
Reparameterizing the Birkhoff Polytope for Variational Permutation Inference

Anonymous Author(s)

Affiliation
Address
email

Abstract

1 How can we efficiently perform posterior inference over the space of permutations
2 when there are $n!$ permutations of a set of n elements? Clearly, estimating a
3 complete probability mass function over this space quickly becomes intractable
4 as n grows. Our goal is to derive a tractable algorithm for performing approx-
5 imate inference over this challenging discrete space. To that end, we consider
6 extensions of the recently proposed Gumbel-softmax method, which leverages
7 continuous relaxations to perform discrete variational inference with reparameteri-
8 zation gradients. While the Gumbel-softmax method is not immediately applicable
9 to permutation inference, we show that two alternative reparameterizations are both
10 comparable to Gumbel-softmax on tractable discrete problems and easily extensi-
11 ble to permutation inference. Specifically, we develop continuous relaxations of
12 permutation matrices to matrices that are either exactly or nearly doubly stochastic,
13 i.e. to points either in or near the Birkhoff polytope. We then derive invertible and
14 differentiable maps from densities on unconstrained space to densities on or near
15 the Birkhoff polytope. These transformations are parameterized by a “temperature”
16 that controls how concentrated the resulting density is at the extrema of the Birkhoff
17 polytope; i.e. at permutation matrices. This relaxation admits variational inference
18 via stochastic gradient ascent over the distributions on doubly stochastic matrices
19 (and in the zero-temperature limit, on permutation matrices) using Monte Carlo
20 estimates of the reparameterized gradient.

21

1 Introduction

22 Permutation inference is central to many modern machine learning problems. Identity management [7]
23 and multiple-object tracking [28, 11] are fundamentally concerned with finding a permutation that
24 maps an observed set of items to a set of canonical labels. Ranking problems, critical to search and
25 recommender systems, require inference over the space of item orderings [19, 15, 1]. Moreover, many
26 probabilistic models, like preferential attachment network models [2] and repulsive point process
27 models [24], incorporate a latent permutation into their generative processes; inference over model
28 parameters requires integrating over the set of permutations that could have given rise to the observed
29 data. In many of these settings, permutation inference is but one component of a larger estimation
30 problem involving unknown model parameters and hierarchical structure.

31 The task of computing optimal point estimates of permutations under various loss functions has been
32 well studied in the combinatorial optimization literature [12, 21, 14]. However, many probabilistic
33 tasks require reasoning about uncertainty regarding permutation matrices. A variety of Bayesian
34 permutation inference algorithms have been proposed, leveraging Markov chain Monte Carlo methods
35 [5], Fourier representations [11, 8], as well as convex [16] and continuous [22] relaxations for
36 approximating the posterior distribution. Given recent advances in scaling variational Bayesian

37 inference, largely driven by efficient Monte Carlo estimators of gradients of the variational lower
38 bound [10, 25], we revisit the problem of permutation inference from a variational perspective.

39 Motivated by the recently proposed Gumbel-softmax method for discrete variational inference [9, 18],
40 we consider a variety of continuous relaxations of permutations that enable gradient-based inference.
41 The Gumbel-softmax method is based on the following observation: discrete distributions may be
42 viewed as atomic densities on the vertices of the simplex; by relaxing this to a continuous density on
43 the interior of the simplex we can approximate the discrete inference problem with a continuous one
44 and thereby capitalize on reparameterization gradients [10, 25] to optimize a variational lower bound
45 on the marginal likelihood. Critically, the Gumbel-softmax method has a temperature parameter that
46 tunes the degree to which the continuous density concentrates around the vertices, and recovers truly
47 discrete inference in the zero-temperature limit.

48 Just as one-hot vectors (discrete random variables) are the vertices of the simplex, permutation matrices
49 are the vertices of the Birkhoff polytope, i.e. the set of doubly stochastic matrices. Analogously
50 to the Gumbel-softmax method, we seek temperature-controlled relaxations of atomic densities on
51 permutation matrices to continuous densities on the interior of the Birkhoff polytope. Unfortunately,
52 due to the dual constraints of row- and column-normalization required of doubly stochastic matrices,
53 the Gumbel-softmax method does not immediately extend to this more challenging domain. How-
54 ever, we derive a variety of alternative continuous relaxations for the simplex and show that: (i) these
55 relaxations achieve comparable performance to the Gumbel-softmax on tractable discrete inference
56 tasks; and (ii) they naturally extend to relaxations of permutation inference problems.

57 The remainder of this paper is structured as follows: Section ?? discusses related work on Bayesian
58 permutation inference, and Section ?? introduces the Gumbel-softmax relaxation upon which our
59 approach builds. Section ?? introduces alternative relaxations for discrete variational inference,
60 and Section ?? presents our primary contribution: a set of relaxations for permutation matrices
61 Sections ??-?? detail a variety of experiments that illustrate the value of our variational approach.

62 1.1 Related Work

- 63 • MCMC [5] methods are successful in some cases, but ultimately rely on local updates to
64 randomly explore the high dimensional space of permutations.
- 65 • Fourier [11, 8]
- 66 • Convex relaxations? [16]
- 67 • Other continuous relaxations [22]

68 1.2 The Gumbel-softmax relaxation for discrete variational inference

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

$$\theta^* = \arg \max_{\theta} -\text{KL}(q(x; \theta) \| p(x | y)), \quad (1)$$

77 where

$$-\text{KL}(q(x; \theta) \| p(x | y)) = \mathbb{E}_q [\log p(x | y) - \log q(x; \theta)] \quad (2)$$

$$\geq \mathbb{E}_q [\log p(x, y) - \log q(x; \theta)] \quad (3)$$

$$= \mathcal{L}(\theta). \quad (4)$$

78 The objective function, $\mathcal{L}(\theta)$, is known as the evidence lower bound, or ELBO. Stochastic gradient
79 ascent is perhaps the simplest method of optimizing the ELBO with respect to the parameters θ .
80 However, computing $\nabla_{\theta}\mathcal{L}(\theta)$ requires some care, since the ELBO contains an expectation with
81 respect to a distribution that depends on these parameters.

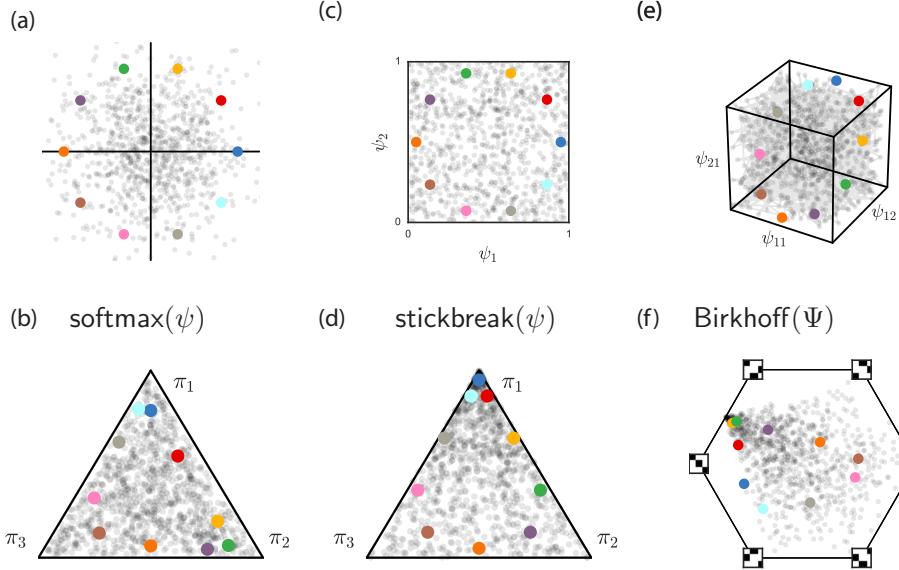


Figure 1

82 When x is a continuous random variable, we can often go one step further and leverage the “reparam-
 83 eterization trick” [27, 10, 23, 3]. Specifically, in some cases we can simulate from q via the following
 84 procedure,

$$\xi \sim p(\xi), \quad (5)$$

$$x = g(\theta, \xi), \quad (6)$$

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

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

89 This gradient can be estimated with Monte Carlo, and, in practice, this leads to lower variance
 90 estimates of the gradient than, for example, the score function estimator [29, 6].

91 Recently, there have been a number of proposals for extending these reparameterization tricks to
 92 high dimensional discrete problems¹ by relaxing them to analogous continuous problems [18, 9, 13].
 93 These approaches are based on the following observation: if $x \in \{0, 1\}^k$ is a one-hot vector drawn
 94 from a categorical distribution, then the support of $p(x)$ is the set of vertices of the $k - 1$ dimensional
 95 simplex. Thus, we can represent the distribution of x as an atomic density on the simplex. Sampling x
 96 is equivalent to sampling this atomic measure. That is,

$$x \sim \text{Cat}(x | \theta) \iff \pi \sim p(\pi | \theta) \\ x \triangleq \pi, \quad (8)$$

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

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

$$\xi \sim p(\xi), \quad (9)$$

$$\psi = g(\theta, \xi), \quad g(\theta, \xi) = \log \theta + \xi, \quad (10)$$

$$\pi = \text{softmax}(\psi). \quad (11)$$

$$(12)$$

¹Discrete 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 $p(y) = \sum_x p(y, x)$.

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

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

107 2 Methods

108 2.1 Alternative continuous relaxations

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

113 2.1.1 Stick-breaking

114 Consider the following alternative model for π :

$$\xi \sim p(\xi) \quad (13)$$

$$\psi = g(\theta, \xi), \quad (14)$$

$$z = h(\psi), \quad z \in [0, 1]^{K-1}, \quad (15)$$

$$\pi = \mathcal{SB}(z), \quad (16)$$

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

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

116 In words: the noise distribution is transformed to ψ , which is latter constrained to belong to the unit
117 hypercube through $z = h(\psi)$, assumed invertible. Then, we apply the stick-breaking transformation
118 to z to obtain a point π on the simplex. z_k can seen as the fraction of the remaining “stick” of
119 probability mass assigned to π_k [17].

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

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

134 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

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

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

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

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

$$\xi \sim p(\xi), \quad (19)$$

$$\psi = g(\theta, \xi), \quad (20)$$

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

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

146 2.2 Continuous relaxations for distributions over permutations

147 To extend the above relaxations to also represent distributions over permutations we start by repre-
 148 senting permutations σ (now not the logistic function) of N elements (i.e., the $N!$ bijections from
 149 $\{1, \dots, N\}$ onto itself) as binary matrices $X \in \{0, 1\}^{N \times N}$, whose row and column sums equal one.
 150 Then, X is like a one-hot vector but defined by a richer combinatorial structure. Notice that the
 151 rounding-based relaxation is available here, as the rounding operator can be casted as a matching
 152 problem, solved by the hungarian algorithm in $O(N^3)$ (see appendix). However, the stick-breaking
 153 equivalent for permutations is more involved: just as one-hot vectors are the vertices of the simplex,
 154 the Birkhoff-von Neumann theorem states that permutation matrices are vertices of the convex
 155 hull of doubly stochastic matrices. Then, an analogous relaxation is to consider $X \approx \Pi$, where
 156 $\Pi \in [0, 1]^{N \times N}$ is a doubly stochastic matrix, i.e. the rows and columns both are positive and sum to
 157 one but entries are not necessarily zero or one. This set is known as the Birkhoff polytope, denoted
 158 \mathcal{B}_N . Due to these constraints, \mathcal{B}_N lies within a $(N - 1)^2$ dimensional subspace of all $[0, 1]^{N \times N}$
 159 matrices.

160 We now derive an invertible and differentiable transformation, $f : \mathbb{R}^{(N-1) \times (N-1)} \rightarrow \mathcal{B}_n$, which
 161 can be used to define a density on \mathcal{B}_N , by extending the original stick-breaking transformation,
 162 with minor modifications to accomodate the additional constraints of doubly stochastic matrices.
 163 Imagine transforming a real-valued matrix $\Psi \in \mathbb{R}^{(N-1) \times (N-1)}$ into a doubly stochastic matrix, $\Pi \in$
 164 $[0, 1]^{N \times N}$. We work entry by entry, starting in the top left and raster scanning left to right then top to
 165 bottom. Denote the (i, j) -th entries of Ψ and Π by ψ_{ij} and π_{ij} , respectively.

166 The first entry is given by, $\pi_{11} = \sigma(\psi_{11})$. As we work left to right in the first row, the “remaining
 167 stick” length decreases as we add new entries. This reflects the row normalization constraints. Thus,

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

$$\pi_{1n} = 1 - \sum_{k=1}^{n-1} \pi_{1k} \quad (23)$$

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

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

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

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

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

175 Rearranging terms, we have,

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

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

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

$$\ell_{ij} \leq \pi_{ij} \leq u_{ij} \quad (29)$$

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

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

178 Thus, we define,

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

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

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

181 2.3 Variational Inference

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

185 2.3.1 From continuum to discrete

186 Doing a continuous relaxation requires re-thinking the objective. As in [18], we maximize a relaxed
 187 ELBO, for which we need to specify a new continuous prior $p(\pi)$ over the latents. This prior should
 188 be peaked around the discrete points for which inference is required, so we can conceive the original
 189 discrete ELBO through a limiting process. Good, peaked priors are necessary as the likelihood $p(y|\pi)$
 190 could be large for points that have nothing to do with the original problem, leading to nonsensical
 191 solutions. There is a trade-off, though: priors that are too peaked will lead to multiple local maxima
 192 of the ELBO, preventing optimizers from achieving the right solution. Then, for optimal performance
 193 the prior has to be treated as a hyperparameter. Here, for one-hot vectors we considered mixture of
 194 gaussians around each point (parameterized by the variance), and for permutations, we considered
 195 the product of N one-dimensional mixture of gaussians around zero and one. Although this prior
 196 puts significant mass on points that are not interesting (e.g. $(1, 1, \dots, 1)$), at least it penalizes π that
 197 are too away from \mathcal{B}_N .

Table 1: Summary of results in VAE

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

198 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:

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

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

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

201 For example, for rounding, F is piecewise linear² and $\log |DF(g(\theta, \xi))| = K \log \tau$. Also, if ψ is
202 gaussian its entropy is given by $K \log(\eta^2 2\pi e)/2$.

203 2.3.3 Constraining the parameters

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

208 3 Results

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

210 3.1 Variational Autoencoders (VAE) with categorical latent variables

211 We first demonstrate that our proposed relaxations are sensible for categorical random variables,
212 and can be efficiently implemented within the current dominant computational frameworks. We
213 considered the density estimation task on MNIST digits, as in [18, 9], where observed digits are
214 thought as reconstructed from a latent discrete code. We used the continuous ELBO for training,
215 and evaluated performance based on the marginal likelihood, estimated through the multi-sample
216 variational objective of the discretized model (via rounding the samples π) with $m = 1000$. We
217 trained our models using ADAM in Tensorflow and compared against the method of [9], finding
218 similar results (Table 1). Also, better results were obtained with rounding. Our results suggest our
219 methods provide a viable alternative to the Gumbel-Softmax. In figure 2 we show some reconstructed
220 images using the different approaches.

221 3.2 Synthetic matching experiments

222 To assess the quality of our approximations for distributions over permutations, we considered a toy
223 problem where $N = 6$ points in \mathbb{R}^2 where normally sampled around other fixed $N = 6$ centers, with
224 varying standard deviation σ . This naturally induces a posterior distribution of assignments between
225 samples and the centers they come from³, and as $N! = 720$ is moderate, this posterior can be
226 computed exactly by enumeration. We measured discrepancy through the Battacharya distance (BD)

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

³think 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.

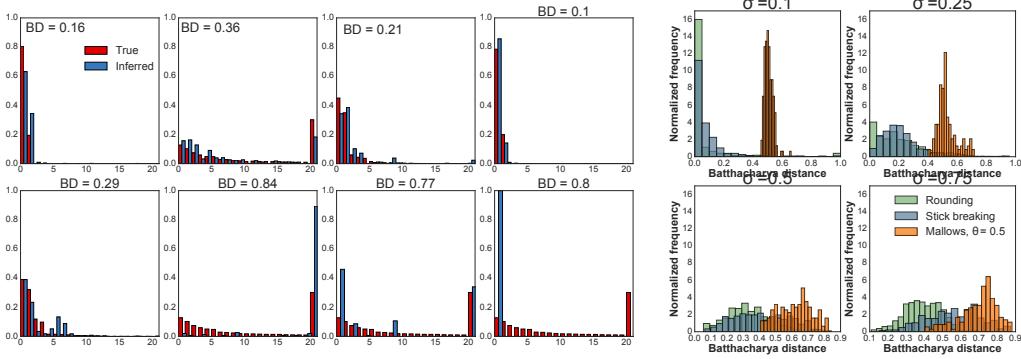


Figure 2

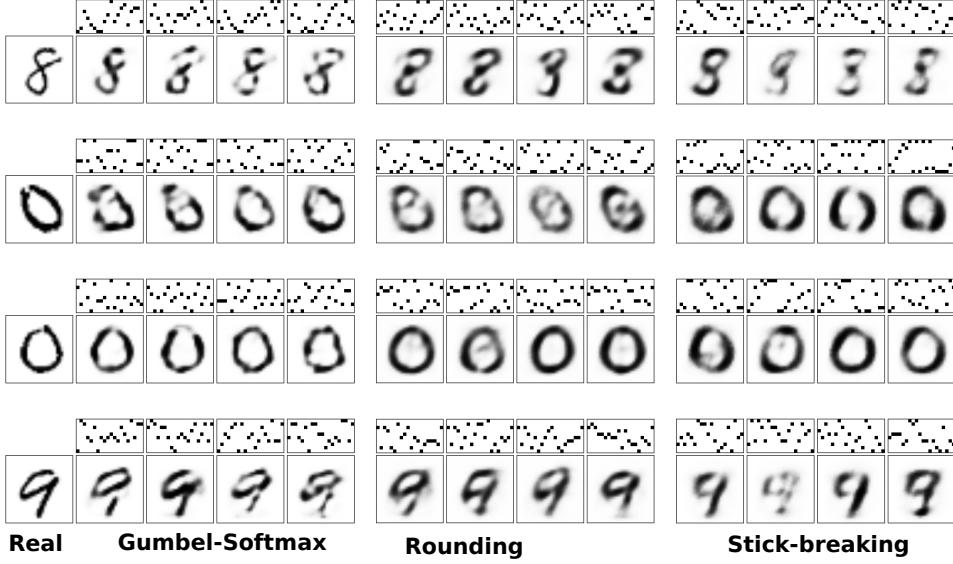


Figure 3: Examples of true and reconstructed digits from their corresponding random codes using with $N = 20$ categorical variables with $K = 10$ possible values. In any case

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

231 **3.3 Hierarchical permutation inference**

232 **References**

- 233 [1] R. P. Adams and R. S. Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.
 234
- 235 [2] B. Bloem-Reddy and P. Orbanz. Random walk models of network formation and sequential
 236 Monte Carlo methods for graphs. *arXiv preprint arXiv:1612.06404*, 2016.
- 237 [3] G. Bonnet. Transformations des signaux aléatoires à travers les systèmes non linéaires sans
 238 mémoire. *Annals of Telecommunications*, 19(9):203–220, 1964.

- 239 [4] A. Browder. *Mathematical analysis: an introduction*. Springer Science & Business Media,
 240 2012.
- 241 [5] P. Diaconis. Group representations in probability and statistics. In S. S. Gupta, editor, *Institute
 242 of Mathematical Statistics Lecture Notes—Monograph Series*, volume 11. 1988.
- 243 [6] P. W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of
 244 the ACM*, 33(10):75–84, oct 1990.
- 245 [7] L. J. Guibas. The identity management problem—A short survey. In *11th International
 246 Conference on Information Fusion*, pages 1–7. IEEE, 2008.
- 247 [8] J. Huang, C. Guestrin, and L. Guibas. Fourier theoretic probabilistic inference over permutations.
 248 *Journal of machine learning research*, 10(May):997–1070, 2009.
- 249 [9] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. *arXiv
 250 preprint arXiv:1611.01144*, 2016.
- 251 [10] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference
 252 on Learning Representations*, 2014.
- 253 [11] R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the
 254 symmetric group. In *AISTATS*, volume 1, page 5, 2007.
- 255 [12] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics
 256 quarterly*, 2(1-2):83–97, 1955.
- 257 [13] M. J. Kusner and J. M. Hernández-Lobato. GANs for sequences of discrete elements with the
 258 Gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.
- 259 [14] E. L. Lawler. The quadratic assignment problem. *Management science*, 9(4):586–599, 1963.
- 260 [15] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *Journal of Machine
 261 Learning Research*, 9(Oct):2401–2429, 2008.
- 262 [16] C. H. Lim and S. Wright. Beyond the Birkhoff polytope: Convex relaxations for vector
 263 permutation problems. In *Advances in Neural Information Processing Systems*, pages 2168–
 264 2176, 2014.
- 265 [17] S. Linderman, M. Johnson, and R. P. Adams. Dependent multinomial models made easy: Stick-
 266 breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing
 267 Systems*, pages 3456–3464, 2015.
- 268 [18] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of
 269 discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
- 270 [19] M. Meilă, K. Phadnis, A. Patterson, and J. Bilmes. Consensus ranking under the exponential
 271 model. In *In Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence
 272 (UAI)*, 2007.
- 273 [20] P. A. Mitnik. New properties of the kumaraswamy distribution. *Communications in Statistics
 274 - Theory and Methods*, 42(5):741–755, 2013. doi: 10.1080/03610926.2011.581782. URL
 275 <http://dx.doi.org/10.1080/03610926.2011.581782>.
- 276 [21] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the society
 277 for industrial and applied mathematics*, 5(1):32–38, 1957.
- 278 [22] S. M. Plis, S. McCracken, T. Lane, and V. D. Calhoun. Directional statistics on permutations.
 279 In *AISTATS*, pages 600–608, 2011.
- 280 [23] R. Price. A useful theorem for nonlinear devices having Gaussian inputs. *IRE Transactions on
 281 Information Theory*, 4(2):69–72, 1958.
- 282 [24] V. Rao, R. P. Adams, and D. D. Dunson. Bayesian inference for Matérn repulsive processes.
 283 *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

- 284 [25] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate
 285 inference in deep generative models. In *Proceedings of The 31st International Conference on*
 286 *Machine Learning*, pages 1278–1286, 2014.
- 287 [26] R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press,
 288 Princeton, N. J., 1970.
- 289 [27] T. Salimans and D. A. Knowles. Fixed-form variational posterior approximation through
 290 stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.
- 291 [28] J. Shin, N. Lee, S. Thrun, and L. Guibas. Lazy inference on object identities in wireless sensor
 292 networks. In *Proceedings of the 4th international symposium on Information processing in*
 293 *sensor networks*, page 23. IEEE Press, 2005.
- 294 [29] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement
 295 learning. *Machine Learning*, 8(3–4):229–256, 1992.

296 A Limit analysis

297 A.1 Stick-breaking

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

302 **Lemma 1.** The following statements are true:

- 303 1. the degenerate case where z_k is deterministic leads to $\pi \sim \delta(\tilde{\pi})$ (i.e, single atom in the point
 304 $\tilde{\pi}$). Also, if z_k can be any in $(0, 1)$ then any deterministic π in the interior of the simplex
 305 can be realized.
- 306 2. the degenerate case where z_k are Bernoulli with parameter $p_k(\theta) \in (0, 1)$ leads to π having
 307 an atomic distribution with atoms in the vertices of Δ^{k-1} ; i.e, π is categorical. We have the
 308 following expression for the probabilities of the atoms $\pi_k = 1$ (one hot vectors):

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

309 Moreover, if for each index k any parameter of the Bernoulli variable z_k can be realized
 310 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):

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

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

313 Now we can state our results:

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

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

318 In both cases the convergence is in distribution

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

$$F_{z_n}(x) = P(\sigma(\psi_n) < x) \quad (35)$$

$$= P(\psi_n < \sigma^{-1}(x)) \quad (36)$$

$$= P(\mu_n + \mu_n/K\xi < \sigma^{-1}(x)), \quad (37)$$

$$= P(\xi < \sigma^{-1}(x)K/\mu_n - K) \quad (38)$$

$$= \Phi(\sigma^{-1}(x)K/\mu_n - K) \quad (39)$$

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

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

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

332 In both cases convergence is in probability. *Proof:* A proof can be found in [20]

333 **Lemma 4.** If $z = \text{Beta}(a, b)$:

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

338 In both cases convergence is in distribution.

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

342 *Proof:* Consider each distribution separately

- 343 1. For the logistic-normal re-parameterization $z_k = \sigma\left(\frac{\mu_k + \eta_k \xi}{\tau}\right)$, in the low temperature case
344 use Lemma 2 (b) by the always available representation $K = \frac{\mu}{\eta^2}$ and conclude by Lemma
345 1(b). In the high temperature case convergence is to the point $\pi = \mathcal{SB}(0.5, 0.5, \dots, 0.5)$.
- 346 2. For Kumaraswamy $z_k = \mathcal{K}(a_k, b_k)$ the argument is similar, but here the temperature can
347 only be defined implicitly through sequences of parameters (a_k, b_k) converging to either ∞
348 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
349 we conclude by Lemma 3(b) and Lemma 1(b). In the hig-temperature case we converge to
350 the point $SB(p_1, \dots, p_{k-1})$
- 351 3. For the Beta $z_k \sim \text{Beta}(\frac{a_k}{\tau}, \frac{b_k}{\tau})$ low-temperature leads to convergence to z_k Bernoulli with
352 parameter $a_k/(a_k + b_k)$ and we conclude from Lemma 4(b) and Lemma 1(b). For high
353 temperatures, convergence is to the point $SB(a_k/(a_k + b_k), \dots, a_{k-1}/(a_{k-1} + b_{k-1}))$.

354 A.2 Rounding

355 Here we have two extremes: at $\tau = 1$ we obtain a continuous distribution in the space (here, gaussian).
356 If $\tau = 0$ the resulting distribution has only atoms in the one-hot vectors p_n , in this proof assumed to

357 be the one-hot vectors. We show that in this case it is possible to represent any arbitrary categorical
 358 distribution through a judicious choice of the parameters.

359 **Proposition:** In the zero temperature case, i.e., $\pi = R^{\mathcal{P}}(\psi)$ it is possible to represent any arbitrary
 360 distribution i.e, for any α in the $N - 1$ simplex there exists gaussian parameters (μ, η) so that
 361 $P(\pi = p_n) = \alpha_n$. Points inside the simplex are realized directly, while distributions with some
 362 $\alpha_k = 0$ 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

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

363 Three conclusions are drawn from the above: first, we see that probabilities are ultimately gaussian
 364 integrals over a new partition, a translation of the Voronoi regions $V_n^{\mathcal{P}}$. Second, the map
 365 $(\mu_1, \dots, \mu_n) \xrightarrow{m} (\alpha_1, \dots, \alpha_n)$ is continuous by virtue of the dominated convergence theorem [4]:
 366 indeed, if $\mu_i \rightarrow \mu$ then $(\alpha_1^i, \dots, \alpha_n^i) \rightarrow (\alpha_1, \dots, \alpha_n)$, as the α_n^i are integrals that can be expressed
 367 using indicator functions in the integrands (which are all bounded by the integrable gaussian density),
 368 and as pointwise convergence of the indicators holds because of the continuity of the translation
 369 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
 370 $\alpha_m = (1 - q)/(n - 1)$. Indeed, by moving μ_n between $-\infty$ and ∞ while keeping the other μ_k fixed
 371 then $V_n^{\mathcal{P}} - \mu$ fluctuates between the empty set ($\alpha_n = 0$) and the entire space ($\alpha_n = 1$). Therefore, by
 372 the continuity of m and the intermediate value theorem, every value in $\alpha_n \in (0, 1)$ is realized, and by
 373 symmetry the other α_k occupy the remaining mass uniformly.

374 To conclude, we use the fact that the image of a convex set through a continuous function is also
 375 convex [26]. We also showed that for each tolerance ϵ the image of m contains these ϵ -one-hot
 376 vectors: that is, the points with $(1 - \epsilon)$ in one coordinate and $\epsilon/(n - 1)$ in the rest. Then, given a
 377 point in the interior of the simplex choose ϵ small enough so it is in the convex hull of the ϵ -one
 378 vectors. Then, by the above theorem there must be a pre-image μ that realizes this point.

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

382 B 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:

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

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

$$q(F(\psi); \theta) = p(F^{-1}(\pi); \theta) |DF^{-1}(\pi)| \tag{40}$$

$$= p(\psi; \theta) |DF(\psi)|^{-1}. \tag{41}$$

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

$$\mathbb{E}_{p(\xi)} [-\log q(F(g(\theta, \xi)); \theta)] = \mathbb{E}_{p(\psi; \theta)} [-\log p(\psi; \theta)] + \mathbb{E}_{p(\psi; \theta)} [\log |DF(\psi)|] \tag{42}$$

$$= Entropy(\psi; \theta) + E_{p(\xi)} [\log |DF(g(\theta, \xi))|]. \tag{43}$$

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

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

389 lower triangular, and its determinant is the product of its diagonal,

$$\left| \frac{d\Psi}{d\Pi} \right| = \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{\partial \psi_{ij}}{\partial \pi_{ij}} \quad (44)$$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{\partial}{\partial \pi_{ij}} \sigma^{-1} \left(\frac{\pi_{ij} - \ell_{ij}}{u_{ij} - \ell_{ij}} \right) \quad (45)$$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \left(\frac{1}{u_{ij} - \ell_{ij}} \right) \left(\frac{u_{ij} - \ell_{ij}}{\pi_{ij} - \ell_{ij}} \right) \left(\frac{u_{ij} - \ell_{ij}}{u_{ij} - \pi_{ij}} \right) \quad (46)$$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{u_{ij} - \ell_{ij}}{(\pi_{ij} - \ell_{ij})(u_{ij} - \pi_{ij})} \quad (47)$$

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

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

$$\Pi = f(\Psi) \quad (49)$$

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

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

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