# Supplementary Material

**Scott W. Linderman*** Columbia University

**Gonzalo E. Mena*** Columbia University

**Hal Cooper** Columbia University

**Liam Paninski** Columbia University

**John P. Cunningham** Columbia University

## A Alternative methods of discrete variational inference

We can gain insight and intuition about the stick-breaking and rounding transformations by considering their counterparts for discrete, or categorical, variational inference. Continuous relaxations are an appealing approach for this problem, affording gradient-based inference with the reparameterization trick. First we review the Gumbel-softmax method [Maddison et al., 2017, Jang et al., 2017, Kusner and Hernández-Lobato, 2016]—a recently proposed method for discrete variational inference with the reparameterization trick—then we discuss analogs of our permutation and rounding transformations for the categorical case. These can be considered alternatives to the Gumbel-softmax method, which we compare empirically in Appendix A.5.

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., 2017, Jang et al., 2017, 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,

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

The output $x$ is now a point on the simplex, and the parameter $\theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}_+^N$ can be optimized via stochastic gradient ascent with the reparameterization trick.

The Gumbel distribution leads to a nicely interpretable model: adding i.i.d. Gumbel noise to $\log \theta$ and taking the argmax yields an exact sample from the normalized probability mass function $\bar{\theta}$, where $\bar{\theta}_n = \theta_n / \sum_{m=1}^N \theta_m$ [Gumbel, 1954]. 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 constructing 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

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

vector,

$$z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1),$$
$$\psi_n = \mu_n + \nu_n z_n, \qquad 1 \le n \le N-1,$$

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

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

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

$$x_1 = \beta_1,$$
$$x_n = \beta_n \left(1 - \sum_{m=1}^{n-1} x_m\right), \qquad 2 \le n \le N-1,$$
$$x_N = 1 - \sum_{m=1}^{N-1} x_m,$$

Here, $\beta_n$ 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 $\tau \to 0$ we can recover any categorical distribution (we will discuss this in detail in Section A.4. 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 is only one possible choice. We could instead let $\beta_n \sim \text{Beta}(\frac{a_n}{\tau}, \frac{b_n}{\tau})$. This would lead to a generalized 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 [Kumaraswamy, 1980], 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. Let $e_n$ denote a one-hot vector with $n$-th entry equal to one. Define the rounding operator,

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

where

$$n^* = \arg\min_n \|e_n - \psi\|^2$$
$$= \arg\max_n \psi_n.$$

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,

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

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

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

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

**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 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. □

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 $\psi \sim \mathcal{N}(\mathsf{proj}(m), \nu)$, 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 $0 \le \tau < 1$ 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 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,

$$q(x; m, \nu) = \tau^{-N} \mathcal{N}\left(\frac{1}{\tau} x - \frac{1-\tau}{\tau} \mathsf{round}(x);\, \mathsf{proj}(m), \mathrm{diag}(\nu)\right) \times \mathbb{I}[x \in \mathcal{X}_\tau].$$

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

### A.4 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.

Let $\beta = \sigma(\psi/\tau)$ with $\psi \sim \mathcal{N}(\mu, \nu^2)$. In the limit $\tau \to 0$ we have $\beta \sim \mathrm{Bern}(\Phi(-\frac{\mu}{\nu}))$, where $\Phi(\cdot)$ denotes the Gaussian cumulative distribution function (cdf). Moreover, when $\beta_n \sim \mathrm{Bern}(\rho_n)$ with $\rho_n \in [0, 1]$ for $n = 1, \ldots, N$, the random variable $x$ obtained from applying the stick-breaking transformation to $\beta$ will have an atomic distribution with atoms in the vertices of $\Delta_N$; i.e, $x \sim \mathrm{Cat}(\pi)$ where

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

These two facts, combined with the invertibility of the stick-breaking procedure, lead to the following proposition

**Proposition 2.** *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,

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

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, any desired categorical distribution $\pi$ implies a set of Bernoulli parameters $\rho$. In the zero temperature limit, any desired $\rho_n$ can be obtained with appropriate choice of Gaussian mean $\mu_n$ and variance $\nu_n^2$. Together these imply that stick-breaking can realize any categorical distribution when $\tau \to 0$. □

### A.5 Variational Autoencoders (VAE) with categorical latent variables

We considered the density estimation task on MNIST digits, as in Maddison et al. [2017], Jang et al. [2017], 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 with the variational objective of the discretized model. We compared against the methods of Jang et al. [2017], Maddison et al. [2017] 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.

## B Variational permutation inference details

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

### B.1 Continuous prior distributions.

Continuous relaxations require re-thinking the objective: the model log-probability is defined with discrete latent variables, but our relaxed posterior is a continuous density. As in Maddison et al. [2017], we instead maximize a relaxed ELBO. We assume the functional form of the likelihood remains unchanged, and simply accepts continuous values instead of discrete. However, we need to specify a new continuous prior $p(X)$ over the relaxed discrete latent variables, here, over relaxations of permutation matrices. It is important that the prior be sensible: ideally, the 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

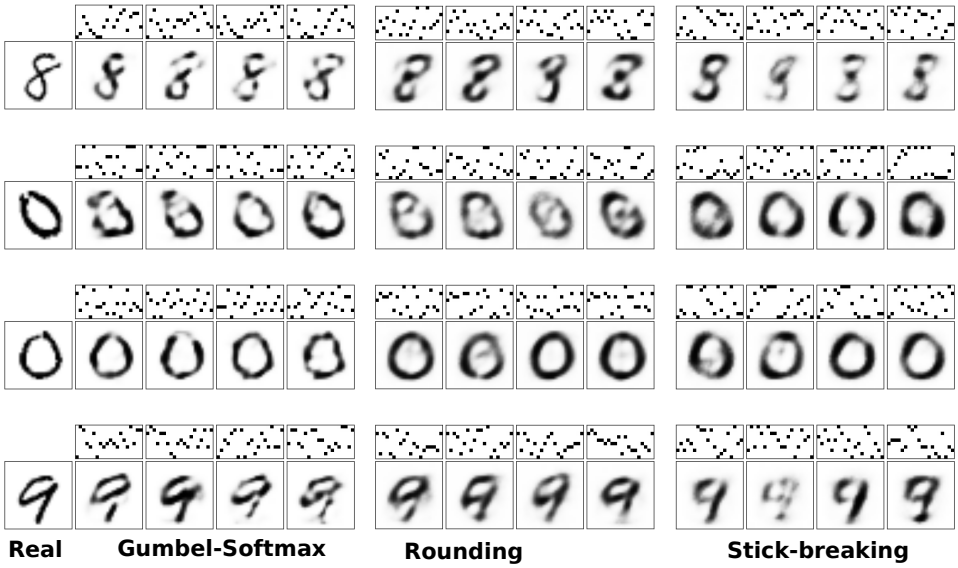


**Figure 1:** *Examples of true and reconstructed digits from their corresponding discrete latent variables.* The real input image is shown on the left, and we show sets of four samples from the posterior predictive distribution for each discrete variational method: Gumbel-softmax, rounding, and stick-breaking. Above each sample we show the corresponding sample of the discrete latent "code." The random codes consist of of $K = 20$ categorical variables with $N = 10$ possible values each. The codes are shown as $10 \times 20$ binary matrices above each image.

Gaussians for each coordinate,

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

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

### B.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 $\Psi = f(z; \theta)$ and $X = h_\tau(\Psi)$. In both cases, $f$ takes in a matrix of independent standard Gaussians ($z$) and transforms it with the means and variances in $\theta$ to output a matrix $\Psi$ with entries $\psi_{mn} \sim \mathcal{N}(\mu_{mn}, \nu^2_{mn})$. 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 $J_{h_\tau}(u) = \frac{\partial h_\tau(U)}{\partial U}\big|_{U=u}$ denote the Jacobian of a function $h_\tau$ evaluated at value $u$. By the change of variables theorem and properties of the determinant,

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

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

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

Since $\Psi$ consists of independent Gaussians with variances $\nu^2_{mn}$, the entropy is simply,

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

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

**Jacobian of the stick-breaking transformation.** Here $h_\tau$ consists of two steps: map $\Psi \in \mathbb{R}^{N-1 \times N-1}$ to $B \in [0, 1]^{N-1 \times N-1}$ with a temperature-controlled, elementwise logistic function, then map $B$ to $X$ in the Birkhoff polytope 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 $\beta$ 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,

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

Since both transformations are bijective, the determinant is,

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

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

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

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

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

It is important to note that the transformation that maps $B \to X$ is only piecewise continuous: the function is not differentiable at the points where the bounds change; for example, when changing $B$ causes the active upper bound to switch from the row to the column constraint or vice versa. In practice, we find that our stochastic optimization algorithms still perform reasonably in the face of this discontinuity.

**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,

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

This transformation is piecewise linear with jumps at the boundaries of the "Voronoi cells;" i.e., the points where $\mathsf{round}(X)$ 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,

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

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

# C Experiment details

We used Tensorflow [Abadi et al., 2016] for the VAE experiments, slightly changing the code made available from Jang et al. [2017]. 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 ADAM [Kingma and Ba, 2014] with learning rate 0.1 for optimization. For rounding, the parameter vector $V$ defined in 3.2 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 3.1 were constrained between $10^{-8}$ and 1. In either case, the temperature, along with maximum values for the noise variances were calibrated using a grid search on the interval $[10^{-2}, 1]$. Improvements may be obtained with the use of an annealing schedule, a direction we intend to explore in the future.

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. All results in Figure 4 are averages over five experiment simulations with different sampled matrices $W$. For results in Figure 4b we considered either one or four worms (squares and circles, respectively), and for the x-axis we used the values $\nu \in \{0.0075, 0.01, 0.02, 0.04, 0.05\}$. We fixed the number of known neuron identities to 25 (randomly chosen). For results in Figure 4c we used four worms and considered two values for $\nu$; 0.1 (squares) and 0.05 (circles). Different x-axis values correspond to fixing 110, 83, 55 and 25 neuron identities.

# 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. J. Gumbel. *Statistical theory of extreme values*

*and some practical applications: a series of lectures*. Number 33. US Govt. Print. Office, 1954.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. In *Proceedings of the International Conference on Learning Representations*, 2017.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

P. Kumaraswamy. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1-2):79–88, 1980.

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. Autograd: 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. 2017.

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.