Derivation of Exponential Index Distribution

The goal is to generate indices $i \in \{0, 1, \ldots, \text{len\_seq} - 1\}$ with probability proportional to e^i.

Initial Probability Distribution

Starting with unnormalized probabilities:

$$P(i) \propto e^i$$

To make a proper probability distribution, we normalize by the sum of all possible values:

$$P(i) = \frac{e^i}{\sum_{j=0}^{\text{len\_seq}-1} e^j}$$

Geometric Series Normalization

The denominator is a geometric series. For any geometric series where r is the common ratio:

$$\sum_{j=0}^{n-1} r^j = \frac{r^n - 1}{r - 1}$$

In our case, $r = e$, so:

$$\sum_{j=0}^{\text{len\_seq}-1} e^j = \frac{e^{\text{len\_seq}} - 1}{e - 1}$$

Therefore:

$$P(i) = \frac{e^i}{\frac{e^{\text{len\_seq}}-1}{e-1}} = \frac{(e-1)e^i}{e^{\text{len\_seq}} - 1}$$

Cumulative Distribution Function

For inverse transform sampling, we need the Cumulative Distribution Function (CDF). The CDF up to k represents the probability of obtaining any value up to and including k:

$$F(k) = \sum_{j=0}^{k} P(j) = \sum_{j=0}^{k} \frac{(e-1)e^j}{e^{\text{len\_seq}} - 1}$$

$$F(k) = \frac{e-1}{e^{\text{len\_seq}} - 1} \sum_{j=0}^{k} e^j$$

Using the geometric series formula for the sum from 0 to k:

$$\sum_{j=0}^{k} e^j = \frac{e^{k+1} - 1}{e - 1}$$

Therefore:

$$F(k) = \frac{e - 1}{e^{\text{len\_seq}} - 1} \cdot \frac{e^{k+1} - 1}{e - 1} = \frac{e^{k+1} - 1}{e^{\text{len\_seq}} - 1}$$

Inverse Transform Sampling

For inverse transform sampling, let $u \sim U(0, 1)$ and solve $F(k) = u$:

$$u = \frac{e^{k+1} - 1}{e^{\text{len\_seq}} - 1}$$

$$u(e^{\text{len\_seq}} - 1) = e^{k+1} - 1$$

$$e^{k+1} = 1 + u(e^{\text{len\_seq}} - 1)$$

$$k + 1 = \ln(1 + u(e^{\text{len\_seq}} - 1))$$

$$k = \ln(1 + u(e^{\text{len\_seq}} - 1)) - 1$$

Approximation for Large Sequence Length

For large sequence length, we can approximate:

$$e^{\text{len\_seq}} - 1 \approx e^{\text{len\_seq}},$$

and similarly:

$$e^{k+1} - 1 \approx e^{k+1}.$$

As a result, the approximate CDF becomes:

$$F(k) \approx \frac{e^{k+1}}{e^{\text{len\_seq}}} = e^{k+1-\text{len\_seq}}.$$

The approximate inverse transform sampling can be derived similarly, let $u \sim U(0, 1)$ and solve $F(k) = u$.

$$u = e^{k+1-\text{len\_seq}}$$

Taking the natural logarithm:

$$\ln(u) = k + 1 - \text{len\_seq}$$

Finally:

$$k = \text{len\_seq} + \ln(u) - 1$$