Gradient Estimation with Stochastic Softmax Tricks

Max B. Paulus*, Dami Choi*, Danny Tarlow, Andreas Krause, Chris J. Maddison

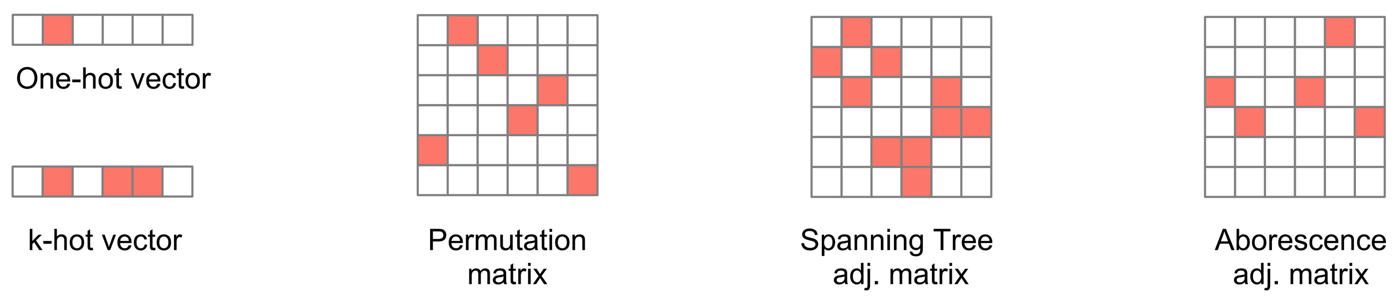
ETH zürich · VECTOR INSTITUTE · UNIVERSITY OF TORONTO · Google Research · DeepMind

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

We learn deep latent variable models over discrete structured domains…

…where the discrete latent variable may be…



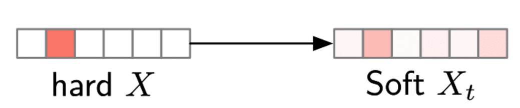
For example, using a k-hot variable, we can learn to identify important words without direct supervision…

Pours a slight tangerine orange and straw yellow. The head is nice and bubbly but fades very quickly with a little lacing. Smells like Wheat and European hops, a little yeast in there too. There is some fruit in there too, but you have to take a good whiff to get it. The taste is of wheat, a bit of malt, and a little fruit flavour in there too. Almost feels like drinking Champagne, medium mouthful otherwise. Easy to drink, but not something I'd be trying every night.

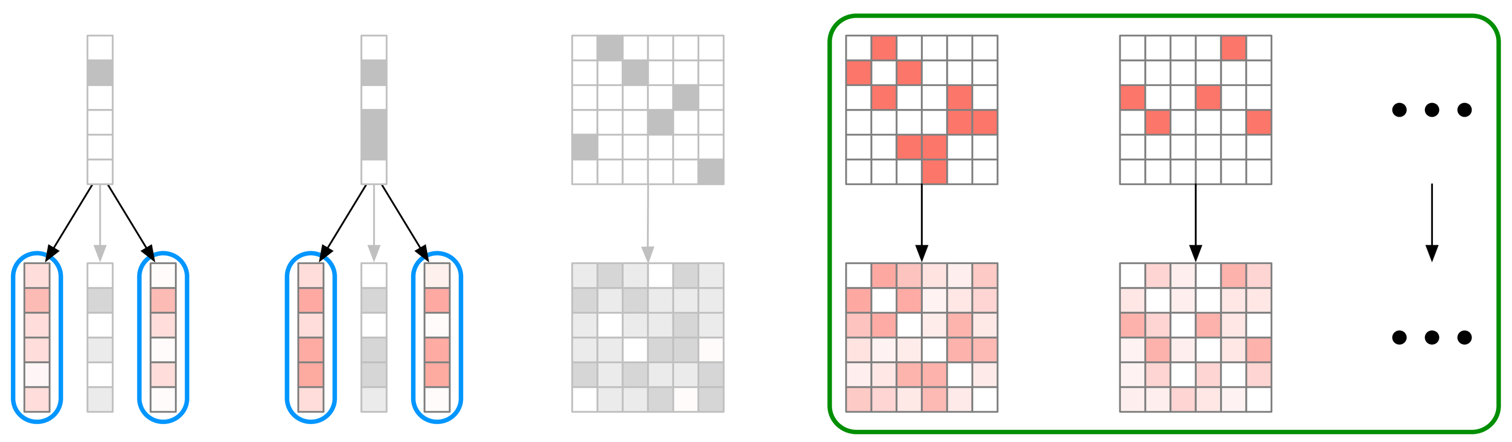
Appearance: 3.5 ***Aroma: 4.0*** *Palate: 4.5 Taste: 4.0 Overall: 4.0*

We generalize the Gumbel-Softmax to combinatorial spaces.

We leverage continuous relaxations to design gradient estimators for structured discrete variables..



Our framework generalizes previous work on relaxations and includes new relaxations and new structured variables..

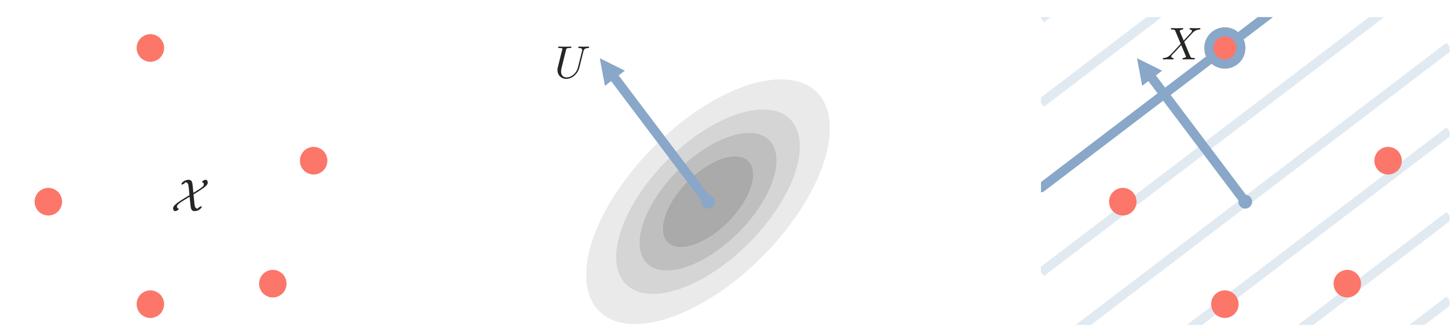


Stochastic Argmax Tricks (SMTs)

Stochastic Argmax Tricks (SMTs) reparameterize X as the solution to a random linear program…

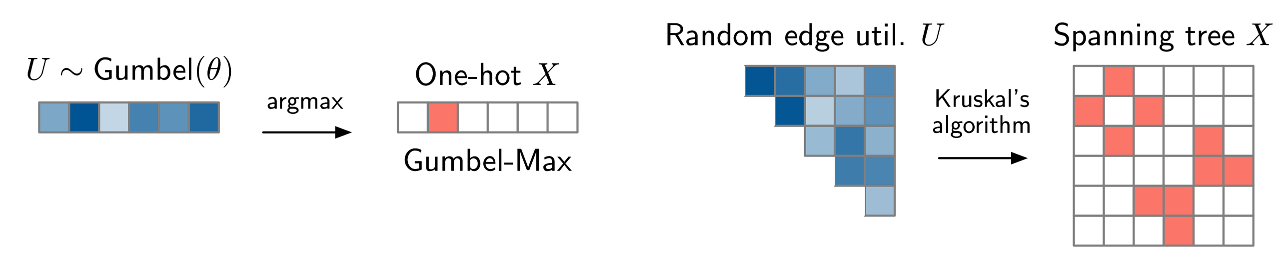
$$X = \arg\max_{x \in \mathcal{X}} U^T x.$$

..where U induces a distribution over $\mathcal{X}$ (Hazan et al., 2016).



Finite Set · Random Utility · SMT

SMTs recover the Gumbel-Max trick in the one-hot case and generalize it to structured X for which efficient linear solvers are available…

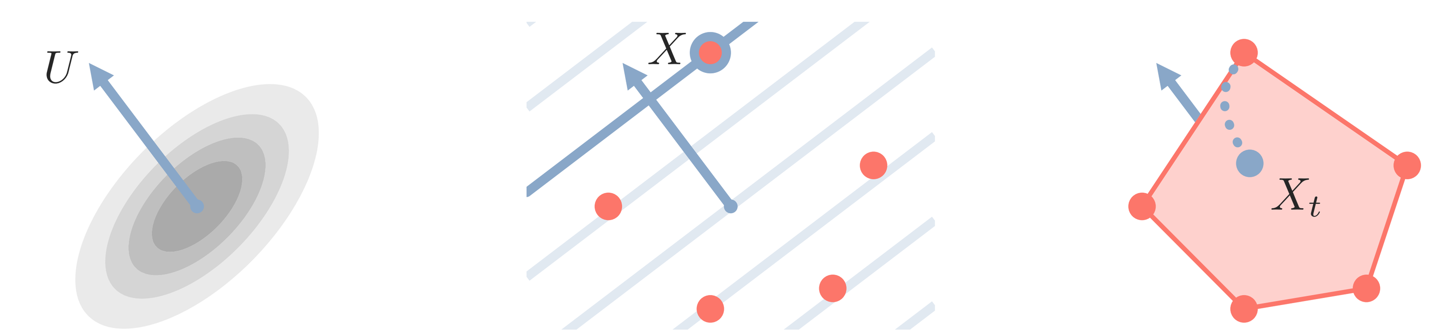


Stochastic Softmax Tricks (SSTs)

Stochastic Softmax Tricks (SSTs) relax a given SMT..

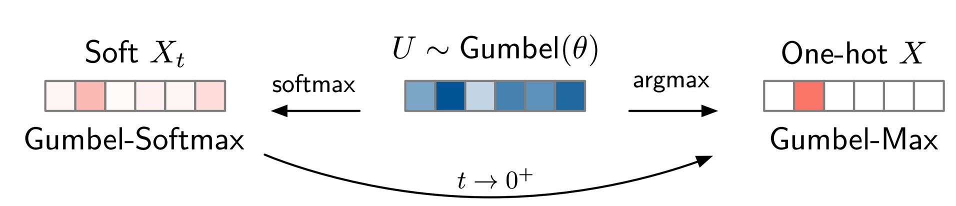
$$X_t = \arg\max_{x \in \text{conv}(\mathcal{X})} U^T x - \underbrace{t\, f(x)}_{\text{strongly convex regularizer}}$$

..to relax discrete X to continuous X_t and admit a reparam. gradient..

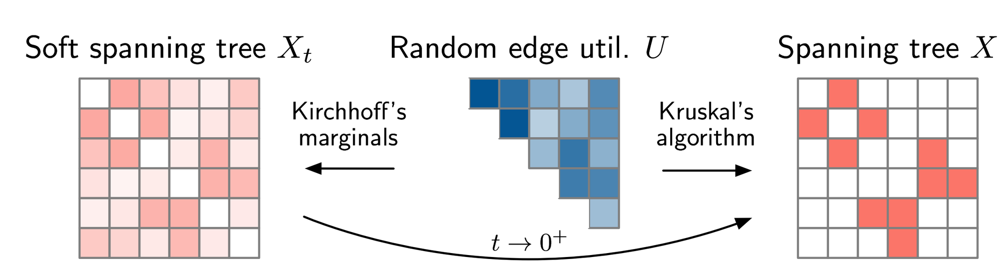


Random Utility · SMT · SST

SSTs recover the Gumbel-Softmax in the one-hot case..

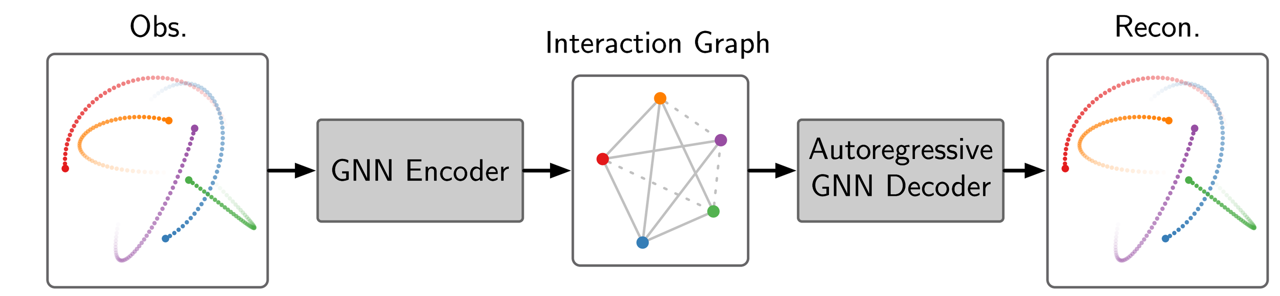


..and generalize it to other structured X when efficient solvers are available..

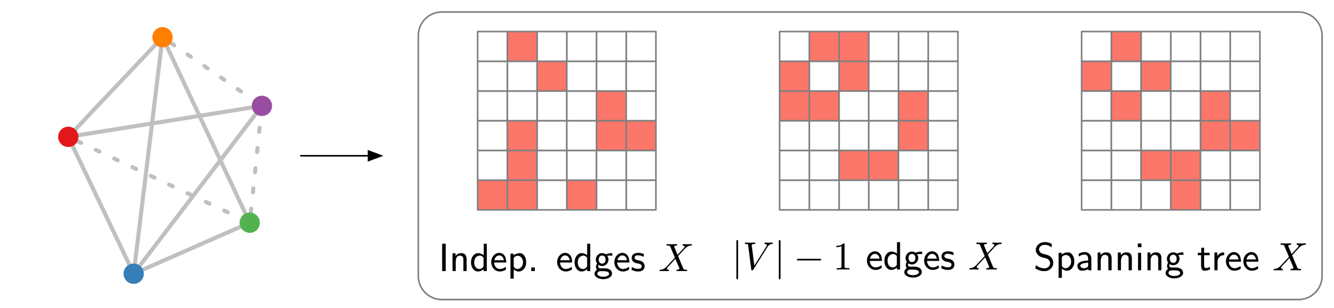


Neural Relational Inference for Graph Layout

NRI (Kipf et al., 2018) is a VAE with a latent graph…

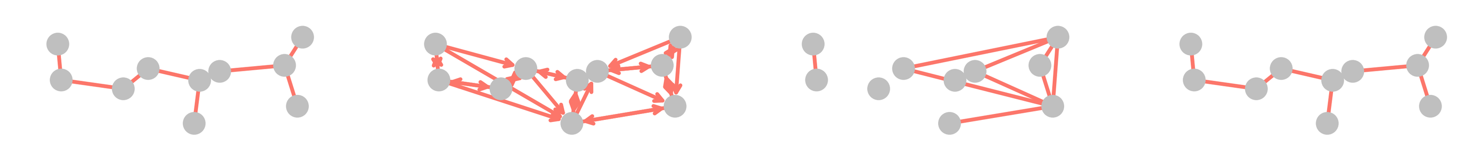


..on which we can impose varying degrees of structure…



For trajectories from a force-directed algorithm (with true latent spanning tree structure) more structured models improve performance..

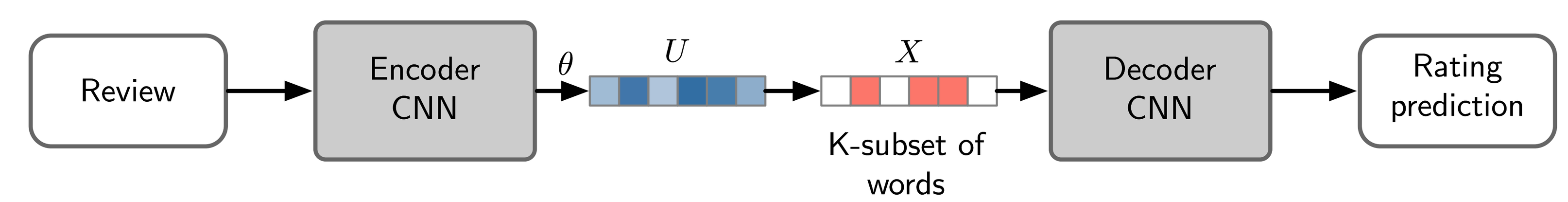
Edge Distribution	ELBO	Edge Prec.	Edge Recall
Indep. Edges	-1370 ± 20	48 ± 2	93 ± 3
|V|-1 edges	-2100 ± 20	41 ± 1	41 ± 1
Spanning Tree	-1080 ± 110	91 ± 3	91 ± 3



Ground Truth · Indep. Edges · |V|-1 edges · Spanning Tree

Learning To Explain (L2X) Aspect Ratings

We use SSTs for subset selection on a sentiment prediction task…



..to select contiguous phrases (see Motivation) and improve performance..

Relaxation	*k* = 5 MSE	*k* = 5 Subset Prec.	*k* = 10 MSE	*k* = 10 Subset Prec.	*k* = 15 MSE	*k* = 15 Subset Prec.
L2X (Chen et al., 2018)	3.6 ± 0.1	28.3 ± 1.7	3.0 ± 0.1	25.5 ± 1.2	2.6 ± 0.1	25.5 ± 0.4
SoftSub (Xie & Ermon, 2019)	3.6 ± 0.1	27.2 ± 0.7	3.0 ± 0.1	26.1 ± 1.1	2.6 ± 0.1	25.1 ± 1.0
E.F. Ent. Top k	3.5 ± 0.1	28.8 ± 1.7	2.7 ± 0.1	32.8 ± 0.5	2.5 ± 0.1	29.2 ± 0.8
Corr. Top k	2.9 ± 0.1	63.1 ± 5.3	2.5 ± 0.1	53.1 ± 0.9	2.4 ± 0.1	45.5 ± 2.7

References

Jang et al. (2016). "Categorical reparameterization with gumbel-softmax."

Maddison et al. (2016). "The concrete distribution: A continuous relaxation of discrete random variables."

Hazan et al. (2016). "Perturbation, Optimization, and Statistics."

Kipf et al. (2018). "Neural relational inference for interacting systems."

Chen et al. (2018). "Learning to explain: An information-theoretic perspective on model interpretation."

Xie & Ermon. (2019). "Reparameterizable subset sampling via continuous relaxations."