Instructions for paper submissions to AISTATS 2018

Author 1
Institution 1

Author 2
Institution 2

Author 3
Institution 3

Abstract

How can we efficiently perform posterior inference over the space of permutations when there are $N!$ permutations of a set of N elements? Clearly, estimating a complete probability mass function over this space quicky becomes intractable as N grows. Our goal is to derive a tractable algorithm for performing approximate inference over this challenging discrete space. To that end, we consider extensions of the recently proposed Gumbel-softmax method, which leverages continuous relaxations to perform discrete variational inference with reparameterization gradients. While the Gumbel-softmax method is not immediately applicable to permutation inference, we show that two alternative reparameterizations are both comparable to Gumbel-softmax on tractable discrete problems and easily extensible to permutation inference. Specifically, we develop continuous relaxations of permutation matrices to matrices that are either exactly or nearly doubly stochastic, i.e. to points either in or near the Birkhoff polytope. We then derive 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. This relaxation admits variational inference via stochastic gradient ascent over the distributions on doubly stochastic matrices (and in the zero-temperature limit, on permutation matrices) using Monte Carlo estimates of the reparameterized gradient.

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

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 but 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 Markov chain Monte Carlo methods [Diaconis, 1988], 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. Given recent advances in scaling variational Bayesian inference, largely driven by efficient Monte Carlo estimators of gradients of the variational lower bound [Kingma and Welling, 2014, Rezende et al., 2014], we revisit the problem of permutation inference from a variational perspective.

Continuous relaxations underlie many approximate algorithms for discrete optimization and inference. After relaxation, we can capitalize on local gradients and curvature information. Indeed, this is the motivation for the recently proposed Gumbel-softmax method for discrete

variational inference [Jang et al., 2016, Maddison et al., 2016]. It is based on the following observation: categorical distributions may be viewed as atomic densities on the vertices of the simplex; by relaxing this to a continuous density on the interior of the simplex we can approximate the discrete inference problem with a continuous one and thereby capitalize on reparameterization gradients [Kingma and Welling, 2014, Rezende et al., 2014] to optimize a variational lower bound on the marginal likelihood. Critically, the Gumbel-softmax method has a temperature parameter that tunes the degree to which the continuous density concentrates around the vertices, and recovers truly discrete inference in the zero-temperature limit.

Just as one-hot vectors (discrete random variables) are the vertices of the simplex, permutation matrices are the vertices of the Birkhoff polytope, i.e. the set of doubly stochastic matrices. Thus, we seek temperature-controlled relaxations of atomic densities on permutation matrices to continuous densities on the interior of the Birkhoff polytope. Unfortunately, the dual constraints of row- and column-normalization required of doubly stochastic matrices present difficulties that are not faced in the categorical setting. However, we derive a variety of alternative continuous relaxations for the simplex and show that: (i) these relaxations achieve comparable performance to the Gumbel-softmax on tractable discrete inference tasks; and (ii) they naturally extend to relaxations of permutation inference problems.

The remainder of this paper is structured as follows: Section 2 discusses related work on Bayesian permutation inference and the continuous relaxations for discrete inference, including the Gumbel-softmax method. Section 3 introduces alternative relaxations for discrete variational inference, and Section 4 presents our primary contribution: a set of relaxations for permutation matrices. Section 5 presents a variety of experiments that illustrate the benefits of the proposed variational approach.

2 Related Work

Bayesian permutation inference. As mentioned above, a number of previous works have considered approximate methods of posterior inference over the space of permutations. When a point estimate will suffice, convex relations are commonly employed [Fogel et al., 2013, Lim and Wright, 2014]. Given noisy measurements of a sum of a small number of permutation matrices, we can recover the underlying coefficients via a convex optimization penalized by the norm induced by the Birkhoff polytope [Chandrasekaran et al., 2012]. For some ranking problems, we can rewrite the objective function in terms of the expected assignment probabilities under a distribution over permutation matrices, which in turn are points in the Birkhoff polytope. Adams and Zemel [2011] leveraged this property to develop stochastic gradient descent algorithms that minimize these objective functions, using Sinkhorn propagation [Knight, 2008] as a differentiable map from the positive orthant to the Birkhoff polytope. We will use the same approach in one of our proposed methods.

When a point estimate is insufficient, it may be possible to turn efficient algorithms for optimizing linear cost functions over the set of permutation matrices into efficient sampling algorithms using Perturb-and-MAP [Li et al., 2013]. For simple problems, Markov chain Monte Carlo (MCMC) algorithms can perform quite well by simply using Metropolis-Hastings proposals to swap assignments at random [Diaconis, 1988]. Such methods ultimately rely on a random walk to explore the high dimensional space of permutations. Harrison and Miller [2013] developed an importance sampling algorithm that fills in count matrices one row at a time, leveraging column- and row-sum constraints, and showed promising results for matrices with $O(100)$ rows and columns. 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. 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.

Variational inference and the reparameterization trick. 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, 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),

$$\mathscr{L}(\theta) \triangleq \mathbb{E}_q\left[\log p(x,y) - \log q(x;\theta)\right]. \tag{1}$$

Perhaps the simplest method of optimizing the ELBO is stochastic gradient ascent. However, computing $\nabla_\theta \mathscr{L}(\theta)$ 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,

$$x \sim q(x;\theta) \quad \Longleftrightarrow \quad \xi \sim r(\xi), \quad x = g(\theta,\xi), \tag{2}$$

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

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{r(\xi)}[\nabla_\theta \log p(g(\theta,\xi)\,|\,y) - \nabla_\theta \log q(g(\theta,\xi);\theta)]. \tag{3}$$

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.

Continuous relaxations for discrete variational inference. Recently, there have been a number of proposals for extending the reparameterization trick to high dimensional discrete problems[1] by relaxing them to analogous continuous problems [Maddison et al., 2016, Jang et al., 2016, Kusner and Hernández-Lobato, 2016]. These approaches are based on the following observation: one-hot vectors $x \in \{0,1\}^N$ can alternatively be viewed as vertices of the simplex Δ_N; likewise, discrete probability mass functions $q(x;\theta)$ can be seen as atomic densities on the vertices of the simplex. This motivates a natural relaxation: let x assume any value in the simplex, not just the vertices, and let $q(x;\theta)$ be a density on the interior of the simplex. One way to define such a density is via the following reparameterization,

$$\xi \sim r(\xi), \tag{4}$$

$$g(\theta,\xi) = \left[\frac{\theta_1+\xi_1}{\sum_{n=1}^N \theta_n+\xi_n}, \ldots, \frac{\theta_N+\xi_N}{\sum_{n=1}^N \theta_n+\xi_n}\right] \triangleq \text{softmax}(\theta+\xi) \tag{5}$$

where $\xi, \theta \in \mathbb{R}^N$.

In the aforementioned papers, the noise is assumed to be a vector of independent Gumbel random variables, i.e. $p(\xi) = \prod_{n=1}^N \text{Gumbel}(\xi_n\,|\,0,1)$. This choice leads to a nicely interpretable model: adding Gumbel noise and taking the *argmax* of $\theta+\xi$ yields an exact sample from $\pi = \text{softmax}(\theta)$, thus the *softmax* of $\theta+\xi$ is a natural relaxation. Ultimately, however, this is just a continuous relaxation of an atomic density to a continuous density.

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

3 Alternative relaxations for categorical random variables

Here we introduce two alternative ways of conceiving a relaxed version of a discrete random variable, both of which are amenable to the reparameterization trick. However, unlike the Gumbel-Softmax, these relaxations enable extensions to more complex combinatorial objects like permutations.

3.1 Stick-breaking transformations

First, let us consider an alternative reparameterization of the simplex via a stick breaking construction. We break this into two steps. First, we transform the noise and parameters to a point in the $N-1$ dimensional unit hypercube,

$$\xi \sim r(\xi), \qquad \psi = f(\theta,\xi), \tag{6}$$

where $\psi \in [0,1]^{N-1}$. Then we transform the hypercube to Δ_N via a stick-breaking transformation,

$$x_n = g_n(\psi) = \begin{cases} \psi_1 & n=1, \\ \psi_n\left(1-\sum_{m=1}^{n-1} x_m\right) & 1<n<N, \\ 1-\sum_{n=1}^{N-1} x_n & n=N. \end{cases} \tag{7}$$

The intermediate values ψ_n can seen as the fraction of the remaining "stick" of probability mass assigned to π_n. In addition to its use in Bayesian nonparametrics, this type of transformation has been used in efficient MCMC algorithms for multinomial and categorical inference [Linderman et al., 2015].

We focus on standard Gaussian noise $r(\xi) = \mathcal{N}l(0,I)$ and we take f to be a logistic transformation $\psi_n = \sigma((\mu_n + \eta_n\xi_n)/\tau)$, where $\sigma(u) = (1+e^{-u})^{-1}$ is the logistic function and τ is a *temperature* parameter. This *logistic-normal stick breaking* transformation is parameterized by $\theta = \{\mu_n, \eta_n\}_{n=1}^{N-1}$, and it enjoys following properties: i) the density of x can be expressed in closed form as a function of μ_n and η_n^2; ii) the temperature τ controls how concentrated $p(x)$ is at the vertices of the simplex; iii) with appropriate choices of parameters, in the limit $\tau \to 0$ we can recover any categorial distribution, i.e., the density becomes concentrated on atoms at the N vertices; and iv) as $\tau \to \infty$, the density concentrates on a point in the interior of the simplex determined by the parameters. For

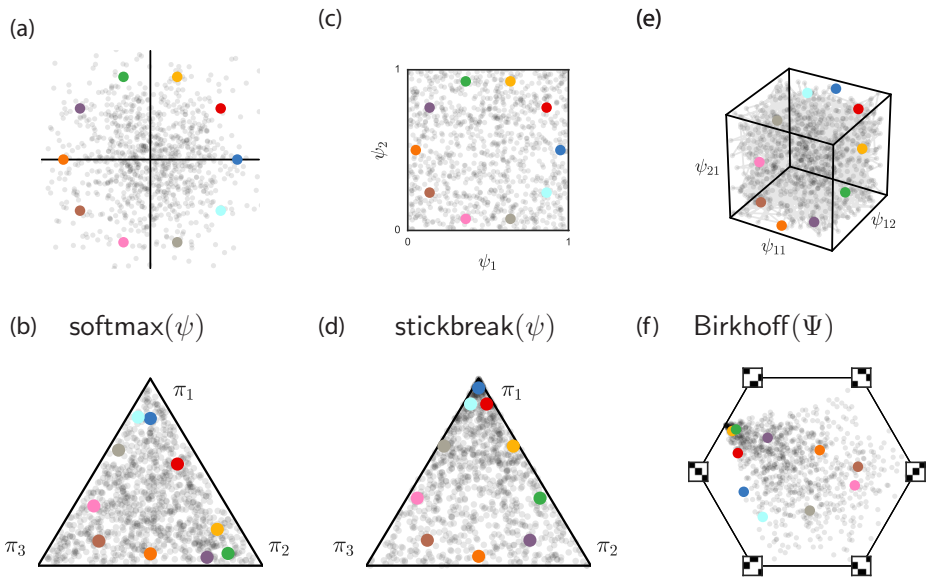

Figure 1: Reparameterizations of discrete polytopes. (a,b) The Gumbel-softmax, or "Concrete" transformation maps points $\psi \in \mathbb{R}^N$ to points $x \in \Delta_N$ by adding noise and applying the softmax. Here we show a slice for $N = 3$ with $\psi_3 = 0$. Colored points are aids to visualize 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, $\mathscr{B}_3$ 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).

all intermediate temperatures, the density is continuous on the simplex.

Note that the logistic-normal stick breaking transformation one of many available. For example, we could take r and f to be a reparameterization of the Kumaraswamy of beta distributions on the unit interval. The former is easily reparameterizable and the latter—which leads to the generalized Dirichlet distribution on the simplex—can be reparameterized following Naesseth et al. [2017]. We include proofs of points (i-iv) and details of the Kumaraswamy and beta stick breaking constructions in the appendix.

3.2 Rounding transformations

Both the Gumbel-softmax and stick-breaking relaxations consider distributions on the simplex, and while this offers an intuitive interpretation, it is not strictly required. For example, first consider a distribution on $\psi \in \mathbb{R}^N$. These points are the rounded to the nearest vertex of the simplex via the operator,

$$\text{round}(\psi) = \arg\min_{e_n} \|\psi - e_n\|. \tag{8}$$

Unfortunately this rounding operator is non-invertible and non-differentiable. Thus, we instead consider a map that pulls a point towards its rounded value, by taking a convex combination between both. Specifically, we consider the following reparameterization:

$$\xi \sim p(\xi), \tag{9}$$
$$\psi = f(\theta, \xi), \tag{10}$$
$$x = \tau\psi + (1-\tau)\cdot \text{round}(\psi). \tag{11}$$

In the zero-temperature limit we recover a discrete distribution on the vertices. For $\tau > 0$, the distribution is continuous on $\mathbb{R}^N$. If the distribution of ψ is concentrated near the simplex—e.g. if θ is a point on the simplex and ξ is small, additive Gaussian noise—the rounded points will lie close to the simplex as well. Moreover, this technique is easily generalized to more complex discrete polytopes.

4 Continuous relaxations of permutation distributions

Just as one-hot vectors are the vertices of the simplex, the Birkhoff-von Neumann theorem states that permutation matrices X are vertices of the convex hull of doubly stochastic matrices. By analogy, it is natural to relax to $X \in \mathscr{B}_N \subset [0,1]^{N\times N}$, the Birkhoff polytope of doubly

stochastic matrices defined by,

$$\mathscr{B}_N = \left\{ X : \; x_{m,n} \geq 0 \, \forall m,n \in [N]; \; \sum_{n=1}^{N} x_{m,n} = 1 \, \forall m \in [N]; \; \sum_{m=1}^{N} x_{m,n} = 1 \, \forall n \in [N] \right\}. \tag{12}$$

Due to these linear row- and column-normalization constraints, $\mathscr{B}_N$ lies within a $(N-1)^2$ dimensional subspace. Unfortunately, these constraints also present difficulties to reparameterization. Next we show how the stick-breaking and rounding reparameterizations can be extended to the Birkhoff polytope.

4.1 Stick-breaking transformations of the Birkhoff polytope

We now derive an invertible and differentiable transformation, $g : \mathbb{R}^{(N-1)\times(N-1)} \to \mathscr{B}_n$ by extending the original stick-breaking transformation with minor modifications to accomodate the additional constraints of doubly stochastic matrices. This can be used to define a density on $\mathscr{B}_N$. Let Ψ be an 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 working 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.

The first entry 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. Thus,

$$x_{1n} = \psi_{1n}\left(1 - \sum_{k=1}^{n-1} x_{1k}\right) \quad \text{for } n = 2,\ldots,N-1 \tag{13}$$

$$x_{1N} = 1 - \sum_{n=1}^{N-1} x_{1n} \tag{14}$$

So far, this is exactly as in the stick breaking construction above. However, the remaining rows must now conform to both row- and column-constraints. That is,

$$x_{mn} \leq 1 - \sum_{k=1}^{n-1} x_{mk} \qquad \text{(row sum)} \tag{15}$$

$$x_{mn} \leq 1 - \sum_{k=1}^{m-1} x_{kn} \qquad \text{(column sum)}. \tag{16}$$

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,

$$\underbrace{1 - \sum_{k=1}^{n} x_{mk}}_{\text{remaining stick}} \leq \underbrace{\sum_{j=n+1}^{N} (1 - \sum_{k=1}^{m-1} x_{kj})}_{\text{remaining upper bounds}}. \tag{17}$$

Rearranging terms, we have,

$$x_{mn} \geq 1 - \sum_{k=1}^{n-1} x_{mk} - \sum_{j=n+1}^{N} (1 - \sum_{k=1}^{m-1} x_{kj}) \tag{18}$$

$$= 1 - N + n - \sum_{k=1}^{n-1} x_{mk} + \sum_{k=1}^{m-1} \sum_{j=n+1}^{N} x_{kj} \tag{19}$$

Of course, this bound is only relevant if the right hand side is greater than zero. Taken together, x_{mn} is bounded by,

$$\ell_{mn} \leq x_{mn} \leq u_{mn} \tag{20}$$

$$\ell_{mn} \triangleq \max\left\{0, \, 1 - N + n - \sum_{k=1}^{n-1} x_{mk} + \sum_{k=1}^{m-1} \sum_{j=n+1}^{N} x_{kj}\right\} \tag{21}$$

$$u_{mn} \triangleq \min\left\{1 - \sum_{k=1}^{n-1} x_{mk}, \, 1 - \sum_{k=1}^{m-1} x_{kn}\right\}. \tag{22}$$

Thus, we define,

$$x_{mn} = \ell_{mn} + (\psi_{mn}(u_{mn} - \ell_{mn}). \tag{23}$$

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,

$$\psi_{mn} = \frac{x_{mn} - \ell_{mn}}{u_{mn} - \ell_{mn}}. \tag{24}$$

4.2 Rounding toward permutation matrices

The rounding-based relaxation immediately extends to the permutation case. Now we simulate matrices $\Psi \in \mathbb{R}^{N\times N}$ and round them to the nearest permutation matrix by solving a matching problem. This can be done in $O(N^3)$ time with the Hungarian algorithm [Kuhn, 1955, Munkres, 1957]. As before, if Ψ is close to the Birkhoff polytope, its rounded value X will be as well. Fortunately, it is easy to generate a distribution that concentrates near $\mathscr{B}_N$. We use the reparameterization $\Psi = \widetilde{\Theta} + \Xi$, where $r((\Xi)) = \mathcal{N}(0, I)$, and then we define $\widetilde{\Theta}$ to be the result of applying a fixed number of Sinkhorn propagation [Knight, 2008] steps to the unconstrained matrix $\Theta \in \mathbb{R}^{N\times N}$. Since the Sinkhorn algorithm is differentiable, we can backpropagate gradients through this procedure [c.f. Adams and Zemel, 2011]. Note that even though Sinkhorn is non-invertible, we can still evaluate the density of $q(X;\theta)$ since it applies *before* the introduction of the random noise Ξ.

5 Results

We are interested in two principal questions: (i) how sensitive are categorical relaxations to the choice of Gumbel-softmax, stick-breaking, or categorical reparameterization? (ii) 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 (iii) when, if ever, do our proposed continuous relaxations offer advantages over alternative approximate Bayesian permutation inference algorithms? We will address these questions in turn, but first we discuss some practical details of our experimental protocol.

5.1 Experimental protocol

Continuous prior distributions. 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, $p(x) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{N}(x \mid e_k, \eta^2)$. For permutations, we use a mixture of Gaussians for each dimension,

$$p(X) = \prod_{m=1}^{N}\prod_{n=1}^{N}\frac{1}{2}\left(\mathcal{N}(x_{mn} \mid 0, \eta^2) + \mathcal{N}(x_{mn} \mid 1, \eta^2\right). \tag{25}$$

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

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:

$$\mathbb{E}_{r(\xi)}[-\log q(g(f(\theta,\xi));\theta)] = \mathbb{H}(\psi;\theta) + \mathbb{E}_{r(\xi)}\left[\log\left|\frac{\partial}{\partial\psi}g(f(\theta,\xi))\right|\right], \tag{26}$$

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 [2] and $\log|\frac{\partial}{\partial\psi}g(f(\theta,\xi))| = N\log\tau$. Also, if ψ is Gaussian its entropy is given by $N\log(\eta^2 2\pi e)/2$.

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

5.2 Variational Autoencoders (VAE) with categorical latent variables

We first demonstrate that our proposed relaxations are sensible for categorical random 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 (via rounding the samples π) with $m = 1000$. We trained our models using ADAM in Tensorflow and compared against the method of Jang et al. [2016], finding similar results (Table 1). Our best results were obtained with rounding. Our results suggest our methods provide a viable alternative to the Gumbel-Softmax. Fig 2 shows a random selection of reconstructed images using the different approaches. By eye, the reconstructed images and the latent codes seem very comparable.

5.3 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.

We measured the discrepancy using the Battacharya distance (BD) between true posterior and an empirical estimate of the inferred posterior constructed by sampling from $q(X;\theta)$ and 'rounding' to the nearest permutation using the Hungarian algorithm. We found that our methods provide reasonable approximations to the true posterior, allowing us to represent more complex distributions over permutations than, e.g., simple Mallows distribution around the MAP estimate. Fig 5 shows examples of true posteriors (ranked) and their approximations, and quantifies the discrepancies by the distribution of the BD.

5.4 Hierarchical permutation inference

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 stereotyp-

Table 1: Summary of results in VAE

	Gumbel-Softmax	Concrete	Rounding	Stick-breaking
- $\log p(x)$	106.7	111.5	121.1	119.8

Table 2: Battacharya distances in the synthetic matching experiment

	Rounding	Stickbreaking	Mallows				
			$\theta = 0.1$	$\theta = 1$	$\theta = 2$	$\theta = 5$	$\theta = 10$
$\sigma = 0.1$	.06	.09	.93	.51	.23	.08	.08
$\sigma = 0.25$	.21	.23	.92	.53	.33	.27	.27
$\sigma = 0.5$	.32	.41	.89	.61	.53	.54	.54
$\sigma = 0.75$	.38	.55	.85	.71	.69	.72	.72

ical 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 ultimately 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 $\mathscr{C}$.

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 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.

5.4.1 Modeling details

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 $\mathscr{C}_{i,j} = 0$. 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 appendix 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) a relaxation based on the rounding approximation, and choose the prior defined in equation 25.

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 4.2.

Finally, we use neural position along the worm's body to constrain the number of possible neural identities for

[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

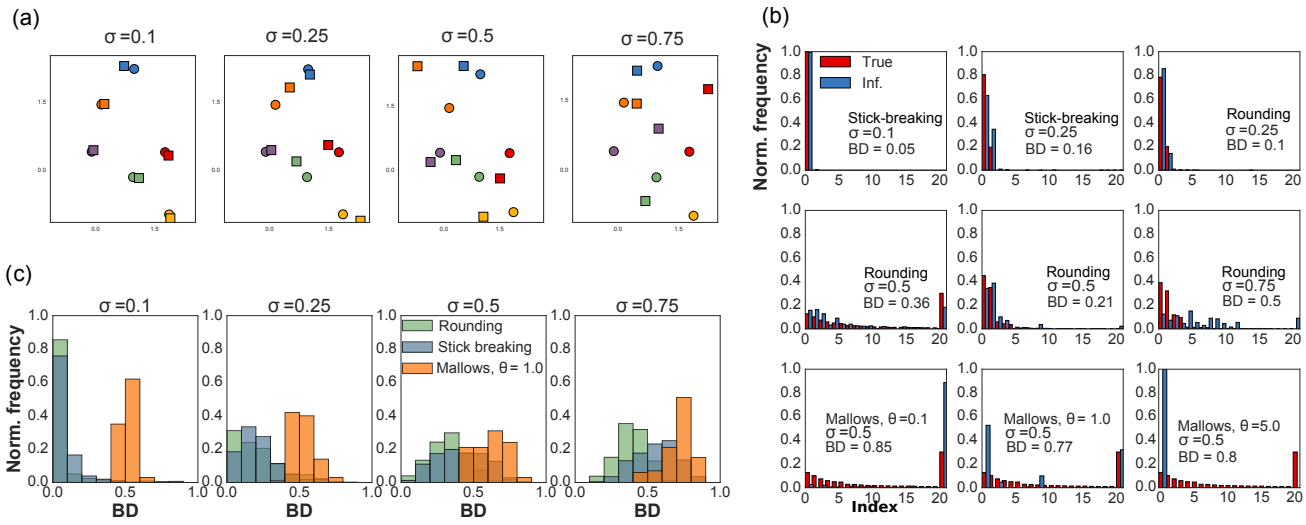

Figure 2: Matching experiment results. (a)Examples of center locations (circles) and noisy samples (squares), at different noise variances. (b) For illustration, histograms of the true and inferred posterior distribution of identities along the corresponding Bhattacharya distance (BD), for selected cases. 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. .

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 4.2. 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.

5.4.2 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 řeffig:elegantresultsb shows a clear dominance in our method when varying the number of neurons. Likewise, Fig řeffig: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.

5.4.3 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 non-linear 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.

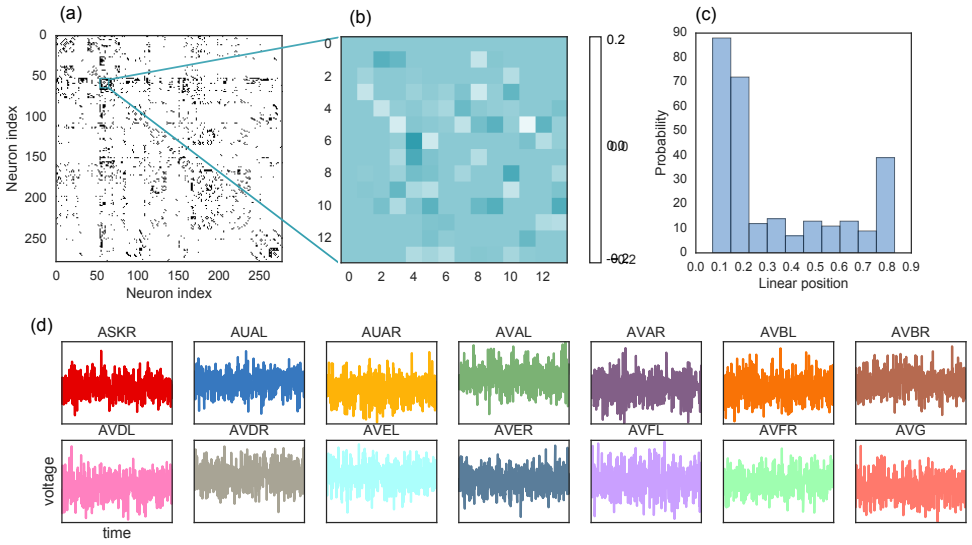

Figure 3: Problem setup. (a) Hermaphrodite C.elegans reference connectome (from Varshney et al. [2011], Lints et al. [2005]) consisting of 278 somatic neurons, merging two distinct types of synapses: chemical and electrical (gap junctions). (b) Example of matrix W consistent with the connectome information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

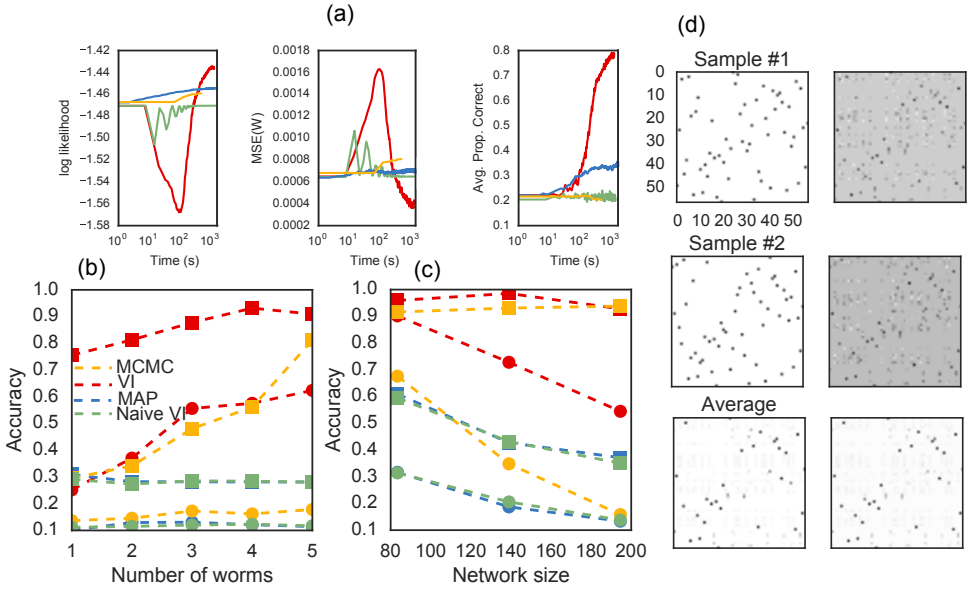

Figure 4: Results on the C.elegans inference example. (a) An example of convergence of the algorithm, and the baselines. (b) Accuracy on identity inference as a function of number of worms, for two values of ν ($\nu =$ for circles and $\nu =$ for squares). (c) Same as in (b), but using sub-networks of 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.

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. URL https://www.biorxiv.org/content/early/2017/08/31/182923.

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.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

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.

F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont. Convex relaxations for permutation problems. In *Advances in Neural Information Processing Systems*, pages 1016–1024, 2013.

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 problemâĂŤa short survey. In *11th International Conference on Information Fusion*, pages 1–7. IEEE, 2008.

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

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. SchrÃűdel, 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. URL http://www.sciencedirect.com/science/article/pii/S0092867415011964.

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

P. A. Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

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 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.

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. URL http://www.wormbase.org/db/misc/paper?name=WBPaper00025697.

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.

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.

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. URL http://dx.doi.org/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.

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.

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.

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

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

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

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

A Supplemental result on MNIST

B Limit analysis

B.1 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):

$$P(\pi_k = 1) = \prod_{i=1}^{k-1}(1 - p_i(\theta))p_k(\theta) \text{ for } k = 2, \ldots, K-1, \quad P(\pi_K = \tag{27}$$

   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 (27):

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

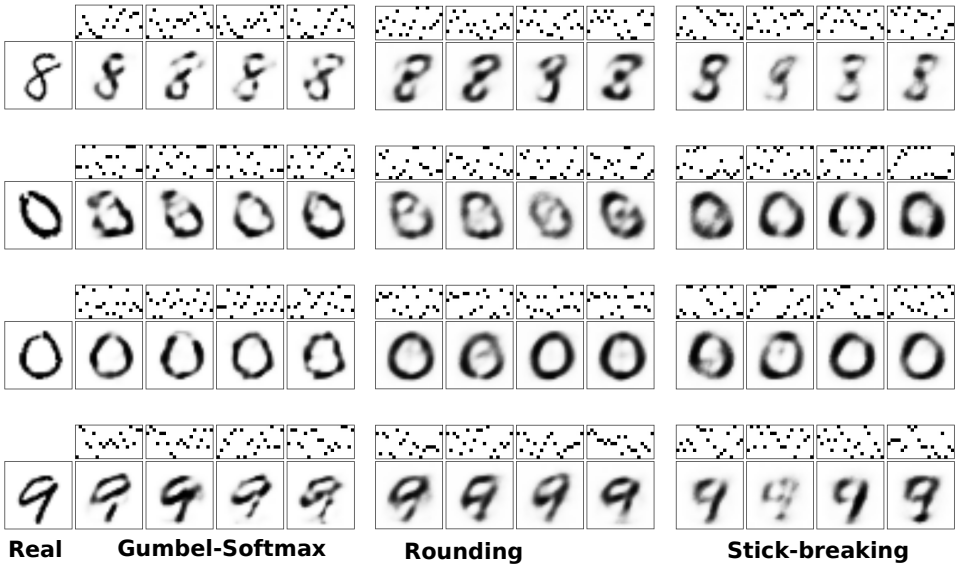

Figure 5: Examples of true and reconstructed digits from their corresponding random codes using with $N = 20$ categorical variables with $K = 10$ possible values.

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).

$$\begin{aligned} F_{z_n}(x) &= P(\sigma(\psi_n) < x) && (28)\\ &= P(\psi_n < \sigma^{-1}(x)) && (29)\\ &= P(\mu_n + \mu_n/K\xi < \sigma^{-1}(x)), && (30)\\ &= P(\xi < \sigma^{-1}(x)K/\mu_n - K) && (31)\\ &= \Phi(\sigma^{-1}(x)K/\mu_n - K) && (32) \end{aligned}$$

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 sequence with negative μ instead.

Lemma 3. If $z = \mathcal{K}(a, b)$:

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

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 $\pi = \mathscr{SB}(0.5, 0.5, \ldots, 0.5)$.
2. For Kumaraswamy $z_k = \mathscr{K}(a_k, b_k)$ 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 $\mathscr{SB}(a_k/(a_k+b_k), \ldots, a_{k-1}/(a_{k-1}+b_{k-1}))$.

C Deriving the approximation for the ELBO

Here we show that

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

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

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

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

$$q(F(\psi);\theta) = p(F^{-1}(\pi);\theta)|DF^{-1}(\pi)| \tag{33}$$
$$= p(\psi;\theta)|DF(\psi)|^{-1}. \tag{34}$$

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

$$\mathbb{E}_{p(\xi)}[-\log q(F(g(\theta,\xi));\theta)] = \mathbb{E}_{p(\psi;\theta)}[-\log p(\psi;\theta)] + \mathbb{E}_{p(\psi;\theta)}[\log|DF(\psi)|] \tag{35}$$
$$= Entropy(\psi;\theta) + E_{p(\xi)}[\log|DF(g(\theta,\xi))|]. \tag{36}$$

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

$$\left|\frac{\mathrm{d}\Psi}{\mathrm{d}\Pi}\right| = \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\frac{\partial\psi_{ij}}{\partial\pi_{ij}} \tag{37}$$
$$= \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\frac{\partial}{\partial\pi_{ij}}\sigma^{-1}\left(\frac{\pi_{ij}-\ell_{ij}}{u_{ij}-\ell_{ij}}\right) \tag{38}$$
$$= \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\left(\frac{1}{u_{ij}-\ell_{ij}}\right)\left(\frac{u_{ij}-\ell_{ij}}{\pi_{ij}-\ell_{ij}}\right)\left(\frac{u_{ij}-\ell_{ij}}{u_{ij}-\pi_{ij}}\right) \tag{39}$$
$$= \prod_{i=1}^{n-1}\prod_{j=1}^{n-1}\frac{u_{ij}-\ell_{ij}}{(\pi_{ij}-\ell_{ij})(u_{ij}-\pi_{ij})} \tag{40}$$

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

$$\mathrm{vec}(\Psi) \sim \mathcal{N}(\mu, \mathrm{diag}(\eta^2)) \tag{41}$$
$$\Pi = f(\Psi) \tag{42}$$
$$\implies p(\Pi\,|\,\mu, \mathrm{diag}(\eta^2)) = \left|\frac{\mathrm{d}\Psi}{\mathrm{d}\Pi}\right| \mathcal{N}(f^{-1}(\Pi)\,|\,\mu, \mathrm{diag}(\eta^2)) \tag{43}$$

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.