A note on the density of Gumbel-softmax

Tomonari MASADA @ Nagasaki University

May 29, 2019

This note explicates some details of the discussion given in Appendix B of [1].

The Gumbel-softmax trick gives a k-dimensional sample vector $\boldsymbol{y} = (y_1, \ldots, y_k) \in \Delta^{k-1}$ whose entries are obtained as

$$y_i = \frac{\exp((\log(\pi_i) + g_i)/\tau)}{\sum_{j=1}^k \exp((\log(\pi_j) + g_j)/\tau)} \quad \text{for } i = 1, \ldots, k, \tag{1}$$

by using $g_1, \ldots, g_k$, which are i.i.d samples drawn from Gumbel$(0, 1)$.

Define $x_i = \log(\pi_i)$. Then $\boldsymbol{y}$ is rewritten as

$$y_i = \frac{\exp((x_i + g_i)/\tau)}{\sum_{j=1}^k \exp((x_j + g_j)/\tau)} \quad \text{for } i = 1, \ldots, k, \tag{2}$$

Divide both numerator and denominator by $\exp((x_k + g_k)/\tau)$.

$$y_i = \frac{\exp((x_i + g_i - (x_k + g_k))/\tau)}{\sum_{j=1}^k \exp((x_j + g_j - (x_k + g_k))/\tau)} \quad \text{for } i = 1, \ldots, k, \tag{3}$$

Define $u_i = x_i + g_i - (x_k + g_k)$ for $i = 1, \ldots, k-1$, where $g_i \sim$ Gumbel$(0, 1)$. When g_k is given, $g_i = u_i - x_i + (x_k + g_k)$ and u_i can thus be regarded as a sample from the Gumbel whose mean is $x_i - (x_k + g_k)$ and scale parameter is 1. Therefore, $p(u_i|g_k) = e^{-\{(u_i - x_i + (x_k + g_k)) + e^{-(u_i - x_i + (x_k + g_k))}\}}$. Consequently, the density $p(u_1, \ldots, u_{k-1})$ is given as follows:

$$\begin{aligned}
p(u_1, \ldots, u_{k-1}) &= \int_{-\infty}^{\infty} dg_k p(u_1, \ldots, u_{k-1}|g_k) p(g_k) \\
&= \int_{-\infty}^{\infty} dg_k p(g_k) \prod_{i=1}^{k-1} p(u_i|g_k) \\
&= \int_{-\infty}^{\infty} dg_k e^{-g_k - e^{-g_k}} \prod_{i=1}^{k-1} e^{x_i - u_i - x_k - g_k - e^{x_i - u_i - x_k - g_k}}
\end{aligned} \tag{4}$$

Perform a change of variables with $v = e^{-g_k}$. Then $\frac{dv}{dg_k} = -e^{-g_k}$. Therefore, $dv = -e^{-g_k} dg_k$ and $dg_k = -dv e^{g_k} = -dv/v$.

$$\begin{aligned}
p(u_1, \ldots, u_{k-1}) &= \int_{\infty}^{0} (-dv) e^{-v} \prod_{i=1}^{k-1} v e^{x_i - u_i - x_k - v e^{x_i - u_i - x_k}} \\
&= \prod_{i=1}^{k-1} e^{x_i - u_i - x_k} \int_0^{\infty} dv e^{-v} v^{k-1} \prod_{i=1}^{k-1} e^{-v e^{x_i - u_i - x_k}} \\
&= e^{-(k-1)x_k} \prod_{i=1}^{k-1} e^{x_i - u_i} \int_0^{\infty} dv v^{k-1} e^{-v(1 + e^{-x_k} \sum_{i=1}^{k-1} (e^{x_i - u_i}))}
\end{aligned} \tag{5}$$

Recall the following fact related to Gamma integral:

$$\int_0^{\infty} x^{z-1} e^{-ax} dx = \int_0^{\infty} \left(\frac{y}{a}\right)^{z-1} e^{-y} \frac{dy}{a} = \left(\frac{1}{a}\right)^z \int_0^{\infty} y^{z-1} e^{-y} dy = a^{-z} \Gamma(z) \tag{6}$$

Therefore,

$$\begin{aligned}
p(u_1,\ldots,u_{k-1}) &= e^{-(k-1)x_k}\prod_{i=1}^{k-1} e^{x_i-u_i}\int_0^\infty dv v^{k-1} e^{-v(1+e^{-x_k}\sum_{i=1}^{k-1}(e^{x_i-u_i}))} \\
&= e^{-kx_k}e^{x_k}\prod_{i=1}^{k-1} e^{x_i-u_i}\left(1+e^{-x_k}\sum_{i=1}^{k-1}(e^{x_i-u_i})\right)^{-k}\Gamma(k) \\
&= e^{x_k}\prod_{i=1}^{k-1} e^{x_i-u_i}\left(e^{x_k}+\sum_{i=1}^{k-1}(e^{x_i-u_i})\right)^{-k}\Gamma(k) \\
&= \exp\left(x_k+\sum_{i=1}^{k-1}(x_i-u_i)\right)\left(e^{x_k}+\sum_{i=1}^{k-1}(e^{x_i-u_i})\right)^{-k}\Gamma(k) \qquad (7)
\end{aligned}$$

Define $u_k = 0$. Then

$$p(u_1,\ldots,u_{k-1}) = \Gamma(k)\left(\prod_{i=1}^{k}\exp(x_i-u_i)\right)\left(\sum_{i=1}^{k}\exp(x_i-u_i)\right)^{-k} \qquad (8)$$

A k-dimensional sample vector $\boldsymbol{y} = (y_1,\ldots,y_k) \in \Delta^{k-1}$ is obtained from $u_1,\ldots,u_{k-1}$ by applying a deterministic transformation $\boldsymbol{h}$:

$$h_i(u_1,\ldots,u_{k-1}) = \frac{\exp(u_i/\tau)}{1+\sum_{j=1}^{k-1}\exp(u_j/\tau)} \quad \text{for } i=1,\ldots,k-1 \qquad (9)$$

as follows:

$$y_i = h_i(u_1,\ldots,u_{k-1}) \quad \text{for } i=1,\ldots,k-1 \qquad (10)$$

Note that y_k is fixed given $y_1,\ldots,y_{k-1}$:

$$y_k = \frac{1}{1+\sum_{j=1}^{k-1}\exp(u_j/\tau)} = 1-\sum_{j=1}^{k-1} y_j \qquad (11)$$

By using the change of variables we can obtain the density function for $\boldsymbol{y}$:

$$p(\boldsymbol{y}) = p(\boldsymbol{h}^{-1}(y_1,\ldots,y_{k-1}))\det\left(\frac{\partial(h_1^{-1}(y_1,\ldots,y_{k-1}),\ldots,h_{k-1}^{-1}(y_1,\ldots,y_{k-1}))}{\partial(y_1,\ldots,y_{k-1})}\right) \qquad (12)$$

The inverse of $\boldsymbol{h}$ is obtained as follows:

$$\begin{aligned}
y_i &= \frac{\exp(u_i/\tau)}{1+\sum_{j=1}^{k-1}\exp(u_j/\tau)} \\
y_i &= y_k\exp(u_i/\tau) \quad \text{from } y_k = \tfrac{1}{1+\sum_{j=1}^{k-1}\exp(u_j/\tau)} \\
\log y_i &= \log y_k + u_i/\tau \\
\therefore h_i^{-1}(y_1,\ldots,y_{k-1}) &= u_i = \tau\times(\log y_i - \log y_k) = \tau\times\left(\log y_i - \log\left(1-\sum_{j=1}^{k-1} y_j\right)\right) \qquad (13)
\end{aligned}$$

Therefore, we obtain the Jacobian:

$$\begin{aligned}
\frac{\partial h_i^{-1}(y_1,\ldots,y_{k-1})}{\partial y_i} &= \tau\times\left(\frac{1}{y_i}-\frac{1}{y_k}\frac{\partial y_k}{\partial y_i}\right) = \tau\times\left(\frac{1}{y_i}+\frac{1}{y_k}\right) \qquad (14) \\
\frac{\partial h_i^{-1}(y_1,\ldots,y_{k-1})}{\partial y_j} &= \tau\times\left(-\frac{1}{y_k}\frac{\partial y_k}{\partial y_j}\right) = \tau\times\frac{1}{y_k} \quad \text{for } j\neq i \qquad (15)
\end{aligned}$$

Eqs. from (24) to (28) are easy to understand.

References

[1] E. Jang, S. Gu, and B. Poole. Categorical representation with Gumbel-softmax. ICLR, 2017.