On the Gumbel-Softmax Trick

Abstract

The reparameterization trick enables optimizing stochastic computation graphs via gradient descent [2]. The essence of the trick is to re-factor each stochastic node into a differentiable function of its parameters and a random variable with fixed distribution. After refactoring, the gradients of the loss propagated by the chain rule through the graph are low variance unbiased estimators of the gradients of the expected loss. While many continuous random variables have such reparameterizations, discrete random variables lack useful reparameterizations due to the discontinuous nature of discrete states. As a remedy, the gumbel soft-max trick [1] or concrete distribution [3] serves as a continuous relaxation of discrete random variables.

1 Reparameterization Trick

Let us shed some light on the reparameterization trick used by Kingma and Welling [2] to train their variational auto-encoders.

First, the law of the unconscious statistician states (LOTUS) states that for a random variable ϵ with pdf f_ϵ and a measurable function g, it holds:

$$\mathbb{E}(g(\epsilon)) = \int g(X) dF_\epsilon(x). \tag{1}$$

In other words, to compute the expectation of $z = g(\epsilon)$ we only need to know the mapping g and the distribution of ϵ, but we do not need the explicit distribution of z:

$$\mathbb{E}_{\epsilon \sim p(\epsilon)}(g(\epsilon)) = \mathbb{E}_{z \sim p(z)}(z). \tag{2}$$

Now, suppose z has a distribution that depends on a parameter ϕ, i.e. $z \sim p_\phi(z)$. Moreover, assume one can express $z = g_\phi(\epsilon)$ for known function g and a certain noise distribution, e.g. $\epsilon \sim \mathcal{N}(0, 1)$. Then LOTUS states that for any measurable function f:

$$\mathbb{E}_{z \sim p_\phi(z)}(f(z)) = \mathbb{E}_{\epsilon \sim p(\epsilon)}(f(g_\phi(\epsilon))). \tag{3}$$

In auto-encoding variational bayes formulations we encounter the gradient of some expectation with respect to a parameter ϕ and may use the following equality:

$$\nabla_\phi \mathbb{E}_{z\sim p(z)} = \nabla_\phi \mathbb{E}_{\epsilon\sim p(\epsilon)}(f(g_\phi(\epsilon))) = \mathbb{E}_{\epsilon\sim p(\epsilon)}(\nabla_\phi f(g_\phi(\epsilon))). \quad (4)$$

Further, we have conveniently expressed z so that expectations of functions of z can be expressed as integrals with respect to a density that does not depend on the parameter. Therefore, we can exchange the expectation and gradient.

Finally, the reparameterization gives rise to an unbiased estimate of the above gradient via MCMC:

$$\nabla_\phi \mathbb{E}_{z\sim p(z)} \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_\phi f(g_\phi(\epsilon_i)). \quad (5)$$

For reasons not yet completely understood, empirically it is seen that this reparameterization based estimate of the gradient exhibits much less variance than competing estimators.

Let us look at a trivial example. Assume we have a normal distribution q parameterized by ϕ, specifically $q_\phi(x) = \mathcal{N}(\phi, 1)$. We want to solve:

$$\phi^* = \arg\min_\phi \mathbb{E}_q(x^2). \quad (6)$$

First, we calculate $\nabla_\phi \mathbb{E}_q(x^2)$ as

$$\begin{aligned}
\nabla_\phi \mathbb{E}_q(x^2) &= \nabla_\phi \int q_\phi(x) x^2 dx && (7)\\
&= \int x^2 \nabla_\phi q_\phi(x) \frac{q_\phi(x)}{q_\phi(x)} dx && (8)\\
&= \int x^2 q_\phi(x) \nabla_\phi \log q_\phi(x) dx && (9)\\
&= \mathbb{E}_q(x^2 \nabla_\phi \log q_\phi(x)). && (10)
\end{aligned}$$

For $q_\phi(x) = \mathcal{N}(\phi, 1)$, this method gives

$$\nabla_\phi \mathbb{E}_q(x^2) = \mathbb{E}_q(x^2(x-\phi)). \quad (11)$$

Second, we use the reparameterization to factor out the stochastic element in q and make it independent of ϕ:

$$x = \phi + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{12}$$

Then

$$\nabla_\phi \mathbb{E}_q(x^2) = \mathbb{E}_p((\phi + \epsilon)^2). \tag{13}$$

where $p = \mathcal{N}(0, 1)$ and $\epsilon \sim p(\epsilon)$. Now, the expectation is independent of ϕ and we rewrite the gradient:

$$\nabla_\phi \mathbb{E}_q(x^2) = \nabla_\phi \mathbb{E}_p((\phi + \epsilon)^2) = \mathbb{E}_p(2(\phi + \epsilon)). \tag{14}$$

Finally, note that empirically the variance of estimator (14) is an order of magnitude lower compared to (11).

2 Variational Inference

Consider variational inference in a latent variable model: we want to evaluate the posterior $p_\theta(z \mid x)$, which usually is intractable due to the evidence $p_\theta(x)$ in the denominator. In a standard variational setup, we find a variational approximation $q_\phi(z \mid x)$ such that it miminizes its "distance" with the true posterior. Or, equivalently, it maximizes a lower bound on the log evidence:

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{z \sim q_\phi(z|x)}(\log(p_\theta(x, z))) - \log(q_\phi(z \mid x)). \tag{15}$$

Then, we perform learning by a double maximization of this surrogate function (the ELBO): first, with respect to ϕ to bridge the gap with the true posterior, and, second, with respect to θ to maximize the evidence.

The most direct approach for learning is to do gradient descent. However, unfortunately, gradients are not available in closed form because they suffer from the often intractable expectation with respect to $z \sim q_\phi(z \mid x)$. One solution is, then, to replace the true gradients by noisy, unbiased estimators.

We reparameterize $q_\phi(z \mid x)$ with respect to a noise distribution, $s(\epsilon)$, and we will be able to construct an unbiased estimator of the ELBO based on MCMC samples, as we'll have:

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{\epsilon \sim s(\epsilon)}(\log(p_\theta(x, g_\phi(\epsilon)))) - \log(q_\phi(g_\phi(\epsilon)). \tag{16}$$

3 Gumbel

Unfortunately, for discrete random variables we cannot apply the reparameterization trick directly: A non-dengenerate mapping from a continuous set onto a discrete set is not differentiable. That is, the functional relation $z = g_\phi(\epsilon)$, we can not conceive $\frac{\partial z}{\partial \phi}$. We can, however, construct alternative estimates of the gradient, that may not enjoy the low-variance property of the reparameterization based ones.

The Gumbel-softmax trick [1] is a remedy to overcome the inability of applying the reparameterization trick to discrete random variables. It relies on two insights: (1) the Gumbel distribution provides a nice parameterization for a discrete distribution, and (2) the corresponding non-continuous function can be made continuous by applying an approximation that depends on a temperature parameter, which in the zero-temperature case degenerates to the discontinuous, original expression.

First, we consider the Gumbel distribution. Let $U \sim Unif(0, 1)$, then the random variable G is said to have a Gumbel distribution, if $G = -\log(-\log(U))$. Now, we can parameterize any discrete distribution in terms of Gumbel random variables. Let X denote a discrete random variable with $P(X = k) \propto \alpha_k$. Let $\{G_k\}_{k \leq K}$ be an i.i.d. sequence of Gumbel random variables. Then:

$$X = \arg\max_k (\log \alpha_k + G_k). \tag{17}$$

Thus, a recipe for sampling from a discrete distribution is: (1) draw from a uniform and apply the Gumbel transform, (2) add $\log \alpha_k$, (3) take the value k that yields the maximum.

Second, since the $\arg\max$ in (17) is not continuous, we want to relax the discrete set by considering random variables taking values in a larger set. To construct the relaxation, we recognize that: (1) any discrete random variable can always be expressed in the form of an one-hot vector by mapping the realization k to the index of the non-zero entry in the vector, and (2) that the convex hull of the set

of one-hot vectors is a probability simplex:

$$\Delta^{K-1} = \left\{ x \in \mathbb{R}_+^K, \sum_{k=1}^{K} x_k = 1 \right\}. \tag{18}$$

Therefore, a natural way to relax a discrete random variable by extending its codomain to the probability simplex. Both [3] and [1] propose softmax indexed by a temperature parameter τ:

$$f_\tau(x)_k = \frac{\exp(x_k/\tau)}{\sum_{k'=1}^{K} \exp(x_{k'}/\tau)}. \tag{19}$$

Then, we define the sequence of simplex-valued random variables as:

$$X_k^\tau = f_\tau(\log \alpha + G)_k = \left(\frac{\exp((\log \alpha_k + G_k)/\tau)}{\sum_{k'=1}^{K} \exp((\log \alpha_{k'} + G_{k'})/\tau)} \right). \tag{20}$$

The random variable X^τ is said to have a concrete distribution, denoted by $x^\tau \sim Concrete(\alpha, \tau)$ (a portmanteau between continuous and discrete). Its density is given by:

$$p_{\alpha,\tau} = (n-1)!\tau^{n-1} \prod_{k=1}^{K} \left(\frac{\alpha_k x_k^{-\tau-1}}{\sum_{k'=1}^{K} \alpha_{k'} x_{k'}^{-\tau}} \right), \; x \in \Delta^{K-1}. \tag{21}$$

Note, the expression is in closed-form and can be evaluated exactly for x, α and τ. Indeed, in the definition of the ELBO we need to evaluate the entropy term $\mathbb{E}_{z \sim p_\theta(z|x)}(-\log(q_\theta(z \mid x)))$ which explicitly depends on the density.

The nature of this relaxation can be better understood by four properties:

1. (Zero temperature) $P(\lim_{\tau \to 0} X_k^\tau = 1) = \alpha_k / \sum_{k'=1}^{K} \alpha_{k'}$.
2. (Convex eventually) if $\tau \leq (n-1)^{-1}$, then $p_{\alpha,\tau}(x)$ is log-convex in x.
3. (Rounding) $P(X_k^\tau > X_i^\tau, \forall i \neq k) = \alpha_k / \sum_{k'=1}^{K} \alpha_{k'}$
4. (Tradeoff) For learning, there is a tradeoff between small and large temperatures.

First, in the zero-temperature limit we recover the discrete distribution. Think of a logistic function modulated by a slope parameter which for high slopes will become the Heaviside step function.

Second, recall we optimize a function that involves the log density. By convexity, we will be better off as long as the temperature is low enough.

Third, even with a non-zero temperature relaxation we can map points x in the simplex to a one-hot vector (extreme-point of the simplex), with non-zero component k, so that x_k is closest to one.

Fourth, for small temperatures, the samples resemble a one-hot vector closely, but the variance of the estimate of the gradient is high. For high temperatures, the samples are smooth, but the variance is small.

References

[1] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2016. cite arxiv:1611.01144.

[2] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[3] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712, 2016.