A. Proof of the Gumbel-Top-k trick

Theorem 1. *For $k \leq n$, let $I_1^*, ..., I_k^* = \arg\operatorname{top} k \, G_{\phi_i}$. Then $I_1^*, ..., I_k^*$ is an (ordered) sample without replacement from the Categorical $\left(\frac{\exp \phi_i}{\sum_{j \in N} \exp \phi_j}, i \in N\right)$ distribution, e.g. for a realization $i_1^*, ..., i_k^*$ it holds that*

$$P\left(I_1^* = i_1^*, ..., I_k^* = i_k^*\right) = \prod_{j=1}^{k} \frac{\exp \phi_{i_j^*}}{\sum_{\ell \in N_j^*} \exp \phi_\ell} \tag{15}$$

where $N_j^* = N \setminus \{i_1^*, ..., i_{j-1}^*\}$ is the domain (without replacement) for the j-th sampled element.

Proof. First note that

$$\begin{aligned}
&P\left(I_k^* = i_k^* \middle| I_1^* = i_1^*, ..., I_{k-1}^* = i_{k-1}^*\right) \\
&= P\left(i_k^* = \arg\max_{i \in N_k^*} G_{\phi_i} \middle| I_1^* = i_1^*, ..., I_{k-1}^* = i_{k-1}^*\right) \\
&= P\left(i_k^* = \arg\max_{i \in N_k^*} G_{\phi_i} \middle| \max_{i \in N_k^*} G_{\phi_i} < G_{\phi_{i_{k-1}^*}}\right) \quad (16) \\
&= P\left(i_k^* = \arg\max_{i \in N_k^*} G_{\phi_i}\right) \quad (17) \\
&= \frac{\exp \phi_{i_k^*}}{\sum_{\ell \in N_k^*} \exp \phi_\ell}. \quad (18)
\end{aligned}$$

The step from (16) to (17) follows from the independence of the $\max$ and $\arg\max$ (Section 2.3) and the step from (17) to (18) uses the Gumbel-Max trick. The proof follows by induction on k. The case $k = 1$ is the Gumbel-Max trick, while if we assume the result (15) proven for $k - 1$, then

$$\begin{aligned}
&P\left(I_1^* = i_1^*, ..., I_k^* = i_k^*\right) \\
&= P\left(I_k^* = i_k^* \middle| I_1^* = i_1^*, ..., I_{k-1}^* = i_{k-1}^*\right) \\
&\quad \cdot P\left(I_1^* = i_1^*, ..., I_{k-1}^* = i_{k-1}^*\right) \\
&= \frac{\exp \phi_{i_k^*}}{\sum_{\ell \in N_k^*} \exp \phi_\ell} \cdot \prod_{j=1}^{k-1} \frac{\exp \phi_{i_j^*}}{\sum_{\ell \in N_j^*} \exp \phi_\ell} \quad (19) \\
&= \prod_{j=1}^{k} \frac{\exp \phi_{i_j^*}}{\sum_{\ell \in N_j^*} \exp \phi_\ell}.
\end{aligned}$$

In (19) we have used Equation (18) and Equation (15) for $k - 1$ by induction. □

B. Sampling set of Gumbels with maximum T

B.1. The truncated Gumbel distribution

A random variable G' has a *truncated* Gumbel distribution with location ϕ and maximum T (e.g. $G' \sim \text{TruncatedGumbel}(\phi, T)$) with CDF $F_{\phi,T}(g)$ if:

$$\begin{aligned}
&F_{\phi,T}(g) \\
&= P(G' \leq g) \\
&= P(G \leq g | G \leq T) \\
&= \frac{P(G \leq g \cap G \leq T)}{P(G \leq T)} \\
&= \frac{P(G \leq \min(g, T))}{P(G \leq T)} \\
&= \frac{F_\phi(\min(g, T))}{F_\phi(T)} \\
&= \frac{\exp(-\exp(\phi - \min(g, T)))}{\exp(-\exp(\phi - T))} \\
&= \exp(\exp(\phi - T) - \exp(\phi - \min(g, T))). \quad (20)
\end{aligned}$$

The inverse CDF is:

$$F_{\phi,T}^{-1}(u) = \phi - \log(\exp(\phi - T) - \log u). \tag{21}$$

B.2. Sampling set of Gumbels with maximum T

In order to sample a set of Gumbel variables $\{\tilde{G}_{\phi_i} | \max_i \tilde{G}_{\phi_i} = T\}$, e.g. with their maximum being *exactly* T, we can first sample the $\arg\max$, i^* and then sample the Gumbels conditionally on both the $\max$ and $\arg\max$:

1. Sample $i^* \sim \text{Categorical}\left(\frac{\exp \phi_i}{\sum_j \exp \phi_j}\right)$. We do not need to condition on T since the $\arg\max$ i^* is independent of the $\max$ T (Section 2.3).
2. Set $\tilde{G}_{\phi_{i^*}} = T$, since this follows from conditioning on the $\max$ T and $\arg\max$ i^*.
3. Sample $\tilde{G}_{\phi_i} \sim \text{TruncatedGumbel}(\phi_i, T)$ for $i \neq i^*$. This works because, conditioning on the $\max$ T and $\arg\max$ i^*, it holds that:
$$\begin{aligned}
&P(\tilde{G}_{\phi_i} < g | \max_i \tilde{G}_{\phi_i} = T, \arg\max_i \tilde{G}_{\phi_i} = i^*, i \neq i^*) \\
&= P(\tilde{G}_{\phi_i} < g | \tilde{G}_{\phi_i} < T).
\end{aligned}$$

Equivalently, we can let $G_{\phi_i} \sim \text{Gumbel}(\phi_i)$, let $Z = \max_i G_{\phi_i}$ and define

$$\begin{aligned}
\tilde{G}_{\phi_i} &= F_{\phi_i,T}^{-1}(F_{\phi_i,Z}(G_{\phi_i})) \\
&= \phi_i - \log(\exp(\phi_i - T) \\
&\qquad - \exp(\phi_i - Z) + \exp(\phi_i - G_{\phi_i})) \\
&= -\log(\exp(-T) - \exp(-Z) + \exp(-G_{\phi_i})).
\end{aligned} \tag{22}$$

Here we have used (20) and (21). Since the transformation (22) is monotonically increasing, it preserves the $\arg\max$ and it follows from the Gumbel-Max trick (3) that

$$\arg\max_i \tilde{G}_{\phi_i} = \arg\max_i G_{\phi_i} \sim \text{Categorical}\left(\frac{\exp \phi_i}{\sum_j \exp \phi_j}\right).$$

We can think of this as using the Gumbel-Max trick for step 1 (sampling the argmax) in the sampling process described above. Additionally, for $i = \arg\max_i G_{\phi_i}$:

$$\tilde{G}_{\phi_i} = F^{-1}_{\phi_i,T}(F_{\phi_i,Z}(G_{\phi_i})) = F^{-1}_{\phi_i,T}(F_{\phi_i,Z}(Z)) = T$$

so here we recover step 2 (setting $\tilde{G}_{\phi_{i^*}} = T$). For $i \neq \arg\max_i G_{\phi_i}$ it holds that:

$$\begin{aligned}
&P(\tilde{G}_{\phi_i} \leq g | i \neq \arg\max_i G_{\phi_i}) \\
=&\mathbb{E}_Z(P(\tilde{G}_{\phi_i} \leq g | Z, i \neq \arg\max_i G_{\phi_i})) \\
=&\mathbb{E}_Z(P(\tilde{G}_{\phi_i} \leq g | Z, G_{\phi_i} < Z)) \\
=&\mathbb{E}_Z(P(F^{-1}_{\phi_i,T}(F_{\phi_i,Z}(G_{\phi_i})) \leq g | Z, G_{\phi_i} < Z)) \\
=&\mathbb{E}_Z(P(G_{\phi_i} \leq F^{-1}_{\phi_i,Z}(F_{\phi_i,T}(g)) | Z, G_{\phi_i} < Z)) \\
=&\mathbb{E}_Z(F_{\phi_i,Z}(F^{-1}_{\phi_i,Z}(F_{\phi_i,T}(g)))) \\
=&\mathbb{E}_Z(F_{\phi_i,T}(g)) = F_{\phi_i,T}(g).
\end{aligned}$$

This means that $\tilde{G}_{\phi_i} \sim \text{TruncatedGumbel}(\phi_i, T)$, so this is equivalent to step 3 (sampling $\tilde{G}_{\phi_i} \sim \text{TruncatedGumbel}(\phi_i, T)$ for $i \neq i^*$).

B.3. Numeric stability of truncated Gumbel computation

Direct computation of (22) can be unstable as large terms need to be exponentiated. Instead, we compute:

$$v_i = T - G_{\phi_i} + \text{log1mexp}(G_{\phi_i} - Z) \tag{23}$$

$$\tilde{G}_{\phi_i} = T - \max(0, v_i) - \text{log1pexp}(-|v_i|) \tag{24}$$

where we have defined

$$\begin{aligned}
\text{log1mexp}(a) &= \log(1 - \exp(a)), \quad a \leq 0 \\
\text{log1pexp}(a) &= \log(1 + \exp(a)).
\end{aligned}$$

This is equivalent as

$$\begin{aligned}
&T - \max(0, v_i) - \log(1 + \exp(-|v_i|)) \\
=&T - \log(1 + \exp(v_i)) \\
=&T - \log\left(1 + \exp\left(T - G_{\phi_i} + \log\left(1 - \exp\left(G_{\phi_i} - Z\right)\right)\right)\right) \\
=&T - \log\left(1 + \exp\left(T - G_{\phi_i}\right)\left(1 - \exp\left(G_{\phi_i} - Z\right)\right)\right) \\
=&T - \log\left(1 + \exp\left(T - G_{\phi_i}\right) - \exp\left(T - Z\right)\right) \\
=&-\log\left(\exp(-T) + \exp(-G_{\phi_i}) - \exp(-Z)\right) \\
=&\tilde{G}_{\phi_i}
\end{aligned}$$

The first step can be easily verified by considering the cases $v_i < 0$ and $v_i \geq 0$. log1mexp and log1pexp can be computed accurately using $\text{log1p}(a) = \log(1 + a)$ and $\text{expm1}(a) = \exp(a) - 1$ (Mächler, 2012):

$$\text{log1mexp}(a) = \begin{cases} \log(-\text{expm1}(a)) & a > -0.693 \\ \text{log1p}(-\exp(a)) & \text{otherwise} \end{cases}$$

$$\text{log1pexp}(a) = \begin{cases} \text{log1p}(\exp(a)) & a < 18 \\ x + \exp(a) & \text{otherwise} \end{cases}$$

C. Numerical stability of importance weights

We have to take care computing the importance weights as depending on the entropy the terms in the quotient $\frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x})}{q_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x})}$ can become very small, and in our case the computation of $P(G_{\phi_i} > \kappa) = 1 - \exp(-\exp(\phi_i - \kappa))$ can suffer from catastrophic cancellation. We can rewrite this expression using the more numerically stable implementation $\text{exp1m}(x) = \exp(x) - 1$ as $p(G_{\phi_i} > \kappa) = -\text{exp1m}(-\exp(\phi_i - \kappa))$ but in some cases this still suffers from instability as $\exp(\phi_i - \kappa)$ can underflow if $\phi_i - \kappa$ is small. Instead, for $\phi_i - \kappa < -10$ we use the identity

$$\log(1 - \exp(-z)) = \log(z) - \frac{z}{2} + \frac{z^2}{24} - \frac{z^4}{2880} + \mathcal{O}(z^6)$$

to directly compute the log importance weight using $z = \exp(\phi_i - \kappa)$ and $\phi_i = \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x})$ (we assume ϕ_i is normalized):

$$\begin{aligned}
&\log\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x})}{q_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x})}\right) = \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}) \\
=& \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}) - \log\left(1 - \exp(-\exp(\phi_i - \kappa))\right) \\
=& \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}) - \log\left(1 - \exp(-z)\right) \\
=& \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}) - \left(\log(z) - \frac{z}{2} + \frac{z^2}{24} - \frac{z^4}{2880} + \mathcal{O}(z^6)\right) \\
=& \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}) - \left(\phi_i - \kappa - \frac{z}{2} + \frac{z^2}{24} - \frac{z^4}{2880} + \mathcal{O}(z^6)\right) \\
=& \kappa + \frac{z}{2} - \frac{z^2}{24} + \frac{z^4}{2880} + \mathcal{O}(z^6)
\end{aligned}$$

If $\phi_i - \kappa < -10$ then $0 < z < 10^{-6}$ so this computation will not lose any significant digits.

D. Proof of unbiasedness of priority sampling estimator

The following proof is adapted from the proofs by Duffield et al. (2007) and Vieira (2017). For generality of the proof, we write $f(i) = f(\boldsymbol{y}^i)$, $p_i = p_{\boldsymbol{\theta}}(\boldsymbol{y}^i|\boldsymbol{x})$ and $q_i(\kappa) = q_{\boldsymbol{\theta},\kappa}(\boldsymbol{y}^i|\boldsymbol{x})$, and we consider general keys h_i (not necessarily Gumbel perturbations).

We assume we have a probability distribution over a finite domain $1, ..., n$ with normalized probabilities p_i, e.g. $\sum_{i=1}^n p_i = 1$. For a given function $f(i)$ we want to estimate the expectation

$$\mathbb{E}[f(i)] = \sum_{i=1}^n p_i f(i).$$

Each element i has an associated random key h_i and we define $q_i(a) = P(h_i > a)$. This way, if we *know* the threshold a it holds that $q_i(a) = P(i \in S)$ is the probability that element i is in the sample S. As was noted by Vieira (2017), the actual distribution of the key does not influence the unbiasedness of the estimator but does determine the effective sampling scheme. Using the Gumbel perturbed log-probabilities as keys (e.g. $h_i = G_{\phi_i}$) is equivalent to the PPSWOR scheme described by Vieira (2017).

We define shorthand notation $h_{1:n} = \{h_1, ..., h_n\}$, $h_{-i} = \{h_1, ..., h_{i-1}, h_{i+1}, ..., h_n\} = h_{1:n} \setminus \{h_i\}$. For a given sample size k, let κ be the $(k+1)$-th largest element of $h_{1:n}$, so κ is the *empirical threshold*. Let κ'_i be the k-th largest element of h_{-i} (the k-th largest of all *other* elements).

Similar to Duffield et al. (2007) we will show that every element i in our sample contributes an unbiased estimate of $\mathbb{E}[f(i)]$, so that the total estimator is unbiased. Formally, we will prove that

$$\mathbb{E}_{h_{1:n}}\left[\frac{\mathbb{1}_{\{i \in S\}}}{q_i(\kappa)}\right] = 1 \tag{25}$$

from which the result follows:

$$\begin{aligned}
&\mathbb{E}_{h_{1:n}}\left[\sum_{i \in S} \frac{p_i}{q_i(\kappa)} f(i)\right] \\
=&\mathbb{E}_{h_{1:n}}\left[\sum_{i=1}^n \frac{p_i}{q_i(\kappa)} f(i) \mathbb{1}_{\{i \in S\}}\right] \\
=&\sum_{i=1}^n p_i f(i) \cdot \mathbb{E}_{h_{1:n}}\left[\frac{\mathbb{1}_{\{i \in S\}}}{q_i(\kappa)}\right] \\
=&\sum_{i=1}^n p_i f(i) \cdot 1 = \sum_{i=1}^n p_i f(i) = \mathbb{E}[f(i)]
\end{aligned}$$

To prove (25), we make use of the observation (slightly rephrased) by Duffield et al. (2007) that conditioning on h_{-i}, we know κ'_i and the event $i \in S$ implies that $\kappa = \kappa'_i$ since i will only be in the sample if $h_i > \kappa'_i$ which means that κ'_i is the $k+1$-th largest value of $h_{-i} \cup \{h_i\} = h_{1:n}$. The reverse is also true (if $\kappa = \kappa'_i$ then h_i must be larger than κ'_i since otherwise the $k+1$-th largest value of $h_{1:n}$ will be smaller than κ'_i).

$$\begin{aligned}
&\mathbb{E}_{h_{1:n}}\left[\frac{\mathbb{1}_{\{i \in S\}}}{q_i(\kappa)}\right] \\
=&\mathbb{E}_{h_{-i}}\left[\mathbb{E}_{h_i}\left[\frac{\mathbb{1}_{\{i \in S\}}}{q_i(\kappa)}\middle| h_i\right]\right] \\
=&\mathbb{E}_{h_{-i}}\left[\mathbb{E}_{h_i}\left[\frac{\mathbb{1}_{\{i \in S\}}}{q_i(\kappa)}\middle| h_{-i}, i \in S\right] P(i \in S | h_{-i})\right. \\
&\quad \left. + \mathbb{E}_{h_i}\left[\frac{\mathbb{1}_{\{i \in S\}}}{q_i(\kappa)}\middle| h_{-i}, i \notin S\right] P(i \notin S | h_{-i})\right] \\
=&\mathbb{E}_{h_{-i}}\left[\mathbb{E}_{h_i}\left[\frac{1}{q_i(\kappa)}\middle| h_{-i}, i \in S\right] P(i \in S | h_{-i}) + 0\right] \\
=&\mathbb{E}_{h_{-i}}\left[\mathbb{E}_{h_i}\left[\frac{1}{q_i(\kappa)}\middle| h_{-i}, i \in S\right] q_i(\kappa'_i)\right] \\
=&\mathbb{E}_{h_{-i}}\left[\mathbb{E}_{h_i}\left[\frac{1}{q_i(\kappa)}\middle| \kappa = \kappa'_i\right] q_i(\kappa'_i)\right] \\
=&\mathbb{E}_{h_{-i}}\left[\mathbb{E}_{h_i}\left[\frac{1}{q_i(\kappa'_i)}\right] q_i(\kappa'_i)\right] \\
=&\mathbb{E}_{h_{-i}}\left[\frac{1}{q_i(\kappa'_i)} q_i(\kappa'_i)\right] = \mathbb{E}_{h_{-i}}[1] = 1
\end{aligned}$$