The problem with modeling sequences

- For most interesting sequence problems (language, genes), learning the joint distribution of observed sequences is intractable.

$$P(sentence) = P(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$$

  - Consider a sequence of length $T = 10$ generated from a vocabulary containing only $N = 1000$ words, the number of possible sentences is 10^{30}. The large branching factor, N, makes estimating the probability of each possible outcome intractable.

Can’t we just use the chain rule?

- Can’t we just use the chain rule of probability to factor the joint distribution into a distribution over its suffix conditioned on its prefix?

$$P(sentence) = P(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}) = \underset{\text{yes}}{P(\mathbf{x}^{(1)})} \cdot \underset{\text{yes}}{P(\mathbf{x}^{(2)} \mid \mathbf{x}^{(1)})} \cdot \underset{\text{yes}}{P(\mathbf{x}^{(3)} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)})} \ldots \underset{\text{No!}}{P(\mathbf{x}^{(T)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T-1)})}$$

- No because we end up with (essentially) the same problem, now it’s too many possible sequences over $T - 1$ tokens.

- We need to make a simplifying assumption regarding the independence of words in a sentence.