Differentiable Top-k: From One-Hot to k-Hot

Klas Wijk[1,2] Ricardo Vinuesa[2,3] Hossein Azizpour[1,2,4]

[1]KTH Royal Institute of Technology [2]Swedish e-Science Research Institute [3]University of Michigan [4]Science for Life Laboratory

Motivation

- Top-k and k-hot vectors are a fundamental modeling tool.
- Top-k is non-differentiable, hindering gradient-based optimization.
- Existing solutions are not as simple and scalable as those for sampling one-hot vectors.

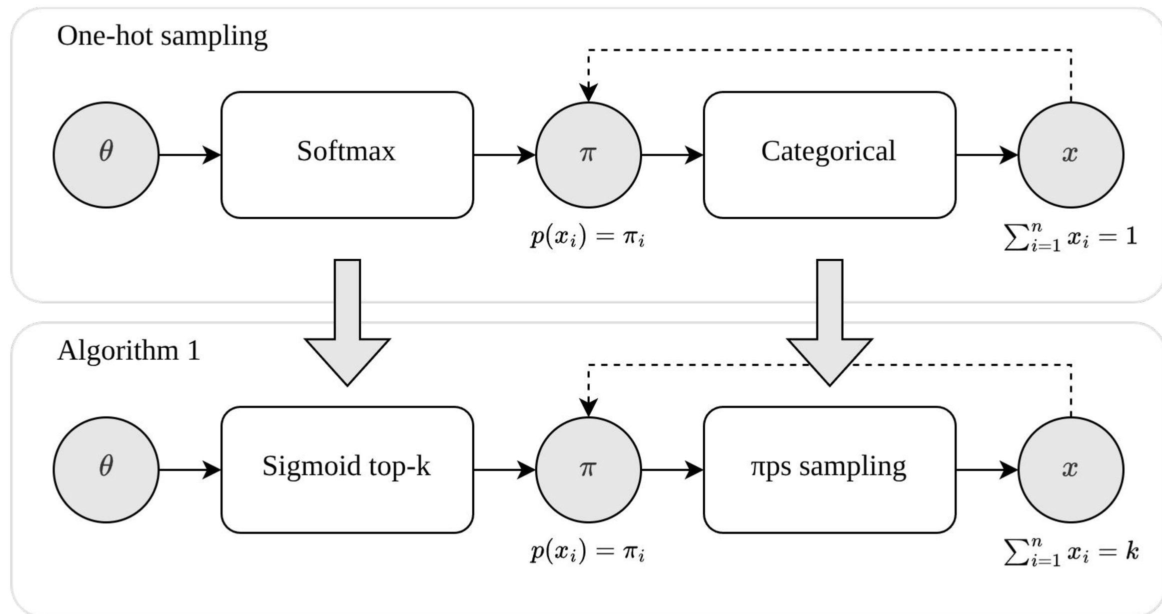
Overview

Algorithm 1 Sample k-hot

Require: $\boldsymbol{\theta} \in \mathbb{R}^n, k \in \mathbb{N}, 1 < k < n$
1: $\boldsymbol{\pi} \leftarrow \sigma_k(\boldsymbol{\theta})$
2: Sample $\boldsymbol{x} \in \{0,1\}^n_k$ such that $p(x_i) = \pi_i$
3: **return** $\boldsymbol{x}$

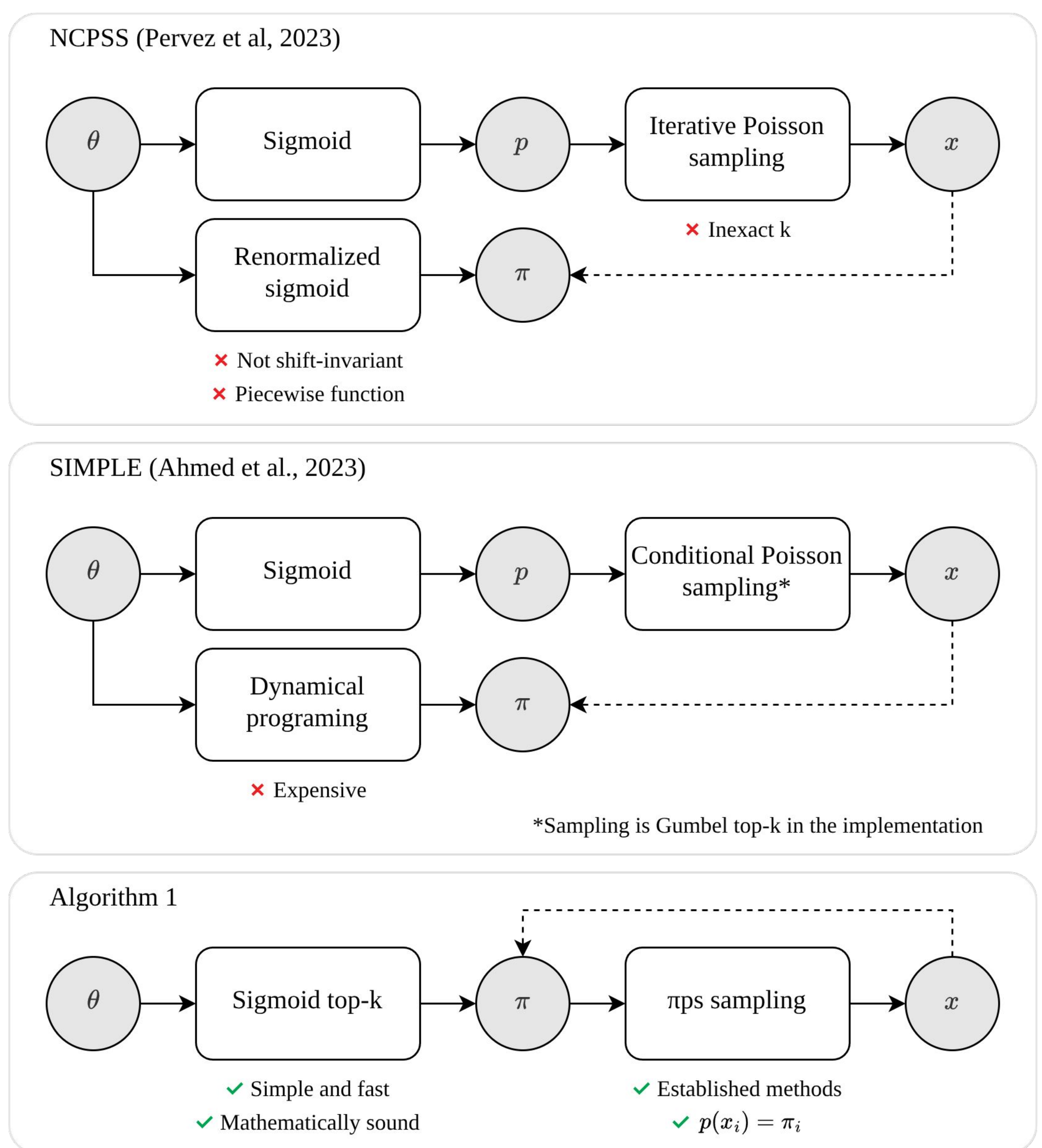
Algorithm 2 Sample relaxed k-hot

Require: $\boldsymbol{\theta} \in \mathbb{R}^n, k \in \mathbb{N}, 1 < k < n$
1: Sample $g_i \sim \text{Gumbel}(0,1)$ for $i = 1, \ldots, n$
2: $\boldsymbol{x} \leftarrow \sigma_k(\boldsymbol{\theta} + \boldsymbol{g})$
3: **return** $\boldsymbol{x}$



A simplified and improved approach

- Fast, correct, and uses a simple straight-through gradient estimate.



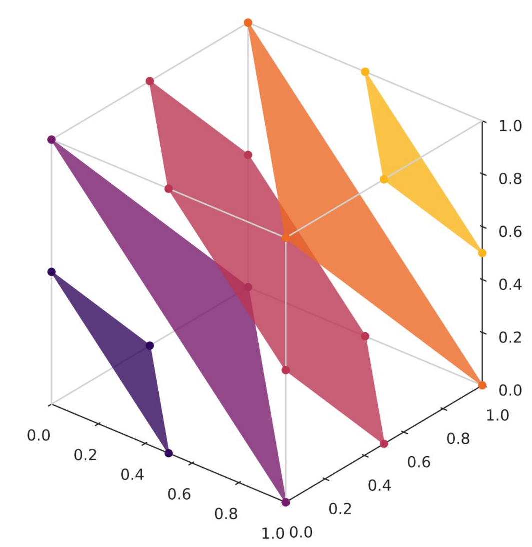
Relaxed top-k

The k-capped simplex

Generalized simplex

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

- Probabilities that sum to k.
- Codomain of sigmoid top-k.
- Parameter space of πps sampling.



Sigmoid top-k

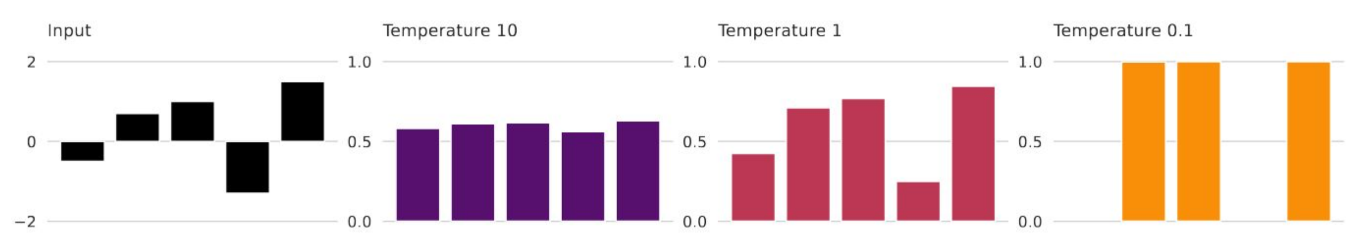
Generalized softmax

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

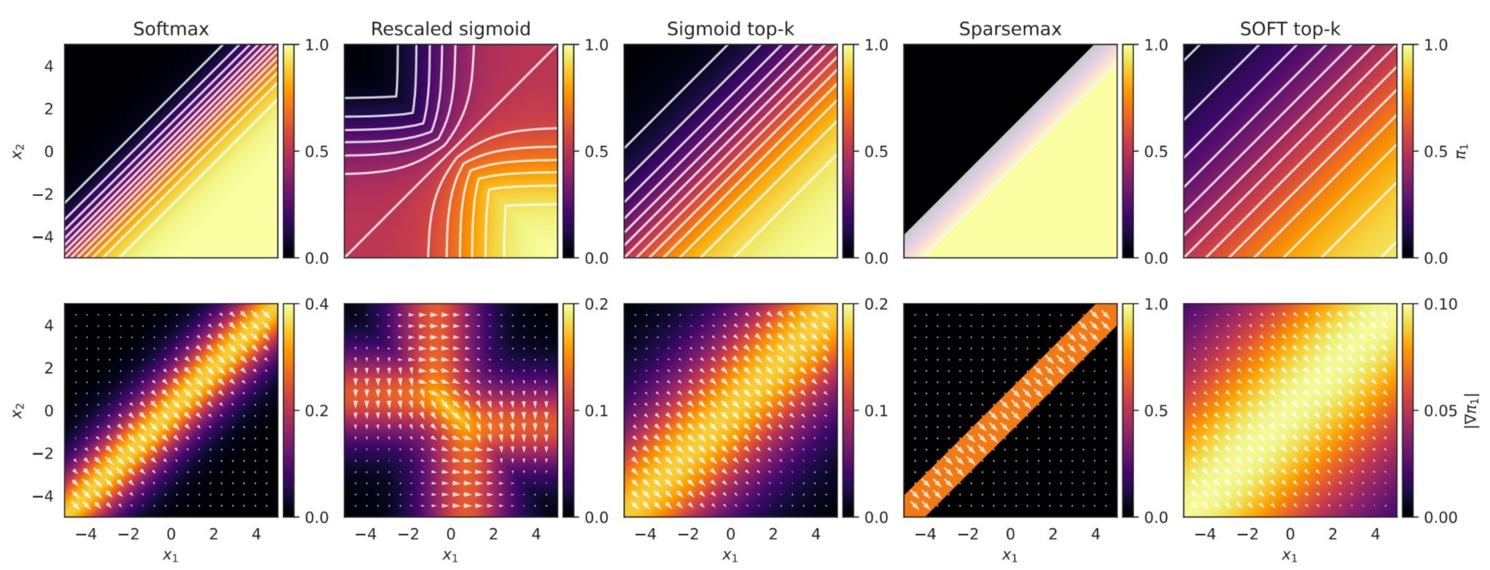
- Top-k relaxation.
- Shares many properties with softmax.
- Forward: **scalar root-finding**.
- Backward: **implicit differentiation**.

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

- Sigmoid top-k solves an entropy-regularized optimization problem.
- It is fully differentiable, with respect to both x and k.
- It can be tempered like softmax.
- It is inverted by the logit function up to an additive constant.



Sigmoid top-k can be tempered like softmax.



Value and gradients of top-k relaxations for n = 2 and k = 1.

Sampling

πps sampling

Generalized categorical

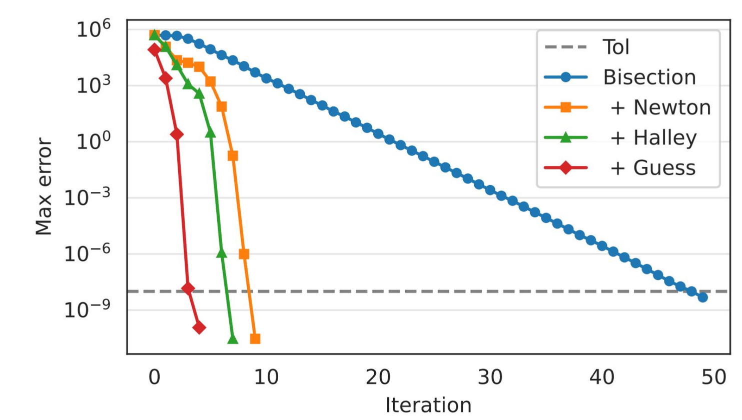
- Methods from the field of sampling design.
- Sampling proportional to size without replacement.
- Parameterized by desired **inclusion probabilities**.
- Many possible distributions (designs) and algorithms (procedures).

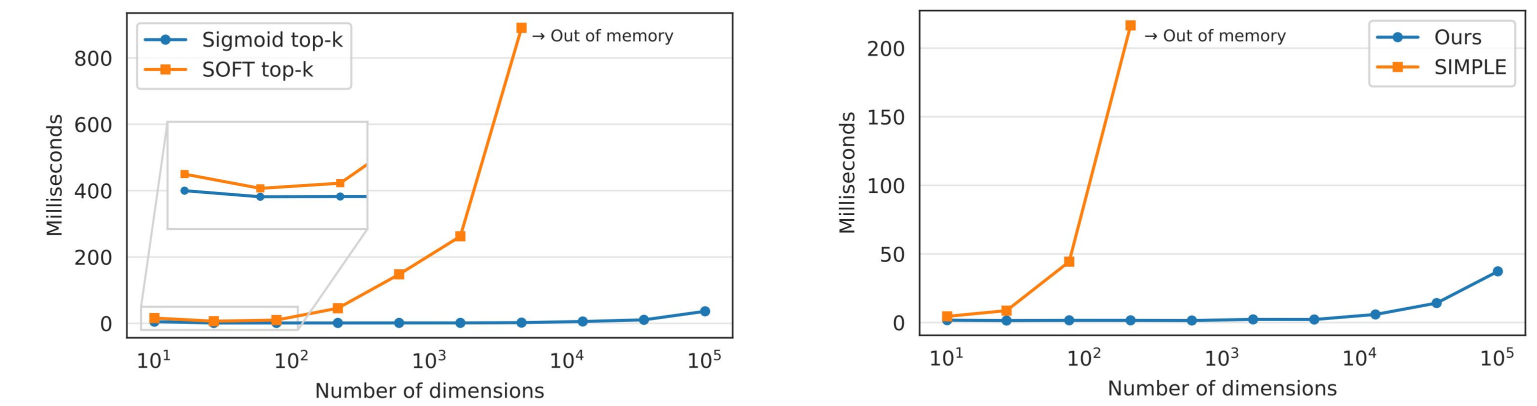
$$p(x_i) = \pi_i$$

- Exact inclusion probabilities.
- We prove that straight-through estimation is a first-order approximation of the true gradient in expectation.

Scalability

- Scalar root-finding is scalable.
- Guaranteed convergence.
- Scalable to high dimensions.
- A benchmark in 10^6 dim is shown to the right.
- Comparisons against two methods from prior work below.





Applications

Top-k appears in many models.
Our methods can be used to make it **differentiable**, **stochastic**, or **both**.

- Routing
  - Mixture of experts (MoE)
- Sparse coding
  - Dictionary learning (k-SVD)
  - Sparse system identification (SINDy)
  - Sparse autoencoders (SAE)
- Search
  - k-nearest neighbors
  - Beam search
- Sparse networks
- Feature selection