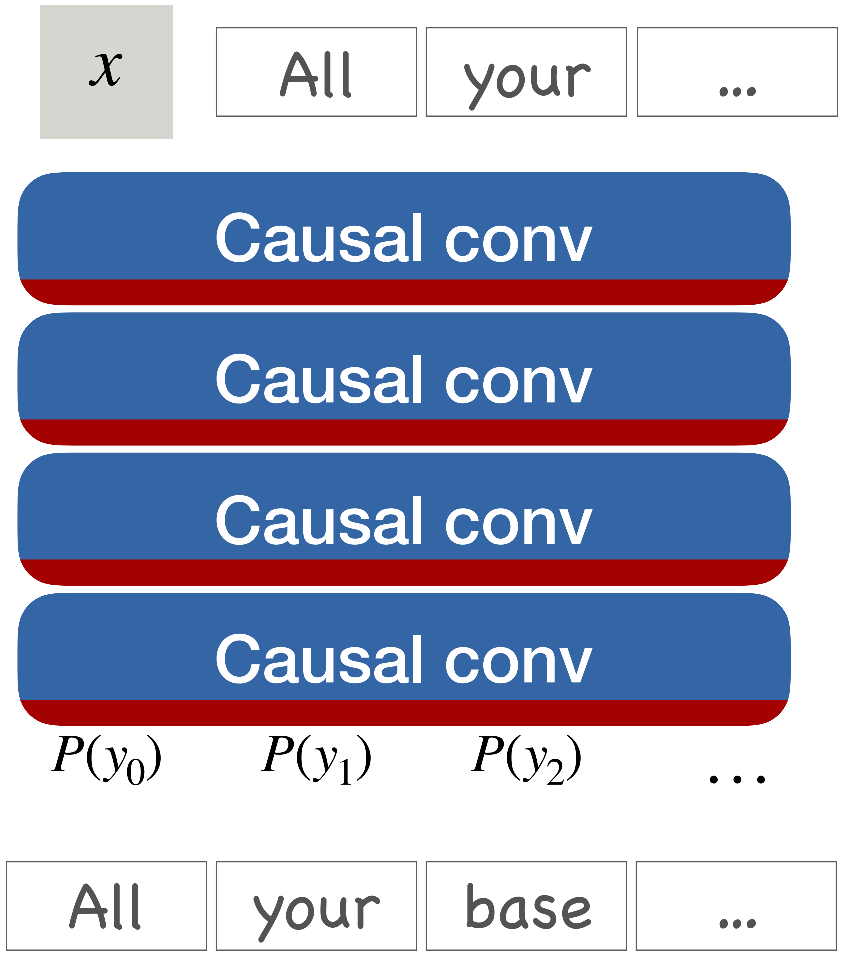
Sampling in sequence models

Sampling

- Example temporal convolutional network
  - Autoregressive

$$P(y_0, y_1, y_2, \ldots) = P(y_0 \mid x) \cdot P(y_1 \mid x, y_0) \cdot P(y_2 \mid x, y_0, y_1) \cdot \ldots$$

- Objective find
  - $\hat{y} = \arg\max_y P(y_0, y_1, y_2, \ldots)$



Greedy sampling

$$P(y_0, y_1, y_2, \ldots) = P(y_0 \,|\, x) \cdot$$
$$P(y_1 \,|\, x, y_0) \cdot$$
$$P(y_2 \,|\, x, y_0, y_1) \cdot$$
$$\ldots$$

- Pick sequentially

  $$\hat{y}_t = \arg\max_{y_t} P(y_t \,|\, x, \hat{y}_0, \hat{y}_1, \ldots)$$

  - Single sample
  - Not optimal

Sequential sampling

$$P(y_0, y_1, y_2, \ldots) = P(y_0 \mid x) \cdot$$
$$P(y_1 \mid x, y_0) \cdot$$
$$P(y_2 \mid x, y_0, y_1) \cdot$$
$$\ldots$$

- For n iterations
  - Sample sequentially $\hat{y}_t \sim P(y_t \mid x, \hat{y}_0, \hat{y}_1, \ldots)$
- Unbiased sampling
  - Not sample efficient

Beam search

- Biased sampling
  - High likelihood samples
- Generally works best

- Keep top k samples S
  - Largest $P(y_0)$
- For t steps
  - For each $\hat{y}_0, \hat{y}_1, \ldots \in S$
    - Compute $P(x, \hat{y}_0, \hat{y}_1, \ldots, y_t)$
  - Keep top k samples S