University of Amsterdam
ORTEC — OPTIMIZE YOUR WORLD
AMLAB Amsterdam Machine Learning Lab

STOCHASTIC BEAMS AND WHERE TO FIND THEM
The Gumbel-Top-k Trick for Sampling Sequences Without Replacement

Get the code!

Get the paper!

Wouter Kool, Herke van Hoof, Max Welling

- *"Stochastic Beam Search is the method you've been looking for if you want **unique samples** from a sequence model!"*
- *"You will never have duplicate samples again!"*
- *"If you want a principled way to randomize a beam search, you should probably use Stochastic Beam Search!"*

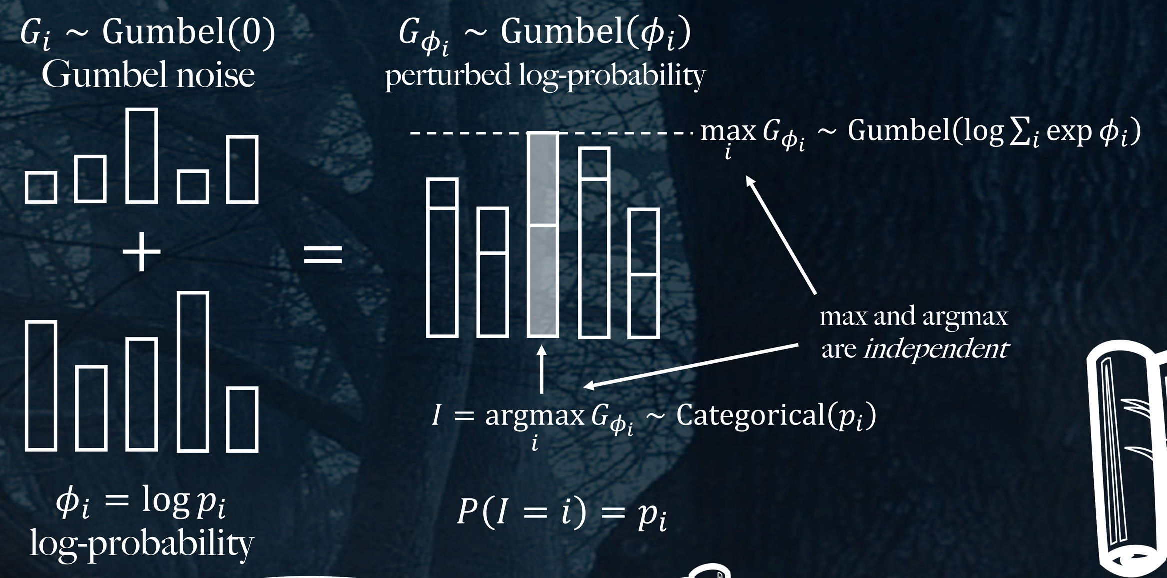
TL;DR

Stochastic Beam Search finds a set of unique samples (without replacement) from a sequence model.

Preliminaries

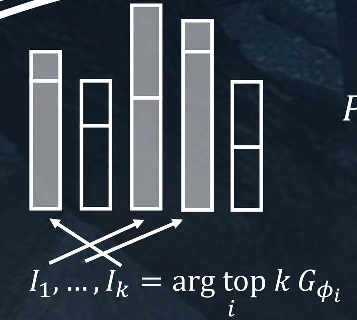
The Gumbel-Max Trick

"Mathemagic with prof. Gumbeldore: sample from the Categorical distribution"



The 'Gumbel-Top-k' Trick

"More magic: sample without replacement (also known as Plackett-Luce)"



$$P(I_1 = i_1, \dots, I_k = i_k) = \prod_{j=1}^{k} \frac{p_{i_j}}{1 - \sum_{\ell=1}^{j-1} p_{i_\ell}}$$

Beam Search

"Method to find high probability sequences from a sequence model"

Sequence model: $p_{\boldsymbol{\theta}}(y_{1:t}) = \prod_{t'=1}^{t} p_{\boldsymbol{\theta}}(y_{t'} | y_{1:t'-1})$

Beam Search: expand k most probable sequences at each step

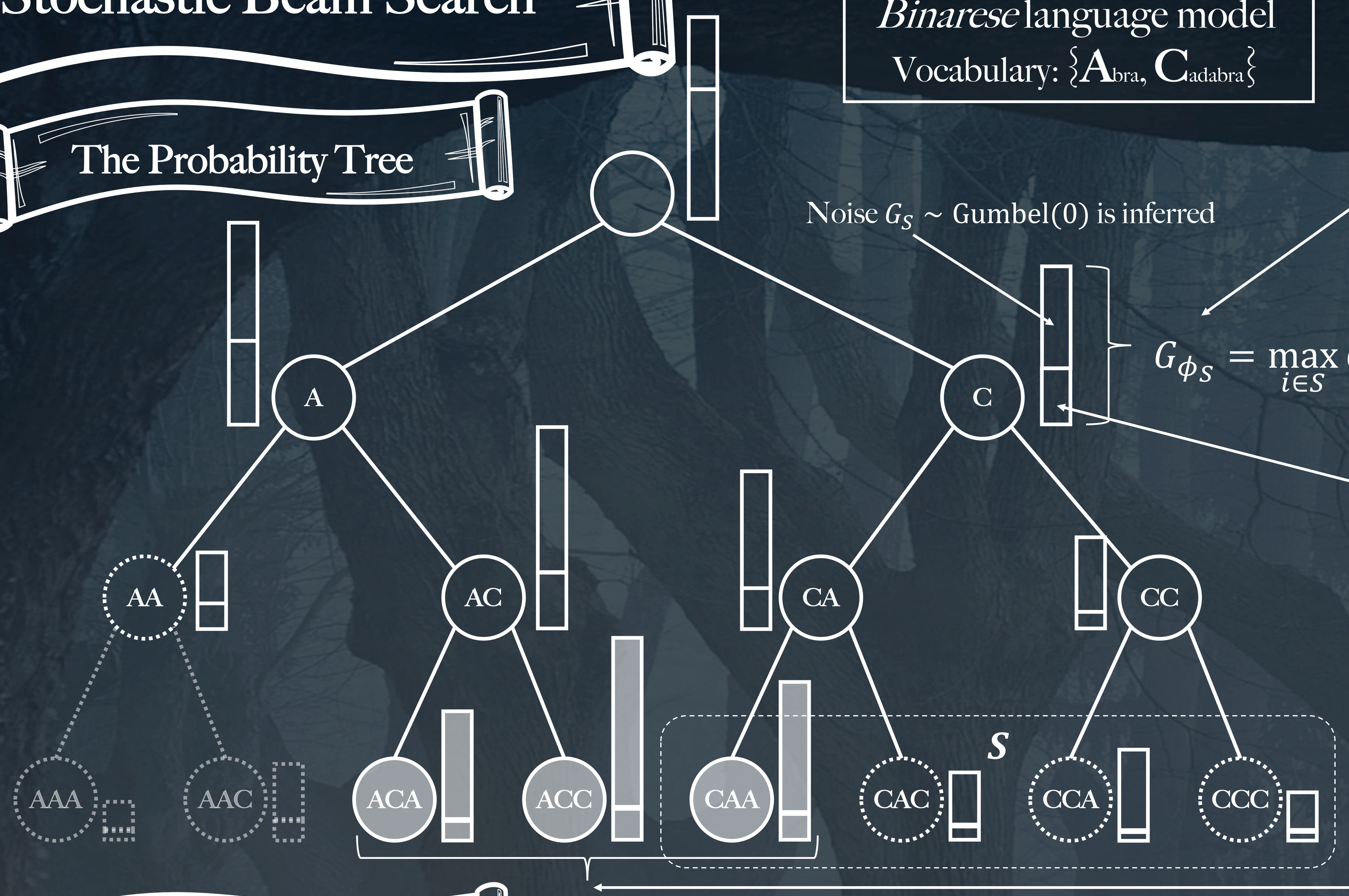
Stochastic Beam Search

Example

Binarese language model

Vocabulary: {Abra, Cadabra}

The Probability Tree



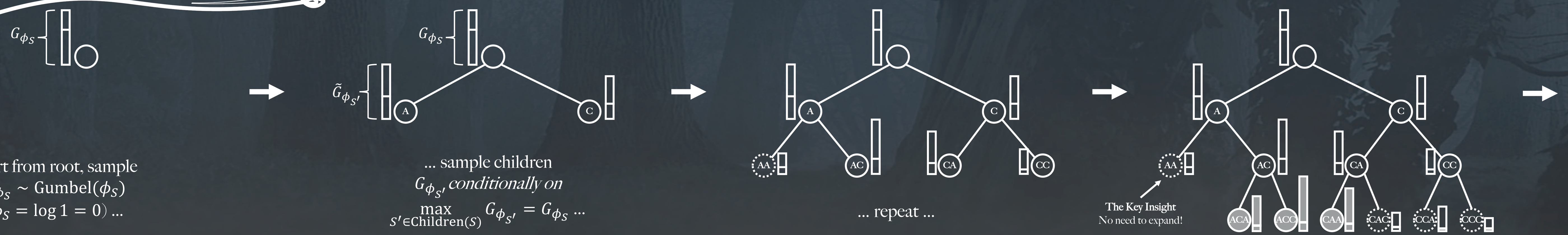
perturbed log-probability of partial sequence ("C") = maximum of perturbed log-probabilities in subtree

$$G_{\phi_S} = \max_{i \in S} G_{\phi_i} \sim \text{Gumbel}\Big(\underbrace{\log \sum_{i \in S} \exp \phi_i}\Big)$$

ϕ_S = log-probability of partial sequence ("C")

*Idea: obtain sample from leaf nodes using Gumbel-Top-k trick **without constructing complete tree***

Top-down Sampling



Start from root, sample $G_{\phi_S} \sim \text{Gumbel}(\phi_S)$ ($\phi_S = \log 1 = 0$) ...

... sample children $G_{\phi_{S'}}$ *conditionally on* $\max_{S' \in \text{Children}(S)} G_{\phi_{S'}} = G_{\phi_S}$ ...

... repeat ...

... the result is **equivalent** to sampling G_{ϕ_i} for leaves directly!

... even if we would continue, we would **only** have to expand k nodes at each level!

1. sample $G_{\phi_{S'}}$ independently, compute $Z = \max_{S'} G_{\phi_{S'}}$
2. 'shift' Gumbels: $\tilde{G}_{\phi_{S'}} = -\log\left(\exp(-G_{\phi_S}) - \exp(-Z) + \exp\left(-G_{\phi_{S'}}\right)\right)$

The Algorithm

Algorithm 1 StochasticBeamSearch($p_{\boldsymbol{\theta}}, k$)

```
Input: one-step probability distribution p_θ, beam/sample size k
Initialize BEAM empty
add (y^N = ∅, φ_N = 0, G_{φ_N} = 0) to BEAM
for t = 1, ..., steps do
    Initialize EXPANSIONS empty
    for (y^S, φ_S, G_{φ_S}) ∈ BEAM do
        Z ← −∞
        for S' ∈ Children(S) do
            φ_{S'} ← φ_S + log p_θ(y^{S'} | y^S)
            G_{φ_{S'}} ~ Gumbel(φ_{S'})
            Z ← max(Z, G_{φ_{S'}})
        end for
        for S' ∈ Children(S) do
            G̃_{φ_{S'}} ← −log(exp(−G_{φ_S}) − exp(−Z) + exp(−G_{φ_{S'}}))
            add (y^{S'}, φ_{S'}, G̃_{φ_{S'}}) to EXPANSIONS
        end for
    end for
    BEAM ← take top k of EXPANSIONS according to G̃
end for
Return BEAM
```

The Key Insight

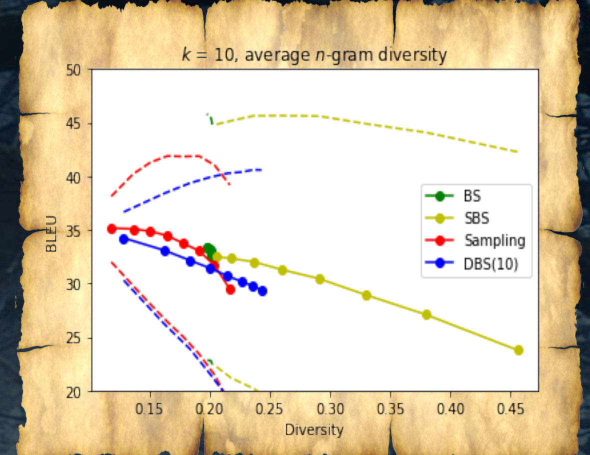
*If we use the Gumbel-Top-k trick with Top-down Sampling, we **only** need to expand the top k nodes at each level in the tree*

- Each node in the top k yields (at least) one leaf with the maximum perturbed log-probability
- Don't expand nodes not in the top k, since the maximum of their leaves is lower than the top k leaves from expanding the top k nodes

Experiments

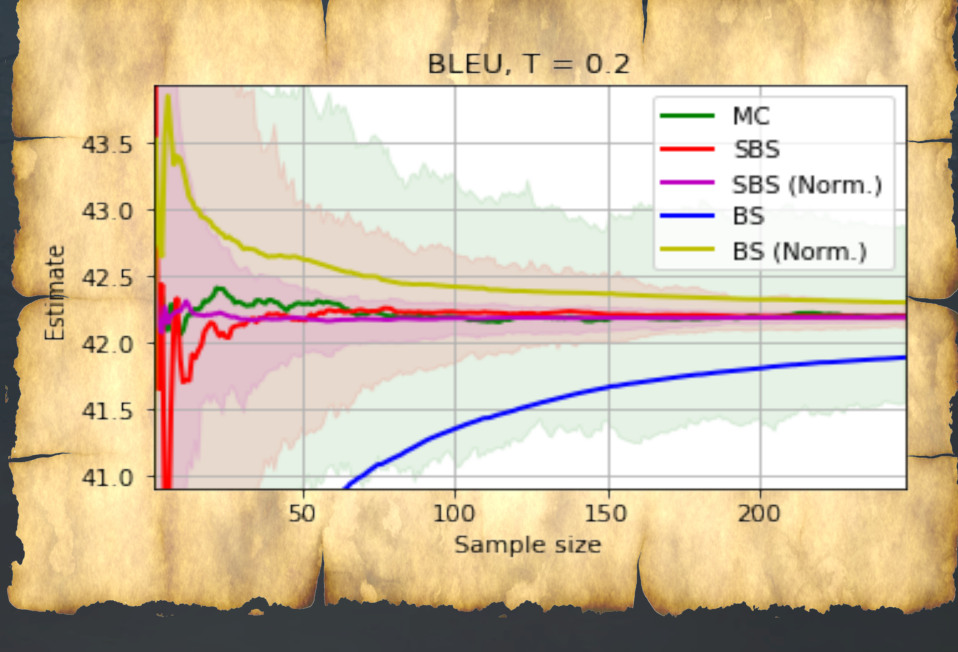
Translation Diversity

- Generate k translations
- Plot min, mean and max BLEU score against n-gram diversity
- Vary (local) softmax temperature from 0.1 (low diversity) to 0.8
- Compare:
  - Beam Search
  - Stochastic Beam Search
  - Sampling
  - Diverse Beam Search



BLEU Score Estimation

- Estimate sentence-level BLEU
- Plot mean and 95% interval vs. number of samples
- Compare:
  - Monte Carlo Sampling
  - Stochastic Beam Search with (normalized) Importance Weighted estimator
  - Beam Search with deterministic estimate



References

- Gumbel, E. J. Statistical theory of extreme values and some practical applications: a series of lectures. Number 33. US Govt. Print. Office, 1954.
- Maddison, C. J., Tarlow, D., and Minka, T. A* sampling. In *Advances in Neural Information Processing Systems*, pp. 3086–3094, 2014.
- Vieira, T. Gumbel-max trick and weighted reservoir sampling (blog post).
- Vieira, T. Estimating means in a finite universe (blog post).
- Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., and Batra, D. Diverse beam search for improved description of complex scenes. In *AAAI*, 2018.