

---

# Reparameterizing the Birkhoff Polytope for Variational Permutation Inference: Supplementary Material

---

Anonymous Authors  
Anonymous Institutions

## A Alternative methods of discrete variational inference

Recently there have been a number of proposals for extending the reparameterization trick [Rezende et al., 2014, Kingma and Welling, 2014] to high dimensional discrete problems<sup>1</sup> 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: if  $x \in \{0, 1\}^N$  is a one-hot vector drawn from a categorical distribution, then the support of  $p(x)$  is the set of vertices of the  $N - 1$  dimensional simplex. We can represent the distribution of  $x$  as an atomic density on the simplex.

### A.1 The Gumbel-softmax method

Viewing  $x$  as a vertex of the simplex motivates a natural relaxation: rather than restricting ourselves to atomic measures, consider continuous densities on the simplex. To be concrete, suppose the density of  $x$  is defined by the transformation,

$$\begin{aligned} \xi_n &\stackrel{\text{iid}}{\sim} \text{Gumbel}(0, 1) \\ \psi_n &= \log \theta_n + \xi_n \\ x &= \text{softmax}(\psi/\tau) \\ &= \left( \frac{e^{\psi_1/\tau}}{\sum_{n=1}^N e^{\psi_n/\tau}}, \dots, \frac{e^{\psi_N/\tau}}{\sum_{n=1}^N e^{\psi_n/\tau}} \right). \end{aligned}$$

The output  $x$  is now a point on the simplex, and the parameter  $\theta = (\theta_1, \dots, \theta_N)$  can be optimized via stochastic gradient ascent with the reparameterization trick.

The Gumbel distribution leads to a nicely interpretable model: when  $\theta$  is a probability mass function, adding Gumbel noise and taking the argmax yields an exact sample from  $\theta$ ; the softmax is a natural relaxation. As

---

<sup>1</sup>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 temperature  $\tau$  goes to zero, the softmax converges to the argmax function. Ultimately, however, this is just a continuous relaxation of an atomic density to a continuous density.

Stick-breaking and rounding offer two alternative ways of conceiving a relaxed version of a discrete random variable, and both are amenable to reparameterization. However, unlike the Gumbel-Softmax, these relaxations enable extensions to more complex combinatorial objects, notably, permutations.

### A.2 Stick-breaking

The stick-breaking transformation to the Birkhoff polytope presented in the main text contains a recipe for stick-breaking on the simplex. In particular, as we filled in the first row of the doubly-stochastic matrix, we were transforming a real-valued vector  $\psi \in \mathbb{R}^{N-1}$  to a point in the simplex. We present this procedure for discrete variational inference again here in simplified form. Start with a reparameterization of a Gaussian vector,

$$\begin{aligned} \xi_n &\stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1), \\ \psi_n &= \mu_n + \eta_n \xi_n, \quad 1 \leq n \leq N - 1, \end{aligned}$$

parameterized by  $\theta = (\mu_n, \eta_n)_{n=1}^{N-1}$ . Then map this to the unit hypercube in a temperature-controlled manner with the logistic function,

$$z_n = \sigma(\psi_n/\tau),$$

where  $\sigma(u) = (1 + e^{-u})^{-1}$  is the logistic function. Finally, transform the unit hypercube to a point in the simplex:

$$\begin{aligned} x_1 &= z_1, \\ x_n &= z_n \left( 1 - \sum_{m=1}^{n-1} x_m \right), \quad 2 \leq n \leq N - 1, \\ x_N &= 1 - \sum_{m=1}^{N-1} x_m, \end{aligned}$$

Here,  $z_n$  is the fraction of the remaining “stick” of probability mass assigned to  $x_n$ . This transformation is invertible, the Jacobian is lower-triangular, and the determinant of the Jacobian is easy to compute. [Linderman et al. \[2015\]](#) compute the density of  $x$  implied by a Gaussian density on  $\psi$ .

The temperature  $\tau$  controls how concentrated  $p(x)$  is at the vertices of the simplex, and with appropriate choices of parameters, in the limit  $\tau \rightarrow 0$  we can recover any categorical distribution. In the other limit, as  $\tau \rightarrow \infty$ , 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, note that the logistic-normal construction is only one possible choice. We could instead let  $z_n \sim \text{Beta}(\frac{a_n}{\tau}, \frac{b_n}{\tau})$ . This would lead to the Dirichlet distribution on the simplex. The beta distribution is slightly harder to reparameterize since it is typically simulated with a rejection sampling procedure, but [Naesseth et al. \[2017\]](#) have shown how this can be handled with a mix of reparameterization and score-function gradients. Alternatively, the beta distribution could be replaced with the Kumaraswamy distribution, which is quite similar to the beta distribution but is easily reparameterizable.

### A.3 Rounding

Rounding transformations also have a natural analog for discrete variational inference. Define the rounding operator,

$$\text{round}(\psi) = \arg \min_{e_n} \|e_n - \psi\|^2,$$

which maps  $\psi \in \mathbb{R}^N$  to the one-hot vectors  $e_n$ ; i.e. the vectors in  $\{0, 1\}^N$  with  $n$ -th entry equal to one and all other entries equal zero. This is equivalent to defining  $\text{round}(\psi) = e_{n^*}$  where

$$\begin{aligned} n^* &= \arg \min_n \|e_n - \psi\|^2 \\ &= \arg \min_n \sum_{m \neq n} \psi_m^2 + (1 - \psi_n)^2 \\ &= \arg \min_n \sum_{m \neq n} \psi_m^2 + \psi_n^2 - 2\psi_n + 1 \\ &= \arg \min_n \|\psi\|^2 - 2\psi_n + 1 \\ &= \arg \max_n \psi_n. \end{aligned}$$

In the case of a tie, let  $n^*$  be the smallest index  $n$  such that  $\psi_n > \psi_m$  for all  $m < n$ . Rounding effectively partitions the space into  $N$  disjoint “Voronoi”

cells,

$$V_n = \left\{ \psi \in \mathbb{R}^N : \psi_n \geq \psi_m \forall m \wedge \psi_n > \psi_m \forall m < n \right\}.$$

By definition,  $\text{round}(\psi) = e_{n^*}$  for all  $\psi \in V_{n^*}$ .

We define a map that pulls points toward their rounded values,

$$x = \tau\psi + (1 - \tau)\text{round}(\psi). \quad (1)$$

**Proposition 1.** *For  $\tau \in [0, 1]$ , the map defined by (1) moves points strictly closer to their rounded values so that  $\text{round}(\psi) = \text{round}(x)$ .*

*Proof.* Note that the Voronoi cells are intersections of halfspaces and, as such, are convex sets. Since  $x$  is a convex combination of  $\psi$  and  $e_{n^*}$ , both of which belong to the convex set  $V_{n^*}$ ,  $x$  must belong to  $V_{n^*}$  as well.  $\square$

Similarly,  $x$  will be a point on the simplex if and only if  $\psi$  is on the simplex as well. By analogy to the rounding transformations for permutation inference, in categorical inference we use a Gaussian distribution  $\psi \sim \mathcal{N}(\text{proj}(m), H)$ , where  $\text{proj}(m)$  is the projection of  $m \in \mathbb{R}_+^N$  onto the simplex. Still, the simplex has zero measure under the Gaussian distribution. It follows that the rounded points  $x$  will almost surely not be on the simplex either. The supposition of this approach is that this is not a problem: relaxing to the simplex is nice but not required.

In the zero-temperature limit we obtain a discrete distribution on the vertices of the simplex. For  $\tau \in (0, 1]$  we have a distribution on  $\mathcal{X}_\tau \subseteq \mathbb{R}^N$ , the subset of the reals to which the rounding operation maps. (For  $0 \leq \tau < 1$  this is a strict subset of  $\mathbb{R}^N$ .) To derive the density  $q(x)$ , we need the inverse transformation and the determinant of its Jacobian. From Proposition 1, it follows that the inverse transformation is given by,

$$\psi = \frac{1}{\tau}x - \frac{1 - \tau}{\tau}\text{round}(x).$$

As long as  $\psi$  is in the interior of its Voronoi cell, the  $\text{round}$  function is piecewise constant and the Jacobian is  $\frac{\partial \psi}{\partial x} = \frac{1}{\tau}I$ , and its determinant is  $\tau^{-N}$ . Taken together, we have,

$$\begin{aligned} q(x; m, H) &= \\ &\tau^{-N} \mathcal{N}\left(\frac{1}{\tau}x - \frac{1 - \tau}{\tau}\text{round}(x); \text{proj}(m), H\right) \\ &\quad \times \mathbb{I}[x \in \mathcal{X}_\tau]. \end{aligned}$$

Compare this to the density of the rounded random variables for permutation inference.

## B Limit analysis for stick-breaking

We show that stick-breaking for discrete variational inference can converge to any categorical distribution in the zero-temperature limit. We do so with a sequence of propositions: first we show that in the zero-temperature limit, the distribution of  $\sigma(\psi_n/\tau)$  converges to a Bernoulli distribution. Then we show that when  $\sigma(\psi_n/\tau)$  is Bernoulli (rather than a continuous density on the unit interval), the distribution on  $x$  obtained by applying the stick-breaking transformation to  $\psi$  is categorical.

**Proposition 2.** *Let  $z = \sigma(\psi/\tau)$  with  $\psi \sim \mathcal{N}(\mu, \eta^2)$ . In the limit  $\tau \rightarrow 0$  we have  $z \sim \text{Bern}(\Phi(-\frac{\mu}{\eta}))$ , where  $\Phi(\cdot)$  denotes the Gaussian cumulative distribution function (cdf).*

*Proof.* Let  $F_z$  be the cdf of the random variable  $z$ . Since  $z$  is a random variable on the unit interval,  $F_z$  is a non-decreasing function on  $[0, 1]$  with  $F_z(0) = 0$  and  $F_z(1) = 1$ . Reparameterize  $\psi = \mu + \eta\xi$  where  $\xi \sim \mathcal{N}(0, 1)$ . Then we have,

$$\begin{aligned} F_z(u) &= \Pr(\sigma(\psi/\tau) < u) \\ &= \Pr(\psi < \tau\sigma^{-1}(u)) \\ &= \Pr(\xi < \frac{\tau}{\eta}\sigma^{-1}(u) - \frac{\mu}{\eta}) \\ &= \Phi(-\frac{\tau}{\eta}\sigma^{-1}(u) - \frac{\mu}{\eta}). \end{aligned}$$

By the continuity of  $\Phi$  we have,

$$\lim_{\tau \rightarrow 0} F_z(u) = \Phi(-\frac{\mu}{\eta}) \quad \text{for } u \in (0, 1).$$

This is the cdf of a Bernoulli random with probability  $\rho = \Phi(-\frac{\mu}{\eta})$ .  $\square$

**Proposition 3.** *As above, let  $z_n = \sigma(\psi_n/\tau)$ . When  $z_n \sim \text{Bern}(\rho_n)$  with  $\rho_n \in [0, 1]$  for  $n = 1, \dots, N$ , the random variable  $x$  obtained from applying the stick-breaking transformation to  $z$  will have an atomic distribution with atoms in the vertices of  $\Delta_N$ ; i.e.,  $x \sim \text{Cat}(\pi)$  where*

$$\begin{aligned} \pi_1 &= \rho_1 \\ \pi_n &= \rho_n \prod_{m=1}^{n-1} (1 - \rho_m) \quad n = 2, \dots, N-1, \\ \pi_N &= \prod_{m=1}^{N-1} (1 - \rho_m). \end{aligned}$$

*Proof.* From the stick-breaking definition,  $x_1 = z_1$ ,  $x_n = z_n(1 - \sum_{m < n} x_m)$ , and  $x_N = 1 - \sum_{m < N} x_m$ . When  $z_n \in \{0, 1\}$  for all  $n = 1, \dots, N-1$ , we have the following equivalencies. For the first element,

$$x_1 = 1 \iff z_1 = 1;$$

for  $1 < n < N-1$ :

$$x_n = 1 \iff (z_n = 1) \bigwedge_{m=1}^{n-1} (z_m = 0);$$

and for the last element,

$$x_N = 1 \iff \bigwedge_{m=1}^{N-1} (z_m = 0).$$

These events are mutually exclusive, implying that  $x$  will necessarily be a one-hot vector, i.e. a categorical random variable. Since  $z_1, \dots, z_{N-1}$  are independent Bernoulli random variables, the probabilities of these events are given by the  $\pi, \dots, \pi_N$  stated in the proposition.  $\square$

These two propositions, combined with the invertibility of the stick-breaking procedure, lead to our main result.

**Lemma 1.** *In the zero-temperature limit, stick-breaking of logistic-normal random variables can realize any categorical distribution on  $x$ .*

*Proof.* There is a one-to-one correspondence between  $\pi \in \Delta_N$  and  $\rho \in [0, 1]^{N-1}$ . Specifically,

$$\begin{aligned} \rho_1 &= \pi_1 \\ \rho_n &= \frac{\pi_n}{\prod_{m=1}^{n-1} (1 - \rho_m)} \quad \text{for } n = 2, \dots, N-1. \end{aligned}$$

Since these are recursively defined, we can substitute the definition of  $\rho_m$  to obtain an expression for  $\rho_n$  in terms of  $\pi$  only. Thus, by Proposition 3, any desired categorical distribution  $\pi$  implies a set of Bernoulli parameters  $\rho$ . From Proposition 2, in the zero temperature limit, any desired  $\rho_n$  can be obtained with appropriate choice of Gaussian mean  $\mu_n$  and variance  $\eta_n^2$ . Thus, stick-breaking can realize any categorical distribution when  $\tau \rightarrow 0$ .  $\square$

## C Variational Autoencoders (VAE) with categorical latent variables

We considered the density estimation task on MNIST digits, as in Maddison et al. [2016], Jang et al. [2016], where observed digits are reconstructed from a latent discrete code. We used the continuous ELBO for training, and evaluated performance based on the marginal likelihood, estimated through the multi-sample variational objective of the discretized model. We compared against the methods of Jang et al. [2016], Maddison et al. [2016] and obtained the results in Table 1. While stick-breaking and rounding fare slightly worse than

the Gumbel-softmax method, they are readily extensible to more complex discrete objects, as shown in the main paper.

**Table 1:** Summary of results in VAE

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

Figure 1 shows MNIST reconstructions using Gumbel-Softmax, stick-breaking and rounding reparameterizations. In all the three cases reconstructions are reasonably accurate, and there is diversity in reconstructions.

## D Variational permutation inference details

Here we discuss more of the subtleties of variational permutation inference and present the mathematical derivations in more detail.

### D.1 Continuous prior distributions.

Continuous relaxations require re-thinking 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 relaxed discrete latent variables, here, over relaxations of permutation matrices. Moreover, it is critical to design sensible priors for relaxed permutations. Ideally, this prior should penalize values of  $X$  that are far from permutation matrices.

For our categorical experiment on MNIST we use a mixture of Gaussians around each vertex,  $p(x) = \frac{1}{N} \sum_{n=1}^N \mathcal{N}(x | e_k, \eta^2)$ . This can be extended to permutations, where we use a mixture of Gaussians for each dimension,

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

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

### D.2 Computing the ELBO

Here we show how to evaluate the ELBO. Note that the stick-breaking and rounding transformations are compositions of invertible functions,  $g_\tau = h_\tau \circ f$  with  $\Psi = f(\Xi; \theta)$  and  $X = h_\tau(\Psi)$ . In both cases,  $f$  takes in

a matrix of independent standard Gaussians ( $\Xi$ ) and transforms it with the means and variances in  $\theta$  to output a matrix  $\Psi$  with entries  $\psi_{mn} \sim \mathcal{N}(\mu_{mn}, \eta_{mn}^2)$ . Stick-breaking and rounding differ in the temperature-controlled transformations  $h_\tau(\Psi)$  they use to map  $\Psi$  toward the Birkhoff polytope.

To evaluate the ELBO, we must compute the density of  $q_\tau(X; \theta)$ . Let  $J_h(u) = \frac{\partial h(U)}{\partial U} \big|_{U=u}$  denote the Jacobian of a function  $h$  evaluated at value  $u$ . By the change of variables theorem and properties of the determinant,

$$\begin{aligned} q_\tau(X; \theta) &= p(h_\tau^{-1}(X); \theta) \times |J_{h_\tau^{-1}}(X)| \\ &= p(h_\tau^{-1}(X); \theta) \times |J_{h_\tau}(h_\tau^{-1}(X))|^{-1}. \end{aligned}$$

Now we appeal to the law of the unconscious statistician to compute the entropy of  $q_\tau(X; \theta)$ ,

$$\begin{aligned} \mathbb{E}_{q_\tau(X; \theta)} [-\log q(X; \theta)] &= \mathbb{E}_{p(\Psi; \theta)} [-\log p(\Psi; \theta) + \log |J_{h_\tau}(\Psi)|] \\ &= \mathbb{H}(\Psi; \theta) + \mathbb{E}_{p(\Psi; \theta)} [\log |J_{h_\tau}(\Psi)|]. \end{aligned} \quad (3)$$

Since  $\Psi$  consists of independent Gaussians, the entropy is simply,

$$\mathbb{H}(\Psi; \theta) = \frac{1}{2} \sum_{m,n} \log(2\pi e \eta_{mn}^2).$$

We estimate the second term of equation (3) using Monte-Carlo samples. For both transformations, the Jacobian has a simple form.

**Jacobian of the rounding transformation.** The rounding transformation is given in matrix form in the main text, and we restate it here in coordinate-wise form for convenience,

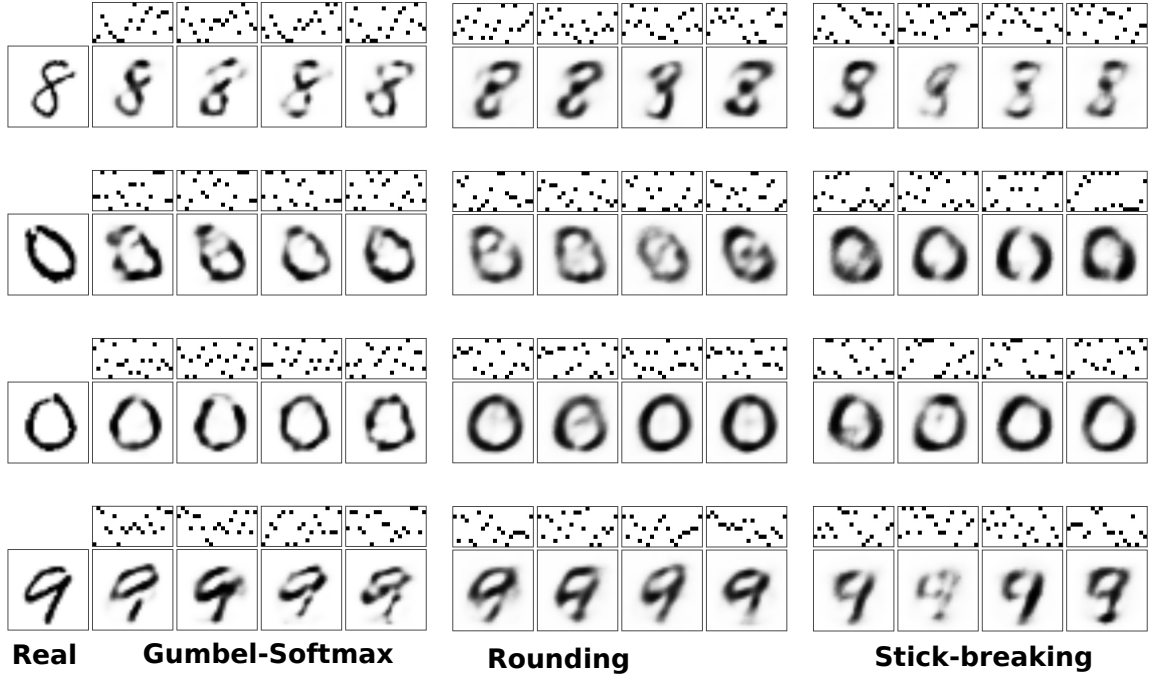
$$x_{mn} = [h_\tau(\Psi)]_{mn} = \tau \psi_{mn} + (1 - \tau) [\text{round}(\Psi)]_{mn}.$$

This transformation is piecewise linear with jumps at the boundaries of the ‘‘Voronoi cells,’’ i.e., the points where  $\text{round}(X)$  changes. The set of discontinuities has Lebesgue measure zero so the change of variables theorem still applies. Within each Voronoi cell, the rounding operation is constant, and the Jacobian is,

$$\log |J_{h_\tau}(\Psi)| = \sum_{m,n} \log \tau = N^2 \log \tau.$$

For the rounding transformation with given temperature, the Jacobian is constant.

**Jacobian of the stick-breaking transformation.** Here  $h_\tau$  consists of two steps: map  $\Psi \in \mathbb{R}^{N-1 \times N-1}$



**Figure 1:** Examples of true and reconstructed digits from their corresponding random codes using with  $K = 20$  categorical variables with  $N = 10$  possible values.

to  $Z \in [0, 1]^{N-1 \times N-1}$  with a temperature-controlled, elementwise logistic function, then map  $Z$  to  $X \in \mathcal{B}_N$  with the stick-breaking transformation.

As with the standard stick-breaking transformation to the simplex, our transformation to the Birkhoff polytope is feed-forward; i.e. to compute  $x_{mn}$  we only need to know the values of  $z$  up to and including the  $(m, n)$ -th entry. Consequently, the Jacobian of the transformation is triangular, and its determinant is simply the product of its diagonal.

We derive an explicit form in two steps. With a slight abuse of notation, note that the Jacobian of  $h_\tau(\Psi)$  is given by the chain rule,

$$J_{h_\tau(\Psi)} = \frac{\partial X}{\partial \Psi} = \frac{\partial X}{\partial Z} \frac{\partial Z}{\partial \Psi}.$$

Since both transformations are bijective, the determinant is,

$$|J_{h_\tau(\Psi)}| = \left| \frac{\partial X}{\partial Z} \right| \left| \frac{\partial Z}{\partial \Psi} \right|.$$

the product of the individual determinants. The first determinant is,

$$\left| \frac{\partial X}{\partial Z} \right| = \prod_{m=1}^{N-1} \prod_{n=1}^{N-1} \frac{\partial x_{mn}}{\partial z_{mn}} = \prod_{m=1}^{N-1} \prod_{n=1}^{N-1} (u_{mn} - \ell_{mn}).$$

The second transformation, from  $\Psi$  to  $Z$ , is an element-wise, temperature-controlled logistic transformation

such that,

$$\begin{aligned} \left| \frac{\partial Z}{\partial \Psi} \right| &= \prod_{m=1}^{N-1} \prod_{n=1}^{N-1} \frac{\partial z_{mn}}{\partial \psi_{mn}} \\ &= \prod_{m=1}^{N-1} \prod_{n=1}^{N-1} \frac{1}{\tau} \sigma(\psi_{mn}/\tau) \sigma(-\psi_{mn}/\tau). \end{aligned}$$

double check

It is important to note that the transformation that maps  $Z \rightarrow X$  is only piecewise continuous: the function is not differentiable at the points where the bounds change; for example, when changing  $Z$  causes the active upper bound to switch from the row to the column constraint or vice versa.

conclude this section.

## E Experiment details

We used Tensorflow [Abadi et al., 2016] for the VAE experiments, slightly changing the code made available from Jang et al. [2016]. For experiments on synthetic matching and the C. elegans example we used Autograd [Maclaurin et al., 2015], explicitly avoiding propagating gradients through the non-differentiable round operation, which requires solving a matching problem.

We used the ADAM with learning rate 0.1 for opti-

mization. For rounding, the parameter vector  $H$  defined in ??(iii) was constrained to lie in the interval  $[0.1, 0.5]$ . Also, for rounding, we used ten iterations of the Sinkhorn-Knopp algorithm, to obtain points in the Birkhoff polytope. For stick-breaking the variances  $\nu$  defined in ?? were constrained between  $1e-8$  and  $1.0$ . In either case, the temperature, along with maximum values for the noise variances were calibrated using a grid search.

In the C. elegans example we considered the symmetrized version of the adjacency matrix described in [Varshney et al., 2011]; i.e. we used  $A' = (A + A^\top)/2$ , and the matrix  $W$  was chosen antisymmetric, with entries sampled randomly with the sparsity pattern dictated by  $A'$ . To avoid divergence, the matrix  $W$  was then re-scaled by 1.1 times its spectral radius. This choice, although not essential, induced a reasonably well behaved linear dynamical system, rich in non-damped oscillations. We used a time window of  $T = 1000$  time samples, and added spherical standard noise at each time.

## References

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.
- E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.
- 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.
- 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.
- D. Maclaurin, D. Duvenaud, and R. P. Adams. Auto-grad: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, 2015.
- C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.
- 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.
- 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.
- 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.