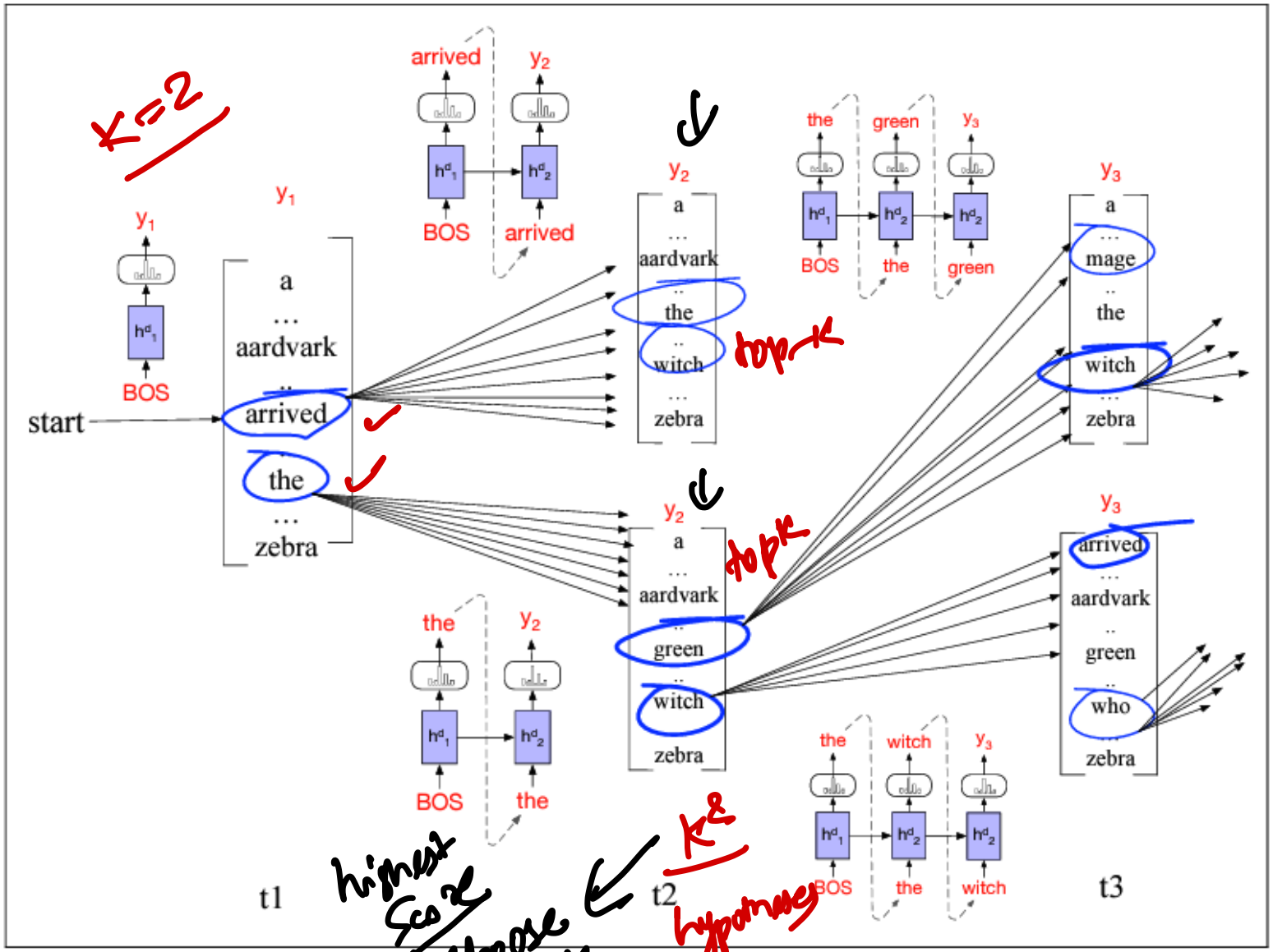
Beam Search: Example



Beam Search: Scoring

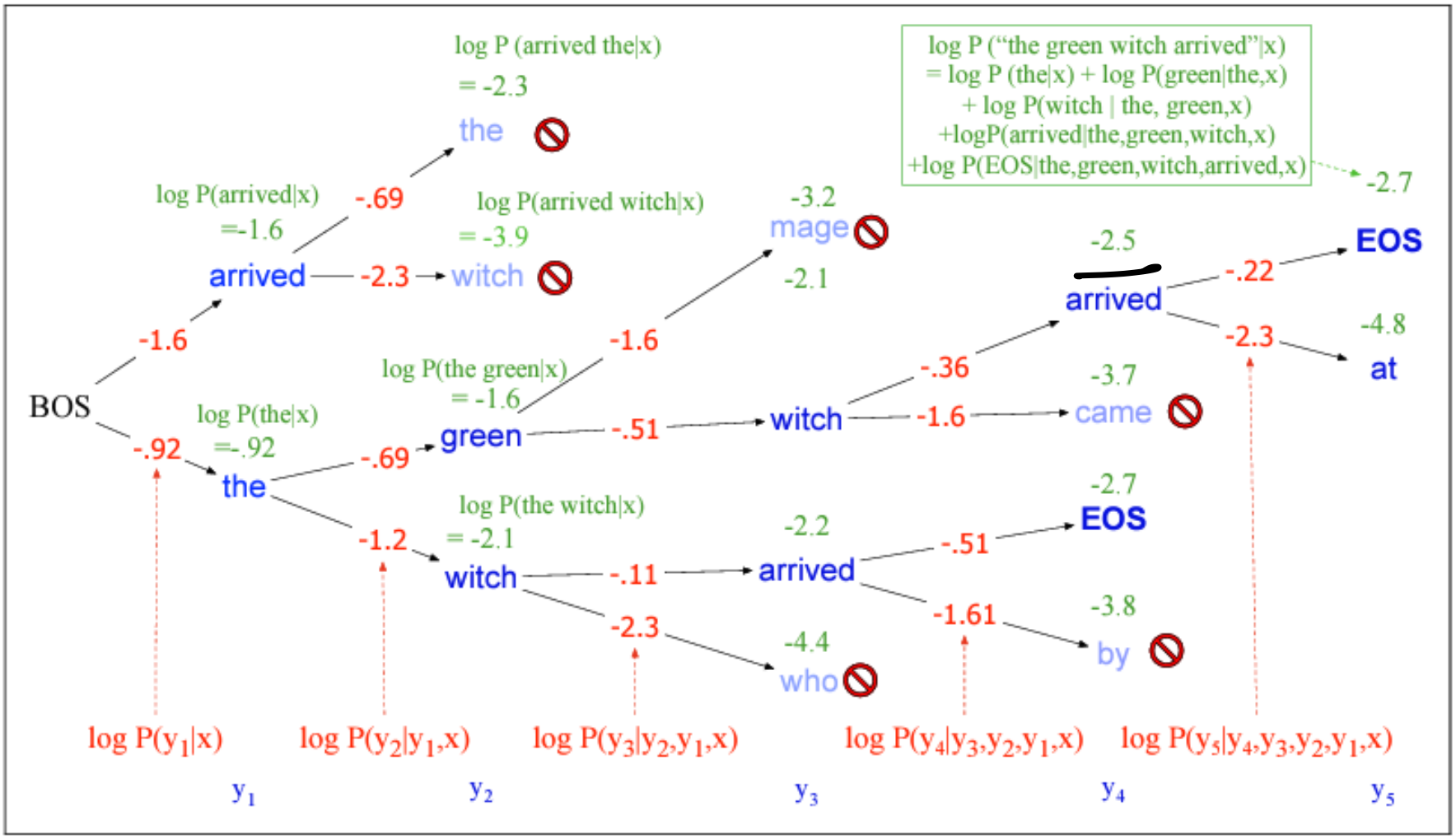
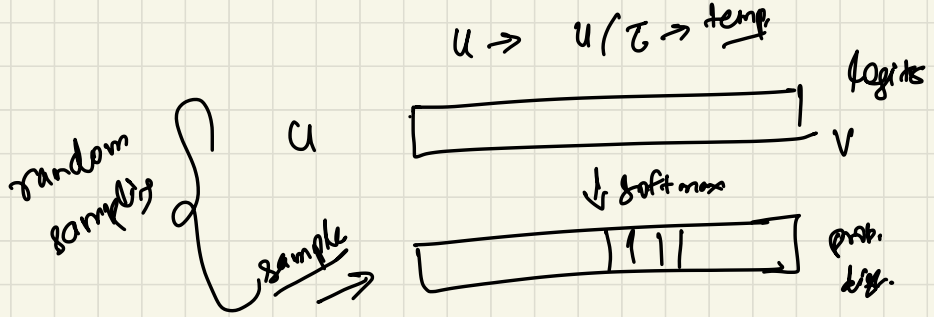


Figure 13.9 Scoring for beam search decoding with a beam width of $k = 2$. We maintain the log probability of each hypothesis in the beam by incrementally adding the logprob of generating each next token. Only the top k paths are extended to the next step.

1. temperature ✓

2. Top-K

3. Top-P



Random Sampling with Temperature

Intuition from thermodynamics

A system at a *high temperature* is flexible and can explore various states, while a system at a *low temperature* is likely to explore a subset of lower energy (better) states

How is this implemented

Divide the logits by a temperature parameter $\tau \in (0, 1]$ before passing it through softmax

Random sampling: $y = softmax(u)$

Random sampling with temperature: $y = softmax(u/\tau)$

Why does that work?

Lower the value of τ, larger are the scores being passed through softmax

This results in pushing high values towards 1 and low values towards 0

High-temperature sampling: $\tau > 1$ can be used to flatten the distribution

Top-k Sampling

Only Top k tokens are considered for generation, so the less probable words would not have any chance

1. Choose in advance a number of words k
2. For each word in the vocabulary V, use the language model to compute the likelihood of this word given the context $p(w_t|\mathbf{w}_{<t})$
3. Sort the words by their likelihood, and throw away any word that is not one of the top k most probable words.
4. Renormalize the scores of the k words to be a legitimate probability distribution.
5. Randomly sample a word from within these remaining k most-probable words according to its probability.

Nucleus Sampling or top-p sampling

Issues with Top-k Sampling

Shape of the probability distribution differs in different contexts. Top-k may include most of the probability mass in some cases, and very small mass in other cases.

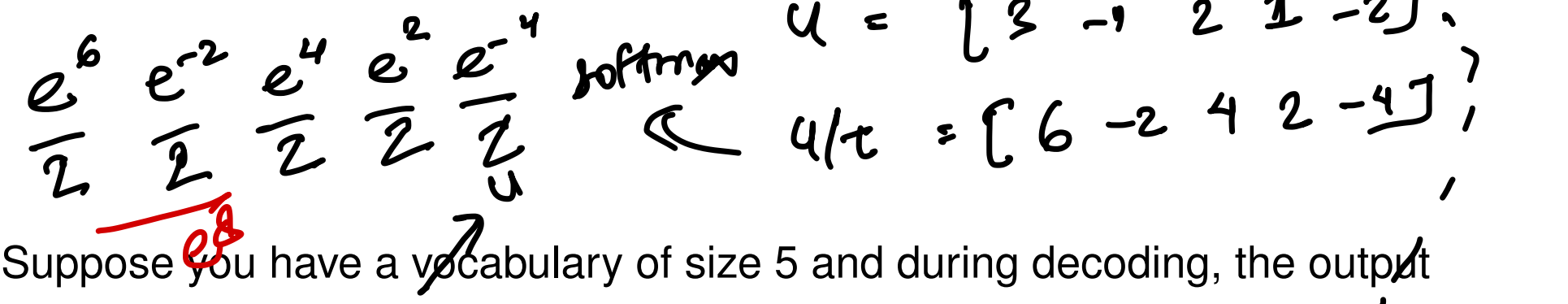
Nucleus Sampling or top-p sampling

Keep not the top k words but top p percent of the probability mass

Given a distribution $P(w_t|w_{<t})$, top-p vocabulary $V^{(p)}$ is the smallest set of words such that

$$\sum_{w \in V^{(p)}} P(w|w_{<t}) \geq p$$

Sample Problem



Suppose you have a vocabulary of size 5 and during decoding, the output vector is [3, -1, 2, 1, -2]. Write down the effective probability distribution when you use the following sampling strategies.

- Random sampling with temperature 0.5
- Top-2 sampling
- Nucleus sampling with p = 0.5