On the Gumbel-Softmax Trick for Inference of Discrete Variables

Abstract

The reparameterization trick enables optimizing stochastic computation graphs via gradient descent. 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 (or concrete distribution) 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 [1] 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 $z = g(\epsilon, \phi)$ 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:

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

In black-box variational inference formulations we encounter the gradient of some expectation with respect to a parameter ϕ and may use the following equality:

$$\nabla_\phi \mathbb{E}_{z \sim p(z)} = \nabla_\phi \mathbb{E}_{\epsilon \sim p(\epsilon)}(f(g(\epsilon, \phi))) = \mathbb{E}_{\epsilon \sim p(\epsilon)}(\nabla_\phi f(g(\epsilon, \phi))). \tag{4}$$

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:

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

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

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

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

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

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

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

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

References

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