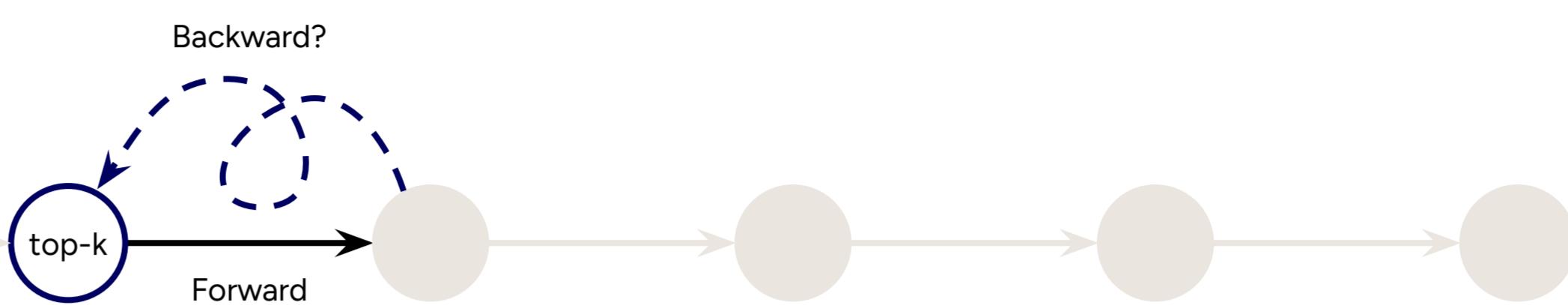
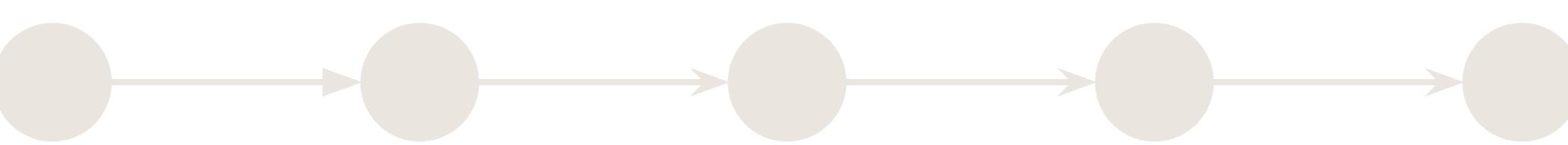


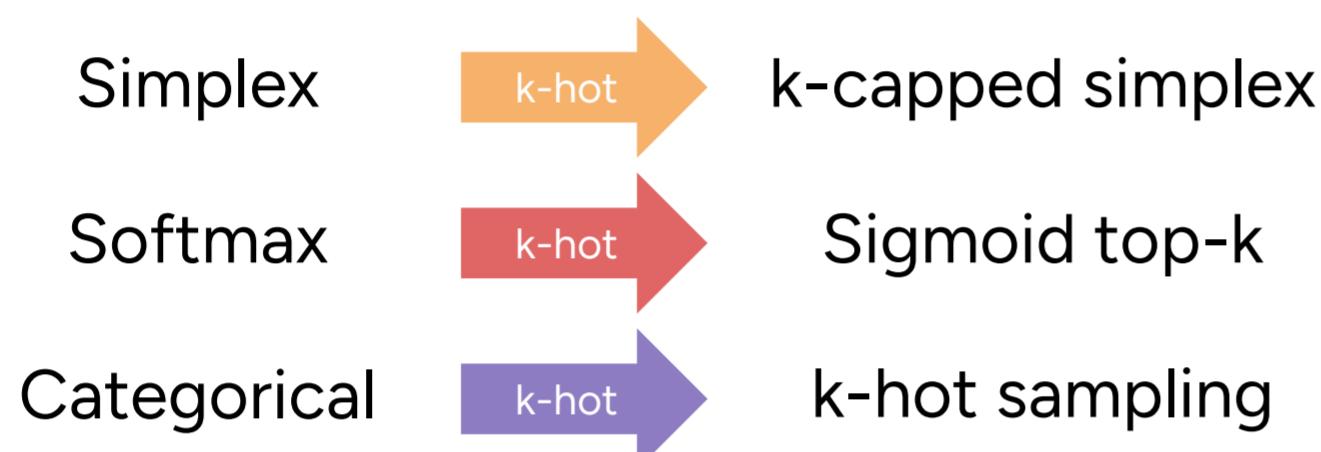
## Discrete modeling without breaking backprop

Using relaxations, sampling, and gradient estimates



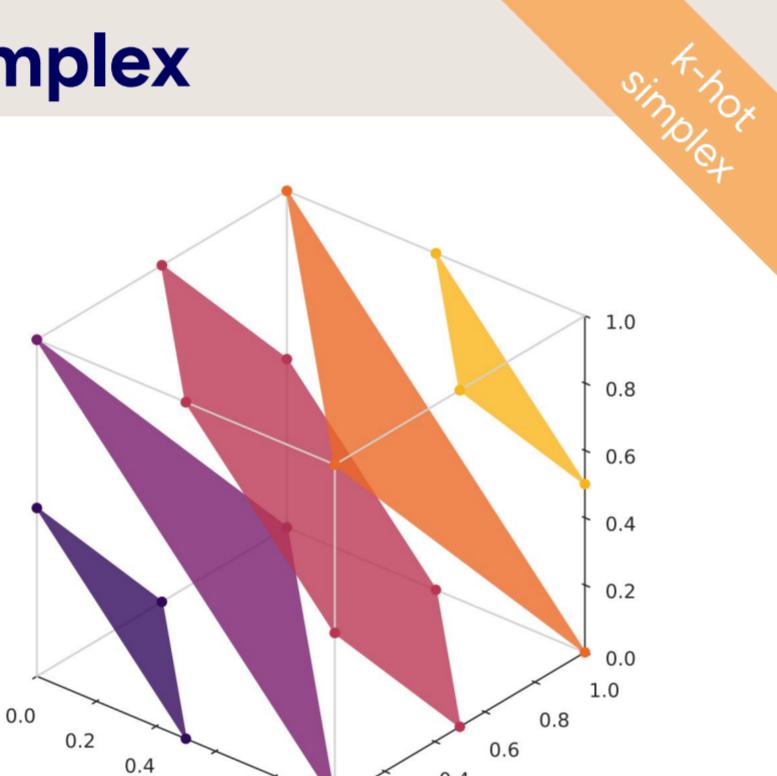
### The one-hot case

- There are simple and functional approaches to dealing with the non-differentiability of argmax and the categorical distribution.
- The k-hot case remains less developed.
- We propose a simple and modular approach similar to the one-hot case.



### The k-capped simplex

- The set of k-hot probabilities, or relaxed k-hot vectors.
- Probabilities that sum to k.
- Codomain of sigmoid top-k.
- Parameter space of πps sampling.



$$\Delta_k^{n-1} := \{\pi \in [0, 1]^n \mid \sum_{i=1}^n \pi_i = k\}$$

### Sampling k-hot vectors

k-hot  
categorical

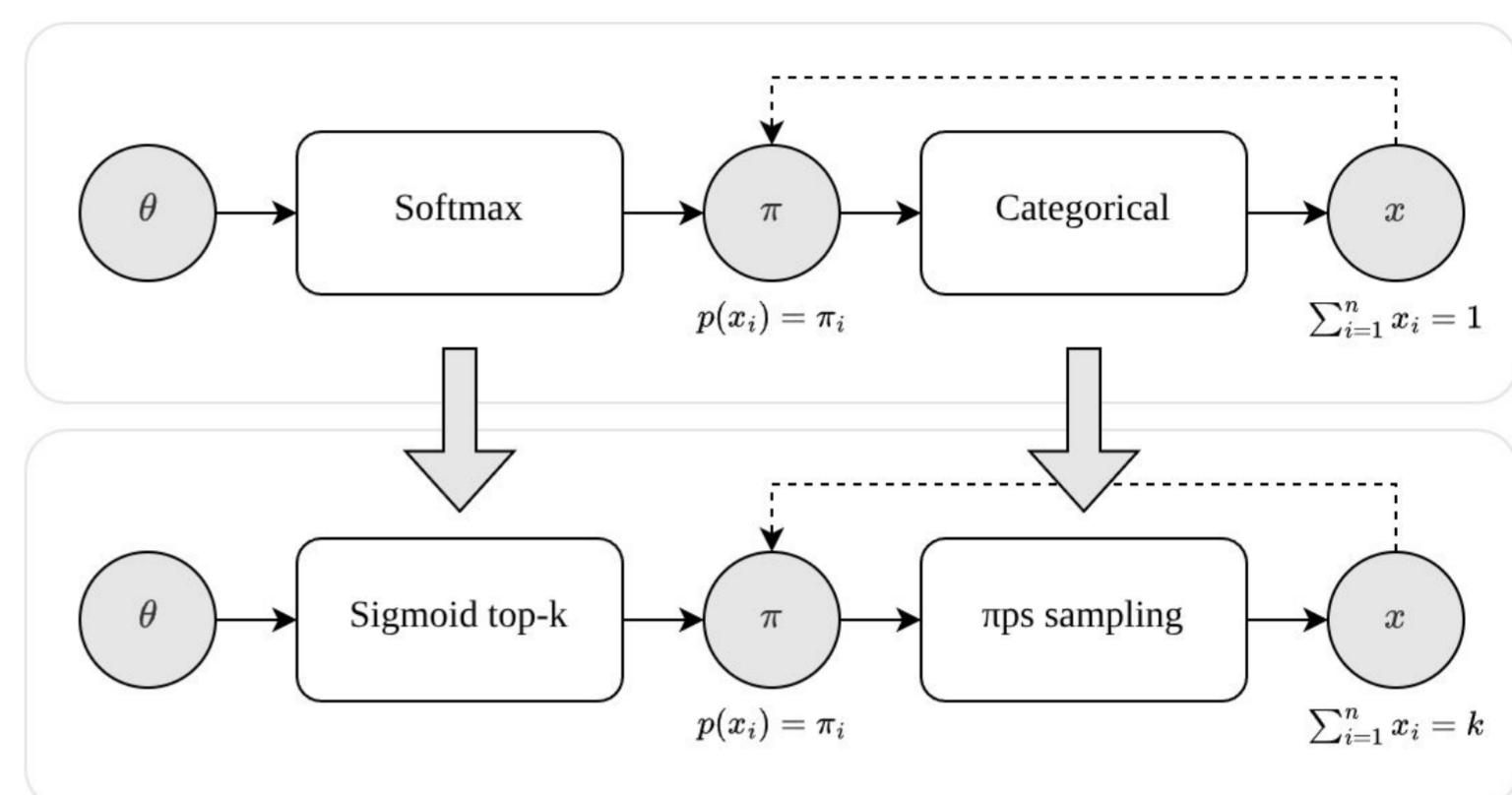
- Use ideas from the field of sampling design.
- Sampling without replacement (πps sampling).
- Distribution over all k-hot vectors.
- Parameterized by marginal (inclusion) probabilities.

$$p(x_i) = \pi_i$$

- Many distributions and algorithms to choose from.

### Gradient estimates

- Gradient estimates for differentiable k-hot sampling.
- Three main approaches:
  - Approximate pathwise gradient, e.g., straight-through.
  - Relaxed sampling, e.g., Gumbel-softmax.
  - Score function estimators, e.g., REINFORCE.
- The components above (sigmoid top-k, πps sampling) are useful in all three cases.
- Have been combined in previous works on one-hot sampling.



### Sigmoid top-k

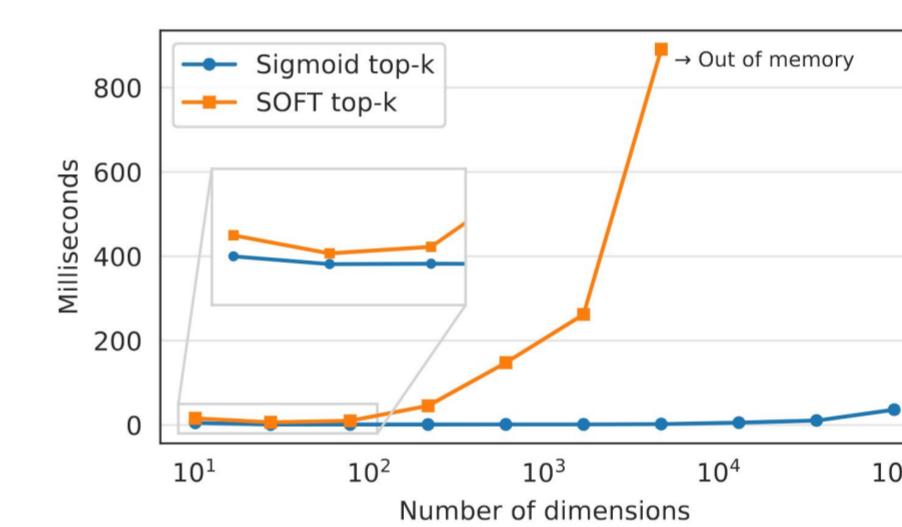
k-hot softmax

- Differentiable relaxation of top-k.

$$\sigma_k(\mathbf{x}) := \sigma(\mathbf{x} + c\mathbf{1}), \text{ where } c \in \mathbb{R} \text{ solves } \sum_{i=1}^n \sigma(x_i + c) = k$$

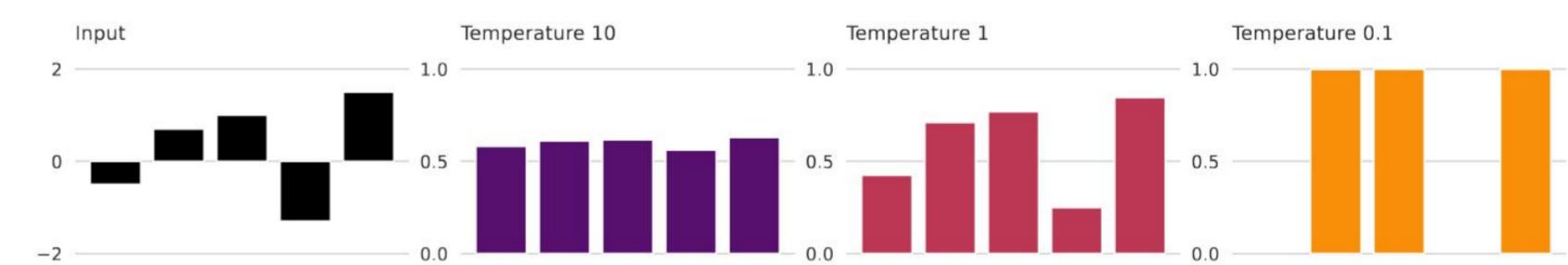
- Forward: solve a scalar root-finding problem.
- Backward: use implicit derivative.

Scalable!



- Defined for non-integer k.
- Differentiable w.r.t. both x and k.
- Solves the following optimization problem:

$$\sigma_k(\mathbf{x}) = \arg \max_{\pi \in \Delta_k^{n-1}} \mathbf{x}^\top \pi + \sum_{i=1}^n H(\pi_i)$$



Sigmoid top-k can be tempered like softmax

### Potential applications

kNN Sparsity Feature selection

Mixture of Experts Dictionary learning

Beam search

### Read more

SFESS: Score Function Estimators for k-Subset Sampling  
Klas Wijk, Ricardo Vinuesa, Hossein Azizpour  
ICLR 2025

Differentiable Top-k: From One-Hot to k-Hot  
Klas Wijk, Ricardo Vinuesa, Hossein Azizpour  
EurIPS 2025 Workshop on Differentiable Systems and Scientific Machine Learning