Notes on Concrete Distribution

Wilker Aziz

April 4, 2018

Here I revisit a continuous relaxation to discrete random variables based on the Concrete distribution (or Gumbel-Softmax distribution) (Maddison et al., 2017; Jang et al., 2017). My presentation follows that of Balog et al. (2017) who used an *exponential clocks* (Buchbinder et al., 2013) formulation.

Preliminaries Suppose X a discrete random variable taking on values in $\mathcal{X}$. Let $P(X|\phi) = \text{Cat}(\text{softmax}(\phi))$ for $\phi \in \mathbb{R}^{|\mathcal{X}|}$, where ϕ can be interpreted as a vector of log-potentials indexed by elements of $\mathcal{X}$. We use $\lambda \in \mathbb{R}^{|\mathcal{X}|}_{\geq 0}$ to denote potentials, otherwise seen as unnormalised probabilities, that is, $\lambda_x = \exp(\phi_x) \propto P_\phi(x)$, and thus $Z = \sum_{x \in \mathcal{X}} \lambda_x$ is the constant that normalises the components of λ as to yield the pmf $p_\phi(x) = P(X = x|\phi)$.

Exponential clocks Consider we associate with each $x \in |\mathcal{X}|$ an independent clock such that each clock rings after a random time $T_x \sim \text{Exp}(\lambda_x)$, where λ_x is the rate parameter. Assume we start all clocks simultaneously. Then, it is easy to show that the time until some clock rings is exponentially distributed with rate parameter Z, i.e.,

$$\min_{x \in \mathcal{X}} \{T_x\} \sim \text{Exp}(Z) \tag{1}$$

and that the probability of a clock ringing first is proportional to its rate, and therefore, to the pmf $p_\phi(x)$, i.e.,

$$\arg\min_{x \in \mathcal{X}} \{T_x\} \sim P(X|\phi) \ . \tag{2}$$

Two observations are due. First, because the mean of $\text{Exp}(\mu)$ is $1/\mu$, solving the minimisation problem on the left-hand side of (1), for sampled times, yields an MC estimate of $1/Z$. Second, and again given sampled times, the solution to the minimisation problem on the left-hand side of (2) yields a sample from the distribution $P(X|\phi)$.[1]

[1] While this algorithm bypasses the need to compute the cdf and its inverse, it still requires assessing all of the potentials (as each parameterises an exponential clock) so that ringing times can be sampled.

Gumbel reparameterisation It is known that if $Y \sim \text{Exp}(\eta)$ and $g(y) = -\ln y - c$, then $g(Y) \sim \text{Gumbel}(-c + \ln \eta)$ where c is the Euler-Mascheroni constant and $\text{Gumbel}(-c + \ln \eta)$ is a Gumbel distribution with location $-c + \ln \eta$, scale 1, and mean $\ln \eta$. Applying g to both sides of the equality in distribution expressed in (1) and (2) yields

$$\max_{x \in \mathcal{X}} \{\phi_x + \gamma_x\} \sim \text{Gumbel}(-c + \ln Z) \tag{3}$$

and

$$\arg\max_{x \in \mathcal{X}} \{\phi_x + \gamma_x\} \sim P(X|\phi) \tag{4}$$

where $\gamma_x \sim \text{Gumbel}(-c)$ is a noise variable sampled from a fixed distribution. In direct correspondence to exponential clocks, we simulate random Gumbel noise γ, and solve the left-hand side of (3) to obtain an MC estimate of $\ln Z$ and the left-hand side of (7) to obtain a sample from $P(X|\phi)$. If we are not interested in estimating $\ln Z$ through (3), but instead all we care about is to sample from $P_\phi(X)$ through (7), we can represent X in terms of the standard Gumbel distribution $\text{Gumbel}(0, 1)$, that is

$$X \sim \arg\max_{x \in \mathcal{X}} \{\phi_x + \gamma_x\} \quad \text{with} \quad \gamma_x \sim \text{Gumbel}(0, 1) \tag{5}$$

which is equivalent to shifting the Gumbel samples by c—an operation that does not affect the arg max.

Continuous relaxation For each $x \in \mathcal{X}$, we can think of $\phi_x + \gamma_x$ in (5) as a random log-potential Φ_x, where the randomness comes from standard noise $\gamma_x \sim \text{Gumbel}(0, 1)$. Then, for a temperature $\alpha \in \mathbb{R}_{>0}$, the transformation

$$\theta_x = \frac{\exp(\ln \lambda_x + \gamma_x)/\alpha}{\sum_{x' \in \mathcal{X}} \exp(\ln \lambda_{x'} + \gamma_{x'})/\alpha} \tag{6}$$

defines the components of a random vector Θ taking values in the simplex $\Delta^{|\mathcal{X}|-1}$. Each vector in this simplex can be thought of as indexing a member of a family of pmfs over $\mathcal{X}$. Then note that the argmax in (7) returns a discrete value associated with a vertex of the simplex. If we relax this computation as to allow points in the interior of the simplex, we obtain something Maddison et al. (2017) called a *concrete* variable (in allusion to the words *con*tinuous and dis*crete*—but note that concrete variables are indeed continuous). Crucially, the tempered softmax transformation in (6) does not affect the relative order of the random log-potentials. The equivalence in distribution

$$\arg\max_{x \in \mathcal{X}} \{\theta_x\} \sim P(X|\phi) \tag{7}$$

for $\Theta \sim \text{Concrete}(\phi, \alpha)$ still holds at temperature 1.0, but it changes to

$$\arg\max_{x \in \mathcal{X}} \{\theta_x\} \sim P(X|\phi)^\alpha \tag{8}$$

in the general case.[2] Moreover, as $\alpha \to 0$, sampling from the concrete pdf tends to return one-hot encodings of events in $\mathcal{X}$, as if their corresponding discrete events had been sampled from the discrete distribution $P(X|\phi)^\alpha$. Maddison et al. (2017) put forward some desiderata for a relaxation of this kind and argue that the concrete distribution meets all of them. In particular, the temperature parameter gives a handle on the degree to which the relaxation allows points in the interior of the simplex.[3]

Implications Maddison et al. (2017) and Jang et al. (2017) used the Gumbel reparameterisation to push stochasticity out of a computation graph in a way similar to the standard Gaussian reparameterisation of deep Gaussian models. They also employed the continuous relaxation above to obtain biased gradient estimates of objectives defined over large discrete spaces. Maddison et al. (2017) changes the ELBO with concrete variables and use its log-density in the objective making the gradient estimates biased with respect to the original objective (based on discrete variables), but unbiased with respect to the relaxation. Jang et al. (2017) instead use the argmax in the forward pass and the relaxation in the backward pass ignoring its density in the objective. This results in a technique similar to *straight-through* (Bengio et al., 2013) whose downside is that the objective being optimised is unknown. Balog et al. (2017) used the results above to produce an efficient sequential sampler for undirected graphical models. They also prove a whole family of reparameterisation tricks by changing the function $g(y)$ in the formulation above.

References

Balog, M., Tripuraneni, N., Ghahramani, Z., and Weller, A. (2017). Lost relatives of the gumbel trick. In *34th International Conference on Machine Learning*.

Bengio, Y., Léonard, N., and Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Buchbinder, N., Naor, J. S., and Schwartz, R. (2013). Simplex partitioning via exponential clocks and the multiway cut problem. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 535–544, New York, NY, USA. ACM.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with gumbel-softmax. *International Conference on Learning Representations*.

[2] Refer to (Maddison et al., 2017) for the closed-form density, but note that we simulate concrete variables rather easily with a tempered softmax transformation $\theta \stackrel{\text{def}}{=} \text{softmax}_\alpha(\phi + \gamma)$ of log-potentials perturbed by additive Gumbel noise.

[3] But note that setting α too close to zero leads to numerical instability, exploding gradients, and high variance (Maddison et al., 2017; Jang et al., 2017).

Maddison, C. J., Mnih, A., and Teh, Y. W. (2017). The concrete distribution: A continous relaxation of discrete random variables. *International Conference on Learning Representations*.