# 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[1] by relaxing them to analogous continuous problems [Maddison et al., 2016, Jang et al., 2016, Kusner and Hernández-Lobato, 2016]. These approaches are based on the following observation: 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,

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

The output $x$ is now a point on the simplex, and the parameter $\theta = (\theta_1, \ldots, \theta_N)$ 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 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,

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

parameterized by $\theta = (\mu_n, \eta_n)_{n=1}^{N-1}$. Then map this to a point in the simplex:

$$\begin{aligned}
x_1 &= \sigma(\psi_1/\tau), \\
x_n &= \sigma(\psi_n/\tau) \left( 1 - \sum_{m=1}^{n-1} x_m \right), \qquad 2 \leq n \leq N-1, \\
x_N &= 1 - \sum_{m=1}^{N-1} x_m,
\end{aligned}$$

where $\sigma(u) = (1 + e^{-u})^{-1}$ is the logistic function. Here, $\sigma(\psi_n/\tau)$ is the fraction of the remaining "stick" of probability mass assigned to $x_n$. This procedure is invertible, the Jacobian $\frac{\mathrm{d}x}{\mathrm{d}\psi}$ 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 $\psi$.

[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 temperature $\tau$ controls how concentrated $p(\pi)$ is at the vertices of the simplex, and with appropriate choices of parameters, in the limit $\tau \to 0$ we can recover any categorial distribution. In the other limit, 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, note that the logistic-normal construction only one possible choice. We could instead let $\psi_n \sim \mathrm{Beta}(\frac{a_n}{\tau}, \frac{b_n}{\tau})$ and $x_n = \psi_n$. 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,

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

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 $\mathsf{round}(\psi) = e_{n^*}$ 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,

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

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

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

$$x = \tau \psi + (1-\tau)\mathsf{round}(\psi). \qquad (1)$$

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

*Proof.* Note that the Voronoi cells are defined by linear inequalities, making them 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. □

Similarly, $x$ will be a point on the simplex if an 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 $\psi \sim \mathcal{N}(\mathsf{proj}(m), H)$, where $\mathsf{proj}(m)$ 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,

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

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

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

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

## B Limit analysis for stick-breaking

Here we state and prove that for the stick-breaking we consider here, we can arrive to either arbitrary points in the i) simplex or ii) to any categorical distribution as limiting cases (in the temperature). 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 $\pi$ 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 $\pi$ having an atomic distribution with atoms in the vertices of $\Delta^{k-1}$; i.e, $\pi$ is categorical. We have the following expression for the probabilities of the atoms $\pi_k = 1$ (one hot vectors):

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

   Moreover, if for each index $k$ any parameter of the Bernoulli variable $z_k$ can be realized through appropriate choice of $\theta$, 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 (2):

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

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 $\mu$ 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 $\mu_n$ and study the cdf $F$ of $z_n$ on the interval (0,1) (it evaluates zero below zero and one above one).

$$\begin{aligned} F_{z_n}(x) &= P(\sigma(\psi_n) < x) && (2)\\ &= P(\psi_n < \sigma^{-1}(x)) && (3)\\ &= P(\mu_n + \mu_n/K\xi < \sigma^{-1}(x)), && (4)\\ &= P(\xi < \sigma^{-1}(x)K/\mu_n - K) && (5)\\ &= \Phi(\sigma^{-1}(x)K/\mu_n - K) && (6)\end{aligned}$$

Therefore, by continuity of $\Phi$ 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 $\mu$ instead.

**Proposition.** For the stick-breaking construction, any arbitrary distribution can be realized in the low-temperature limit. Also, in the high-temperature limit convergence is to certain point(s) in the interior of the simplex.

*Proof*: Consider each distribution separately

We have $z_k = \sigma\left(\frac{\mu_k + \eta_k \xi}{\tau}\right)$, 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 $\pi = \mathcal{SB}(0.5, 0.5, \ldots, 0.5)$.

## 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], finding similar (although slightly worse) results (Table 1). This difference may be interpreted as the price to be paid in order to enable an extension of a relaxed distribution over categories, to permutations.

### MNIST reconstructions

In figure 1 we show some MNIST reconstructions using Gumbel-Softmax, stick-breaking and rounding repa-

**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 |

rameterizations. In all the three cases reconstructions are reasonably accurate, and there is diversity in reconstructions.

# D Variational inference for Permutation details

### Continuous prior distributions.

Continuous relaxations requires re-thinking of the objective. As in Maddison et al. [2016], we maximize a relaxed ELBO, for which we need to specify a new continuous prior $p(x)$ over the latent variables. Moreover, it is critical to conceive sensible priors for permutations, that could serve in a variational inference routine to penalize configurations that are away from permutation matrices (i.e. close to the barycenter of the Birkhoff polytope).

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 use a mixture of Gaussians for each dimension,

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

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

### Deriving an expression for the ELBO

Here we show that if $X = G(\Xi;\theta)$ with $G$ differentiable one can evaluate the second term in equation (**??**) [2]. Moreover, here, to exploit the similarities between both methods (stick-breaking and rounding), we further factor $G$ into two functions: $G = H \circ F$, $X = H(\Psi)$ and $\Psi = F(\Xi;\theta)$ (both $H, F$ invertible). This means all dependency of $X$ in the parameters is through $\Psi$. Under this assumption, and denoting $\Psi \sim p(\Psi,\theta)$, the second term in equation (**??**) (without gradient) can be computed as: then

$$\begin{aligned}\mathbb{E}_{r(\Xi)}\left[-\log q(G(\Xi;\theta));\theta)\right] =& \mathbb{H}(\Psi;\theta) + \\ & E_{r(\Xi)}\left[\log|DH(F(\Xi,\theta))|\right].\end{aligned}$$

**Proof**: Indeed, first, it is obvious that

$$\mathbb{E}_{r(\Xi)}\left[-\log q(G(\Xi;\theta));\theta)\right] = \mathbb{E}_{r(\Xi)}\left[-\log q(H(F(\Xi,\theta));\theta)\right]$$

Then, by the 'Law of the Unconscious Statistician' we have:

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

Now, by the change of variable theorem and derivative and determinant inversion rules, we obtain ($D$ means the Jacobian, the matrix of derivatives) :

$$\begin{aligned}q(H(\Psi);\theta) &= p(H^{-1}(X);\theta)|DH^{-1}(X)| \\ &= p(\Psi;\theta)|DH(\Psi)|^{-1}.\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}_{r(\Xi)}\left[-\log q(G(\Xi;\theta));\theta)\right] &= \mathbb{E}_{p(\Psi;\theta)}\left[-\log p(\Psi;\theta)\right] + \\ &\quad \mathbb{E}_{p(\psi;\theta)}\left[\log|DH(\psi)|\right] \\ &= \mathbb{H}(\Psi;\theta) + \\ &\quad E_{r(\xi)}\left[\log|DH(F(\Xi;\theta))|\right].\end{aligned} \tag{7}$$

### Estimating the ELBO

Here we describe how to compute each of terms of equation (7), needed for ELBO computations. First, as $\Psi$ is Gaussian for both rounding and stick-breaking, the entropy term is straightforward and equal to $N\log(\eta^2 2\pi e)/2$ ($\eta$ may depend on the temperature and depends on the method).

Notice that to state $\Psi$ is Gaussian in the stick-breaking case we slightly deviate from **??**. Specifically, here we call $\Psi = \frac{\mu_{mn}+\eta_{mn}\Xi_{mn}}{\tau}$ and define $\Psi' = \sigma(\Psi)$.

The second term of equation (7) is estimated using Monte-Carlo samples, and its derivation depends on the method.

### Rounding

Here $H$ is piecewise linear: the set of discontinuities (border of the 'Voronoi cells' associated to each permutation) has Lebesgue measure zero. So we can still apply the change of variables theorem. Therefore, $\log|DH(F(\Xi;\theta))| = N\log\tau$. This means we don't even need to take samples to compute this term.

### Stick-breaking

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

Notice that these bounds only depend on values of $X$ that have already been computed; i.e., those that are

[2] Notice that we uppercase the variables in (**??**) this is in consistency to our notation in section **??**

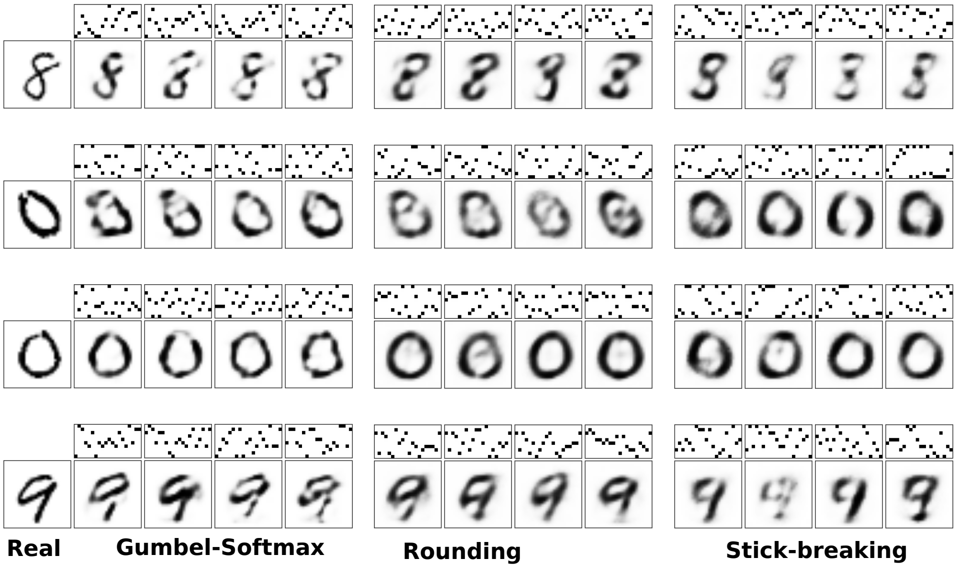

**Figure 1:** Examples of true and reconstructed digits from their corresponding random codes using with $N = 20$ categorical variables with $K = 10$ possible values.

above or to the left of the $(i, j)$-th entry. Thus, the transformation from $\Psi'$ to $X$ is feed-forward according to this ordering. Consequently, the Jacobian of the inverse transformation $H^{-1}$ , $\mathrm{d}\Psi'/\mathrm{d}X$, is lower triangular, and its determinant is the product of its diagonal,

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

To compute the gradient of the forward transformation $H$ one simply needs to invert the above (or put a negative sign, in the logarithm scale). Finally, to incorporate the effect of $\sigma$ ($\Psi' = \sigma(\Psi)$, by the chain rule, one only needs to add a term corresponding to this derivative, $d\sigma(x)/dx = \sigma(x)\sigma(-x)$.

### Experiment details

Experiments were run on a High Performance Computing (HPC) cluster, allowing the execution of hundreds of processes in parallel to efficiently determine best hyperparameter configurations.

For experiments with Variational Auto-encoder we used Tensorflow [Abadi et al., 2016], slightly changing the code made available in conjunction with **?**. For experiments on synthetic matching and the C. elegans example we used Autograd [**?**], explicitly avoiding propagating gradients through the non-differentiable operation of solving a matching problem (the round in **??**).

In all experiment we used the ADAM as optimizer, with learning rate 0.1. 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 **??** was constrained between $1e{-}8$ and 1.0. In either case, the temperature parameter was 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.

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.