Gumbel-max trick derivation

Notations

- $P(A)$: The probability of event A.
- $P(A, B)$: The probability of events A and B both occurring.
- $P(A|B)$: The conditional probability of event A given the occurrence of event B.
- $f(x) = \exp\left[-\left(x + \exp(-x)\right)\right]$: The probability density function of the standard Gumbel distribution.
- $F(x) = \exp\left[-\exp(-x)\right]$: The cumulative distribution function of the standard Gumbel distribution.
- $p_1, p_2, \ldots, p_N$: The unnormalized event probabilities of the target categorical distribution, assuming there are N categories.
- $x_1, x_2, \ldots, x_N$: Independent and identically distributed random variables from the standard Gumbel distribution.
- $\arg\max_m (x_m + \ln p_m)$: A function that finds the index m which maximizes $x_n + \ln p_n$.

Derivation

Consider calculating the probability of n being the index that maximizes $x_n + \ln p_n$, *i.e.*, $P(n = \arg\max_m (x_m + \ln p_m))$. In order to do so, we first express it in terms of a marginal probability:

$$\begin{aligned} & P\left(n = \arg\max_m (x_m + \ln p_m)\right) \\ &= \int_{-\infty}^{\infty} \mathrm{d}x_1 \int_{-\infty}^{\infty} \mathrm{d}x_2 \cdots \int_{-\infty}^{\infty} \mathrm{d}x_N P\left(n = \arg\max_m (x_m + \ln p_m), x_1, x_2, \ldots, x_N\right). \end{aligned} \tag{1}$$

The integrand in eq. (1) can be expanded using conditional probability

$$\begin{aligned} & P\left(n = \arg\max_m (x_m + \ln p_m), x_1, x_2, \ldots, x_N\right) \\ &= P\left(n = \arg\max_m (x_m + \ln p_m) \middle| x_1, x_2, \ldots, x_N\right) P(x_1, x_2, \ldots, x_N). \end{aligned} \tag{2}$$

As $x_1, x_2, \ldots, x_N$ are independent of each other, both terms on the right-hand side of eq. (2) can be further expanded as products of probabilities:

$$P\left(n = \arg\max_m (x_m + \ln p_m) \middle| x_1, x_2, \ldots, x_N\right) = \prod_{m=1}^{N} P(x_n + \ln p_n \geq x_m + \ln p_m | x_n, x_m), \tag{3}$$

$$P(x_1, x_2, \ldots, x_N) = \prod_{m=1}^{N} P(x_m). \tag{4}$$

Plugging eqs. (2) to (4) into eq. (1) yields

$$\begin{aligned}
&P\left(n = \arg\max_m \left(x_m + \ln p_m\right)\right) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x_n P\left(x_n + \ln p_n \geq x_n + \ln p_n | x_n\right) P\left(x_n\right) \\
&\times \prod_{\substack{m=1 \\ m\neq n}}^{N} \int_{-\infty}^{\infty} \mathrm{d}x_m P\left(x_n + \ln p_n \geq x_m + \ln p_m | x_n, x_m\right) P\left(x_m\right).
\end{aligned} \tag{5}$$

Now that we have fully expanded the integrand in eq. (1), we are ready to proceed by specifying the mathematical form of each integrand of eq. (5). For each $m = 1, 2, \ldots, N$, the event of $x_n + \ln p_n \geq x_m + \ln p_m$ is a true-or-false question, so $P\left(x_n + \ln p_n \geq x_m + \ln p_m | x_n, x_m\right)$ can only take 1 or 0 as its value. Therefore, we can express it in terms of the Heaviside step function:

$$P\left(x_n + \ln p_n \geq x_m + \ln p_m | x_n, x_m\right) = H(x_n + \ln p_n - x_m - \ln p_m). \tag{6}$$

Note that $P\left(x_n + \ln p_n \geq x_n + \ln p_n | x_n\right)$ is always equal to 1. Moreover, as x_m's are independently sampled from the standard Gumbel distribution, by definition,

$$P\left(x_m\right) = f\left(x_m\right). \tag{7}$$

Consequently, for any $m \neq n$,

$$\begin{aligned}
&\int_{-\infty}^{\infty} \mathrm{d}x_m P\left(x_n + \ln p_n \geq x_m + \ln p_m | x_n, x_m\right) P\left(x_m\right) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x_m H(x_n + \ln p_n - x_m - \ln p_m) f\left(x_m\right) = \int_{-\infty}^{x_n + \ln p_n - \ln p_m} \mathrm{d}x_m f\left(x_m\right) \\
&= F\left(x\right)\Big|_{-\infty}^{x_n + \ln p_n - \ln p_m} = \exp\left[-\exp\left(-x_n - \ln p_n + \ln p_m\right)\right] \\
&= \exp\left[-\frac{p_m}{p_n}\exp\left(-x_n\right)\right].
\end{aligned} \tag{8}$$

It follows that

$$\prod_{\substack{m=1 \\ m\neq n}}^{N} \int_{-\infty}^{\infty} \mathrm{d}x_m P\left(x_n + \ln p_n \geq x_m + \ln p_m | x_n, x_m\right) P\left(x_m\right) = \exp\left[-\frac{\sum_{m=1, m\neq n}^{N} p_m}{p_n}\exp\left(-x_n\right)\right] \tag{9}$$

Therefore,

$$\begin{aligned}
P\left(n = \arg\max_m \left(x_m + \ln p_m\right)\right) &= \int_{-\infty}^{\infty} \mathrm{d}x_n f\left(x_n\right) \exp\left[-\frac{\sum_{m=1, m\neq n}^{N} p_m}{p_n}\exp\left(-x_n\right)\right] \\
&= \int_{-\infty}^{\infty} \mathrm{d}x_n \exp\left(-x_n\right) \exp\left[-\exp\left(-x_n\right) - \frac{\sum_{m=1, m\neq n}^{N} p_m}{p_n}\exp\left(-x_n\right)\right] \\
&= \int_{-\infty}^{\infty} \mathrm{d}x_n \exp\left(-x_n\right) \exp\left[-\frac{\sum_{m=1}^{N} p_m}{p_n}\exp\left(-x_n\right)\right].
\end{aligned} \tag{10}$$

At first this integral may seem intimidating, however, it can be evaluated through the simple substitution of $u = \exp\left(-x_n\right)$, and the final result is what we are looking for:

$$P\left(n = \arg\max_m \left(x_m + \ln p_m\right)\right) = \int_{0}^{\infty} \mathrm{d}u \exp\left(-\frac{\sum_{m=1}^{N} p_m}{p_n} u\right) = \frac{p_n}{\sum_{m=1}^{N} p_m}. \tag{11}$$