Reparameterizing the Birkhoff Polytope for Variational Permutation Inference

Anonymous Authors
Anonymous Institutions

Abstract

How can we efficiently perform posterior inference over the space of permutations when there are $N!$ permutations of a set of N elements? Combinatorial optimization algorithms may enable efficient point estimation, but fully Bayesian inference poses a severe challenge in this high-dimensional, discrete space. We begin with a common maneuver: we relax the discrete set of permutation matrices—the vertices of the Birkhoff polytope—to the continuous set of doubly-stochastic matrices—the interior of the polytope. Our primary contribution is a pair of invertible and differentiable maps from densities on unconstrained space to densities on or near the Birkhoff polytope. These transformations are parameterized by a "temperature" that controls how concentrated the resulting density is at the extrema of the Birkhoff polytope; i.e. at permutation matrices. With these transformations, we perform variational inference over distributions on doubly stochastic matrices (and in the zero-temperature limit, on permutation matrices), leveraging reparameterization gradients to guide our optimization.

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]. Moreover, 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 many of these settings, permutation inference is just one component of a larger estimation 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 require reasoning about uncertainty regarding 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]. Sections 4 and 5 presents a variety of experiments that illustrate the benefits of the proposed variational approach.

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

2 Background

2.1 Definitions and notation.

A permutation is a bijective mapping of a set $\mathcal{X}$ onto itself. When $\mathcal{X} = \{x_1, \ldots, x_N\}$, this mapping is conveniently represented as a binary matrix $X \in \{0,1\}^{N \times N}$ where $X_{m,n} = 1$ implies that x_m is mapped to x_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 permutation matrices are vertices of the convex hull 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] projects 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 Variational inference and the reparameterization trick

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). 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 often go one step further and 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,

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

where r is a distribution on the "noise" ξ and where $g(\xi; \theta)$ is a deterministic and differentiable function. For example, if $q(x; \theta) = \mathcal{N}(x \,|\, \theta, 1)$, we can reparameterize by setting the noise distribution to $r(\xi) = \mathcal{N}(\xi \,|\, 0, 1)$ and using the transformation $g(\xi; \theta) = \xi + \theta$. The reparameterization trick effectively "factors out" the randomness of q. With this transformation, we can bring the gradient inside the expectation as follows,

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

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]. However, for g to be differentiable x needs to be continuous.

Recently, there have been a number of proposals for extending the reparameterization trick to high-dimensional discrete inference problems via continuous relaxation [Maddison et al., 2016, Jang et al., 2016, Kusner and Hernández-Lobato, 2016]. The key here is a "temperature" knob that controls the degree of relaxation. In Section 3.3 we discuss how our approach extends these ideas to discrete inference problems with more complex structure, like permutation matrices.

2.3 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. It may also be possible to turn the Hungarian algorithm into an efficient sampling algorithms using Perturb-and-MAP [Li et al., 2013]. 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].

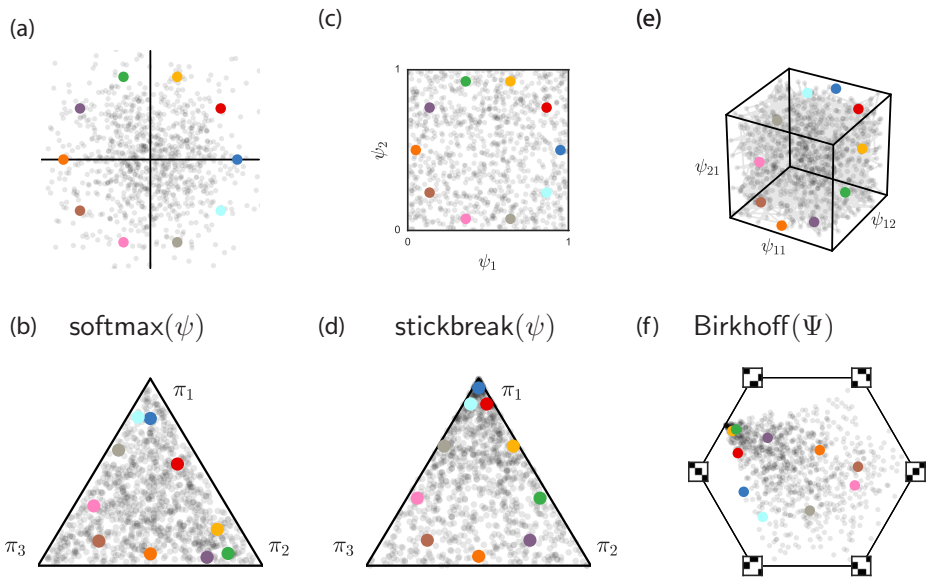

Figure 1: Reparameterizations of discrete polytopes. (a,b) The Gumbel-softmax, or "Concrete" transformation maps points $\psi \in \mathbb{R}^N$ to points $x \in \Delta_N$ by adding noise and applying the softmax. Here we show a slice for $N = 3$ with $\psi_3 = 0$. Colored points are aids to visualize the transformation. (c,d) Stick-breaking offers and alternative transformation, here from points $\psi \in [0, 1]^{N-1}$ to Δ_N. The ordering of the stick-breaking induces an asymmetry in the transformation. (e,f) We extend this stick-breaking transformation to reparameterize the Birkhoff polytope, i.e. the set of doubly stochastic matrices. Here, $\mathcal{B}_3$ is reparameterized in terms of matrices $\Psi \in [0, 1]^{2\times 2}$, of which three coordinates are shown in (e). These points are mapped to doubly stochastic matrices, which we have projected onto $\mathbb{R}^2$ in panel (f).

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. While the vMF distribution does have a concentration parameter, as the concentration goes to infinity, the distribution converges to a point on the sphere. By contrast, we will derive temperature-controlled densities over points inside or near the Birkhoff polytope such that as the temperature goes to zero, the distribution converges to an atomic density on permutation matrices.

3 Variational permutation inference via reparameterization

The key to stochastic variational inference with the reparameterization trick is an invertible and differentiable mapping $x = g(\xi; \theta)$. With this mapping, a simple noise density $r(\xi)$ is transformed into a variational posterior density $q(x; \theta)$ that can be sampled and evaluated pointwise—the necessary ingredients for computing the stochastic gradients of the ELBO (1). For discrete variational inference via continuous relaxation, we further require a temperature control. Here we develop two mappings for permutation inference.

3.1 Stick-breaking transformations of the Birkhoff polytope

Let Ψ be an arbitrary 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. Formally, the stick-breaking transformation for the first row is given by,

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

$$x_{1N} = 1 - \sum_{n=1}^{N-1} x_{1n}.$$

However, 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.

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 a point $M \in \mathbb{R}_+^{N\times N}$;

(ii) Project M onto the Birkhoff polytope (approximately) using the Sinkhorn-Knopp algorithm;

(iii) Sample a Gaussian random variable Ψ with mean $\text{proj}(M)$ and variance Σ;

(iv) Find the permutation matrix $P^*(\Psi)$ nearest to Ψ using the Hungarian algorithm; and

(v) Return $X = \tau\Psi + (1-\tau)P^*(\Psi)$.

This procedure implicitly defines a distribution over matrices X. Steps (i) and (ii) involve differentiable transformations of parameter M to set the mean close to the Birkhoff polytope; the only challenge in computing the density $p(X; M, \Sigma)$ stems from step (iv), since the rounding operation is not differentiable. However, this operation is piecewise constant with discontinuities only at points that are equidistant from two or more permutation matrices—a set of measure zero. In practice, we find that we can safely ignore these discontinuities and treat $P^*(\Psi)$ as constant with respect to Ψ. Furthermore, note that $P^*(\Psi) \equiv P^*(X)$ so that the inverse transformation is $\Psi = \tau^{-1}X - \tau^{-1}(1-\tau)P^*(X)$. Taken together, X is a linear function of a Gaussian random variable and its density is,

$$p(X; M, \Sigma) = \frac{1}{\tau}\mathcal{N}\left(\frac{1}{\tau}X - \frac{1-\tau}{\tau}P^*(X);\, \text{proj}(M), \Sigma\right).$$

In the zero-temperature limit we recover a discrete distribution on permutation matrices, and for $\tau \in (0, 1]$, the distribution is continuous on $\mathbb{R}^{N\times N}$, with density concentrating near the vertices as $\tau \to 0$.

work in progress

- Stick-breaking relaxes to $\mathcal{B}_N$ whereas rounding is to reals; Birkhoff is intuitively nice.
- Stick-breaking is $O(N^2)$ whereas rounding requires $O(N^3)$ Hungarian call.
- Stick-breaking admits exact density on Birkhoff whereas rounding has weird measure-zero concerns.
- Rounding easily handles constraints, hard to do this with stick breaking.
- Rounding is more "symmetric" whereas stick-breaking has an implicit dependence on ordering.

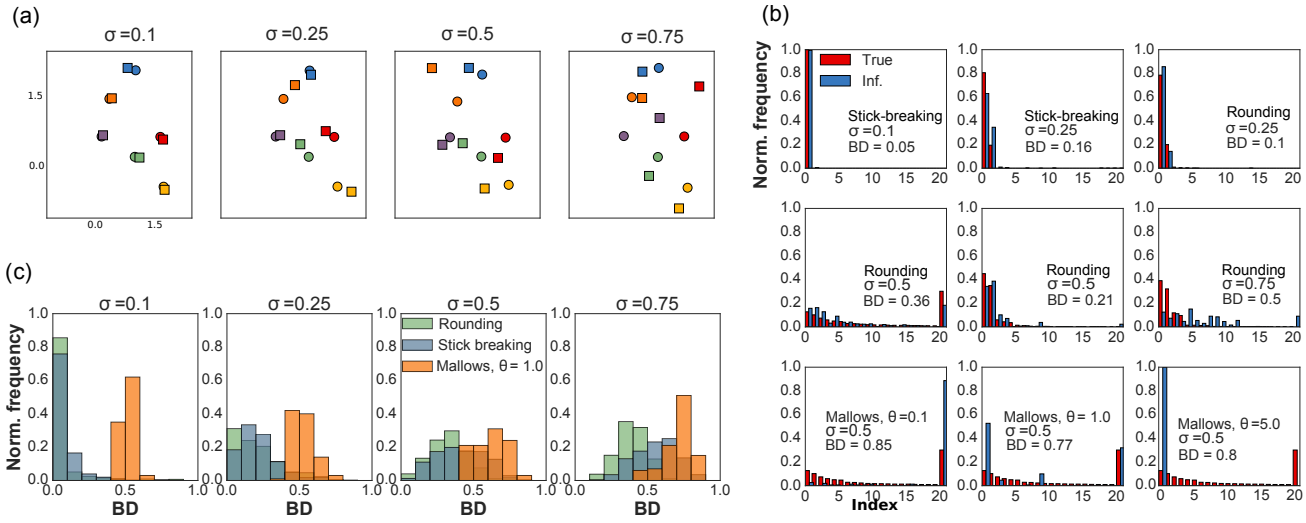

Figure 2: Synthetic matching experiment results. (a)Examples of center locations (circles) and noisy samples (squares), at different noise variances. (b) For illustration, histograms of the true and inferred posterior distribution of identities along the corresponding BD, for selected cases. Histogram indexes are sorted from the highest to lowest actual posterior probability. Only the 20 most likely configurations are shown, and the 21st bar collapses the mass of all remaining configurations. (c) Population results (histograms) across 200 experiment repetitions of each parameter configuration.

This can have some pathological effects on the resulting distribution.

- Neither admits simple pmf on permutations (contrast with Gumbel-softax).

3.3 Extending relaxation-based variational inference to permutations

With the above ingredients we can now conceive a variational inference routine for permutations, based on the analogy that the one hot vectors are to the probability simplex as permutation matrices are to the Birkhoff polytope, and by extending the relaxation-based framework presented in Maddison et al. [2016].

Namely, we consider a variational distribution $q(X;\theta)$ over 'relaxed' permutations, and dependent on a temperature parameter. As our constructions are reparameterizable; i.e. $X = g(\Psi)$ and $\Psi = f(\theta, \Xi)$, one has:

$$\mathbb{E}_{r(\Xi)}\left[-\log q(g(f(\theta,\Xi));\theta)\right] = \mathbb{H}(\Psi;\theta) + \mathbb{E}_{r(\Xi)}\left[\log\left|\frac{\partial}{\partial\Psi} g(f(\theta,\Xi))\right|\right]. \quad (2)$$

The entropy term in the r.h.s. can be computed explicitly, while the expectation can be approximated using Monte-Carlo samples. Also, in either relaxation equation 2 is differentiable w.r.t. θ

Regarding the prior $p(X;\theta)$, we use the product of N mixtures of two gaussians, around $[0,1]$. This prior weights high at points close to the vertices of the hypercube $[0,1]^N$; in particular, at permutation matrices. We benefit from this simple prior to penalize points that might be rather close to the center of the Birkhoff polytope than to permutations.

We refer the reader to the supplement for proofs of the claims stated here.

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, if ever, do our proposed continuous relaxations offer advantages over alternative Bayesian permutation inference algorithms?

Before addressing those questions we start by comparing how the categorical counterparts [1] of our proposed distributions over permutations perform on a simple VAE task. Results of this task may shed light on the usefulness of our proposed relaxations.

[1] That is, simple stick breaking and rounding in the probability simplex.

Table 1: Summary of results in VAE

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

4.1 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], finding similar (although slightly worse) results (Table 1). This difference may be intepreted as the price to be paid in order to enable an extension of a relaxed distribution over categories, to permutations. In the supplement more results on this task are available.

4.2 Synthetic matching experiments

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 is a distance between permutations [2] and θ controls the spread around ϕ_0. This is perhaps the most popular exponential family model for permutations; however, it is too simple and might fail to capture complex features of distributions.

Using the Battacharya distance (BD) we measured the discrepancy between true posterior and an empirical estimate of the inferred posteriors: in our relaxations, by sampling from $q(X;\theta)$ and 'rounding' to the nearest permutation using the Hungarian algorithm. Likewise, for the Mallows distribution, we set ϕ_0 to the MAP estimate, also through the Hungarian algorithm, and sampled using MCMC.

[2] Here, $d(\phi, \phi_0) = \sum_{i=1}^{N} |\phi(i) - \pi_0(i)|$

Table 2: 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 found our method outperforms the simple Mallows distribution, suggesting it might reasonably approximate non-trivial distributions over permutations. Fig 5 illustrates our findings by showing experiment configurations (a), examples of inferred posteriors (b) and histogram of BD's histograms (c). These histograms are complemented by Table 2.

5 Inferring neuron identities in *C. Elegans*

We conclude by showing an application of our method to the problem of inference of identity in a dynamical system. This example is motivated by the study of the neural dynamics in the *Caenorhabditis elegans* (C.elegans) Kato et al. [2015], a nematode (worm) of particular interest for neuroscience, as its neural network is stereotypical from animal to animal. Recent efforts have focused on establishing a self-consistent, accurate and complete neural wiring diagram from anatomical data [Varshney et al., 2011]. This diagram — the connectome — is utimately represented as a graph whose nodes are neurons (there are 278 somatic neurons for the hermaphrodyte C.elegans) and whose edges are synapses. Fig 3a shows the corresponding adjacency matrix, that we refer to as $\mathcal{C}$.

The C.elegans, then, is particularly suited from investigating how patterns of neural activity gives rise to behaviour, a question that has been recently rigorously addressed Kato et al. [2015]. However, there, intensive manual data curation was needed in order to match neural recordings from calcium imaging techniques to actual neurons. This manual analysis was based on the study of joint patterns of neural activity, and the comparison of observed linear position of recorded neurons to a reference worm. In some cases, identity could not be exactly resolved, and only putative candidates were

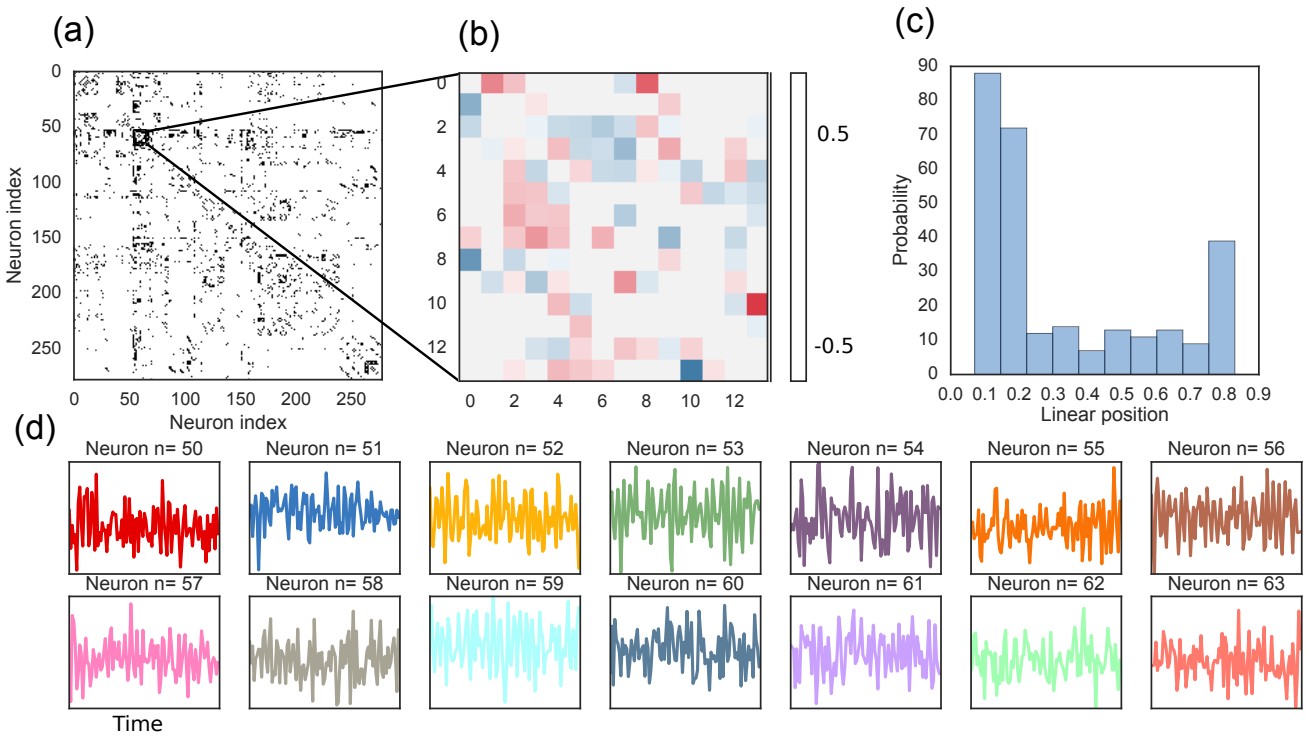

Figure 3: Problem setup. (a) Hermaphrodite C.elegans reference connectome (from Varshney et al. [2011], Lints et al. [2005]) consisting of 278 somatic neurons, merging two distinct types of synapses: chemical and electrical (gap junctions). (b) Example of matrix W consistent with the connectome information (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). Examples of the dynamical system sampled from matrix W

inferred. Unfortunately, besides this lack of certainty, this manual method does not scale if one requires to do inference in real time, or perhaps in experimental protocols that includes neural stimulation (e.g., using optogenetics Grosenick et al. [2015]).

This difficulty offers fertile ground for the development of new methods. Recently, promising approaches Aoki et al. [2017] have illustrated the plausibility of using the Brainbow technology Livet et al. [2007] for such purposes, by genetically engineering worms to express fluorescent proteins. Then, neural identification is greatly facilitated in combination with standard microscopy techniques.

We prototype an alternative solution that bypasses the need for such sophisticated genetic engineering. Our method, in essence, embodies the criteria of manual data curation into an algorithm: the assumption is that neural identity could be resolved if enough information were available from the connectome, some covariates (e.g. position) and neural dynamics. Moreover, given the neural system changes little from worm to worm, one should be able to combine recording from many individuals to resolves identity in hard cases, based on a hierarchical bayesian model.

5.1 Probabilistic Model

We consider $n = 1, \dots, M$ linear (for simplicity) dynamical systems recorded during $t = 1, \dots, T$ time-steps $Y_t^m = P_m W P_m^\top Y_{t-1} + \varepsilon_t$ (Fig 3d). Each of the Y^m is a $N = 278$ dimensional vector representing the recorded activity of the entire nervous system. These recordings are a permutation (represented by P_n) of the dynamics in a canonical order. Entries of W[3] are chosen consistently with the connectome: i.e., $W_{i,j} = 0$ if $\mathcal{C}_{i,j} = 0$. The remaining non-zero entries are then independently sampled from a normal distribution, and scaled by a factor of the spectral radius to ensure stability (see Fig 3b for an example of W, and see supplement for further details).

We perform variational inference on this model for the joint estimation of the posterior probability of P_m and W given Y_m [4]. For W we use a gaussian prior $p(W) \sim \mathcal{N}(0, I)$. Also, for P_m we consider (at training)

[3] Alternatively, one could have chosen a hierarchical model of $W_m \sim p(W)$, a direction that we avoided here for the sake of simplicity.

[4] ε is assumed known for simplicity, but could otherwise be included in the posterior, or be directly estimated from data

a relaxation based on the rounding approximation, and choose the prior defined in equation 20.

The true posterior $p(W, P_m|Y) \propto p(Y|W, P_m) \times p(W)\prod_{m=1}^{M} p(P_m)$ is then approximated by a variational family q of the form $q(W, P_m) \equiv q(W)\prod_{m}^{M} q(P_m)$, where $q(W)$ is also gaussian and $q(P_m)$ has the distribution described in 3.2.

Finally, we use neural position along the worm's body to constrain the number of possible neural identities for a given neuron: specifically, relative positions of each neuron have been documented as numbers between zero and one White et al. [1986], Lints et al. [2005] , under the abstraction that a worm can be represented as one-dimensional object (Fig 3c). Then, given this stablished data, the estimated position of all (or some) neurons, and a tolerance ν, we can conceive a binary *confusion* matrix D^m so that $D^m_{i,j} = 1$ if (observed) neuron i is close enough to (canonical) neuron j; i.e., if their distance is smaller than ν. We then enforce that constrain during inference, by ensuring that $P_{m_{i,j}} = 0$ if $D^m_{i,j} = 0$. This can be easily done by multiplying by zero such entries in the parameter matrix $\widetilde{\Theta}$ described in 3.2. Besides ease in inference, this modeling choice greatly reduces the number of effective parameters of the model, promoting scalability. Also, we allow for a certain number of neural identities to be known beforehand, easily encoded in D^m as well.

5.2 Results

We compared against three methods: i) naive variational inference, where we don't enforce the constraint that P is a permutation but allow many neurons to be mapped to the same one, ii) MCMC, where one alternates between sampling from the conditionals of W (gaussian) and P_m, from which one can sample by proposing local swipes, as described in Diaconis [2009], and iii) MAP estimator, which can be understood as a 'hard' version of ii); instead of iteratively sampling, we alternate between the MAP estimate of W (a ridge regression-like expression) and the MAP of the P_m's. For the P_m's we notice the objective is a quadratic assignment problem (QAP) in P_m, that is, it can be expressed as $Trace(APBP^\top)$ for some matrices A, B. We used the QAP solver proposed in Vogelstein et al. [2015].

Results show that in our data our method outperforms each of the three baselines. This is illustrated in Fig 4: Fig 4a depicts convergence to a better solutions, for a certain parameter configuration. More conclusively, Fig ẽffig:elegantresultsb shows a clear dominance in our method when varying the number of neurons. Likewise, Fig ẽffig:elegantresultsc depicts a similar finding when varying the size of the network. Here, variational inference and MCMC perform equally well in a regime where there is enough certainty about neural identity (squares), but when location information is more imprecise variational inference does better. This suggests our method might be particularly useful to profit from this kind of side information.

6 Discussion

Our results provide evidence that permutation variational inference might provide a helpful tool for the inference of neural identity, as it allows to properly represent shared information across animals, and different degrees of certainty based on covariates. In order to apply it to real data it is necessary to consider more realistic models of neural dynamics, which are nonlinear but might be well characterized, for example, by a set of atomic low-dimensional linear dynamical systems, each of one corresponding to a certain behavioral state Kato et al. [2015]. The methodology developed in Linderman et al. [2016] seems particularly suitable to harness that increased level of complexity.

References

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

W. Aoki, H. Matsukura, Y. Yamauchi, H. Yokoyama, K. Hasegawa, R. Shinya, and M. Ueda. Cellomics approach for high-throughput functional annotation of caenorhabditis elegans neural network. *bioRxiv*, 2017. doi: 10.1101/182923.

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. Grosenick, J. H. Marshel, and K. Deisseroth. Closed-loop and activity-guided optogenetic control. *Neuron*, 86(1):106–139, 2015.

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

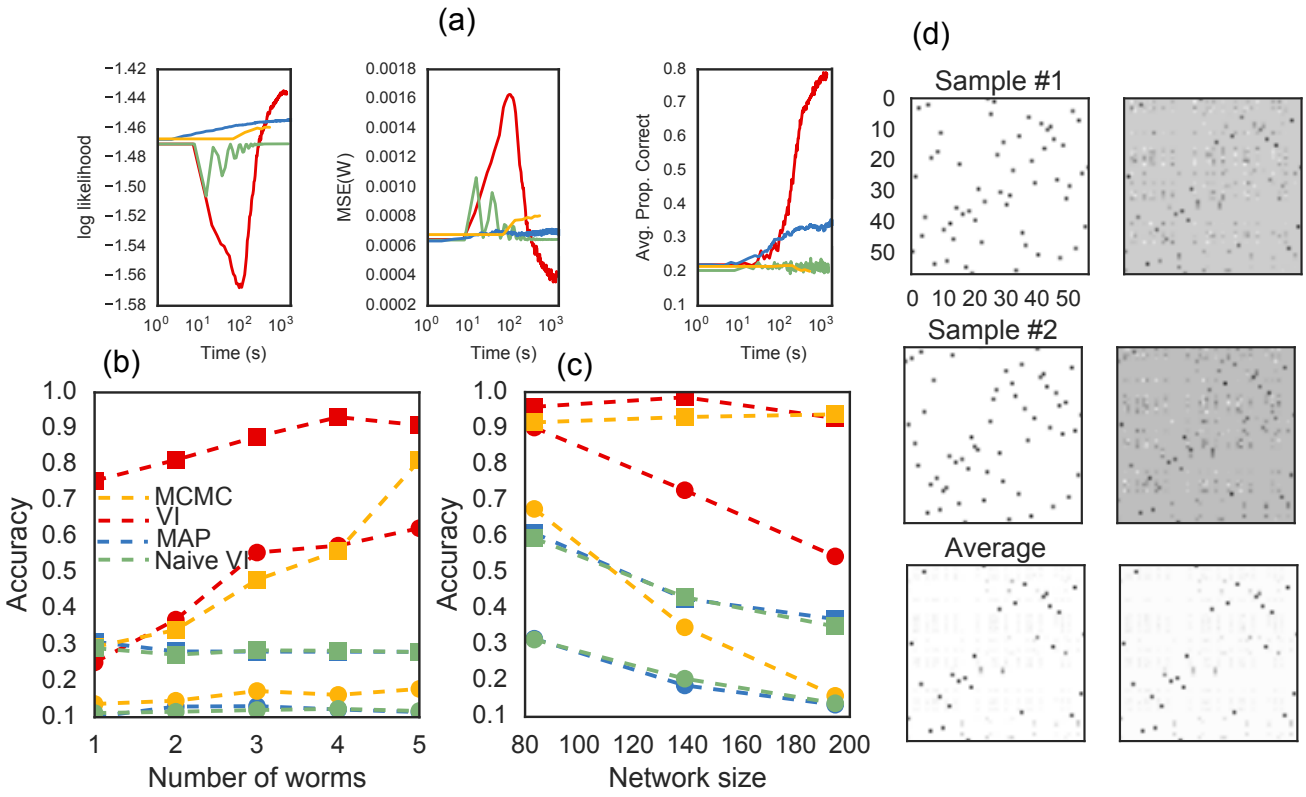

Figure 4: Results on the C.elegans inference example. (a) An example of convergence of the algorithm, and the baselines. (b) Accuracy on identity inference as a function of number of worms, for two values of ν ($\nu =$ for circles and $\nu =$ for squares). (c) Same as in (b), but using sub-networks of different size and $M = 5$ worms. (d) Two samples of permutation matrices (left) and their noisy, non-rounded version (right) during the execution of the algorithm. The average of many samples is also shown, and existence of grey spots indicate that the sampling procedure is indeed non-deterministic.

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*, volume 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 Informa-*

tion 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. W. Linderman, A. C. Miller, R. P. Adams, D. M. Blei, L. Paninski, and M. J. Johnson. Recurrent switching linear dynamical systems. *arXiv preprint arXiv:1610.08466*, 2016.

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

J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56, 2007.

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.

P. A. Mitnik. New properties of the kumaraswamy distribution. *Communications in Statistics - Theory and Methods*, 42(5):741–755, 2013. doi: 10.1080/03610926.2011.581782.

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

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. 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.

Supplement

MNIST reconstructions

Limit analysis for Stick-breaking

Here we state and prove that for all the stick-breaking based distributions in the simplex we consider here; based on the Logistic-gaussian, Kumaraswamy, and Beta distributions, we can arrive to any point in the interior of the simple or any categorical distribution as limiting cases (in τ). First, we need some lemmas.

Lemma 1. The following statements are true:

1. the degenerate case where z_k is deterministic leads to $\pi \sim \delta(\tilde{\pi})$ (i.e, single atom in the point $\tilde{\pi}$). Also, if z_k can be any in $(0, 1)$ then any deterministic π in the interior of the simplex can be realized.
2. the degenerate case where z_k are Bernoulli with parameter $p_k(\theta) \in (0, 1)$ leads to π having an atomic distribution with atoms in the vertices of Δ^{k-1}; i.e, π is categorical. We have the following expression for the probabilities of the atoms $\pi_k = 1$ (one hot vectors):

$$P(\pi_k = 1) = \prod_{i=1}^{k-1}(1 - p_i(\theta))p_k(\theta) \text{ for } k = 2, \dots, K-1, \quad P(\pi_K = 1) = \prod_{i=1}^{K-1}(1 - p_i(\theta)). \tag{3}$$

   Moreover, if for each index k any parameter of the Bernoulli variable z_k can be realized through appropriate choice of θ, then any categorical distribution can be realized.

Proof: (a) both claims are obvious and come from the invertibility of the function $\mathcal{SB} \circ h(\cdot)$. (b) the formulae for $P(\pi_k = 1)$ comes from expressing the event $\pi_k = 1$ equivalently as $\pi_k = 1, \pi_i = 0, i < k$ and then, conditioning backwards successively. The second statement comes from the following expression, which easily follows from (3):

$$p_k(\theta) = \frac{P(\pi_k = 1)}{P(\pi_{k-1} = 1)} \frac{p_{k-1}(\theta)}{1 - p_{k-1}(\theta)}, \quad k = 1, \dots, K-1.$$

The recursive nature of the above equation gives a recipe to iteratively determine the required $p_k(\theta)$, given $P(\pi_k = 1), P(\pi_{k-1} = 1)$ and the already computed $p_{k-1}(\theta)$.

Now we can state our results:

Lemma 2. If $z = \sigma(\psi), \psi \sim \mathcal{N}(\mu, \eta^2)$, then

1. the limit $\eta \to 0$ and μ fixed leads to the deterministic $z = \sigma(\mu)$.
2. the limit $\mu \to \infty, \eta^2 = \mu/K$ with K constant leads to $z \sim$ Bernoulli$(\Phi(K))$, with $\Phi(\cdot)$ denoting the standard normal cdf.

In both cases the convergence is in distribution

Proof. The first convergence is obvious. To see the second, let's index μ_n and study the cdf F of z_n on the interval (0,1) (it evaluates zero below zero and one above one).

$$\begin{aligned} F_{z_n}(x) &= P(\sigma(\psi_n) < x) && (4)\\ &= P(\psi_n < \sigma^{-1}(x)) && (5)\\ &= P(\mu_n + \mu_n/K\xi < \sigma^{-1}(x)), && (6)\\ &= P(\xi < \sigma^{-1}(x)K/\mu_n - K) && (7)\\ &= \Phi(\sigma^{-1}(x)K/\mu_n - K) && (8) \end{aligned}$$

Therefore, by continuity of Φ we obtain $F_{\Psi_n}(x) \to \Phi(-K)$ for all points $x \in (0, 1)$. On the other hand, the cdf of a bernoulli random F variable is given by a step function that abruptly changes at zero, from zero to $1 - p$, and at one, from $1 - p$ to 1. As convergence occurs at all continuity points (the interval $(0, 1)$), we conclude (recall, $1 - p = \Phi(-K) \to \Phi(K) = p$). Notice that the above representation only allows to converge to $p > 0.5$, as K has to be positive. This can be fixed by choosing sequences with negative μ instead.

Lemma 3. If $z = Kw(a, b)$:

1. in the limit $a, b \to$ we converge to deterministc p, provided that $p = bB\left(1 + \frac{1}{a}, b\right)$ along the limiting sequence.
2. In the limit $a, b \to 0$ we obtain convergence to a Bernoulli random variable with parameter p, provided the same condition involving p, a, b holds.

In both cases convergence is in probability. *Proof*: A proof can be found in Mitnik [2013]

Lemma 4. If $z = Beta(a, b)$:

1. in the limit $a, b \to \infty$ we converge to deterministc p, provided that $p = bB\left(1 + \frac{1}{a}, b\right)$ along the limiting sequence.
2. In the limit $a, b \to 0$ we obtain convergence to a Bernoulli random variable with parameter p, provided the same condition involving p, a, b holds.

In both cases convergence is in distribution.

Proposition. In all the discussed cases of reparameterizations of the simplex via stick-breaking, arbitrary categorical distributions can be obtained in the low-temperature limit. Also, in the high-temperature convergence is to certain point(s) in the interior of the simplex.

Proof: Consider each distribution separately

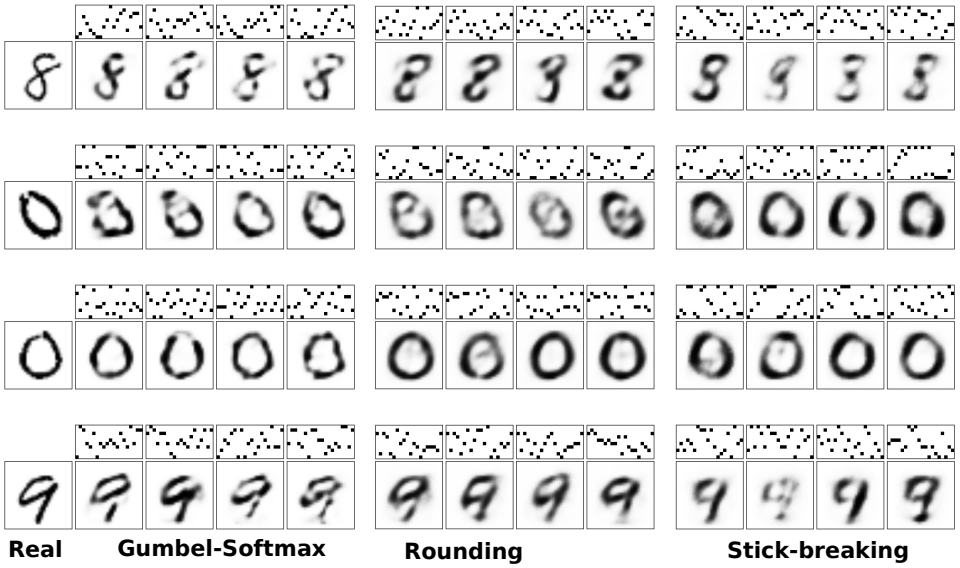

Figure 5: Examples of true and reconstructed digits from their corresponding random codes using with $N = 20$ categorical variables with $K = 10$ possible values.

1. For the logistic-normal re-parameterization $z_k = \sigma\left(\frac{\mu_k+\eta_k\xi}{\tau}\right)$, in the low temperature case use Lemma 2 (b) by the always available representation $K = \frac{\mu}{\eta^2}$and conclude by Lemma 1(b). In the high temperature case convergence is to the point $\pi = \mathcal{SB}(0.5, 0.5, \ldots, 0.5)$.
2. For Kumaraswamy $z_k = \mathcal{K}(a_k, b_k)$ the argument is similar, but here the temperature can only be defined implicitly through sequences of parameters (a_k, b_k) converging to either ∞ or 0 along a sequence with fixed $p_k = b_k B\left(1 + \frac{1}{a_k}, b_k\right)$. Then in the low temperature case we conclude by Lemma 3(b) and Lemma 1(b). In the hig-temperature case we converge to the point $SB(p_1, \ldots p_{k-1}$
3. For the Beta $z_k \sim Beta(\frac{a_k}{\tau}, \frac{b_k}{\tau})$ low-temperature leads to convergence to z_k Bernoulli with parameter $a_k/(a_k + b_k)$ and we conclude from Lemma 4(b) and Lemma 1(b). For high temperatures, convergence is to the point $\mathcal{SB}(a_k/(a_k + b_k), \ldots, a_{k-1}/(a_{k-1} + b_{k-1}))$.

Deriving the approximation for the ELBO

Here we show that

$$\mathbb{E}_{p(\xi)}\left[-\log q(F(g(\theta,\xi));\theta)\right] = \mathbb{H}(\psi;\theta) + E_{p(\xi)}\left[\log|DF(g(\theta,\xi))|\right].$$

Indeed, first, by the 'Law of the Unconscious Statistician' we have:

$$\mathbb{E}_{p(\xi)}\left[-\log q(F(g(\theta,\xi));\theta)\right] = \mathbb{E}_{p(\psi;\theta)}\left[-\log q(F(\psi);\theta)\right].$$

Now, by the change of variable theorem and derivative and determinant inversion rules, we obtain:

$$q(F(\psi);\theta) = p(F^{-1}(\pi);\theta)|DF^{-1}(\pi)| \tag{9}$$
$$= p(\psi;\theta)|DF(\psi)|^{-1}. \tag{10}$$

To conclude we use once more the Law of the Unconscious Statistician:

$$\mathbb{E}_{p(\xi)}\left[-\log q(F(g(\theta,\xi));\theta)\right] = \mathbb{E}_{p(\psi;\theta)}\left[-\log p(\psi;\theta)\right] + \mathbb{E}_{p(\psi;\theta)}\left[\log|\right. \tag{11}$$
$$= Entropy(\psi;\theta) + E_{p(\xi)}\left[\log|DF(g(\theta, \tag{12}$$

Notice R^Z is a piecewise constant function, as maps each $V_m^{\mathcal{P}}$ to p_m

Notice that these bounds only depend on values of Π that have already been computed; i.e., those that are above or to the left of the (i, j)-th entry. Thus, the transformation from Ψ to Π is feed-forward according to this ordering. Consequently, the Jacobian of the inverse transformation, $\mathrm{d}\Psi/\mathrm{d}\Pi$, is lower triangular, and

its determinant is the product of its diagonal,

$$\left|\frac{\mathrm{d}\Psi}{\mathrm{d}\Pi}\right| = \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\frac{\partial\psi_{ij}}{\partial\pi_{ij}} \tag{13}$$

$$= \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\frac{\partial}{\partial\pi_{ij}}\sigma^{-1}\left(\frac{\pi_{ij}-\ell_{ij}}{u_{ij}-\ell_{ij}}\right) \tag{14}$$

$$= \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\left(\frac{1}{u_{ij}-\ell_{ij}}\right)\left(\frac{u_{ij}-\ell_{ij}}{\pi_{ij}-\ell_{ij}}\right)\left(\frac{u_{ij}-\ell_{ij}}{u_{ij}-\pi_{ij}}\right) \tag{15}$$

$$= \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\frac{u_{ij}-\ell_{ij}}{(\pi_{ij}-\ell_{ij})(u_{ij}-\pi_{ij})} \tag{16}$$

With these two ingredients, we can write the density of Π,

$$\mathrm{vec}(\Psi) \sim \mathcal{N}(\mu, \mathrm{diag}(\eta^2)) \tag{17}$$

$$\Pi = f(\Psi) \tag{18}$$

$$\implies p(\Pi \,|\, \mu, \mathrm{diag}(\eta^2)) = \left|\frac{\mathrm{d}\Psi}{\mathrm{d}\Pi}\right| \mathcal{N}(f^{-1}(\Pi) \,|\, \mu, \mathrm{diag}(\eta^2)) \tag{19}$$

Variational inference for Permutation details

The two constructions presented here are re-parameterizable, In both constructions presented here the term can be approximated by Monte-Carlo samples, explicitly and is differentiable with respect to θ $X = g(\Psi)$ and $\Psi = f(\theta, \xi)$. Moreover, both g and f are differentiable and invertible functions. Therefore, by the change of variable theorem and the law of the unconscious statistician:

Discuss the relation to Gumbel-softmax and the analogy between relaxing one hot vectors to simplex and relaxing permutations to Birkhoff.

To perform variational inference in this relaxed setup we proceed as in Maddison et al. [2016], by replacing the original discrete objective by a relaxed one, which is sensible as long as all computations can be extended to the continuum.

Continuous prior distributions. For Stick-breaking Continuous relaxations requires re-thinking of 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 latent variables. For the categorical experiments, we use a mixture of Gaussians around each vertex, $p(x) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}(x \,|\, e_k, \eta^2)$. For permutations, we use a mixture of Gaussians for each dimension,

ec 4.1 can be cut

$$p(X) = \prod_{m=1}^{N}\prod_{n=1}^{N}\frac{1}{2}\left(\mathcal{N}(x_{mn} \,|\, 0, \eta^2) + \mathcal{N}(x_{mn} \,|\, 1, \eta^2\right). \tag{20}$$

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

Estimating the ELBO. Notice in all the relaxations discussed here, $x = g(\psi)$ and $\psi = f(\theta, \xi)$. Moreover, both g and f are differentiable and invertible functions. Therefore, by the change of variable theorem and the law of the unconscious statistician:

$$\mathbb{E}_{r(\xi)}\left[-\log q(g(f(\theta,\xi));\theta)\right] = \mathbb{H}(\psi;\theta) + \mathbb{E}_{r(\xi)}\left[\log\left|\frac{\partial}{\partial\psi}g(f(\theta,\xi))\right|\right] \tag{21}$$

where $\mathbb{H}$ is the entropy and the term inside of the expectation is the (log Jacobian of g evaluated at $\psi = f(\theta, \xi)$. Then, if this Jacobian and the entropy of ψ are available we can consider an unbiased, Monte Carlo estimator for the ELBO. For example, in the rounding transformation, g is piecewise linear [5] and $\log|\frac{\partial}{\partial\psi}g(f(\theta,\xi))| = N\log\tau$. Also, if ψ is Gaussian its entropy is given by $N\log(\eta^2 2\pi e)/2$.

Given the density and a differentiable mapping we can perform variational inference with stochastic optimization of the ELBO. We define a distribution over doubly stochastic matrices as a reparameterization of a multivariate Gaussian distribution over Ψ. We can estimate gradients via the reparameterization trick.

It is important to note that the transformation is only piecewise continuous: the function is not differentiable at the points where the bounds change; for example, when changing Ψ causes the active upper bound to switch from the row to the column constraint or vice versa. I think we can argue that these discontinuities will not have a severe effect on our stochastic gradient algorithm.

[5] The set of discontinuities has Lebesgue measure zero so we can still apply the change of variables theorem.