Roll No. ____________________ **Nov 2024**

**UCS749: Conversational AI: Speech Proc. [...]** Time: **2 hours** MM:–

Faculty: RGB

# Thapar Institute of Engineering & Technology

Computer Science & Engineering Department

EXERCISE (TTS) (2024-25 ODD)

1. If $P_\theta$ models a data $\mathscr{D}$ with an intent of generating novel samples, $\mathbf{x} \sim P_\theta$. Is $P_\theta$ a posterior distribution? Comment. [2 marks]

2. If $X \equiv \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ is a temporal sequence and each time step is independently and identically distributed,
   a) Prove that $\log P(X) = \sum_{i=1}^{T} \log P(x_i)$;
   b) If $X$ is a Bernoulli process, *i.e.* if each $x_i$ is a coin toss, with hit rate $p$, the log likelihood of $k$ hits is $k \log p + (T-k)\log(1-p)$.

2 a) $X \equiv \{x_1 \dots x_T\}$

$$P(X) = P(x_1 \dots x_T)$$

$$= P(x_T | x_1 \dots x_{T-1}) \cdot P(x_1 \dots x_{T-1})$$

... (by Bayes Rule)

$$= \prod_{i=1}^{T} P(x_i | x_1 \dots x_{i-1}) \quad \text{—— } ①$$

... (expanding similarly)

But $x_i \perp x_j \; \forall i \neq j$ because iid.

hence $P(x_i | x_1 \dots x_{i-1}) = P(x_i)$ —— ②

from ① & ②,

$$P(X) = \prod_{i=1}^{T} P(x_i)$$

$$\log P(X) = \log \prod_{i=1}^{T} P(x_i) = \sum_{i=1}^{T} \log P(x_i)$$ ∎

2b) For a Bernoulli Process with hit rate 'p'
The probability of 'k' hits in 'T' trials is
given as,

$$P(X) = P(k, T; p) = p^k (1-p)^{T-k}$$

or $$\log P(X) = k \log p + (T-k) \log (1-p)$$

3. If $X \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is the temporal sequence of a Markov process, prove that $\log P(X) = \sum_{i=1}^{T} \log P(x_i \mid x_{i-1})$.

Speech sample $X$ is a temporal sequence of intensities,

$$X \equiv \{x_1 \dots x_T\}$$

Given speech samples $\mathscr{D}$ as evidence, estimate $P_\theta \approx \mathscr{D}$ so that $X \sim P_\theta$ is a valid speech sample.

Let $G_{\theta,\mathcal{N}} : \mathbb{X}^K \to \mathbb{X}$ represent the model.

where,

- $\mathbb{X}$ is the field of inputs.
  - For a continuous model, $\mathbb{X} \equiv \mathbb{R}$; whereas,
  - For categorical model, $\mathbb{X} \equiv \mathbb{R}^{256}_{[0,1]}$;
  - It may also be a hybrid model,
    *e.g.* $G_{\theta,\mathcal{N}} : \mathbb{R}^K \to \mathbb{R}^{256}_{[0,1]}$.
    Can you think how?

Let $G_{\theta,\mathcal{N}} : \mathbb{X}^K \to \mathbb{X}$ represent the model.

where,

- $\mathcal{N}$ is a noise sampler,
  - Generally, implemented as a normal distribution; or
  - Implemented implicitly as dropouts.

Let $G_{\theta,\mathcal{N}} : \mathbb{X}^K \to \mathbb{X}$ represent the model.

where,

- $\theta$ are parameters of the model; and
- $\mathbf{x}_t = G_{\theta,\mathcal{N}}\left(\left[\mathbf{x}_{t-K} \ \dots \ \mathbf{x}_{t-1}\right]\right)$ models the conditional distribution $P(x_t \mid x_{t-K} \dots x_{t-1})$

In case of Conditional Generation,

$\mathbf{x}_t = G_{\theta,\mathcal{N}}\left(\left[\mathbf{x}_{t-K} \;\dots\; \mathbf{x}_{t-1}\right], \mathbf{h}\right)$
models the conditional distribution
$P(x_t \mid x_{t-K} \dots x_{t-1}, \mathbf{h})$

$\mathbf{h}$ represents the global conditions, *e.g.*

- Text input;
- Speaker;
- Accent;
- and so forth.

Let
$X \equiv \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_T \end{bmatrix} \sim G_{\theta,\mathcal{N}}$ be the output of the auto-regressive model, so that
$\mathbf{x}_t = G_{\theta,\mathcal{N}}\left(\begin{bmatrix} \mathbf{x}_{t-K} & \dots & \mathbf{x}_{t-1} \end{bmatrix}\right)$

And,
$Y \equiv \begin{bmatrix} \mathbf{y}_1 & \dots & \mathbf{y}_T \end{bmatrix} \sim \mathcal{D}$ be a speech sample from dataset. Recall that for each time step, the sound intensity is pre-processed so that the values are remapped, quantised, and converted to one-hot vectors, so that $\mathbf{y}_t \in \{0,1\}^{256}; \|\mathbf{y}_t\|_1 = 1$ .

The training objective is the cross entropy function, given as,

$$\underset{G}{\text{minimise}} \quad \underset{X\sim G, Y\sim\mathcal{D}}{\mathbb{E}}\left[\sum_{i,t} y_{i,t} \log x_{i,t}\right]$$