Deterministic WaveNet Inference

- ## Inverse transform sampling

  Inverse transform sampling is a method for sampling from a probability distribution given the inverse of the distribution's cumulative distribution function (CDF).

  Say X is a random variable with CDF $f(x)$. To sample from X,

  1. Generate a sample u from $\text{unif}(0, 1)$.
  2. Compute $f^{-1}(u)$.

- ## Non-deterministic WaveNet inference

  Given the previously generated audio samples $x_1, x_2, \ldots, x_n$, WaveNet outputs the parameters describing a probability distribution for the next audio sample. The distribution is a mixture of k logistic distributions:

$$\underbrace{\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{bmatrix}}_{\text{logit probs}}, \quad \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}}_{\text{means}}, \quad \underbrace{\begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_k \end{bmatrix}}_{\text{log scales}}.$$

  The usual (non-deterministic) way to sample from this distribution is:

  1. Choose which one of the k distributions to sample from by sampling from the softmax distribution using the Gumbel-max trick:

$$c = \text{argmax}_i(\gamma_i + g_i)$$

     where $g_i = -\log(-\log u_i)$ and u_i are sampled from $\text{unif}(0, 1)$.

2. Sample from the c^{th} logistic distribution using inverse transform sampling.

- ## Deterministic WaveNet inference

Suppose we want to generate N samples of audio. The trick to making the inference deterministic is to generate the u-values ahead of time:

1. Let U be an $N \times k$ matrix whose entries are sampled from $\text{unif}(0, 1)$, and let $\mathbf{v}$ be an N-dimensional vector whose entries are sampled from $\text{unif}(0, 1)$.
2. For each of the N audio samples:
   - Let

$$\underbrace{\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_k \end{bmatrix}}_{\text{logit probs}}, \quad \underbrace{\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}}_{\text{means}}, \quad \underbrace{\begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_k \end{bmatrix}}_{\text{log scales}}$$

   be the distribution parameters generated by WaveNet given the previously generated audio samples $x_1, x_2, \ldots, x_n$, and let

$$\mathbf{p} = \text{softmax}\left(100 \times \begin{bmatrix} \gamma_1 + g_1 \\ \gamma_2 + g_2 \\ \vdots \\ \gamma_k + g_k \end{bmatrix}\right)$$

   where $g_i = -\log(-\log u_{n,i})$.
   - Let $\mu = \sum_{i=1}^{k}(p_i \cdot \mu_i)$ and $s = \exp\left[\sum_{i=1}^{k}(p_i \cdot l_i)\right]$.
   - Let $f(x)$ be the CDF of the logistic distribution that has mean μ and scale s. Generate the next audio sample by computing

$$x_{n+1} = f^{-1}(v_n).$$