SIMPLE: A Gradient Estimator for k-subset Sampling

Kareem Ahmed* UCLA

Zhe Zeng* UCLA

Mathias Niepert Stuttgart University

Guy Van den Broeck UCLA

University of Stuttgart
Germany

tl;dr ***"We propose a gradient estimator for distributions over k-subsets using exact discrete samples, estimating the gradient as the derivative of the conditional marginals in the direction of the downstream gradients"***

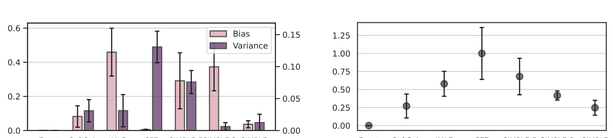
Why

In machine learning, we're often interested in modeling ***distributions over k-subsets of elements*** e.g. in ***L2X***, where we want the k words that ***best explain a classifier's decision***

Truth	Model	Key words
positive	positive	Ray Liotta and Tom Hulce shine in this sterling example of brotherly love and commitment. Hulce plays Dominick, (nicky) a mildly mentally handicapped young man who is putting his 12 minutes younger, twin brother, Liotta, who plays Eugene, through medical school. It is set in Baltimore and deals with the issues of sibling rivalry, the unbreakable bond of twins, child abuse and good always winning out over evil. It is captivating, and filled with laughter and tears. If you have not yet seen this film, please rent it, I promise, you'll be amazed at how such a wonderful film could go unnoticed.
negative	negative	Sorry to go against the flow but I thought this film was unrealistic, boring and way too long. I got tired of watching Gena Rowlands long arduous battle with herself and the crisis she was experiencing. Maybe the film has some cinematic value or represented an important step for the director but for pure entertainment value. I wish I would have skipped it.

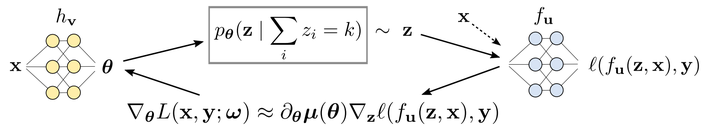
What

Existing approaches [1, 2, 3] generalize Gumbel-Softmax to the k-subset distribution, ***returning relaxed samples which biases the gradient, and cannot be used in many settings such as discrete VAEs***, or otherwise resort to ***approximate sampling and marginals***



SIMPLE uses ***exact, discrete samples***, coupled with ***the derivative of the conditional marginals*** , leading to an estimator with ***lower bias and variance***.

How



We're interested in **sampling** from the distribution

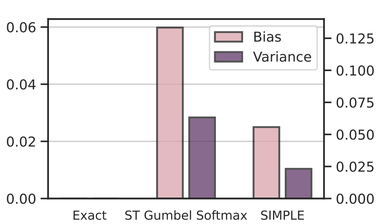
$$p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \textstyle\sum_i z_i = k) = p_{\boldsymbol{\theta}}(\boldsymbol{z}) / p_{\boldsymbol{\theta}}(\textstyle\sum_i z_i = k) \cdot [\![\textstyle\sum_i z_i = k]\!]$$

the key insight here is that **we do not need to care about the order** of variables, but **only their sum**.

Another key insight is **the gradient depends on marginals** of the distribution; the marginals are **the partial derivatives of the log-partition function**

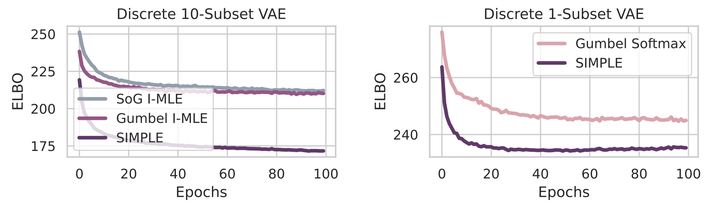
$$\mu(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}\Big(z_i \mid \textstyle\sum_j z_j = k\Big) = \frac{\partial}{\partial \theta_i} \log p_{\boldsymbol{\theta}}(\textstyle\sum_j z_j = k)$$

In the case when $k = 1$, we recover an estimator comparable to Gumbel-Softmax, ***with lower bias and variance***



Experiments

DVAE We show that the KL-divergence between the k-subset distribution and the uniform distribution can be computed exactly.



L2X on BeerAdvocate dataset: To predict ratings for different aspects of beer from free-text reviews.

Method	Appearance		Palate		Taste	
	Test MSE	Precision	Test MSE	Precision	Test MSE	Precision
SIMPLE (Ours)	**2.35 ± 0.28**	**66.81 ± 7.56**	**2.68 ± 0.06**	**44.78 ± 2.75**	**2.11 ± 0.02**	**42.31 ± 0.61**
L2X (t = 0.1)	10.70 ± 4.82	30.02 ± 15.82	6.70 ± 0.63	**50.39 ± 13.58**	6.92 ± 1.61	32.23 ± 4.92
SoftSub (t = 0.5)	**2.48 ± 0.10**	52.86 ± 7.08	2.94 ± 0.08	39.17 ± 3.17	2.18 ± 0.10	**41.98 ± 1.42**
I-MLE (τ = 30)	**2.51 ± 0.05**	**65.47 ± 4.95**	2.96 ± 0.04	40.73 ± 3.15	2.38 ± 0.04	**41.38 ± 1.55**

Method	$k = 5$		$k = 10$		$k = 15$	
	Test MSE	Precision	Test MSE	Precision	Test MSE	Precision
SIMPLE (Ours)	**2.27 ± 0.05**	**57.30 ± 3.04**	**2.23 ± 0.03**	**47.17 ± 2.11**	3.20 ± 0.04	**53.18 ± 1.09**
L2X (t = 0.1)	5.75 ± 0.30	33.63 ± 6.91	6.68 ± 1.08	26.65 ± 9.39	7.71 ± 0.64	23.49 ± 10.93
SoftSub (t = 0.5)	2.57 ± 0.12	**54.06 ± 6.29**	2.67 ± 0.14	44.44 ± 2.27	**2.52 ± 0.07**	37.78 ± 1.71
I-MLE (τ = 30)	2.62 ± 0.05	**54.76 ± 2.50**	2.71 ± 0.10	**47.98 ± 2.26**	2.91 ± 0.18	39.56 ± 2.07

Implementation available at `github.com/UCLA-StarAI/SIMPLE`

References

[1] Jianbo Chen et al. "Learning to Explain: An Information-Theoretic Perspective on Model Interpretation". In: *ICML 2018*.

[2] Mathias Niepert, Pasquale Minervini, and Luca Franceschi. "Implicit mle: Backpropagating through discrete exponential family distributions". In: *Neurips* (2021).

[3] Sang Michael Xie and Stefano Ermon. "Reparameterizable Subset Sampling via Continuous Relaxations". In: *IJCAI-19*.