Gradient Estimation with Stochastic Softmax Tricks

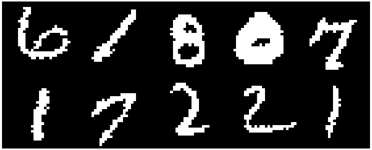
Max B. Paulus*, Dami Choi*,

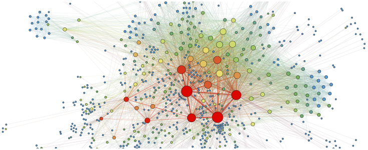
Daniel Tarlow, Andreas Krause, Chris J. Maddison

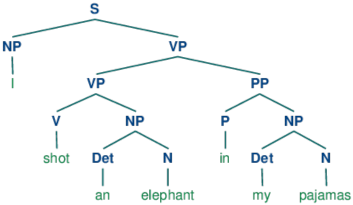
NeurIPS 2020: Oral Presentation

Discrete Data

There is a lot of discrete structure in data...







...that often is unobserved.

Source: MNIST, NLTK, Wikipedia, LabioTech

Why model discrete structure?

By modeling this unobserved structure, we can for example...

- incorporate problem-specific constraints (Mena et al., 2018)
- improve generalization (Graves et al., 2014)
- increase interpretability (Chen et al., 2018)

Why model discrete structure?

By modeling this unobserved structure, we can for example...

- incorporate problem-specific constraints (Mena et al., 2018)
- improve generalization (Graves et al., 2014)
- increase interpretability (Chen et al., 2018)

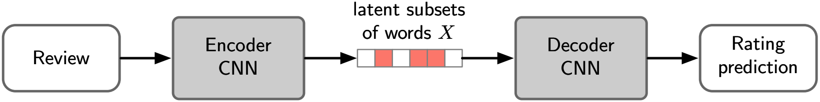
As an example, consider...

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

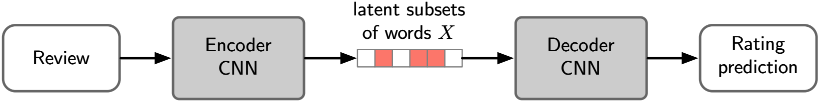
Example: Learning to explain (L2X) aspect ratings

A latent subset variable can be used...



Example: Learning to explain (L2X) aspect ratings

A latent subset variable can be used...



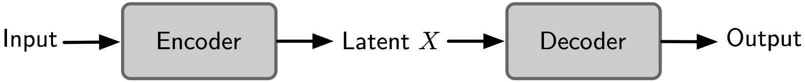
...for an interpretable model (Lei et al., 2016; Chen et al., 2018):

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

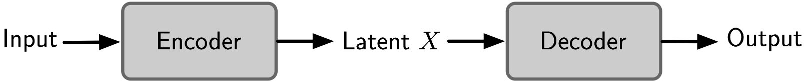
Models with structured latent variables

More generally, we can consider encoder-decoder models...

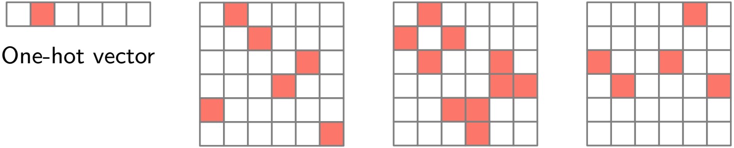


Models with structured latent variables

More generally, we can consider encoder-decoder models...



...where the latent X is another binary array, for example...



One-hot vector

k-hot vector

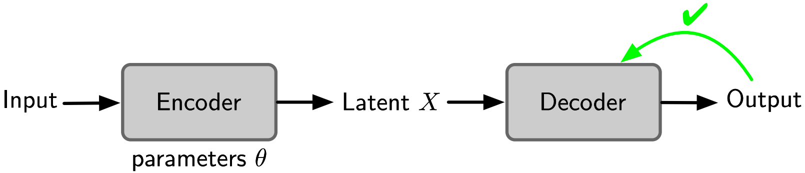
Permutation matrix

Spanning tree adj. matrix

Arborescence adj. matrix

Problem: How to backprop through X?

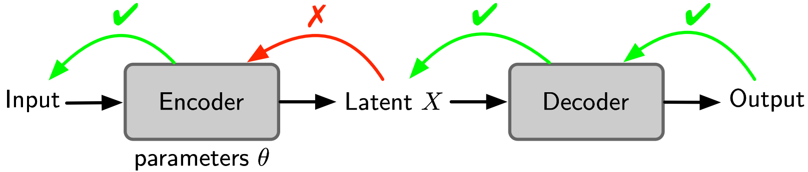
Learning the parameters θ requires backpropagating through X...



This is difficult, because...

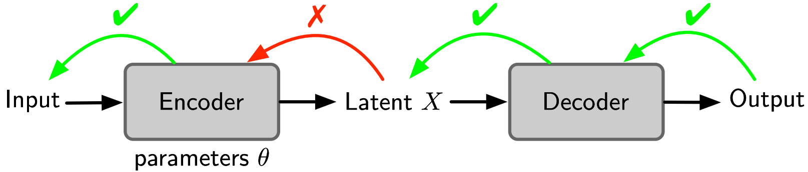
Problem: How to backprop through X?

Learning the parameters θ requires backpropagating through X...



Problem: How to backprop through X?

Learning the parameters θ requires backpropagating through X...

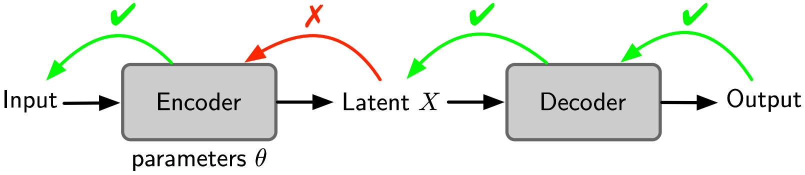


This is difficult, because...

- REINFORCE is high variance.

Problem: How to backprop through X?

Learning the parameters θ requires backpropagating through X...

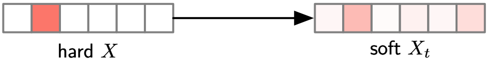


This is difficult, because...

- REINFORCE is high variance.
- No unbiased reparameterization gradient for discrete X.

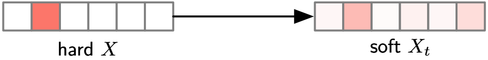
Solution: Grad. estimation with Stochastic Softmax Tricks

Relax discrete X to continuous X_t to admit *biased* gradient...

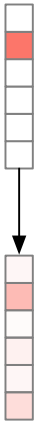


Solution: Grad. estimation with Stochastic Softmax Tricks

Relax discrete X to continuous X_t to admit *biased* gradient...

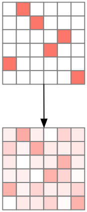


Our framework generalizes previous work on relaxations...



(Jang et al., 2016)
(Maddison et al., 2017)

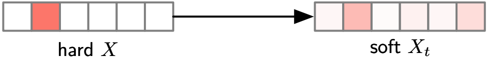
(Chen et al., 2018)
(Xie and Ermon, 2019)



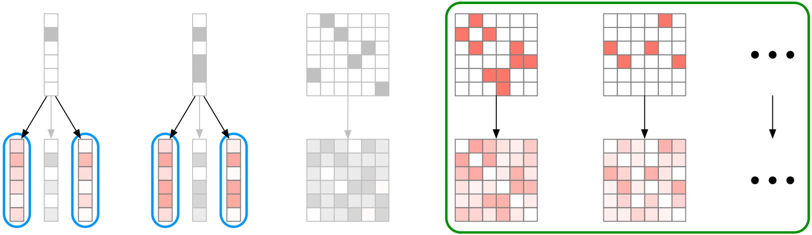
(Mena et al., 2018)

Solution: Grad. estimation with Stochastic Softmax Tricks

Relax discrete X to continuous X_t to admit *biased* gradient...



Our framework generalizes previous work on relaxations...



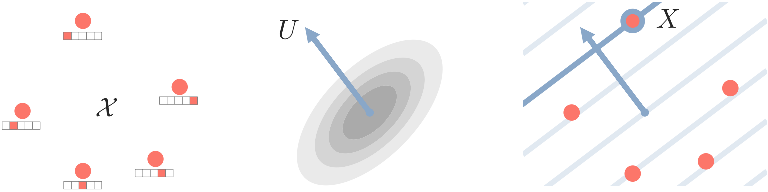
...and includes new relaxations and new structured variables.

Stochastic Argmax Tricks (SMTs)

SMTs reparameterize X as solution to a random linear program...

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

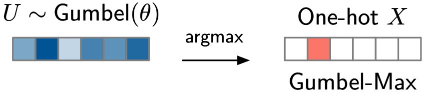
...where the U induces a distribution over $\mathcal{X}$ (Hazan et al., 2016).



Finite set | Random Utility | SMT

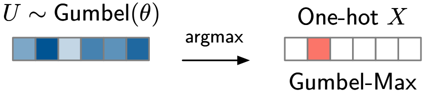
Stochastic Argmax Tricks (SMTs)

SMTs recover the Gumbel-Max trick in the one-hot case...

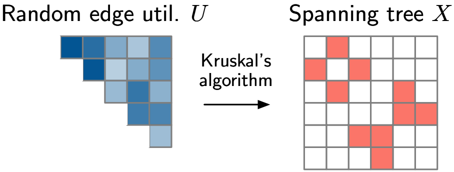


Stochastic Argmax Tricks (SMTs)

SMTs recover the Gumbel-Max trick in the one-hot case...



...and generalize it to other structured X...



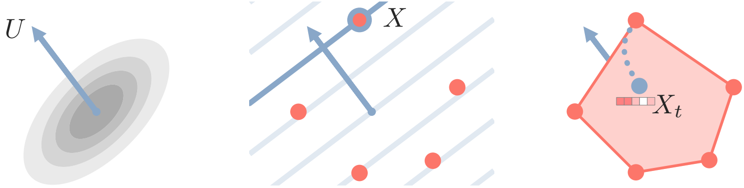
...for which efficient linear solvers are available.

Stochastic Softmax Tricks (SSTs)

SSTs relax a given SMT...

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

...to relax discrete X to continuous X_t...



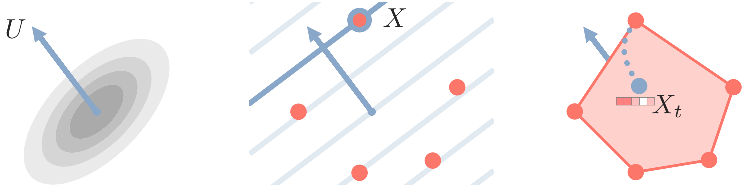
Random Utility SMT SST

Stochastic Softmax Tricks (SSTs)

SSTs relax a given SMT...

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

...to relax discrete X to continuous X_t...

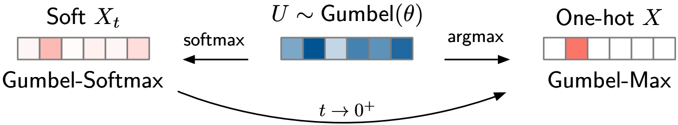


Random Utility SMT SST

... which admits a reparameterization gradient.

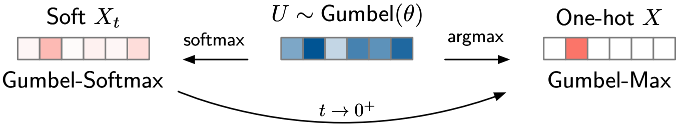
Stochastic Softmax Tricks (SSTs)

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

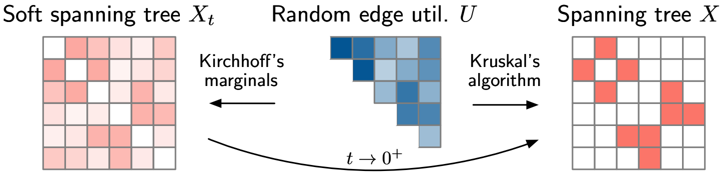


Stochastic Softmax Tricks (SSTs)

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



...to generalize it to other structured X...



...when efficient solvers are available given f and $\mathcal{X}$.

Experiments: Overview

We use SSTs to train deep latent variable models over structured discrete domains...

Experiments: Overview

We use SSTs to train deep latent variable models over structured discrete domains...

- NRI (Kipf et al., 2018) for graph layout

Experiments: Overview

We use SSTs to train deep latent variable models over structured discrete domains...

- NRI (Kipf et al., 2018) for graph layout
- Unsupervised parsing on ListOps (Nangia and Bowman, 2018)

Experiments: Overview

We use SSTs to train deep latent variable models over structured discrete domains...

- NRI (Kipf et al., 2018) for graph layout
- Unsupervised parsing on ListOps (Nangia and Bowman, 2018)
- L2X (Chen et al., 2018) aspect rating

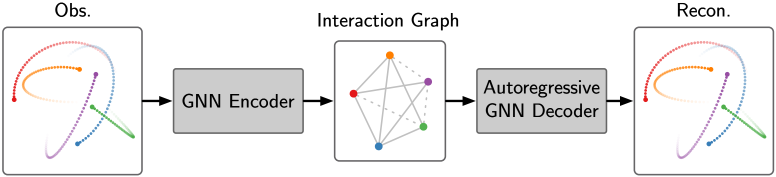
Experiments: Overview

We use SSTs to train deep latent variable models over structured discrete domains...

- **NRI (Kipf et al., 2018) for graph layout**
- Unsupervised parsing on ListOps (Nangia and Bowman, 2018)
- **L2X (Chen et al., 2018) aspect rating**

Neural Relational Inference (NRI)

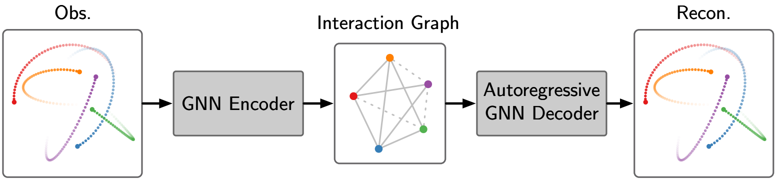
NRI is a VAE with a latent graph...



(Kipf et al., 2018)

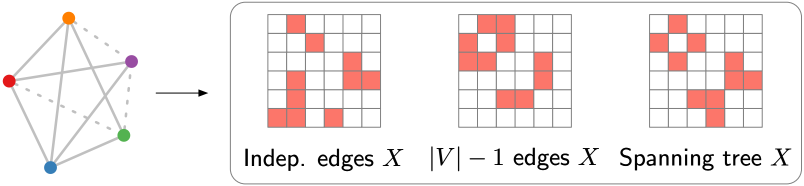
Neural Relational Inference (NRI)

NRI is a VAE with a latent graph...



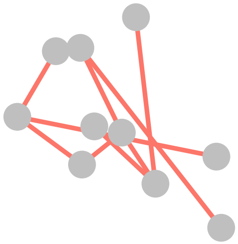
(Kipf et al., 2018)

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



NRI: Data

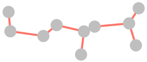
Consider particle trajectories from a force-directed algorithm...



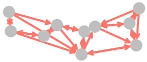
...where the latent graph is a spanning tree.

NRI: Results

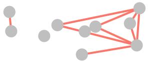
More structured models improve structure recovery...



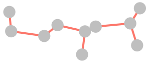
Ground Truth



Indep. Edges



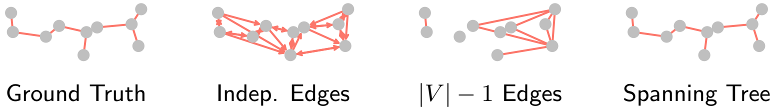
$|V| - 1$ Edges



Spanning Tree

NRI: Results

More structured models improve structure recovery...



...and performance on the task...

Edge Distribution	ELBO	Edge Prec.	Edge Rec.
Indep. Edges	-1370 ± 20	48 ± 2	$\mathbf{93 \pm 1}$
$\lvert V \rvert - 1$ edges	-2100 ± 20	41 ± 1	41 ± 1
Spanning Tree	$\mathbf{-1080 \pm 110}$	$\mathbf{91 \pm 3}$	91 ± 3

L2X: Results

More structured models select contiguous phrases...

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

L2X: Results

More structured models select contiguous phrases...

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

...and select more relevant words to improve performance...

	$k = 5$		$k = 10$		$k = 15$	
Relaxation	MSE	Subs. Prec.	MSE	Subs. Prec.	MSE	Subs. 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 and 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	$\mathbf{2.9 \pm 0.1}$	$\mathbf{63.1 \pm 5.3}$	$\mathbf{2.5 \pm 0.1}$	$\mathbf{53.1 \pm 0.9}$	$\mathbf{2.4 \pm 0.1}$	$\mathbf{45.5 \pm 2.7}$

Conclusion

Gradient estimation with stochastic softmax tricks...

- ...generalizes the Gumbel-Softmax to structured spaces.

Conclusion

Gradient estimation with stochastic softmax tricks...

- ...generalizes the Gumbel-Softmax to structured spaces.
- ...admits novel relaxation for new combinatorial objects.

Conclusion

Gradient estimation with stochastic softmax tricks...

- ...generalizes the Gumbel-Softmax to structured spaces.
- ...admits novel relaxation for new combinatorial objects.
- ...gives a unified perspective on existing reparameterizations and relaxations.

References I

J. Chen, L. Song, M. Wainwright, and M. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, 2018.

A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

T. Hazan, G. Papandreou, and D. Tarlow. *Perturbations, Optimization, and Statistics*. MIT Press, 2016.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2016.

T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, 2018.

T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*, 2016.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

G. Mena, D. Belanger, S. Linderman, and J. Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Byt3oJ-0W`.

N. Nangia and S. R. Bowman. Listops: A diagnostic dataset for latent tree learning. *arXiv preprint arXiv:1804.06028*, 2018.

S. M. Xie and S. Ermon. Reparameterizable subset sampling via continuous relaxations. In *International Joint Conference on Artificial Intelligence*, 2019.