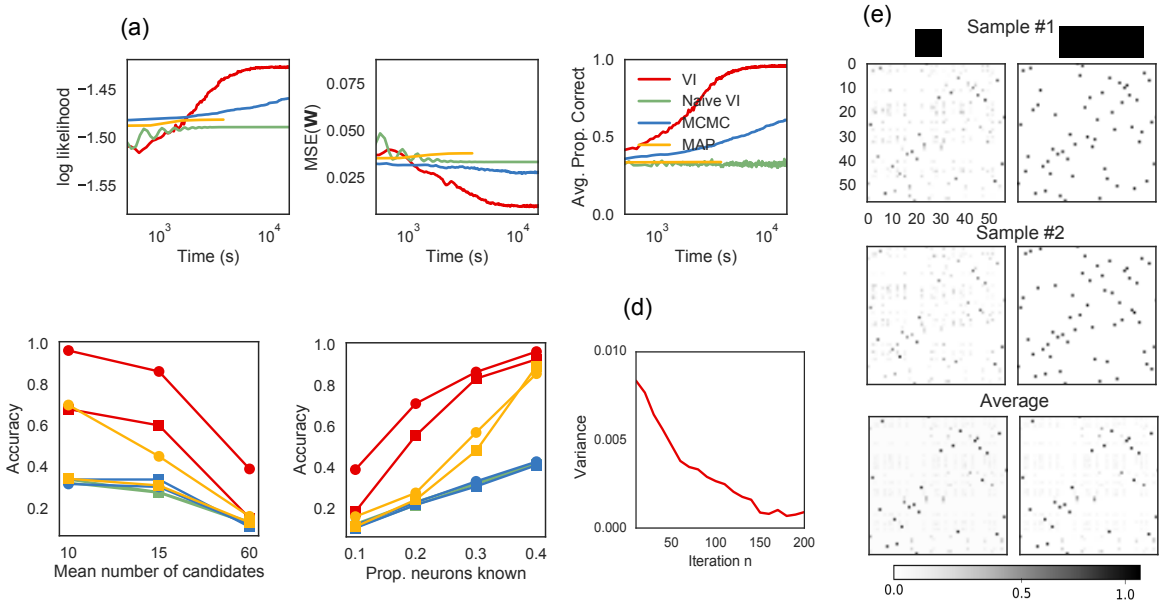

Figure 4: Results on the C.elegans inference example. (a) An example of convergence of the algorithm, and the baselines. (b) Accuracy of identity inference as a function of mean number of candidates (correlated with ν), for $M = 1$ worm (square) and combining information of $M = 5$ worms (circles). (c) Accuracy as a function of the proportion of known networks beforehand, with $\nu = 0.1$ (circles) and $\nu = 0.05$ (squares). (d)Variance of distribution over permutations (vectorized) as a function of the number of iterations. (e) Two samples of permutation matrices $\mathsf{round}(\Psi)$ (right) and their noisy, non-rounded versions Ψ (left) at the twentieth algorithm iteration. The average of many samples is also shown. Presence of grey dots indicate that the sampling procedure is not deterministic.

Our goal is to infer W and $\{X^{(j)}\}$ given $\{Y^{(j)}\}$ using variational permutation inference. We place a standard Gaussian prior on W and a uniform prior on $X^{(j)}$, and we use the rounding transformation to approximate the posterior, $p(W, \{X^{(j)}\} \mid \{Y^{(j)}\}) \propto p(W) \prod_m p(Y^{(j)} \mid W, X^{(j)})\, p(X^{(j)})$.

Finally, we use neural position along the worm's body to constrain the possible neural identities for a given neuron. We use the known positions of each neuron [Lints et al., 2005], approximating the worm as a one-dimensional object with neurons locations distributed as in Fig. 3c. Then, given reported positions of the neurons, we can conceive a binary *constraint* matrix $C^{(j)}$ so that $C^{(j)}_{mn} = 1$ if (observed) neuron m is close enough to (canonical) neuron n; i.e., if their distance is smaller than a tolerance ν. We enforce this constraint during inference by zeroing corresponding entries in the parameter matrix M described in 3.2. This modeling choice greatly reduces the number parameters of the model, and facilitates inference.

Results. We compared against three methods: (i) naive variational inference, where we do not enforce the constraint that $X^{(j)}$ be a permutation and instead treat each row of $X^{(j)}$ as a Dirichlet distributed vector; (ii) MCMC, where we alternate between sampling from the conditionals of W (Gaussian) and $X^{(j)}$, from which one can sample by proposing local swaps, as described in Diaconis [2009], and (iii) maximum a posteriori estimation (MAP). Our MAP algorithm alternates between the optimizing estimate of W given $\{X^{(m)}, Y^{(m)}\}$ using linear regression and finding the optimal $X^{(j)}$. The second step requires solving a quadratic assignment problem (QAP) in $X^{(j)}$; that is, it can be expressed as $\mathrm{Tr}(AXBX^\intercal)$ for matrices A, B. We used the QAP solver proposed by Vogelstein et al. [2015].

We find that our method outperforms each baseline. Fig. 4a illustrates convergence to a better solution for a certain parameter configuration. Moreover, Fig. 4b and Fig. 4c show that our method outperforms alternatives when there are many possible candidates and when only a small proportion of neurons are known with certitude. Fig. 4c also shows that these Bayesian methods benefit from combining information across many worms.

Altogether, these results indicate our method enables a more efficient use of information than its alternatives. This is consistent with other results showing faster convergence of variational inference over MCMC [Blei et al., 2017], especially with simple Metropolis-Hastings proposals. We conjecture that MCMC would eventually obtain similar if not better results, but the local proposals—swapping pairs of labels—leads to slow convergence. On the other hand, Fig 4a shows that our method converges much more quickly while still capturing a distribution over permutations, as shown

by the overall variance of the samples in Fig 4d and the individual samples in Fig 4e.

6 Conclusion

Our results provide evidence that variational permutation variational inference is a valuable tool, especially in complex problems like neural identity inference where information must be aggregated from disparate sources in a hierarchical model. As we apply this to real neural recordings, we must consider more realistic, nonlinear models of neural dynamics. Here, again, we expect variational methods to shine, leveraging automatic gradients of the relaxed ELBO to efficiently explore the space of variational posterior distributions.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

R. P. Adams and R. S. Zemel. Ranking via Sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

M. Balog, N. Tripuraneni, Z. Ghahramani, and A. Weller. Lost relatives of the Gumbel trick. In *Proceedings of the International Conference on Machine Learning*, 2017.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

B. Bloem-Reddy and P. Orbanz. Random walk models of network formation and sequential Monte Carlo methods for graphs. *arXiv preprint arXiv:1612.06404*, 2016.

P. Diaconis. Group representations in probability and statistics. In S. S. Gupta, editor, *Institute of Mathematical Statistics Lecture Notes—Monograph Series*, volume 11. 1988.

P. Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46 (2):179–205, 2009.

P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33 (10):75–84, oct 1990.

L. J. Guibas. The identity management problem: a short survey. In *11th International Conference on Information Fusion*, pages 1–7. IEEE, 2008.

M. T. Harrison and J. W. Miller. Importance sampling for weighted binary random matrices with specified margins. *arXiv preprint arXiv:1301.3928*, 2013.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

J. Huang, C. Guestrin, and L. Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of machine learning research*, 10(May):997–1070, 2009.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

S. Kato, H. Kaplan, T. Schrdel, S. Skora, T. Lindsay, E. Yemini, S. Lockery, and M. Zimmer. Global brain dynamics embed the motor command sequence of caenorhabditis elegans. *Cell*, 163(3):656 – 669, 2015. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2015.09.034.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *AISTATS*, number 1, page 5, 2007.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97, 1955.

M. J. Kusner and J. M. Hernández-Lobato. GANs for sequences of discrete elements with the Gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.

E. L. Lawler. The quadratic assignment problem. *Management science*, 9(4):586–599, 1963.

G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9(Oct):2401–2429, 2008.

K. Li, K. Swersky, and R. Zemel. Efficient feature learning using Perturb-and-MAP. *Neural Information Processing Systems Workshop on Perturbations, Optimization, and Statistics*, 2013.

C. H. Lim and S. Wright. Beyond the Birkhoff polytope: Convex relaxations for vector permutation problems. In *Advances in Neural Information Processing Systems*, pages 2168–2176, 2014.

S. Linderman, M. Johnson, and R. P. Adams. Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464, 2015.

R. Lints, Z. F. Altun, H. Weng, T. Stephney, G. Stephney, M. Volaski, and D. H. Hall. WormAtlas Update. 2005.

D. Maclaurin, D. Duvenaud, and R. P. Adams. Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, 2015.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

C. L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

M. Meilă, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential model. In *In Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

J. W. Miller, M. T. Harrison, et al. Exact sampling and counting for fixed-margin matrices. *The Annals of Statistics*, 41(3):1569–1592, 2013.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

C. Naesseth, F. Ruiz, S. Linderman, and D. Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498, 2017.

J. P. Nguyen, F. B. Shipley, A. N. Linder, G. S. Plummer, M. Liu, S. U. Setru, J. W. Shaevitz, and A. M. Leifer. Whole-brain calcium imaging with cellular resolution in freely behaving Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, 113 (8):E1074–E1081, 2016.

S. M. Plis, S. McCracken, T. Lane, and V. D. Calhoun. Directional statistics on permutations. In *AISTATS*, pages 600–608, 2011.

V. Rao, R. P. Adams, and D. D. Dunson. Bayesian inference for Matérn repulsive processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

T. Salimans and D. A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

J. Shin, N. Lee, S. Thrun, and L. Guibas. Lazy inference on object identities in wireless sensor networks. In *Proceedings of the 4th international symposium on Information processing in sensor networks*, page 23. IEEE Press, 2005.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural properties of the caenorhabditis elegans neuronal network. *PLoS computational biology*, 7(2):e1001066, 2011.

J. T. Vogelstein, J. M. Conroy, V. Lyzinski, L. J. Podrazik, S. G. Kratzer, E. T. Harley, D. E. Fishkind, R. J. Vogelstein, and C. E. Priebe. Fast approximate quadratic programming for graph matching. *PLOS one*, 10(4):e0121002, 2015.

J. G. White, E. Southgate, J. N. Thomson, and S. Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans: the mind of a worm. *Phil. Trans. R. Soc. Lond*, 314:1–340, 1986.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

A Alternative methods of discrete variational inference

Recently there have been a number of proposals for extending the reparameterization trick [Rezende et al., 2014, Kingma and Welling, 2014] to high dimensional discrete problems[4] by relaxing them to analogous continuous problems [Maddison et al., 2016, Jang et al., 2016, Kusner and Hernández-Lobato, 2016]. These approaches are based on the following observation: if $x \in \{0,1\}^N$ is a one-hot vector drawn from a categorical distribution, then the support of $p(x)$ is the set of vertices of the $N-1$ dimensional simplex. We can represent the distribution of x as an atomic density on the simplex.

A.1 The Gumbel-softmax method

Viewing x as a vertex of the simplex motivates a natural relaxation: rather than restricting ourselves to atomic measures, consider continuous densities on the simplex. To be concrete, suppose the density of x is defined by the transformation,

$$\begin{aligned}\xi_n &\overset{\text{iid}}{\sim} \text{Gumbel}(0,1)\\ \psi_n &= \log\theta_n + \xi_n\\ x &= \mathsf{softmax}(\psi/\tau)\\ &= \left(\frac{e^{\psi_1/\tau}}{\sum_{n=1}^N e^{\psi_n/\tau}}, \ldots, \frac{e^{\psi_N/\tau}}{\sum_{n=1}^N e^{\psi_n/\tau}}\right).\end{aligned}$$

The output x is now a point on the simplex, and the parameter $\theta = (\theta_1, \ldots, \theta_N)$ can be optimized via stochastic gradient ascent with the reparameterization trick.

The Gumbel distribution leads to a nicely interpretable model: when θ is a probability mass function, adding Gumbel noise and taking the argmax yields an exact sample from θ; the softmax is a natural relaxation. As the temperature τ goes to zero, the softmax converges to the argmax function. Ultimately, however, this is just a continuous relaxation of an atomic density to a continuous density.

Stick-breaking and rounding offer two alternative ways of conceiving a relaxed version of a discrete random variable, and both are amenable to reparameterization. However, unlike the Gumbel-Softmax, these relaxations enable extensions to more complex combinatorial objects, notably, permutations.

[4]Discrete inference is only problematic in the high dimensional case, since in low dimensional problems we can enumerate the possible values of x and compute the normalizing constant $p(y) = \sum_x p(y,x)$.

A.2 Stick-breaking

The stick-breaking transformation to the Birkhoff polytope presented in the main text contains a recipe for stick-breaking on the simplex. In particular, as we filled in the first row of the doubly-stochastic matrix, we were transforming a real-valued vector $\psi \in \mathbb{R}^{N-1}$ to a point in the simplex. We present this procedure for discrete variational inference again here in simplified form. Start with a reparameterization of a Gaussian vector,

$$\begin{aligned}&\xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\\ &\psi_n = \mu_n + \eta_n\xi_n, \qquad 1 \le n \le N-1,\end{aligned}$$

parameterized by $\theta = (\mu_n, \eta_n)_{n=1}^{N-1}$. Then map this to the unit hypercube in a temperature-controlled manner with the logistic function,

$$z_n = \sigma(\psi_n/\tau),$$

where $\sigma(u) = (1+e^{-u})^{-1}$ is the logistic function. Finally, transform the unit hypercube to a point in the simplex:

$$\begin{aligned}x_1 &= z_1,\\ x_n &= z_n\left(1 - \sum_{m=1}^{n-1} x_m\right), \qquad 2 \le n \le N-1,\\ x_N &= 1 - \sum_{m=1}^{N-1} x_m,\end{aligned}$$

Here, z_n is the fraction of the remaining "stick" of probability mass assigned to x_n. This transformation is invertible, the Jacobian is lower-triangular, and the determinant of the Jacobian is easy to compute. Linderman et al. [2015] compute the density of x implied by a Gaussian density on ψ.

The temperature τ controls how concentrated $p(x)$ is at the vertices of the simplex, and with appropriate choices of parameters, in the limit $\tau \to 0$ we can recover any categorial distribution. In the other limit, as $\tau \to \infty$, the density concentrates on a point in the interior of the simplex determined by the parameters, and for intermediate values, the density is continuous on the simplex.

Finally, note that the logistic-normal construction is only one possible choice. We could instead let $z_n \sim \text{Beta}(\frac{a_n}{\tau}, \frac{b_n}{\tau})$. This would lead to the Dirichlet distribution on the simplex. The beta distribution is slightly harder to reparameterize since it is typically simulated with a rejection sampling procedure, but Naesseth et al. [2017] have shown how this can be handled with a mix of reparameterization and score-function gradients. Alternatively, the beta distribution

could be replaced with the Kumaraswamy distribution, which is quite similar to the beta distribution but is easily reparameterizable.

A.3 Rounding

Rounding transformations also have a natural analog for discrete variational inference. Define the rounding operator,

$$\mathsf{round}(\psi) = \arg\min_{e_n} \|e_n - \psi\|^2,$$

which maps $\psi \in \mathbb{R}^N$ to the one-hot vectors e_n; i.e. the vectors in $\{0,1\}^N$ with n-th entry equal to one and all other entries equal zero. This is equivalent to defining $\mathsf{round}(\psi) = e_{n^*}$ where

$$\begin{aligned} n^* &= \arg\min_n \|e_n - \psi\|^2 \\ &= \arg\min_n \sum_{m \neq n} \psi_m^2 + (1-\psi_n)^2 \\ &= \arg\min_n \sum_{m \neq n} \psi_m^2 + \psi_n^2 - 2\psi_n + 1 \\ &= \arg\min_n \|\psi\|^2 - 2\psi_n + 1 \\ &= \arg\max_n \psi_n. \end{aligned}$$

In the case of a tie, let n^* be the smallest index n such that $\psi_n > \psi_m$ for all $m < n$. Rounding effectively partitions the space into N disjoint "Voronoi" cells,

$$V_n = \Big\{\psi \in \mathbb{R}^N : \psi_n \geq \psi_m \, \forall m \ \wedge \ \psi_n > \psi_m \, \forall m < n\Big\}.$$

By definition, $\mathsf{round}(\psi) = e_{n^*}$ for all $\psi \in V_{n^*}$

We define a map that pulls points toward their rounded values,

$$x = \tau\psi + (1-\tau)\mathsf{round}(\psi). \tag{2}$$

Proposition 1. *For $\tau \in [0,1]$, the map defined by (2) moves points strictly closer to their rounded values so that $\mathsf{round}(\psi) = \mathsf{round}(x)$.*

Proof. Note that the Voronoi cells are intersections of halfspaces and, as such, are convex sets. Since x is a convex combination of ψ and e_{n^*}, both of which belong to the convex set V_{n^*}, x must belong to V_{n^*} as well. □

Similarly, x will be a point on the simplex if an only if ψ is on the simplex as well. By analogy to the rounding transformations for permutation inference, in categorical inference we use a Gaussian distribution $\psi \sim \mathcal{N}(\mathsf{proj}(m), H)$, where $\mathsf{proj}(m)$ is the projection of $m \in \mathbb{R}^N_+$ onto the simplex. Still, the simplex has zero measure under the Gaussian distribution. It follows that the rounded points x will almost surely not be on the simplex either. The supposition of this approach is that this is not a problem: relaxing to the simplex is nice but not required.

In the zero-temperature limit we obtain a discrete distribution on the vertices of the simplex. For $\tau \in (0,1]$ we have a distribution on $\mathcal{X}_\tau \subseteq \mathbb{R}^N$, the subset of the reals to which the rounding operation maps. (For $0 \leq \tau < 1$ this is a strict subset of $\mathbb{R}^N$.) To derive the density $q(x)$, we need the inverse transformation and the determinant of its Jacobian. From Proposition 1, it follows that the inverse transformation is given by,

$$\psi = \frac{1}{\tau}x - \frac{1-\tau}{\tau}\mathsf{round}(x).$$

As long as ψ is in the interior of its Voronoi cell, the round function is piecewise constant and the Jacobian is $\frac{\partial\psi}{\partial x} = \frac{1}{\tau}I$, and its determinant is τ^{-N}. Taken together, we have,

$$\begin{aligned} &q(x; m, H) = \\ &\quad \tau^{-N}\mathcal{N}\left(\frac{1}{\tau}x - \frac{1-\tau}{\tau}\mathsf{round}(x);\, \mathsf{proj}(m), H\right) \\ &\qquad\qquad \times \mathbb{I}[x \in \mathcal{X}_\tau]. \end{aligned}$$

Compare this to the density of the rounded random variables for permutation inference.

B Limit analysis for stick-breaking

We show that stick-breaking for discrete variational inference can converge to any categorical distribution in the zero-temperature limit. We do so with a sequence of propositions: first we show that in the zero-temperature limit, the distribution of $\sigma(\psi_n/\tau)$ converges to a Bernoulli distribution. The we show that when $\sigma(\psi_n/\tau)$ is Bernoulli (rather than a continuous density on the unit interval), the distribution on x obtained by applying the stick-breaking transformation to ψ is categorical.

Proposition 2. *Let $z = \sigma(\psi/\tau)$ with $\psi \sim \mathcal{N}(\mu, \eta^2)$. In the limit $\tau \to 0$ we have $z \sim \mathrm{Bern}(\Phi(-\frac{\mu}{\eta}))$, where $\Phi(\cdot)$ denotes the Gaussian cumulative distribution function (cdf).*

Proof. Let F_z be the cdf of the random variable z. Since z is a random variable on the unit interval, F_z is a non-decreasing function on $[0,1]$ with $F_z(0) = 0$ and $F_z(1) = 1$. Reparame-

terize $\psi = \mu + \eta\xi$ where $\xi \sim \mathcal{N}(0,1)$. Then we have,

$$\begin{aligned} F_z(u) &= \Pr(\sigma(\psi/\tau) < u) \\ &= \Pr(\psi < \tau\sigma^{-1}(u)) \\ &= \Pr(\xi < \tfrac{\tau}{\eta}\sigma^{-1}(u) - \tfrac{\mu}{\eta})) \\ &= \Phi(-\tfrac{\tau}{\eta}\sigma^{-1}(u) - \tfrac{\mu}{\eta}). \end{aligned}$$

By the continuity of Φ we have,

$$\lim_{\tau\to 0} F_z(u) = \Phi(-\tfrac{\mu}{\eta}) \qquad \text{for } u \in (0,1).$$

This is the cdf of a Bernoulli random with probability $\rho = \Phi(-\frac{\mu}{\eta})$. □

Proposition 3. *As above, let $z_n = \sigma(\psi_n/\tau)$. When $z_n \sim \mathrm{Bern}(\rho_n)$ with $\rho_n \in [0,1]$ for $n = 1,\dots,N$, the random variable x obtained from applying the stick-breaking transformation to z will have an atomic distribution with atoms in the vertices of Δ_N; i.e, $x \sim \mathrm{Cat}(\pi)$ where*

$$\begin{aligned} \pi_1 &= \rho_1 \\ \pi_n &= \rho_n \prod_{m=1}^{n-1}(1-\rho_m) \qquad n = 2,\dots,N-1, \\ \pi_N &= \prod_{m=1}^{N-1}(1-\rho_m). \end{aligned}$$

Proof. From the stick-breaking definition, $x_1 = z_1$, $x_n = z_n(1 - \sum_{m<n} x_m)$, and $x_N = 1 - \sum_{m<N} x_m$. When $z_n \in \{0,1\}$ for all $n = 1,\dots,N-1$, we have the following equivalencies. For the first element,

$$x_1 = 1 \iff z_1 = 1;$$

for $1 < n < N-1$:

$$x_n = 1 \iff (z_n = 1) \bigwedge_{m=1}^{n-1} (z_m = 0);$$

and for the last element,

$$x_N = 1 \iff \bigwedge_{m=1}^{N-1} (z_m = 0).$$

These events are mutually exclusive, implying that x will necessarily be a one-hot vector, i.e. a categorical random variable. Since $z_1,\dots,z_{N-1}$ are independent Bernoulli random variables, the probabilities of these events are given by the $\pi,\dots,\pi_N$ stated in the proposition. □

These two propositions, combined with the invertibility of the stick-breaking procedure, lead to our main result.

Lemma 1. *In the zero-temperature limit, stick-breaking of logistic-normal random variables can realize any categorical distribution on x.*

Proof. There is a one-to-one correspondence between $\pi \in \Delta_N$ and $\rho \in [0,1]^{N-1}$. Specifically,

$$\begin{aligned} \rho_1 &= \pi_1 \\ \rho_n &= \frac{\pi_n}{\prod_{m=1}^{n=1} 1-\rho_m} \quad \text{for } n = 2,\dots,N-1. \end{aligned}$$

Since these are recursively defined, we can substitute the definition of ρ_m to obtain an expression for ρ_n in terms of π only. Thus, by Proposition 3, any desired categorical distribution π implies a set of Bernoulli parameters ρ. From Proposition 2, in the zero temperature limit, any desired ρ_n can be obtained with appropriate choice of Gaussian mean μ_n and variance η_n^2. Thus, stick-breaking can realize any categorical distribution when $\tau \to 0$. □

C Variational Autoencoders (VAE) with categorical latent variables

We considered the density estimation task on MNIST digits, as in Maddison et al. [2016], Jang et al. [2016], where observed digits are reconstructed from a latent discrete code. We used the continuous ELBO for training, and evaluated performance based on the marginal likelihood, estimated through the multi-sample variational objective of the discretized model. We compared against the methods of Jang et al. [2016], Maddison et al. [2016] and obtained the results in Table 2. While stick-breaking and rounding fare slightly worse than the Gumbel-softmax method, they are readily extensible to more complex discrete objects, as shown in the main paper.

Table 2: Summary of results in VAE

Method	$-\log p(x)$
Gumbel-Softmax	106.7
Concrete	111.5
Rounding	121.1
Stick-breaking	119. 8

Figure 5 shows MNIST reconstructions using Gumbel-Softmax, stick-breaking and rounding reparameterizations. In all the three cases reconstructions are reasonably accurate, and there is diversity in reconstructions.

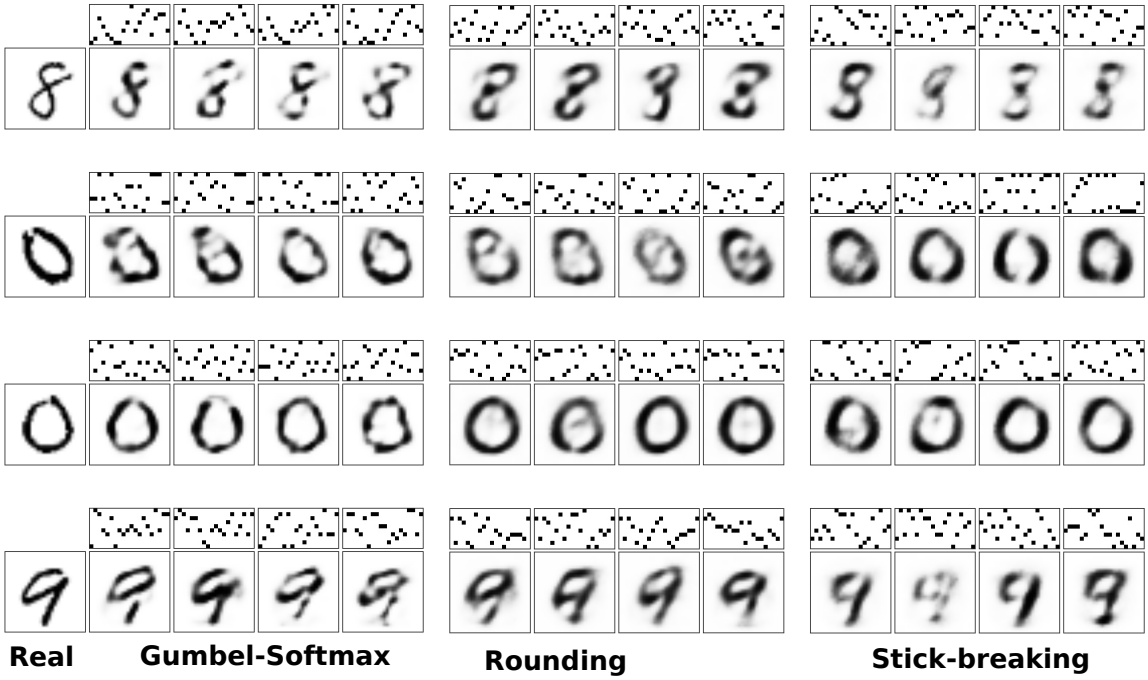

Figure 5: Examples of true and reconstructed digits from their corresponding random codes using with $K = 20$ categorical variables with $N = 10$ possible values.

D Variational permutation inference details

Here we discuss more of the subtleties of variational permutation inference and present the mathematical derivations in more detail.

D.1 Continuous prior distributions.

Continuous relaxations require re-thinking the objective. As in Maddison et al. [2016], we maximize a relaxed ELBO, for which we need to specify a new continuous prior $p(X)$ over the relaxed discrete latent variables, here, over relaxations of permutation matrices. Moreover, it is critical to design sensible priors for relaxed permutations. Ideally, this prior should penalize values of X that are far from permutation matrices.

For our categorical experiment on MNIST we use a mixture of Gaussians around each vertex, $p(x) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}(x \,|\, e_k, \eta^2)$. This can be extended to permutations, where use a mixture of Gaussians for each dimension,

$$p(X) = \prod_{m=1}^{N}\prod_{n=1}^{N}\frac{1}{2}\left(\mathcal{N}(x_{mn} \,|\, 0, \eta^2) + \mathcal{N}(x_{mn} \,|\, 1, \eta^2\right). \tag{3}$$

Although this prior puts significant mass around invalid points (e.g. $\mathbf{1}$), it penalizes X that far from $\mathcal{B}_N$.

D.2 Computing the ELBO

Here we show how to evaluate the ELBO. Note that the stick-breaking and rounding transformations are compositions of invertible functions, $g_\tau = h_\tau \circ f$ with $\Psi = f(\Xi; \theta)$ and $X = h_\tau(\Psi)$. In both cases, f takes in a matrix of independent standard Gaussians (Ξ) and transforms it with the means and variances in θ to output a matrix Ψ with entries $\psi_{mn} \sim \mathcal{N}(\mu_{mn}, \eta^2_{mn})$. Stick-breaking and rounding differ in the temperature-controlled transformations $h_\tau(\Psi)$ they use to map Ψ toward the Birkhoff polytope.

To evaluate the ELBO, we must compute the density of $q_\tau(X; \theta)$. Let $J_h(u) = \frac{\partial h(U)}{\partial U}\big|_{U=u}$ denote the Jacobian of a function h evaluated at value u. By the change of variables theorem and properties of the determinant,

$$\begin{aligned} q_\tau(X; \theta) &= p\big(h_\tau^{-1}(X); \theta\big) \times \left|J_{h_\tau^{-1}}(X)\right| \\ &= p\big(h_\tau^{-1}(X); \theta\big) \times \left|J_{h_\tau}(h_\tau^{-1}(X))\right|^{-1}. \end{aligned}$$

Now we appeal to the law of the unconcious statistician

to compute the entropy of $q_\tau(X;\theta)$,

$$\begin{aligned}\mathbb{E}_{q_\tau(X;\theta)}\Big[-\log q(X;\theta)\Big] &= \mathbb{E}_{p(\Psi;\theta)}\Big[-\log p(\Psi;\theta)+\log|J_{h_\tau}(\Psi)|\Big] \\ &= \mathbb{H}(\Psi;\theta)+\mathbb{E}_{p(\Psi;\theta)}\Big[|J_{h_\tau}(\Psi)|\Big]. \qquad (4)\end{aligned}$$

Since Ψ consists of independent Gaussians, the entropy is simply,

$$\mathbb{H}(\Psi;\theta)=\frac{1}{2}\sum_{m,n}\log(2\pi e\eta_{mn}^2).$$

We estimate the second term of equation (4) using Monte-Carlo samples. For both transformations, the Jacobian has a simple form.

Jacobian of the rounding transformation. The rounding transformation is given in matrix form in the main text, and we restate it here in coordinate-wise form for convenience,

$$x_{mn}=[h_\tau(\Psi)]_{mn}=\tau\psi_{mn}+(1-\tau)[\mathsf{round}(\Psi)]_{mn}.$$

This transformation is piecewise linear with jumps at the boundaries of the "Voronoi cells;" i.e., the points where $\mathsf{round}(X)$ changes. The set of discontinuities has Lebesgue measure zero so the change of variables theorem still applies. Within each Voronoi cell, the rounding operation is constant, and the Jacobian is,

$$\log\left|J_{h_\tau}(\Psi)\right|=\sum_{m,n}\log\tau=N^2\log\tau.$$

For the rounding transformation with given temperature, the Jacobian is constant.

Jacobian of the stick-breaking transformation. Here h_τ consists of two steps: map $\Psi\in\mathbb{R}^{N-1\times N-1}$ to $Z\in[0,1]^{N-1\times N-1}$ with a temperature-controlled, elementwise logistic function, then map Z to $X\in\mathcal{B}_N$ with the stick-breaking transformation.

As with the standard stick-breaking transformation to the simplex, our transformation to the Birkhoff polytope is feed-forward; i.e. to compute x_{mn} we only need to know the values of z up to and including the (m,n)-th entry. Consequently, the Jacobian of the transformation is triangular, and its determinant is simply the product of its diagonal.

We derive an explicit form in two steps. With a slight abuse of notation, note that the Jacobian of $h_\tau(\Psi)$ is given by the chain rule,

$$J_{h_\tau}(\Psi)=\frac{\partial X}{\partial\Psi}=\frac{\partial X}{\partial Z}\frac{\partial Z}{\partial\Psi}.$$

Since both transformations are bijective, the determinant is,

$$\left|J_{h_\tau}(\Psi)\right|=\left|\frac{\partial X}{\partial Z}\right|\left|\frac{\partial Z}{\partial\Psi}\right|.$$

the product of the individual determinants. The first determinant is,

$$\left|\frac{\partial X}{\partial Z}\right|=\prod_{m=1}^{N-1}\prod_{n=1}^{N-1}\frac{\partial x_{mn}}{\partial z_{mn}}=\prod_{m=1}^{N-1}\prod_{n=1}^{N-1}(u_{mn}-\ell_{mn}).$$

The second transformation, from Ψ to Z, is an elementwise, temperature-controlled logistic transformation such that,

$$\begin{aligned}\left|\frac{\partial Z}{\partial\Psi}\right| &= \prod_{m=1}^{N-1}\prod_{n=1}^{N-1}\frac{\partial z_{mn}}{\partial\psi_{mn}} \\ &= \prod_{m=1}^{N-1}\prod_{n=1}^{N-1}\frac{1}{\tau}\sigma\left(\psi_{mn}/\tau\right)\sigma\left(-\psi_{mn}/\tau\right).\end{aligned}$$

double check

It is important to note that the transformation that maps $Z\to X$ is only piecewise continuous: the function is not differentiable at the points where the bounds change; for example, when changing Z causes the active upper bound to switch from the row to the column constraint or vice versa.

conclude this section.

E Experiment details

We used Tensorflow [Abadi et al., 2016] for the VAE experiments, slightly changing the code made available from Jang et al. [2016]. For experiments on synthetic matching and the C. elegans example we used Autograd [Maclaurin et al., 2015], explicitly avoiding propagating gradients through the non-differentiable round operation, which requires solving a matching problem.

We used the ADAM with learning rate 0.1 for optimization. For rounding, the parameter vector H defined in 3.2(iii) was constrained to lie in the interval $[0.1, 0.5]$. Also, for rounding, we used ten iterations of the Sinkhorn-Knopp algorithm, to obtain points in the Birkhoff polytope. For stick-breaking the variances ν defined in **??** were constrained between $1e-8$ and 1.0. In either case, the temperature, along with maximum values for the noise variances were calibrated using a grid search.

In the C. elegans example we considered the symmetrized version of the adjacency matrix described in [Varshney et al., 2011]; i.e. we used $A'=(A+A^\top)/2$,

and the matrix W was chosen antisymmetric, with entries sampled randomly with the sparsity pattern dictated by A'. To avoid divergence, the matrix W was then re-scaled by 1.1 times its spectral radius. This choice, although not essential, induced a reasonably well behaved linear dynamical system, rich in non-damped oscillations. We used a time window of $T = 1000$ time samples, and added spherical standard noise at each time.