Reparameterizing the Birkhoff Polytope for Variational Permutation Inference

Scott W. Linderman* Columbia University

Gonzalo E. Mena* Columbia University

Hal Cooper Columbia University

Liam Paninski Columbia University

John P. Cunningham Columbia University

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 challenge, we start by relaxing the discrete set of permutation matrices to its convex hull the Birkhoff polytope, the set of doubly-stochastic matrices. We then introduce two novel transformations: an invertible and differentiable stick-breaking procedure that maps unconstrained space to the Birkhoff polytope, and a map that rounds points toward the vertices of the polytope. Both transformations include a temperature parameter that, in the limit, concentrates the densities on permutation matrices. We exploit these transformations and reparameterization gradients to introduce variational inference over permutation matrices, and we demonstrate its utility in a series of experiments.

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]. 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 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. We propose an alternative, 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.

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. JMLR: W&CP volume 7X. *Equal contribution.

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., 2017, Jang 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 entry $x_{m,n} = 1$ 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 is called the *Birkhoff polytope*. Let $\mathcal{B}_N$ denote the Birkhoff polytope of $N \times N$ doubly-stochastic matrices. The 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 cost 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. Finally, ranking problems are a special case of a matching problems in which the labels are the ordered set of integers $\{1,\dots,N\}$. The Placket-Luce model is one model for rankings that is parameterized by a "score" for each item, and it admits efficient Bayesian inference algorithms [Guiver and Snelson, 2009]. In general matching problems, however, the output is not ordered, and we instead need scores for each item-label mapping. The methods presented here address general matching problems.

2.3 Variational inference and the reparameterization trick

Given a model with data y, likelihood $p(y \mid x)$, and prior $p(x)$, variational Bayesian inference algorithms aim to approximate the posterior distribution $p(x \mid y)$ with a more tractable distribution $q(x;\theta)$, where "tractable" means that 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,

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

where r is a distribution on the "noise" z and where $g(z;\theta)$ 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

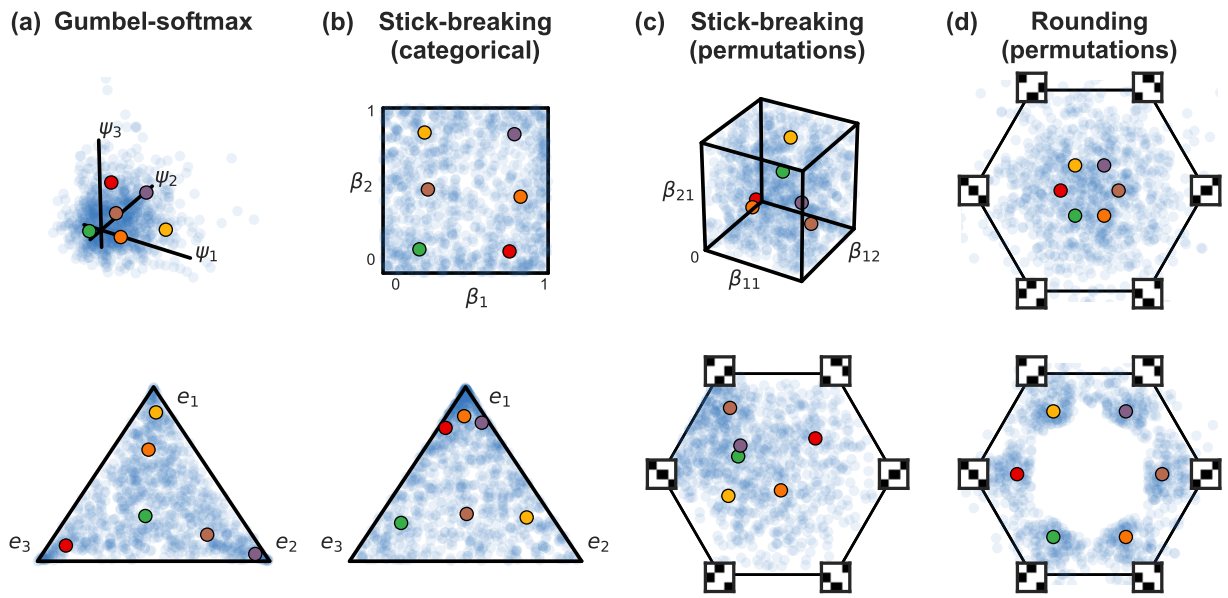

Figure 1: Reparameterizations of discrete polytopes. From left to right: (a) The Gumbel-softmax, or "Concrete" transformation maps Gumbel r.v.'s $\psi \in \mathbb{R}^N$ (blue dots) to points in the simplex $x \in \Delta_N$ by applying the softmax. Colored dots are random variates that aid in visualizing the transformation. (b) Stick-breaking offers and alternative transformation for categorical inference, here from points $\beta \in [0,1]^{N-1}$ 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 $B \in [0,1]^{2\times 2}$ These points are mapped to doubly-stochastic matrices, which we have projected onto $\mathbb{R}^2$ below (stencils show permutation matrices at the vertices). (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,

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

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. [2017] and Jang et al. [2017] proposed the "Gumbel-softmax" method for discrete variational inference. It is based on the following observation: discrete probability mass functions $q(x;\theta)$ can be seen as densities with atoms on the vertices of the simplex; i.e. on the set of one-hot vectors $\{e_n\}_{n=1}^N$, where $e_n = (0,0,\ldots,1,\ldots,0)^\mathsf{T}$ is a length-N binary vector with a single 1 in the n-th position. 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, $g_\tau(\psi) = \left[e^{\psi_1/\tau}/Z, \ldots, e^{\psi_N/\tau}/Z\right]$, 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.

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 implementation details for both methods in the supplement.

[1] While Gumbel-softmax does not immediately extend to permutation inference, the methods presented herein do apply to categorical inference. We explored this direction experimentally and show results 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 B 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 B and X by β_{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 $x_{11} = \beta_{11}$. 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,

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

This is illustrated in Fig. 1b, where points in the unit square map to points in the simplex. Here, the blue dots are two-dimensional $\mathcal{N}(0, 4I)$ variates mapped through a coordinate-wise logistic function.

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 remains 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 $x_{mn} = \ell_{mn} + \beta_{mn}(u_{mn} - \ell_{mn})$. The inverse transformation from X to B is analogous. We start by computing z_{11} and then progressively compute upper and lower bounds and set $\beta_{mn} = (x_{mn} - \ell_{mn})/(u_{mn} - \ell_{mn})$.

To complete the reparameterization, we define a parametric, temperature-controlled density from a standard Gaussian matrix $Z \in \mathbb{R}^{(N-1)\times(N-1)}$ to the unit-hypercube B. Let,

$$\psi_{mn} = \mu_{mn} + \nu_{mn} z_{mn},$$
$$\beta_{mn} = \sigma\left(\psi_{mn}/\tau\right),$$

where $\theta = \{\mu_{mn}, \nu^2_{mn}\}_{m,n=1}^{N}$ are the mean and variance parameters of the intermediate Gaussian matrix Ψ, $\sigma(u) = (1+e^{-u})^{-1}$ is the logistic function, and τ is a temperature parameter. As $\tau \to 0$, the values of β_{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. Here, the blue dots are samples of B with $\mu_{mn} = 0$, $\nu_{mn} = 2$, and $\tau = 1$.

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 $q_\tau(X;\theta)$, is a simple function of the upper and lower bounds and is derived in Appendix B. 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 that supports variational 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 $Z \in \mathbb{R}^{N\times N}$, $M \in \mathbb{R}_+^{N\times N}$, and $V \in \mathbb{R}_+^{N\times N}$;

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

(iii) Set $\Psi = \widetilde{M} + V \odot Z$ 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 defines a mapping $X = g_\tau(Z;\theta)$ with $\theta = \{M, V\}$. When the elements of Z are independently sampled from a standard normal distribution, it implicitly defines a distribution over matrices X parameterized by θ. Furthermore, as τ goes to zero, the density concentrates on permutation matrices. A simple example is shown in Fig. 1d, where $M = \frac{1}{N}\mathbf{1}\mathbf{1}^\mathsf{T}$ with $\mathbf{1}$ a vector of all ones, $V = 0.4^2\mathbf{1}\mathbf{1}^\mathsf{T}$, and $\tau = 0.5$. We use this procedure to define a variational distribution with density $q_\tau(X;\theta)$.

To compute the ELBO and its gradient (1), we need to evaluate $q_\tau(X;\theta)$. 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 Z, the non-invertibility of the Sinkhorn-Knopp algorithm poses no problems. Had we applied this algorithm directly to Z, this would not be true. The challenge in computing the density stems from the rounding in steps (iv) and (v).

To compute $q_\tau(X;\theta)$, we need the inverse $g_\tau^{-1}(X;\theta)$ 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 $g_\tau^{-1}(X;\theta) = \left(\frac{1}{\tau}X - \frac{1-\tau}{\tau}\mathsf{round}(X) - \widetilde{M}\right) \oslash V$, where $\oslash$ denotes elementwise division. A slight wrinkle arises from the fact that step (v) maps to a subset $\mathcal{X}_\tau \subset \mathbb{R}^{N\times N}$ that excludes the center of the Birkhoff polytope (note the "hole" in Fig. 1d), but this inverse is valid for all X in that subset.

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. Thus, the change of variables theorem still applies.

With the inverse and its Jacobian, we have

$$q_\tau(X;\theta) = \prod_{m=1}^{N}\prod_{n=1}^{N} \frac{1}{\tau\nu_{mn}} \mathcal{N}\left(z_{mn};\, 0, 1\right) \times \mathbb{I}[X \in \mathcal{X}_\tau],$$

where $z_{mn} = [g_\tau^{-1}(X;\theta)]_{mn}$ and ν_{mn} are the entries of V. 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 the Birkhoff polytope 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. 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 B. 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 and its inability to easily incorporate constraints. As we show next, these concerns ultimately lead us to favor the rounding based methods in practice.

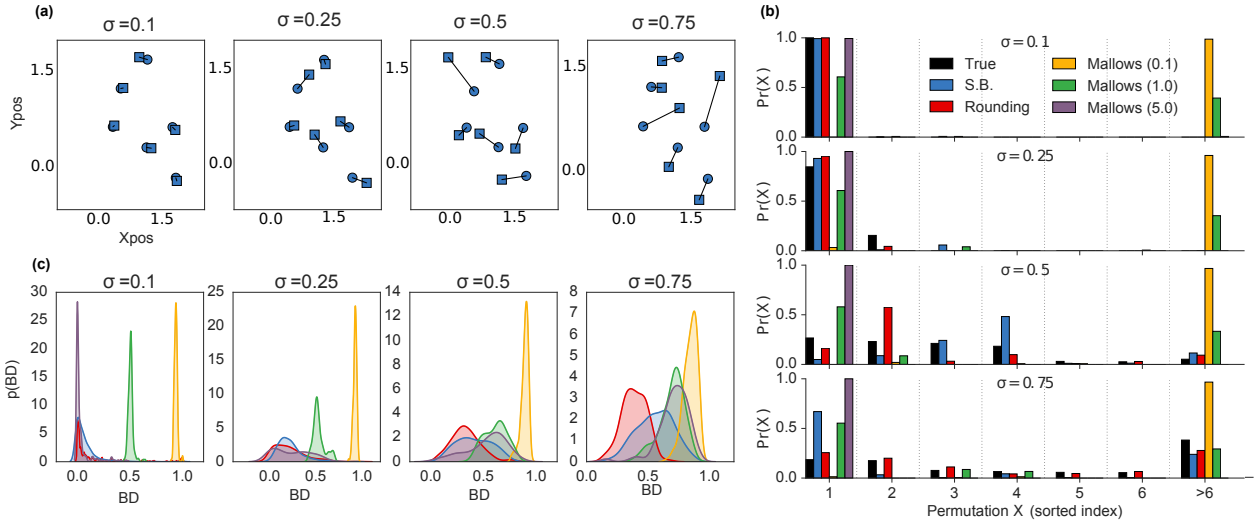

Figure 2: Synthetic matching experiment results. The goal is to infer the lines that match squares to circles. (a) Examples of center locations (circles) and noisy samples (squares), at different noise variances. (b) For illustration, we show the true and inferred probability mass functions for different method (rows) along with the Battacharya distance (BD) 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. For comparison, stick-breaking, rounding, and Mallows ($\theta = 1.0$) have BD's of .36, .35, and .66, respectively, in the $\sigma = 0.5$ row of (b).

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 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 $q_\tau(X; \theta)$ with a fixed τ (a hyperparameter) 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.

Both methods outperform the simple Mallows distribution and reasonably approximate non-trivial distributions over permutations. Fig. 2 shows (a) sample experiment configurations; (b) examples of inferred discrete posteriors for stick breaking, rounding, and Mallows at increasing levels of noise; and (c) histogram

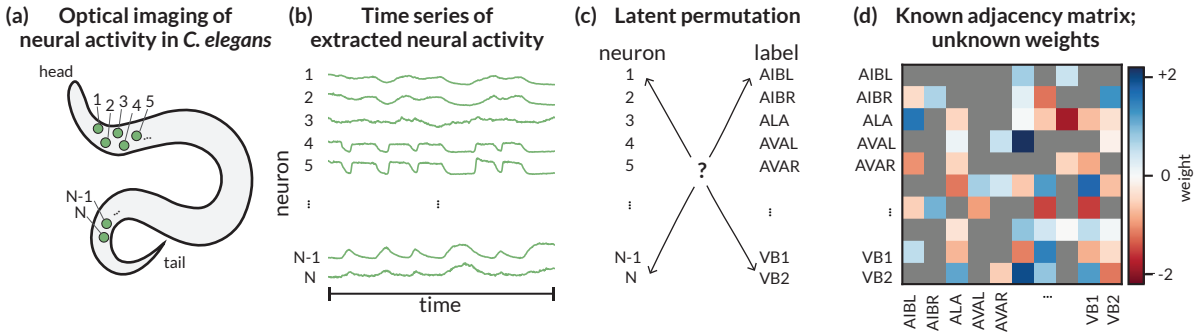

Figure 3: Inferring labels and weights in *C. elegans*. **(a)** Neural activity is optically recorded in genetically modified *C. elegans*. **(b)** The output is a multivariate time series of neural activity of N neurons for each worm. **(c)** The first challenge is to infer a latent permutation that matches observed neuron indices to the known set of neuron names, or labels. **(d)** The second challenge is to infer the weights with which each neuron influences its synaptic neighbors. The connectome (i.e. adjacency matrix) is known, but the weights are not.

of Battacharya distance. The latter are summarized in Table 1.

5 Brain dynamics of *C. elegans*

Finally, we consider an application motivated by the study of the neural dynamics in *C. elegans*. This worm presents many advantages for scientific study. Each hermaphrodite worm has the same $N = 302$ neurons, and each neuron has a label, like AIBL, AVAL, etc. Moreover, the worm's *connectome*—the adjacency matrix that specifies how neurons are connected—is well-characterized [White et al., 1986, Varshney et al., 2011]. Yet while the adjacency matrix is known, the *weights* associated with these connections are not.

Modern recording technologies enable whole-brain recordings in *C. elegans* [Nguyen et al., 2016], presenting an opportunity to learn these weights. Fig. 3a and 3b provide a cartoon illustration: worms are genetically altered so that their neurons fluoresce when active; in each frame of the movie neurons appear as dots in the image, and over time the intensity of these dots provides an optical read-out of the neural activity. However, labeling neurons—i.e. finding the latent permutation that matches observed neurons to known labels, as in Fig. 3c—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 are prone to error. We prototype an alternative solution, leveraging the location of neurons and their activity in a probabilistic model. We infer neural labels by combining information from the connectome, covariates like position, and neural dynamics. Moreover, we combine information from many individual worms to aid in this labeling. The hierarchical nature of this problem and the plethora of prior constraints and observations motivate 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 as a linear autoregressive process $Y_t^{(j)} = X^{(j)} W X^{(j)\mathsf{T}} Y_{t-1}^{(j)} + \varepsilon_t^{(j)}$, where $\varepsilon_t^{(j)}$ is Gaussian. Here, $X^{(j)}$ is a latent permutation that must be inferred for each worm in order to align the per-worm observations with the shared dynamics matrix W. The hierarchical component of the model is that the unknown weight matrix 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 $A \in \{0,1\}^{N \times N}$. 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. A cartoon example is shown in Fig. 3d.

Our goal is to infer W and $\{X^{(j)}\}$ given $\{Y^{(j)}\}$ and A 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, $p(W, \{X^{(j)}\} \mid \{Y^{(j)}\}, A) \propto p(W \mid A) \prod_j p(Y^{(j)} \mid W, X^{(j)}, A)\, p(X^{(j)})$.

Finally, we use neural position along the worm's body [Lints et al., 2005] to constrain the possible neural identities for a given neuron. Given the positions of the neurons, we construct a binary *constraint* matrix $C^{(j)}$ so that $C_{mn}^{(j)} = 1$ if observed neuron m is close to where label n is typically found. We enforce this constraint by zeroing corresponding entries in the parameter matrix M described in Section 3.2. These constraints greatly reduce the number parameters of the model and facilitate inference.

Results. We compared against three methods: (i) naive variational inference, where we do not enforce the

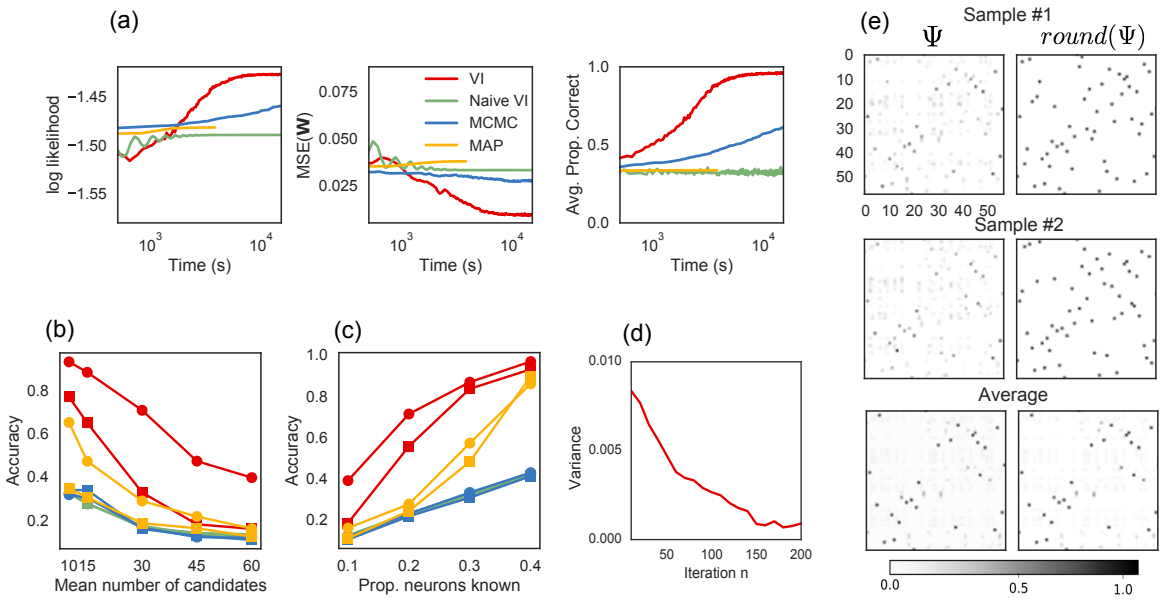

Figure 4: Results on the C.elegans inference example. (a) An example of convergence of the algorithm, and the baselines. (b) Accuracy of identity inference as a function of mean number of candidates (correlated with ν), for $M = 1$ worm (square) and combining information of $M = 5$ worms (circles). (c) 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. These averages take values in $(0, 1)$, indicating uncertainty in the variational posterior.

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 $\mathrm{Tr}(AXBX^\mathsf{T})$ 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 certainty. 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 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.

Acknowledgments. We thank Christian Naesseth for many helpful discussions. SWL is supported by the Simons Collaboration on the Global Brain (SCGB) Postdoctoral Fellowship (418011). HC is supported by Graphen, Inc. LP is supported by ARO MURI W91NF-12-1-0594, the SCGB, DARPA SIMPLEX N66001-15-C-4032, IARPA MICRONS D16PC00003, and ONR N00014-16-1-2176. JPC is supported by the Sloan Foundation, McKnight Foundation, and the SCGB.

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.

J. Guiver and E. Snelson. Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384. ACM, 2009.

E. J. Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*. Number 33. US Govt. Print. Office, 1954.

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. In *Proceedings of the International Conference on Learning Representations*, 2017.

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.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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 *Artificial Intelligence and Statistics*, 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.

P. Kumaraswamy. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1-2):79–88, 1980.

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.

R. D. Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 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. 2017.

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 *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 *Artificial Intelligence and Statistics*, 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.