Reparameterizing the Birkhoff Polytope for Variational Permutation Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

How to perform posterior inference over the space of permutation matrices? By definition, with n nodes, there are $n!$ such matrices. 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.

1 Introduction

Permutation inference is central to many modern machine learning problems. Identity management [6] and multiple-object tracking [24, 10] 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 [16, 12, 1]. Moreover, many probabilistic models, like preferential attachment network models [2] and repulsive point process models [20], 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.

While the problem of finding optimal point estimates of permutations under a variety of cost functions has been the subject of decades of research in combinatorial optimization, many probabilistic tasks require reasoning over uncertainty regarding permutation matrices. Many works have addressed the challenge of Bayesian permutation inference, leveraging Markov chain Monte Carlo methods [4], Fourier representations [10, 7], as well as convex [13] and continuous [18] relaxations for approximating the posterior distribution. Given recent advances in scaling variational Bayesian

cite

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

inference, largely driven by efficient Monte Carlo estimators of gradients of the variational lower bound [9, 21], we revisit the problem of permutation inference from a variational perspective.

Motivated by the recently proposed Gumbel-softmax method for discrete variational inference [8, 15], we consider a variety of continuous relaxations of permutations that enable gradient-based inference. The Gumbel-softmax method is based on the following observation: discrete 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 [9, 21] 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. Analogously to the Gumbel-softmax method, we seek temperature-controlled relaxations of atomic densities on permutation matrices to continuous densities on the interior of the Birkhoff polytope. Unfortunately, due to the dual constraints of row- and column-normalization required of doubly stochastic matrices, the Gumbel-softmax method does not immmediately extend to this more challenging domain. 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 **??** discusses related work on Bayesian permutation inference, and Section **??** introduces the Gumbel-softmax relaxation upon which our approach builds. Section **??** introduces alternative relaxations for discrete variational inference, and Section **??** presents our primary contribution: a set of relaxations for permutation matrices Sections **??**-**??** detail a variety of experiments that illustrate the value of our variational approach.

1.1 Related Work

- MCMC [4] methods are successful in some cases, but ultimately rely on local updates to randomly explore the high dimensional space of permutations.
- Fourier [10, 7]
- Convex relaxations? [13]
- Other continuous relaxations [18]

1.2 The Gumbel-softmax relaxation for discrete variational inference

In Bayesian inference problems, we have a prior distribution $p(x)$ and a likelihood $p(y \mid x)$, and we seek the posterior distribution, $p(x \mid y) = p(x)p(y \mid x)/p(y)$. In general, this problem is intractable since the normalizing constant in Bayes' rule, $p(y)$, involves a high dimensional integral or sum. Variational inference algorithms avoid this problem by limiting their search to a tractable family of distributions, $q(x; \theta)$, parameterized by θ, and searching for the member of this family that best approximates the true posterior. Most commonly, the approximation quality is measured by the Kullback-Leibler (KL) divergence between the variational posterior, $q(x; \theta)$, and the true posterior, $p(x \mid y)$. That is, the optimal variational parameters are given by,

$$\theta^* = \arg\max_\theta -\mathrm{KL}\left(q(x;\theta) \parallel p(x \mid y)\right), \tag{1}$$

where

$$\begin{aligned} -\mathrm{KL}\left(q(x;\theta) \parallel p(x \mid y)\right) &= \mathbb{E}_q\left[\log p(x \mid y) - \log q(x;\theta)\right] && (2)\\ &\geq \mathbb{E}_q\left[\log p(x,\, y) - \log q(x;\theta)\right] && (3)\\ &= \mathcal{L}(\theta). && (4) \end{aligned}$$

The objective function, $\mathcal{L}(\theta)$, is known as the evidence lower bound, or ELBO. Stochastic gradient ascent is perhaps the simplest method of optimizing the ELBO with respect to the parameters θ. 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.

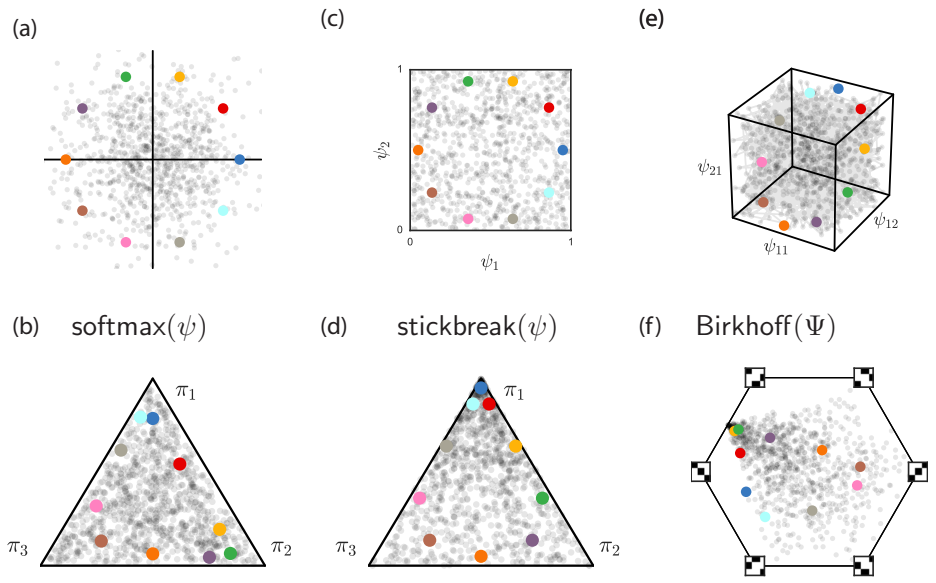

Figure 1

When x is a continuous random variable, we can often go one step further and leverage the "reparameterization trick" [23, 9, 19, 3]. Specifically, in some cases we can simulate from q via the following procedure,

$$\xi \sim p(\xi), \tag{5}$$
$$x = g(\theta, \xi), \tag{6}$$

where $g(\theta, \xi)$ is a deterministic and differentiable (in θ) transformation of the noise distribution ξ (e.g. ξ is a standard normal). Notice for g to be differentiable x needs to be continuous. Thus, we can effectively "factor out" the randomness of q. With this transformation, we can bring the gradient inside the expectation as follows,

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{p(\xi)} \left[\nabla_\theta \left[\log p(g(\theta, \xi) \mid y) - \log q(g(\theta, \xi); \theta) \right] \right]. \tag{7}$$

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 [25, 5].

Recently, there have been a number of proposals for extending these reparameterization tricks to high dimensional discrete problems[1] by relaxing them to analogous continuous problems [15, 8, 11]. These approaches are based on the following observation: if $x \in \{0, 1\}^k$ is a one-hot vector drawn from a categorical distribution, then the support of $p(x)$ is the set of vertices of the $k-1$ dimensional simplex. Thus, we can represent the distribution of x as an atomic density on the simplex. Sampling x is equivalent to sampling this atomic measure. That is,

$$x \sim \mathrm{Cat}(x \mid \theta) \qquad \Longleftrightarrow \qquad \pi \sim p(\pi \mid \theta)$$
$$x \triangleq \pi, \tag{8}$$

where $p(\pi \mid \theta)$ is a density on the simplex with atoms at the k vertices.

Viewing x as a vertex of the simplex motivates a natural relaxation: let us set $x = \pi$ as above, but rather than restricting $p(\pi \mid \theta)$ to be an atomic measure, let it be a continuous density on the simplex. To be concrete, suppose the density of π is defined by the transformation,

$$\xi \sim p(\xi), \tag{9}$$
$$\psi = g(\theta, \xi), \quad g(\theta, \xi) = \log \theta + \xi, \tag{10}$$
$$\pi = \mathrm{softmax}(\psi). \tag{11}$$
$$\tag{12}$$

[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 $p(y) = \sum_x p(y, x)$.

The output x is now a point on the simplex, and the parameters θ can be optimized via stochastic gradient ascent with the reparameterization trick, as discussed above.

In the aforementioned papers, $p(\xi)$ is taken to be the Gumbel distribution. This choice leads to a nicely interpretable model: adding Gumbel noise and taking the argmax yields an exact sample from θ; setting g to the softmax is a natural relaxation. Ultimately, however, this is just a continuous relaxation of an atomic density to a continuous density.

2 Methods

2.1 Alternative continuous relaxations

Here we introduce two alternative ways of conceiving a relaxed version of a discrete random variable that are also amenable for variational inference with Monte Carlo estimates of the gradient based on re-parameterization. However, unlike the Gumbel-Softmax, these relaxations enable extensions to more complex combinatorial objects; notably, permutations.

2.1.1 Stick-breaking

Consider the following alternative model for π:

$$\xi \sim p(\xi) \tag{13}$$
$$\psi = g(\theta, \xi), \tag{14}$$
$$z = h(\psi), \quad z \in [0,1]^{K-1}, \tag{15}$$
$$\pi = \mathcal{SB}(z), \tag{16}$$

where the stick-breaking transformation $\mathcal{SB}(z_i)$ is defined by the assignment:

$$\pi_1 = z_1, \quad \pi_k = z_k \left(1 - \sum_{j=1}^{k-1} \pi_j\right) \quad \text{for } k = 2, \ldots, K-1, \quad \pi_K = 1 - \sum_{j=1}^{K-1} \pi_j. \tag{17}$$

In words: the noise distribution is transformed to ψ, which is latter constrained to belong to the unit hypercube through $z = h(\psi)$, assumed invertible. Then, we apply the stick-breaking transformation to z to obtain a point π on the simplex. z_k can seen as the fraction of the remaining "stick" of probability mass assigned to π_k [14].

In this paper we focus on the choice $\psi_k \sim \mathcal{N}(\mu_k, \eta_k^2)$ (i.e., $\xi \sim \mathcal{N}(0, I), \theta = (\mu, \eta), g(\theta, \xi) = \mu + \eta\xi$), and $h(\cdot) = \sigma(\cdot/\tau)$ with σ the logistic function. We name this the known as a logistic-normal stick-breaking transformation, which enjoys the following properties: i)the density of π can be expressed in closed form as a function of μ_k and η_k^2, ii) the temperature τ controls how concentrated $p(\pi)$ 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 K vertices, and iv) as $\tau \to \infty$, the density concentrates on a point in the interior of the simplex determined by the parameters, and for intermediate values, the density is continuous on the simplex.

Finally, we stress the above logistic-normal construction is not the only available: by taking $h(\cdot)$ the identity function we conceive two other reparameterizations, by letting z have either a Kumaraswamy or Beta distribution on the unit interval. While the former is easily reparameterizable, the latter — which leads to the generalized Dirichlet distribution on the simplex — is not, requiring an additional rejection sampling correction term on its implementation in variational inference. See the appendix for proofs of our claims and details on alternative constructions.

2.1.2 Rounding

This relaxation is based on the notion of partitioning the space into regions around central points. Specifically, consider M points $\mathcal{P} = \{p_m\}$ in $\mathbb{R}^K$ (e.g., but not restricted to, one-hot vectors). Now, given $\psi \in \mathbb{R}^K$ we define the rounding operator

$$R^P(\psi) = \arg\min_{p_n} ||p_n - \psi||^2.$$

This operator partitions R^K into M different regions, the "Voronoi" cells with center p_m,

$$V_m^{\mathcal{P}} = \{\psi \in \mathbb{R}^K : \ ||\psi - p_m|| \leq ||\psi - p_i||, \forall i\}. \tag{18}$$

We might simply conceive a discrete distribution by the application of the above map to some continuous random variable. However, that is not differentiable. Then, instead we consider a map that pulls a point towards its rounded value, by taking a convex combination between both. Specifically, we consider the following sampling scheme:

$$\xi \sim p(\xi), \tag{19}$$

$$\psi = g(\theta, \xi), \tag{20}$$

$$\pi = \tau\psi + (1 - \tau)R^{\mathcal{P}}(\psi) \tag{21}$$

Clearly, in the temperature limit $\tau = 0$ we recover a discrete distribution, but for intermediate values the obtained R.V. is still continuous. Then, this gives us a way to approximate discrete values by continuous ones, but differing with previous approaches in that the approximating values z are not required to lie in the simplex, but anywhere. Notably, if ψ is gaussian parameterized by its mean and standard deviation, by appropriate (limit) choices of parameters we can represent any arbitrary categorical distribution over one-hot vectors. This is shown in the appendix.

2.2 Continuous relaxations for distributions over permutations

To extend the above relaxations to also represent distributions over permutations we start by representing permutations σ (now not the logistic function) of N elements (i.e., the $N!$ biyections from $\{1, \ldots, N\}$ onto itself) as binary matrices $X \in \{0, 1\}^{N \times N}$, whose row and column sums equal one. Then, X is like a one-hot vector but defined by a richer combinatorial structure. Notice that the rounding-based relaxation is available here, as the rounding operator can be casted as a matching problem, solved by the hungarian algorithm in $O(N^3)$ (see appendix). However, the stick-breaking equivalent for permutations is more involved: just as one-hot vectors are the vertices of the simplex, the Birkhoff-von Neumann theorem states that permutation matrices are vertices of the convex hull of doubly stochastic matrices. Then, an analogous relaxation is to consider $X \approx \Pi$, where $\Pi \in [0, 1]^{N \times N}$ is a doubly stochastic matrix, i.e. the rows and columns both are positive and sum to one but entries are not necessarily zero or one. This set is known as the Birkhoff polytope, denoted $\mathcal{B}_N$. Due to these constraints, $\mathcal{B}_N$ lies within a $(N - 1)^2$ dimensional subspace of all $[0, 1]^{N \times N}$ matrices.

We now derive an invertible and differentiable transformation, $f : \mathbb{R}^{(N-1)\times(N-1)} \to \mathcal{B}_n$, which can be used to define a density on $\mathcal{B}_N$, by extending the original stick-breaking transformation, with minor modifications to accomodate the additional constraints of doubly stochastic matrices. Imagine transforming a real-valued matrix $\Psi \in \mathbb{R}^{(N-1)\times(N-1)}$ into a doubly stochastic matrix, $\Pi \in [0, 1]^{N \times N}$. We work entry by entry, starting in the top left and raster scanning left to right then top to bottom. Denote the (i, j)-th entries of Ψ and Π by ψ_{ij} and π_{ij}, respectively.

The first entry is given by, $\pi_{11} = \sigma(\psi_{11})$. 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,

$$\pi_{1j} = \sigma(\psi_{1j})(1 - \sum_{k=1}^{j-1} \pi_{1k}) \qquad \text{for } j = 2, \ldots, n - 1 \tag{22}$$

$$\pi_{1n} = 1 - \sum_{k=1}^{n-1} \pi_{1k} \tag{23}$$

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

$$\pi_{ij} \leq 1 - \sum_{k=1}^{j-1} \pi_{ik} \qquad \text{(row sum)} \tag{24}$$

$$\pi_{ij} \leq 1 - \sum_{k=1}^{i-1} \pi_{kj} \qquad \text{(column sum).} \tag{25}$$

Moreover, there is also a lower bound on π_{ij}. 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,

$$\underbrace{1 - \sum_{k=1}^{j} \pi_{ik}}_{\text{remaining stick}} \leq \underbrace{\sum_{m=j+1}^{n} (1 - \sum_{k=1}^{i-1} \pi_{km})}_{\text{remaining upper bounds}}. \tag{26}$$

Rearranging terms, we have,

$$\pi_{ij} \geq 1 - \sum_{k=1}^{j-1} \pi_{ik} - \sum_{m=j+1}^{n} (1 - \sum_{k=1}^{i-1} \pi_{km}) \tag{27}$$

$$= 1 - n + j - \sum_{k=1}^{j-1} \pi_{ik} + \sum_{k=1}^{i-1} \sum_{m=j+1}^{n} \pi_{km} \tag{28}$$

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

$$\ell_{ij} \leq \pi_{ij} \leq u_{ij} \tag{29}$$

$$\ell_{ij} \triangleq \max \left\{ 0,\, 1 - n + j - \sum_{k=1}^{j-1} \pi_{ik} + \sum_{k=1}^{i-1} \sum_{m=j+1}^{n} \pi_{km} \right\} \tag{30}$$

$$u_{ij} \triangleq \min \left\{ 1 - \sum_{k=1}^{j-1} \pi_{ik},\, 1 - \sum_{k=1}^{i-1} \pi_{kj} \right\}. \tag{31}$$

Thus, we define,

$$\pi_{ij} = \ell_{ij} + \sigma(\psi_{ij})(u_{ij} - \ell_{ij}). \tag{32}$$

The inverse transformation from Π to Ψ is analogous. We start by computing ψ_{11} and then progressively compute upper and lower bounds and set,

$$\psi_{ij} = \sigma^{-1} \left(\frac{\pi_{ij} - \ell_{ij}}{u_{ij} - \ell_{ij}} \right). \tag{33}$$

2.3 Variational Inference

To approximate an intractable posterior distribution over latent discrete or permutation valued variables we choose the best (in the KL sense) among the parametric families that we have described. Now we describe the most relevant details.

2.3.1 From continuum to discrete

Doing a continuous relaxation requires re-thinking the objective. As in [15], we maximize a relaxed ELBO, for which we need to specify a new continuous prior $p(\pi)$ over the latents. This prior should be peaked around the discrete points for which inference is required, so we can conceive the original discrete ELBO through a limiting process. Good, peaked priors are necessary as the likelihood $p(y|\pi)$ could be large for points that have nothing to do with the original problem, leading to nonsensical solutions. There is a trade-off, though: priors that are too peaked will lead to multiple local maxima of the ELBO, preventing optimizers from achieving the right solution.Then, for optimal performance the prior has to be treated as a hyperparameter. Here, for one-hot vectors we considered mixture of gaussians around each point (parameterized by the variance), and for permutations, we considered the product of N one-dimensional mixture of gaussians around zero and one. Although this prior puts significant mass on points that are not interesting (e.g. (1,1,..., 1)), at least it penalizes π that are too away from $\mathcal{B}_N$.

Table 1: Summary of results in VAE

	Gumbel-Softmax	Rounding	Stick-breaking
- $\log p(x)$	113.9	$\sim$106.9	12

2.3.2 Estimating the ELBO

Notice in all the relaxations discussed here, $\pi = F(\psi)$ where F is a differentiable and invertible function (see the appendix for details). Therefore, by the change of variable theorem and the law of the unconscious statistician:

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

where the term inside of the expectation is the (log) Jacobian of F evaluated at $\psi = f(\theta,\xi)$. Then, if this Jacobian and the entropy of ψ are available we can consider an unbiased estimator for the ELBO:

$$\hat{\mathcal{L}}(\theta) = Entropy(\psi) + \frac{1}{L}\sum_{l=1}^{L}\left[\log p(y, g(\theta,\xi_l)) + \log|DF(g(\theta,\xi_l)|\right].$$

For example, for rounding, F is piecewise linear [2] and $\log|DF(g(\theta,\xi))| = K\log\tau$. Also, if ψ is gaussian its entropy is given by $K\log(\eta^2 2\pi e)/2$.

2.3.3 Constraining the parameters

For rounding-based permutation inference, we found that an efficient way of avoiding evaluation of the likelihood in non-relevant points is to constrain μ to belong to $\mathcal{B}_N$. This can be done applying Sinkhorn propagation, i.e., successively normalizing the rows and columns of the matrix μ. The resulting μ is guaranteed to converge to a point in $\mathcal{B}_N$, which will remain close if η is small.

3 Results

Here we summarize our main results. See the appendix for implementation details.

3.1 Variational Autoencoders (VAE) with categorical latent variables

We first demonstrate that our proposed relaxations are sensible for categorical random variables, and can be efficiently implemented within the current dominant computational frameworks. We considered the density estimation task on MNIST digits, as in [15, 8], where observed digits are thought as 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 agains the method of [8], finding similar results (Table 1). Also, better results were obtained with rounding. Our results suggest our methods provide a viable alternative to the Gumbel-Softmax. In figure 2 we show some reconstructed images using the different approaches.

3.2 Synthetic matching experiments

To assess the quality of our approximations for distributions over permutations, we considered a toy problem where $N = 6$ points in $\mathbb{R}^2$ where normally sampled around other fixed $N = 6$ centers, with varying standard deviation σ. This naturally induces a posterior distribution of assignments between samples and the centers they come from [3], and as $N! = 720$ is moderate, this posterior can be computed exactly by enumeration. We measured discrepancy through the Battacharya distance (BD)

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

[3] think of this as a variation of the mixture of gaussians models, where now each center is associated with a unique sample

we show examples of true posteriors (ranked) and their approximations, and quantify the discrepancies by the distribution of the BD.

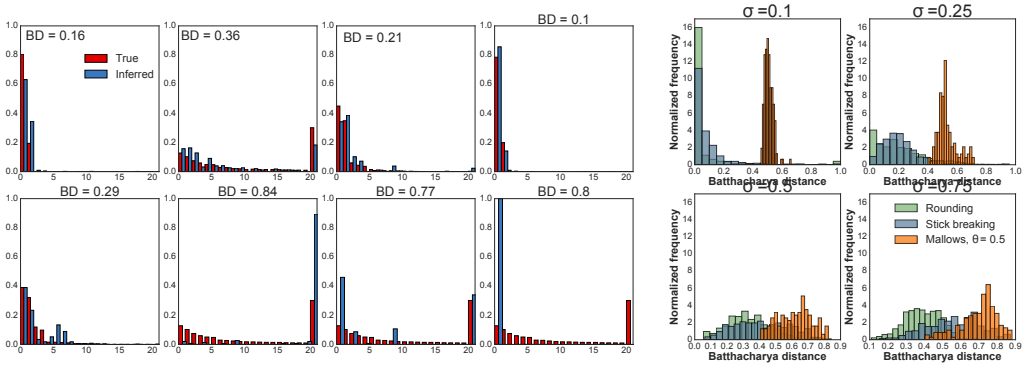

Figure 2

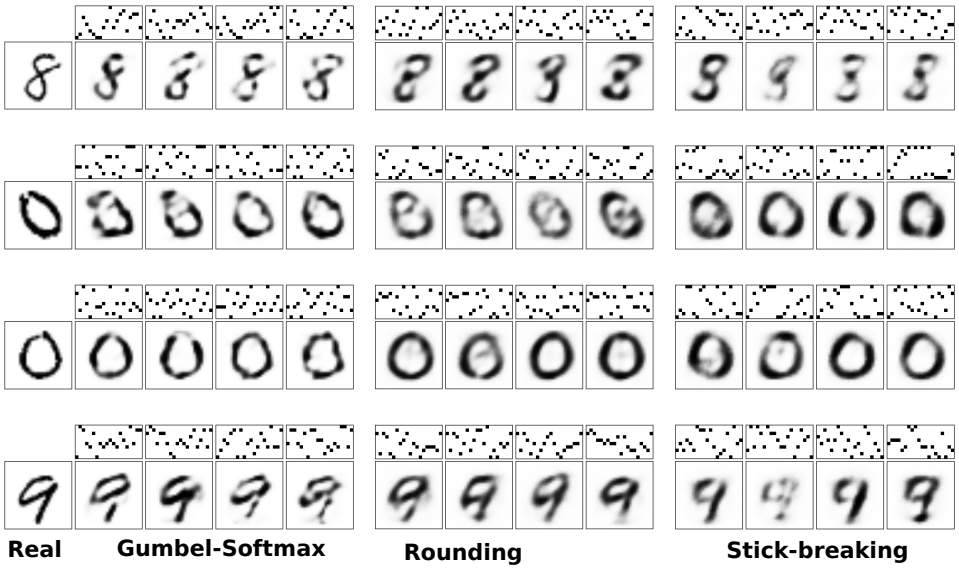

Figure 3: Examples of true and reconstructed digits from their corresponding random codes using with $N = 20$ categorical variables with $K = 10$ possible values. In any case

between true histogram and the histogram of the approximated posterior constructed by sampling from $q(\pi)$ and 'rounding' to the nearest permutation using the Hungarian algorithm. We found our methods provide reasonable fit to the true posterior, allowing to represent more complex distributions over permutations than, e.g., simple Mallows distribution around the MAP estimate. In figure 3

3.3 Hierarchical permutation inference

References

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

[2] 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.

[3] G. Bonnet. Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire. *Annals of Telecommunications*, 19(9):203–220, 1964.

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

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

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

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

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

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

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] R. Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

[20] 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.

[21] 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.

[22] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

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

[24] 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.

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

A Limit analysis

A.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):

$$P(\pi_k = 1) = \prod_{i=1}^{k-1}(1 - p_i(\theta))p_k(\theta) \text{ for } k = 2, \dots, K-1, \quad P(\pi_K = 1) = \prod_{i=1}^{K-1}(1 - p_i(\theta)). \tag{34}$$

   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 (34):

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

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).

$$\begin{aligned} F_{z_n}(x) &= P(\sigma(\psi_n) < x) && (35)\\ &= P(\psi_n < \sigma^{-1}(x)) && (36)\\ &= P(\mu_n + \mu_n/K\xi < \sigma^{-1}(x)), && (37)\\ &= P(\xi < \sigma^{-1}(x)K/\mu_n - K) && (38)\\ &= \Phi(\sigma^{-1}(x)K/\mu_n - K) && (39) \end{aligned}$$

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 [17]

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 re-parameterizations 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 $\pi = \mathcal{SB}(0.5, 0.5, \ldots, 0.5)$.
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 $\mathcal{SB}(a_k/(a_k+b_k), \ldots, a_{k-1}/(a_{k-1}+b_{k-1}))$.

A.2 Rounding

Here we have two extremes: at $\tau = 1$ we obtain a continuous distribution in the space (here, gaussian). If $\tau = 0$ the resulting distribution has only atoms in the one-hot vectors p_n, in this proof assumed to be the one-hot vectors. We show that in this case it is possible to represent any arbitrary categorical distribution through a judicious choice of the parameters.

Proposition: In the zero temperature case, i.e., $\pi = R^{\mathcal{P}}(\psi)$ it is possible to represent any arbitrary distribution i.e, for any α in the $N-1$ simplex there exists gaussian parameters (μ, η) so that $P(\pi = p_n) = \alpha_n$. Points inside the simplex are realized directly, while distributions with some $\alpha_k = 0$ are realized through a limiting process in the parameters.

Proof: First set $\eta_n = 1$. By representing $\psi = \mu + \xi$ where $\xi \sim \mathcal{N}(0, I)$ we see that

$$\alpha_n = P(R^{\mathcal{P}}(\mu+\xi) = p_n) = P(\mu + \xi \in V_n^{\mathcal{P}}) = P(\xi \in V_n^{\mathcal{P}} - \mu) = \int_{V_n^{\mathcal{P}} - \mu} \frac{1}{(2\pi)^{\frac{N}{2}}} e^{-\frac{||x||^2}{2}} dx.$$

Three conclusions are drawn from the above: first, we see that probabilities are ultimately gaussian integrals over a new partition, a translation of the Voronoi regions $V_n^{\mathcal{P}}$. Second, the map $(\mu_1, \ldots, \mu_n) \xrightarrow{m} (\alpha_1, \ldots, \alpha_n)$ is continuous by virtue of the dominated convergence theorem [**?**]: indeed, if $\mu_i \to \mu$ then $(\alpha_1^i, \ldots, \alpha_n^i) \to (\alpha_1, \ldots, \alpha_n)$, as the α_n^i are integrals that can be expressed using indicator functions in the integrands (which are all bounded by the integrable gaussian density), and as pointwise convergence of the indicators holds because of the continuity of the translation operator $f_\mu(\cdot) = \cdot + \mu$. Third, for each $q \in (0, 1)$ it is possible to choose μ so that $\alpha_n = q$ and $\alpha_m = (1 - q)/(n - 1)$. Indeed, by moving μ_n between $-\infty$ and ∞ while keeping the other μ_k fixed then $V_n^{\mathcal{P}} - \mu$ fluctuates between the empty set ($\alpha_n = 0$) and the entire space ($\alpha_n = 1$). Therefore, by the continuity of m and the intermediate value theorem, every value in $\alpha_n \in (0, 1)$ is realized, and by symmetry the other α_k occupy the remaining mass uniformly.

To conclude, we use the fact that the image of a convex set through a continuous function is also convex [22]. We also showed that for each tolerance ϵ the image of m contains these ϵ-one-hot vectors: that is, the points with $(1 - \epsilon)$ in one coordinate and $\epsilon/(n - 1)$ in the rest. Then, given a point in the interior of the simplex choose ϵ small enough so it is in the convex hull of the ϵ-one vectors. Then, by the above theorem there must be a pre-image μ that realizes this point.

For α with zero entries the above arguments has to be extended to permit a limit process where μ goes to either infinity or infinity. It is easy to see that by that extension it is possible to represent any α in the border of the simplex.

B Deriving the approximation for the ELBO

Here we show that

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

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:

$$\begin{aligned} q(F(\psi); \theta) &= p(F^{-1}(\pi); \theta) |DF^{-1}(\pi)| && (40) \\ &= p(\psi; \theta) |DF(\psi)|^{-1}. && (41) \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}_{p(\xi)} \left[-\log q(F(g(\theta, \xi)); \theta) \right] &= \mathbb{E}_{p(\psi;\theta)} \left[-\log p(\psi; \theta) \right] + \mathbb{E}_{p(\psi;\theta)} \left[\log |DF(\psi)| \right] && (42) \\ &= Entropy(\psi; \theta) + E_{p(\xi)} \left[\log |DF(g(\theta, \xi))| \right]. && (43) \end{aligned}$$

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,

$$\begin{aligned} \left| \frac{\mathrm{d}\Psi}{\mathrm{d}\Pi} \right| &= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{\partial \psi_{ij}}{\partial \pi_{ij}} && (44) \\ &= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{\partial}{\partial \pi_{ij}} \sigma^{-1} \left(\frac{\pi_{ij} - \ell_{ij}}{u_{ij} - \ell_{ij}} \right) && (45) \\ &= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \left(\frac{1}{u_{ij} - \ell_{ij}} \right) \left(\frac{u_{ij} - \ell_{ij}}{\pi_{ij} - \ell_{ij}} \right) \left(\frac{u_{ij} - \ell_{ij}}{u_{ij} - \pi_{ij}} \right) && (46) \\ &= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{u_{ij} - \ell_{ij}}{(\pi_{ij} - \ell_{ij})(u_{ij} - \pi_{ij})} && (47) \end{aligned}$$

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

$$\text{vec}(\Psi) \sim \mathcal{N}(\mu, \text{diag}(\eta^2)) \tag{48}$$

$$\Pi = f(\Psi) \tag{49}$$

$$\implies p(\Pi \mid \mu, \text{diag}(\eta^2)) = \left|\frac{\mathrm{d}\Psi}{\mathrm{d}\Pi}\right| \mathcal{N}(f^{-1}(\Pi) \mid \mu, \text{diag}(\eta^2)) \tag{50}$$

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.