Gumbel-Softmax

André Martins

Unbabel Reading Meeting, January 14, 2019

Today's Paper

Eric Jang, Shixiang Gu, Ben Poole.
"Categorical Reparametrization with Gumbel-Softmax." ICLR 2017
(https://arxiv.org/pdf/1611.01144.pdf)

Related:

- Chris Maddison, Andriy Mnih, Yee Whye Teh.
  "The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables." ICLR 2017
- Blog post: https://blog.evjang.com/2016/11/tutorial-categorical-variational.html

Outline

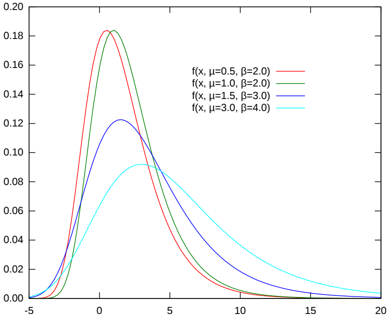
- Gumbel distribution
- Gumbel-max trick
- Gumbel softmax
- ... then we jump to the paper.

Gumbel Distribution

The Gumbel distribution has the following density:

$$p(x; \mu, \beta) = \frac{1}{\beta} \exp\left(\frac{\mu - x}{\beta} - \exp\left(\frac{\mu - x}{\beta}\right)\right)$$

where μ and β are location and scale parameters.



- Useful to model the occurrence of eartquakes, floods, and natural disasters.
- Also called "double-exponential distribution."

Sampling from a Gumbel Distribution

The standard Gumbel distribution Gumbel(0,1) has density:

$$p(x) = e^{-x-e^{-x}}.$$

The cumulative distribution function is

$$F(t) = \mathbb{P}(x \le t) = \int_{-\infty}^{t} p(x) dx = e^{-e^{-t}}.$$

We can sample $g \sim \text{Gumbel}(0,1)$ with inverse transform sampling:

1. Sample $u \sim \text{Uniform}(0,1)$.
2. Compute $g = F^{-1}(u) = -\log(-\log u)$.

This is interesting!

Gumbel Trick

Older than the “Gumbel-softmax”:

- Luce (1959)
- Yellott (1977)
- Papandreou and Yuille (2011)
- Maddison et al. (2014)

See Tim Vieira’s blog post:

- `https://timvieira.github.io/blog/post/2014/07/31/gumbel-max-trick/`

Derivation in Ryan Adams’ blog post:

- `http://lips.cs.princeton.edu/the-gumbel-max-trick-for-discrete-distributions/`

Gumbel Trick

Let $y \sim \text{softmax}(\lambda)$ be a categorical (discrete) random variable

Usually we sample y as follows:

1. Compute the class probabilities $\pi_i = \frac{\exp(\lambda_i)}{\sum_{j=1}^{K} \exp(\lambda_j)}$
2. Compute cumulative distribution function $c_i = \sum_{j \leq i} \pi_i$
3. Sample $u \sim \text{Uniform}(0, 1)$ and return y such that $c_y \leq u < c_{y+1}$.

The Gumbel-max trick offers an alternative:

1. Sample $g_i \sim \text{Gumbel}(0, 1)$, for $i = 1, \ldots, K$
    - Can be done as $u_i \sim \text{Uniform}(0, 1)$ and $g_i = -\log(-\log(u_i))$
2. Compute $y = \arg\max_i(\lambda_i + g_i)$.

The two are equivalent! (The proof requires some math.)

Gumbel Trick

Suppose we have a stochastic neural network with a stochastic node in the middle.

- E.g. a VAE whose encoder computes the parameter λ of a stochastic discrete latent variable $y \sim \text{softmax}(\lambda)$.

Then, the Gumbel trick is an instance of the reparametrization trick:

- Move the stochastic node to the input $u_i \sim \text{Uniform}(0, 1)$
- The part $y = \arg\max_i(\lambda_i - \log(-\log(u_i)))$ is now deterministic.

However this doesn't completely solve the problem: we now have an argmax node (non-differentiable).

Gumbel-Softmax

Key idea: relax the argmax into a softmax, via a temperature parameter τ.

Now y is a continuous random variable in the probability simplex, where each component is defined as:

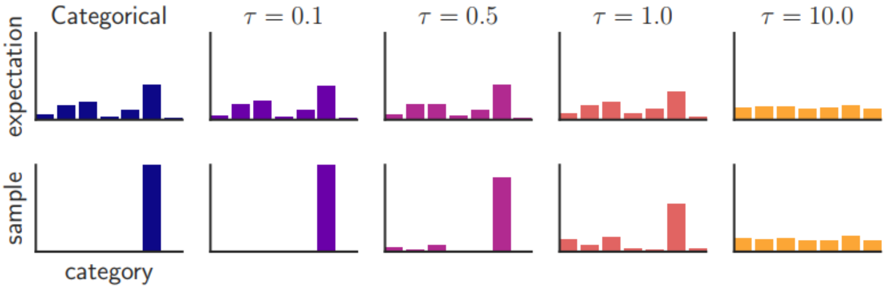
$$y_i = \frac{\exp((\lambda_i + g_i)/\tau)}{\sum_{j=1}^{K} \exp((\lambda_j + g_j)/\tau)},$$

with $g_i = -\log(-\log(u_i))$, $u_i \sim \text{Uniform}(0, 1)$.

- Still easy to sample, with the same reparametrization trick!
- Recovers a discrete categorical distribution with $\tau \to 0^+$.

Jang et al. (2017) derives a closed-form density $p(y)$ (see the appendix for a formal proof).

Some Samples



(From Jang et al. (2017).)

Stochastic Discrete Nodes

Suppose a node in the computation graph is $y \sim \mathrm{softmax}(\lambda)$.

How to compute gradients?

- Reparametrization trick with Gumbel-softmax ($y = g(\phi, \epsilon)$)

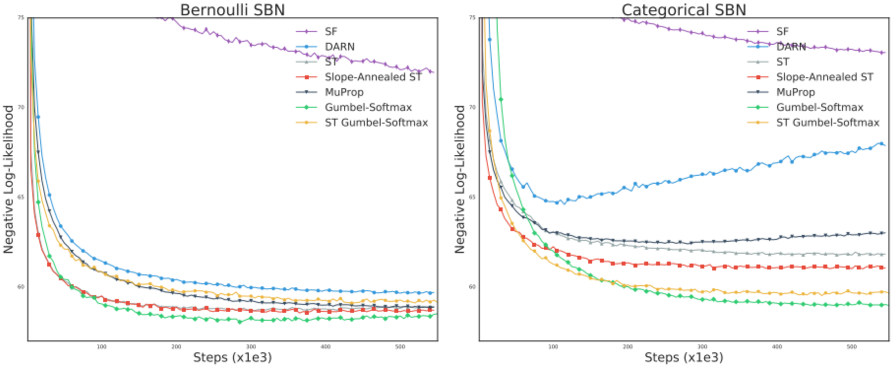
$$\nabla_\phi \mathbb{E}_{y \sim p_\phi}[f(y)] = \nabla_\phi \mathbb{E}_{\epsilon \sim p_\epsilon}[f(g(\phi, \epsilon))] = \mathbb{E}_{\epsilon \sim p_\epsilon}\left[\frac{\partial f}{\partial g}\frac{\partial g}{\partial \phi}\right].$$

- Straight-through estimator: do the argmax in the forward pass, but compute a surrogate gradient using softmax (can also do Straight-through Gumbel-softmax)
- REINFORCE (Williams, 1992):

$$\nabla_\phi \mathbb{E}_{y \sim p_\phi}[f(y)] = \mathbb{E}_{y \sim p_\phi}[f(y)\nabla_\phi \log p_\phi(y)].$$

Unbiased, but high variance. Requires variance reduction techniques (NVIL, DARN, ...)

Structured Prediction



(From Jang et al. (2017).)

References I

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with Gumbel-softmax. In *Proc. of ICLR*.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley and Sons.

Maddison, C. J., Tarlow, D., and Minka, T. (2014). A* sampling. In *Advances in Neural Information Processing Systems*, pages 3086–3094.

Papandreou, G. and Yuille, A. L. (2011). Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, pages 193–200. IEEE.

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

Yellott, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *J. of Mathematical Psychology*, 15(2):109–144.