Long Title

Abstract

How to perform inference over the space of permutation matrices? By definition, with n nodes, there are $n!$ such matrices. Clearly, searching or optimizing over this space quickly becomes intractable as n grows. Our goal is to derive a tractable algorithm for performing approximate inference over this challenging discrete space. To that end, we propose a continuous relaxation of permutation matrices to doubly stochastic matrices, i.e. to points in the Birkhoff polytope. We then derive an invertible and differentiable map from densities on unconstrained space to densities on the Birkhoff polytope. This transformation is parameterized by a "temperature" that controls how concentrated the resulting density is at the extrema of the Birkhoff polytope; i.e. at permutation matrices. This relaxation admits variational inference via stochastic gradient ascent over the distributions on doubly stochastic matrices (and in the zero-temperature limit, on permutation matrices) using Monte Carlo estimates of the reparameterized gradient.

1. Continuous relaxations for discrete inference

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

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

That is, the optimal variational parameters are given by,

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

where

$$-\mathrm{KL}\left(q(x;\theta) \parallel p(x \mid y)\right) = \mathbb{E}_{x\sim q(x;\theta)}\left[\log p(x \mid y) - \log q(x;\theta)\right] \tag{2}$$

$$\geq \mathbb{E}_{x\sim q(x;\theta)}\left[\log p(x,\, y) - \log q(x;\theta)\right] \tag{3}$$

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

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

When x is a continuous random variable, we can often go one step further and leverage the "reparameterization trick" (Salimans & Knowles, 2013; Kingma & Welling, 2014; Price, 1958; Bonnet, 1964). Specifically, in some cases we can simulate from q via the following procedure,

$$\xi \sim p(\xi), \tag{5}$$

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

where $g(\theta, \xi)$ is a deterministic and differentiable transformation (which is only possible if x is continuous). Thus, we can effectively "factor out" the randomness of q. With this transformation, we can bring the gradient inside the expectation as follows,

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

This gradient can be estimated with Monte Carlo, and, in practice, this leads to lower variance estimates of the gradient than, for example, the score function estimator (Williams, 1992; Glynn, 1990).

Recently, there have been a number of proposals for extending these reparameterization tricks to high dimensional discrete problems[1] by relaxing them to analogous continuous problems (Maddison et al., 2016; Jang et al., 2016;

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

Kusner & Hernández-Lobato, 2016). These approaches are based on the following observation: if $x \in \{0,1\}^k$ is a one-hot vector drawn from a categorical distribution, then the support of $p(x)$ is the set of vertices of the $k-1$ dimensional simplex. Thus, we can represent the distribution of x as an atomic density on the simplex. Sampling x is equivalent to sampling this atomic measure. That is,

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

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

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

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

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

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

2. An alternative continuous relaxation

While the Gumbel-softmax has some nice properties, as we will see, it does not lend itself as naturally to more complicated generalizations. Consider the following alternative model for π:

$$\psi_k \sim \mathcal{N}(\mu_k, \tau^{-1}\eta_k^2) \qquad \text{for } k = 1, \ldots, K-1 \tag{12}$$
$$\pi_1 = \sigma(\psi_1) \tag{13}$$
$$\pi_k = \sigma(\psi_k)(1 - \sum_{j=1}^{k-1} \pi_j) \qquad \text{for } k = 2, \ldots, K-1 \tag{14}$$
$$\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j \tag{15}$$

This is known as a logistic stick breaking transformation since $\sigma(\cdot)$ is the logistic function and $\sigma(\psi_k)$ can be seen as the fraction of the remaining "stick" of probability mass assigned to π_k (Linderman et al., 2015). Moreover, the density of π can be expressed in closed form as a function of μ_k and η_k^2. Finally, as with the relaxations above, the temperature τ controls how concentrated $p_\tau(\pi \mid \{\mu_k, \eta_k^2\})$ is at the vertices of the simplex. As $\tau \to 0$, the density becomes concentrated on atoms at the K vertices, and as $\tau \to \infty$, the density concentrates on a point in the interior of the simplex determined by $\{\mu_k\}$. For intermediate values, the density is continuous on the simplex.

2.1. Limit analysis

One nice property of the concrete distribution (Maddison et al., 2016; Jang et al., 2016) is that it can be understood as the 'heating' of the original categorical distribution (equivalently, a probability simplex-valued distribution whose atoms are the vertices of the simplex): specifically, in the high temperature limit ($\tau \to \infty$) the distribution becomes a single atom in the center of mass of the probability simplex, while the zero-temperature limit ($\tau \to 0$) corresponds to the target categorical distribution. Now we show that although the stick-breaking representation may seem 'position biased', in reality, through appropriate choices of the parameters we are able to provide richer representations as compared to the concrete case. Specifically, we will show that degenerate cases can correspond to arbitrary categorical distributions and arbitrary point masses in the simplex.

To see this, consider first the more general stick-breaking representation

$$\psi_k = g(\theta_k, \epsilon_k) \qquad \text{for } k = 1, \ldots, K-1 \tag{16}$$
$$\pi_1 = \psi_1 \tag{17}$$
$$\pi_k = \psi_k \left(1 - \sum_{j=1}^{k-1} \pi_j\right) \qquad \text{for } k = 2, \ldots, K-1 \tag{18}$$
$$\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j. \tag{19}$$

Where g is a $[0,1]$-valued function.

The following lemmas analyze two degenerate but useful cases:

Proposition: the degenerate case where $\psi_k = g(\theta_k)$ (i.e, ψ_k is non-random) leads to $\pi \sim \delta(\tilde{\pi})$ (i.e, single atom in the point $\tilde{\pi}$) with π in Δ^{k-1} is computed from the above formulae. Also, if any point in $[0,1]$ can be realized through $g(\theta_k)$ then any deterministic π can be realized.

Proof: The first part is obvious. The second part is also obvious, once acknowledging the invertibility of the function f that maps $\psi \xrightarrow{f} \pi$.

Proposition: the degenerate case where ψ_k are Bernoulli with some parameter p_{θ_k}) leads to π having an atomic distri-

bution with atoms in the vertices of Δ^{k-1} (i.e, π is categorical). We have the following expression for the probabilities of the atoms $\pi_k = 1$ (one hot vectors):

$$P(\pi_k = 1) = \prod_{i=1}^{k-1}(1 - p_{\theta_i})p_{\theta_k} \quad \text{for } k = 1, \dots, K-1, \tag{20}$$

$$P(\pi_K = 1) = \prod_{i=1}^{K-1}(1 - p_{\theta_i}). \tag{21}$$

Moreover, if any Bernoulli variable can be realized through appropriate choice of the parameters then, any categorical distribution can be realized. *Proof*: the first expression comes from expressing the event $\pi_k = 1$ equivalently as $\pi_k = 1, p_i = 0 i < k$ and then, conditioning backwards successively. The second comes from the following (which is deduced by replacing terms in the above)

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

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

Having stated the above properties it only remains showing that the two re-parameterizations we consider here, based on the Gaussian and Kumaraswamy distributions, lead to the above degenerate distribution as their limits.

Proposition For the choice $\psi = \sigma(\delta), \delta \sim \mathcal{N}(\mu, \eta^2)$, the limit $\eta \to 0$ leads to the non-random $\psi = \sigma(\mu)$. Additionally, the limit $\mu \to \infty, \eta^2 = \mu/K$ with K constant leads to $\psi \sim$ Bernoulli$(\Phi(K))$ (Φ denotes the standard normal cdf). In both cases the convergence is in distribution *Proof.* The first convergence is obvious. To see the second, let's index μ_n and study the cdf of Ψ_n and

$$\begin{aligned} F_{\Psi_n}(x) = & \quad P(\sigma(\delta_n) < x) && (22)\\ = & \quad P(\delta_n < \sigma^{-1}(x)) && (23)\\ = & \quad P(\mu_n + \mu_n/K\epsilon < \sigma^{-1}(x)), && (24)\\ = & \quad P(\epsilon < \sigma^{-1}(x))K/\mu_n - K) && (25)\\ = & \quad \Phi(\sigma^{-1}(x))K/\mu_n - K) && (26) \end{aligned}$$

Therefore, by continuity of Φ we obtain $F_{\Psi_n}(x) \to \Phi(-K)$ for all points $x \in (,1)$. 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 we have obtained convergence to F at all ints continuity points (the interval $(0,1)$), with $1 - p = \Phi(-K) \to \Phi(K) = p$) so we can conclude. Notice that the above representation only allows to converge to $p > 0.5$, as K has to be positive. This can be fixed by choosing sequence with negative μ instead.

Proposition For the choice $\psi = \mathcal{K}(a, b)$: i) in the limit $a, b \to$ we converge to a non-random p, provided that $p = bB\left(1 + \frac{1}{a}, b\right)$ along the limiting sequence. ii) In the limit $a, b \to 0$ we obtain convergence to a Bernoulli random variable with parameter p, provided the same condition involving p, a, b holds. In both cases convergence is in probability.

Proof: A proof can be found in (Mitnik, 2013)

3. Continuous relaxations for permutation matrices

Just as one-hot vectors are the vertices of the simplex, the Birkhoff-von Neumann theorem states that permutation matrices are vertices of the convex hull of doubly stochastic matrices. For permutations of size n, a permutation matrix, $X \in \{0, 1\}^{n \times n}$, is a binary matrix such that every row and every column sums to one. An analogous relaxation to the one above is to consider $X \approx \Pi$, where $\Pi \in [0, 1]^{n \times n}$ is a doubly stochastic matrix, i.e. the rows and columns both sum to one. This set is known as the Birkhoff polytope, which we denote by $\mathcal{B}_n$. Due to these constraints, the Birkhoff polytope lies within a $(n-1)^2$ dimensional subspace of all $[0, 1]^{n \times n}$ matrices.

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

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

$$\pi_{1j} = \sigma(\psi_{1j})(1 - \sum_{k=1}^{j-1} \pi_{1k}) \quad \text{for } j = 2, \dots, n-1 \tag{27}$$

$$\pi_{1n} = 1 - \sum_{k=1}^{n-1} \pi_{1k} \tag{28}$$

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

That is,

$$\pi_{ij} \leq 1 - \sum_{k=1}^{j-1} \pi_{ik} \qquad \text{(row sum)} \tag{29}$$

$$\pi_{ij} \leq 1 - \sum_{k=1}^{i-1} \pi_{kj} \qquad \text{(column sum).} \tag{30}$$

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

$$\underbrace{1 - \sum_{k=1}^{j} \pi_{ik}}_{\text{remaining stick}} \leq \underbrace{\sum_{m=j+1}^{n} (1 - \sum_{k=1}^{i-1} \pi_{km})}_{\text{remaining upper bounds}}. \tag{31}$$

Rearranging terms, we have,

$$\pi_{ij} \geq 1 - \sum_{k=1}^{j-1} \pi_{ik} - \sum_{m=j+1}^{n} (1 - \sum_{k=1}^{i-1} \pi_{km}) \tag{32}$$

$$= 1 - n + j - \sum_{k=1}^{j-1} \pi_{ik} + \sum_{k=1}^{i-1} \sum_{m=j+1}^{n} \pi_{km} \tag{33}$$

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

$$\ell_{ij} \leq \pi_{ij} \leq u_{ij} \tag{34}$$

$$\ell_{ij} \triangleq \max \left\{ 0,\, 1 - n + j - \sum_{k=1}^{j-1} \pi_{ik} + \sum_{k=1}^{i-1} \sum_{m=j+1}^{n} \pi_{km} \right\} \tag{35}$$

$$u_{ij} \triangleq \min \left\{ 1 - \sum_{k=1}^{j-1} \pi_{ik},\, 1 - \sum_{k=1}^{i-1} \pi_{kj} \right\}. \tag{36}$$

Thus, we define,

$$\pi_{ij} = \ell_{ij} + \sigma(\psi_{ij})(u_{ij} - \ell_{ij}). \tag{37}$$

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

$$\psi_{ij} = \sigma^{-1} \left(\frac{\pi_{ij} - \ell_{ij}}{u_{ij} - \ell_{ij}} \right). \tag{38}$$

Notice that these bounds only depend on values of Π that have already been computed; i.e., those that are above or to the left of the (i, j)-th entry. Thus, the transformation from Ψ to Π is feed-forward according to this ordering. Consequently, the Jacobian of the inverse transformation, $\mathrm{d}\Psi/\mathrm{d}\Pi$, is lower triangular, and its determinant is the product of its diagonal,

$$\left| \frac{\mathrm{d}\Psi}{\mathrm{d}\Pi} \right| = \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{\partial \psi_{ij}}{\partial \pi_{ij}} \tag{39}$$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{\partial}{\partial \pi_{ij}} \sigma^{-1} \left(\frac{\pi_{ij} - \ell_{ij}}{u_{ij} - \ell_{ij}} \right) \tag{40}$$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \left(\frac{1}{u_{ij} - \ell_{ij}} \right) \left(\frac{u_{ij} - \ell_{ij}}{\pi_{ij} - \ell_{ij}} \right) \left(\frac{u_{ij} - \ell_{ij}}{u_{ij} - \pi_{ij}} \right) \tag{41}$$

$$= \prod_{i=1}^{n-1} \prod_{j=1}^{n-1} \frac{u_{ij} - \ell_{ij}}{(\pi_{ij} - \ell_{ij})(u_{ij} - \pi_{ij})} \tag{42}$$

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

$$\mathrm{vec}(\Psi) \sim \mathcal{N}(\mu, \mathrm{diag}(\eta^2)) \tag{43}$$

$$\Pi = f(\Psi) \tag{44}$$

$$\implies p(\Pi \mid \mu, \mathrm{diag}(\eta^2)) = \left| \frac{\mathrm{d}\Psi}{\mathrm{d}\Pi} \right| \mathcal{N}(f^{-1}(\Pi) \mid \mu, \mathrm{diag}(\eta^2)) \tag{45}$$

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

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

References

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

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

Jang, Eric, Gu, Shixiang, and Poole, Ben. Categorical reparameterization with Gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

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

Kusner, Matt J and Hernández-Lobato, José Miguel. GANs for sequences of discrete elements with the Gumbel-softmax distribution. *arXiv preprint arXiv:1611.04051*, 2016.

Linderman, Scott, Johnson, Matthew, and Adams, Ryan P. Dependent multinomial models made easy: Stick-breaking with the polya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pp. 3456–3464, 2015.

Maddison, Chris J, Mnih, Andriy, and Teh, Yee Whye. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Mitnik, Pablo A. New properties of the kumaraswamy distribution. *Communications in Statistics - Theory and Methods*, 42(5):741–755, 2013. doi: 10.1080/03610926.2011.581782. URL http://dx.doi.org/10.1080/03610926.2011.581782.

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

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

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