




Latent Structure Models for NLP

Tsvetomila Mihaylova Instituto de Telecomunicações
Vlad Niculae Instituto de Telecomunicações

work with:

André Martins Instituto de Telecomunicações & IST & Unbabel
Nikita Nangia NYU

 deep-spin.github.io/tutorial

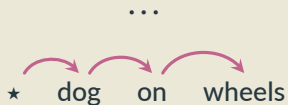
I. Introduction

Structured prediction and NLP

- **Structured prediction:** a machine learning framework for predicting structured, constrained, and interdependent outputs
- **NLP** deals with *structured* and *ambiguous* textual data:
 - machine translation
 - speech recognition
 - syntactic parsing
 - semantic parsing
 - information extraction
 - ...

Examples of structure in NLP

Dependency parsing



...

Examples of structure in NLP

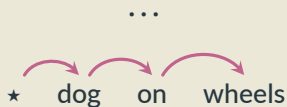
POS tagging

VERB PREP NOUN
dog on wheels

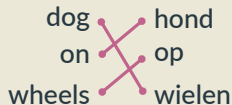
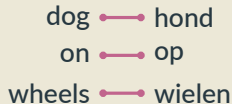
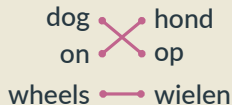
NOUN PREP NOUN
dog on wheels

NOUN DET NOUN
dog on wheels

Dependency parsing

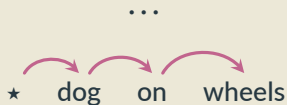


Word alignments



Examples of structure in NLP

Dependency parsing



Exponentially many structures!



...

NLP 5 years ago:

Structured prediction and pipelines



NLP 5 years ago:

Structured prediction and pipelines

- Big pipeline systems, connecting different structured predictors, trained separately
- **Advantages:** fast and simple to train, can rearrange pieces 😊

NLP 5 years ago:

Structured prediction and pipelines

- Big pipeline systems, connecting different structured predictors, trained separately
- **Advantages:** fast and simple to train, can rearrange pieces 😊
- **Disadvantage:** linguistic annotations required for each component 😓

NLP 5 years ago:

Structured prediction and pipelines

- Big pipeline systems, connecting different structured predictors, trained separately
- **Advantages:** fast and simple to train, can rearrange pieces 😊
- **Disadvantage:** linguistic annotations required for each component 😓
- **Bigger disadvantage:** error propagates through the pipeline 💩

NLP today:

End-to-end training



NLP today:

End-to-end training

- Forget pipelines—train everything from scratch!
- No more error propagation or linguistic annotations! 🎉

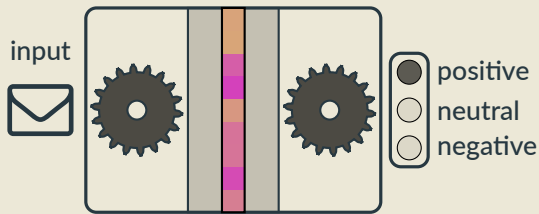
NLP today:

End-to-end training

- Forget pipelines—train everything from scratch!
- No more error propagation or linguistic annotations! 🎉
- Treat everything as *latent*! 🙌

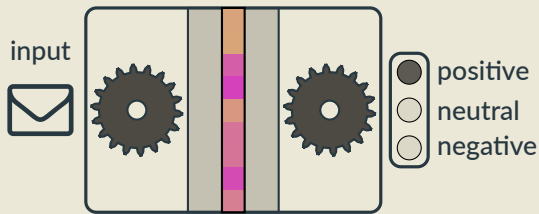
Representation learning

- Uncover hidden representations useful for the *downstream task*.
- Neural networks are well-suited for this: *deep computation graphs*.



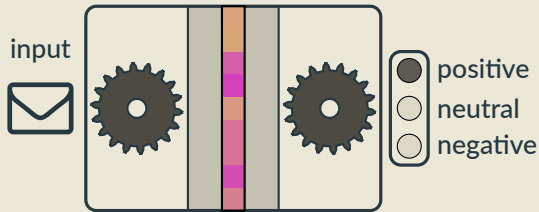
Representation learning

- Uncover hidden representations useful for the *downstream task*.
- Neural networks are well-suited for this: *deep computation graphs*.
- Neural representations are unstructured, inscrutable.
Language data has underlying structure!



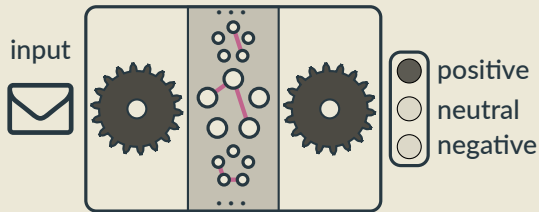
Latent structure models

- Seek *structured* hidden representations instead!



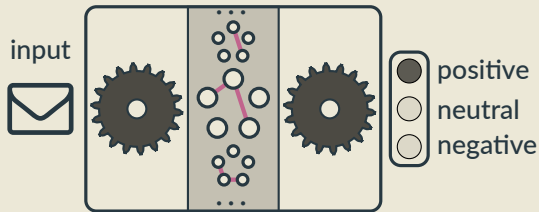
Latent structure models

- Seek *structured* hidden representations instead!



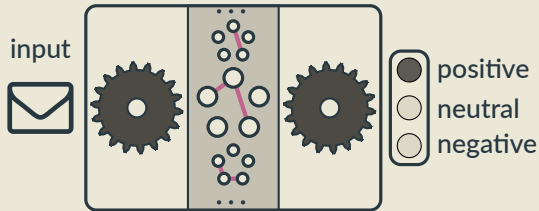
Latent structure models

- Seek *structured* hidden representations instead!
- They can bring us:
 - More interpretability;



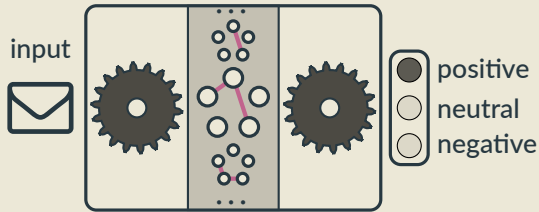
Latent structure models

- Seek *structured* hidden representations instead!
- They can bring us:
 - More interpretability;
 - Better inductive bias;



Latent structure models

- Seek *structured* hidden representations instead!
- They can bring us:
 - More interpretability;
 - Better inductive bias;
 - Hopefully: smaller models.



Latent structure models aren't so new!

- They have a very long history in NLP:
 - IBM Models for SMT (latent word alignments) [Brown et al., 1993]
 - HMMs [Rabiner, 1989]
 - CRFs with hidden variables [Quattoni et al., 2007]
 - Latent PCFGs [Petrov and Klein, 2008, Cohen et al., 2012]
- Trained with EM, spectral learning, method of moments, ...
- Often, very strict assumptions (e.g. strong factorizations)
- Today, neural networks opened up some new possibilities!

What this tutorial is about:

- Discrete, combinatorial latent structures
- Often the structure is inspired by some linguistic intuition
- We'll cover both:
 - RL methods (structure built incrementally, reward coming from downstream task)
 - ... vs end-to-end differentiable approaches (global optimization, marginalization)
 - stochastic computation graphs
 - ... vs deterministic graphs.
- All plugged in *discriminative* neural models.

This tutorial is *not* about:

- It's not about continuous latent variables
- It's not about deep generative learning
- We won't cover GANs, VAEs, etc.
- There are (very good) recent tutorials on deep variational models for NLP:
 - “Variational Inference and Deep Generative Models” (Schulz and Aziz, ACL 2018)
 - “Deep Latent-Variable Models for Natural Language” (Kim, Wiseman, Rush, EMNLP 2018)

Background

Unstructured vs structured

- Simplest example of structure: Just a discrete choice among N categories.
- We call this *unstructured*.
- It will provide an important starting point.

The challenge of discrete choices

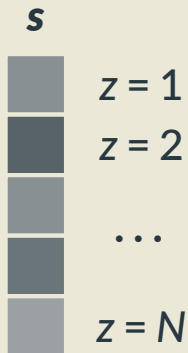
$$z = 1$$

$$z = 2$$

...

$$z = N$$

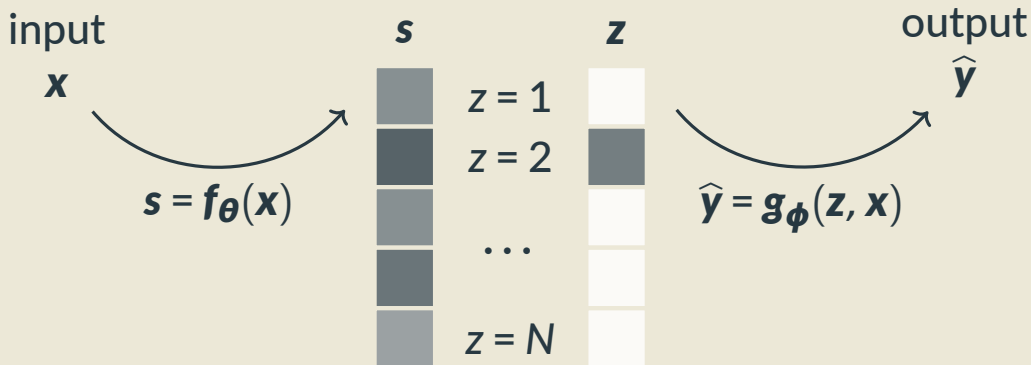
The challenge of discrete choices



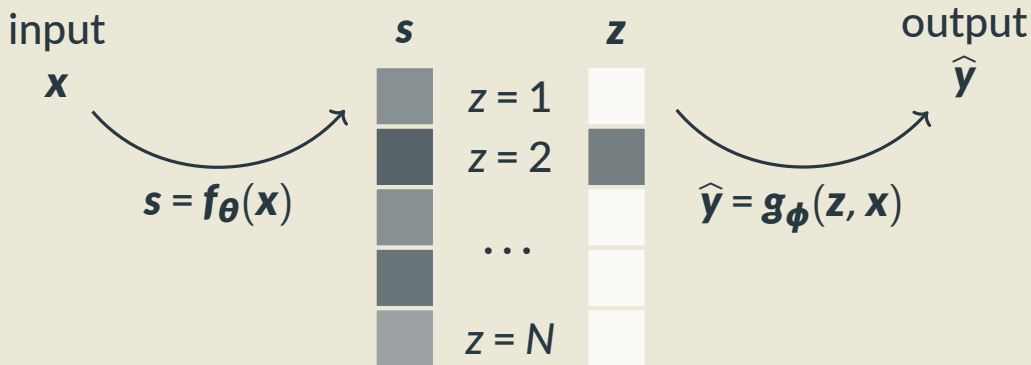
The challenge of discrete choices



The challenge of discrete choices

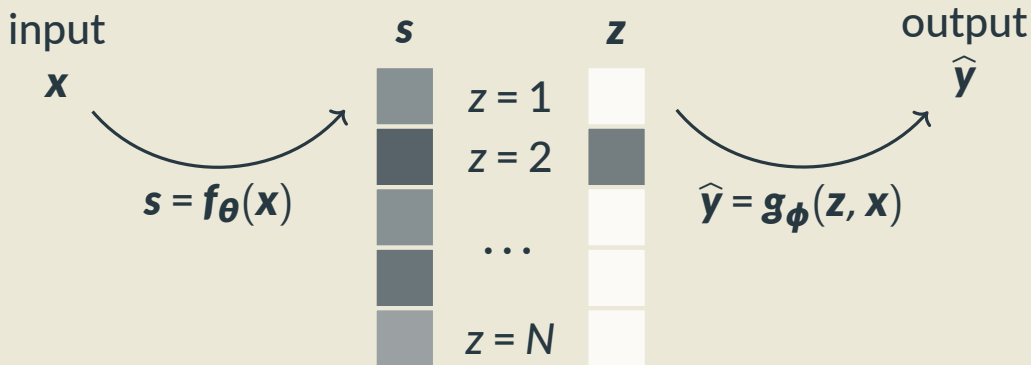


The challenge of discrete choices



$$\frac{\partial L(\hat{y}, y)}{\partial w} = ?$$

The challenge of discrete choices

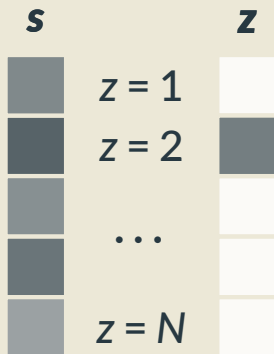


$$\frac{\partial L(\hat{\mathbf{y}}, \mathbf{y})}{\partial \mathbf{w}} = ?$$

or, essentially,

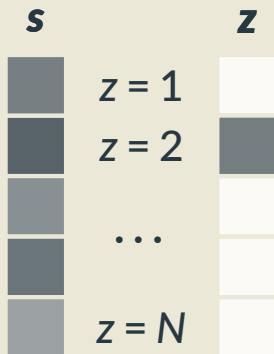
$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



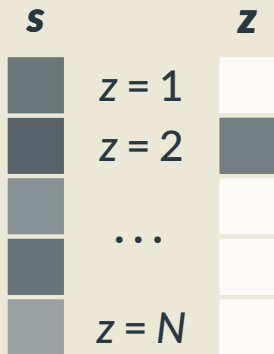
$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



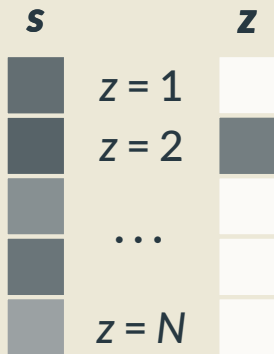
$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”



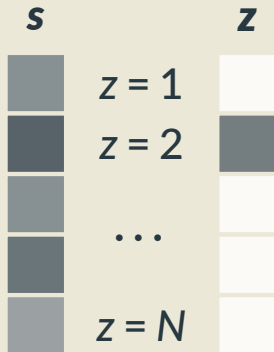
$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Discrete mappings are “flat”

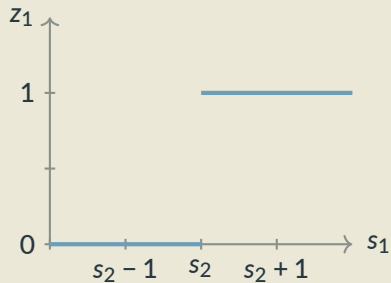


$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = ?$$

Argmax

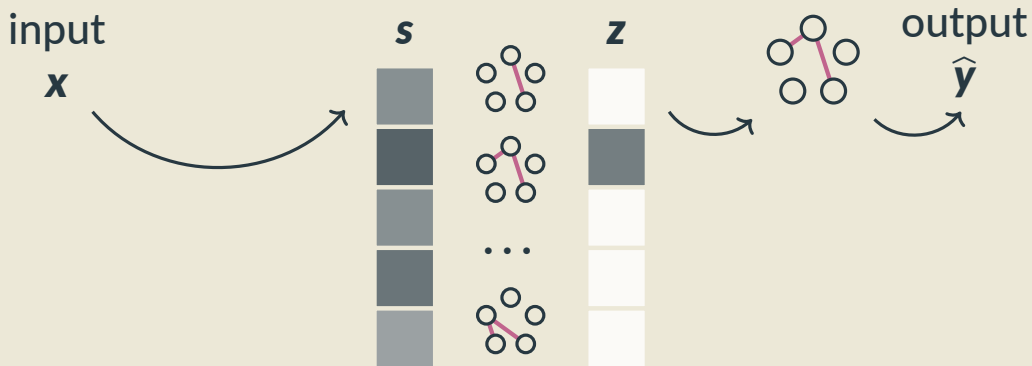


$$\frac{\partial z}{\partial s} = \mathbf{0}$$



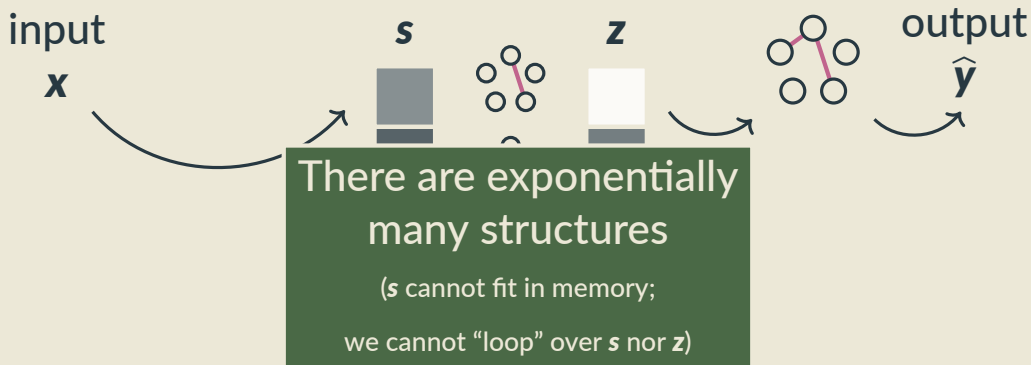
Computing the most likely structure

is a very high-dimensional argmax

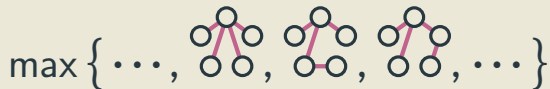
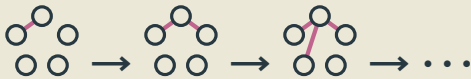


Computing the most likely structure

is a very high-dimensional argmax



Dealing with the combinatorial explosion



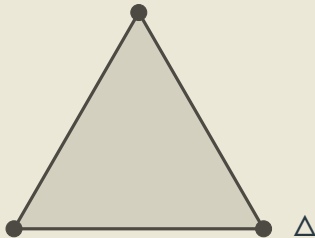
1. Incremental structures

- Build structure **greedily**, as sequence of discrete choices (e.g., shift-reduce).
- Scores (partial structure, action) tuples.
- **Advantages:** flexible, rich histories.
- **Disadvantages:** greedy, local decisions are suboptimal, error propagation.

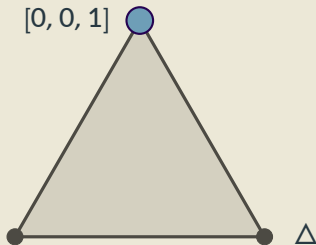
2. Factorization into parts

- Optimizes **globally** (e.g. Viterbi, Chu-Liu-Edmonds, Kuhn-Munkres).
- Scores smaller parts.
- **Advantages:** optimal, elegant, can handle hard & global constraints.
- **Disadvantages:** strong assumptions.

The unstructured case: Probability simplex



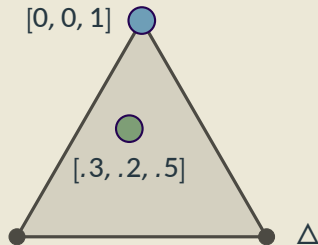
The unstructured case: Probability simplex



- Each vertex is an *indicator vector*, representing one class:

$$\mathbf{z}_c = [0, \dots, 0, \underbrace{1}_{c^{\text{th}} \text{ position}}, 0, \dots, 0].$$

The unstructured case: Probability simplex



- Each vertex is an *indicator vector*, representing one class:

$$\mathbf{z}_c = [0, \dots, 0, \underbrace{1}_{c^{\text{th}} \text{ position}}, 0, \dots, 0].$$

- Points inside are *probability vectors*, a convex combination of classes:

$$\mathbf{p} \geq \mathbf{0}, \quad \sum_c p_c = 1.$$

What's the analogous of Δ for a structure?

- A structured object \mathbf{z} can be represented as a *bit vector*.

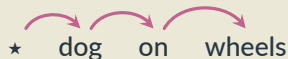
What's the analogous of Δ for a structure?

- A structured object \mathbf{z} can be represented as a *bit vector*.
- Example:
 - a dependency tree can be represented a $O(L^2)$ vector indexed by arcs
 - each entry is 1 iff the arc belongs to the tree
 - **structural constraints:** not all bit vectors represent valid trees!

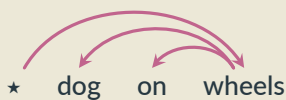
What's the analogous of Δ for a structure?

- A structured object \mathbf{z} can be represented as a *bit vector*.
- Example:
 - a dependency tree can be represented a $O(L^2)$ vector indexed by arcs
 - each entry is 1 iff the arc belongs to the tree
 - **structural constraints:** not all bit vectors represent valid trees!

$$\mathbf{z}_1 = [1, 0, 0, 0, 1, 0, 0, 0, 1]$$



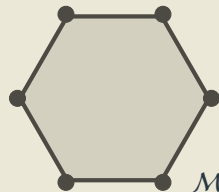
$$\mathbf{z}_2 = [0, 0, 1, 0, 0, 1, 1, 0, 0]$$



$$\mathbf{z}_3 = [1, 0, 0, 0, 1, 0, 0, 1, 0]$$

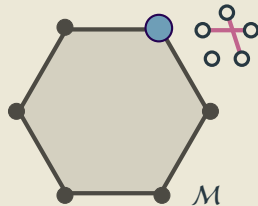


The structured case: Marginal polytope



The structured case: Marginal polytope

- Each vertex corresponds to one such *bit* vector \mathbf{z}



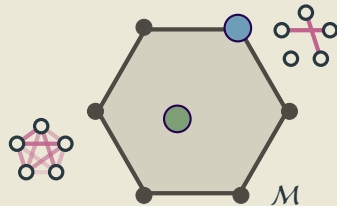
The structured case: Marginal polytope

- Each vertex corresponds to one such *bit vector* \mathbf{z}
- Points inside correspond to *marginal distributions*: convex combinations of structured objects

$$\boldsymbol{\mu} = \underbrace{p_1 \mathbf{z}_1 + \dots + p_N \mathbf{z}_N}_{\text{exponentially many terms}}, \quad \mathbf{p} \in \Delta.$$

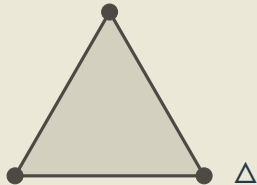
$$\begin{aligned} p_1 &= 0.2, & \mathbf{z}_1 &= [1, 0, 0, 0, 1, 0, 0, 0, 1] \\ p_2 &= 0.7, & \mathbf{z}_2 &= [0, 0, 1, 0, 0, 1, 1, 0, 0] \\ p_3 &= 0.1, & \mathbf{z}_3 &= [1, 0, 0, 0, 1, 0, 0, 1, 0] \end{aligned}$$

$$\Rightarrow \boldsymbol{\mu} = [.3, 0, .7, 0, .3, .7, .7, .1, .2].$$

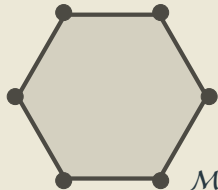


Unstructured vs Structured

- Unstructured case: simplex Δ

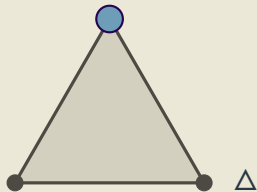


- Structured case: marginal polytope \mathcal{M}

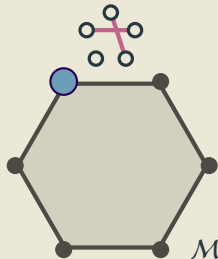


Unstructured vs Structured

- Unstructured case: simplex Δ

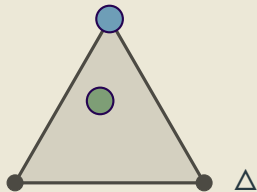


- Structured case: marginal polytope \mathcal{M}

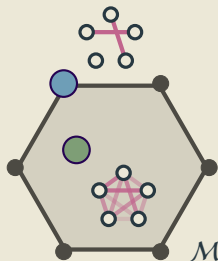


Unstructured vs Structured

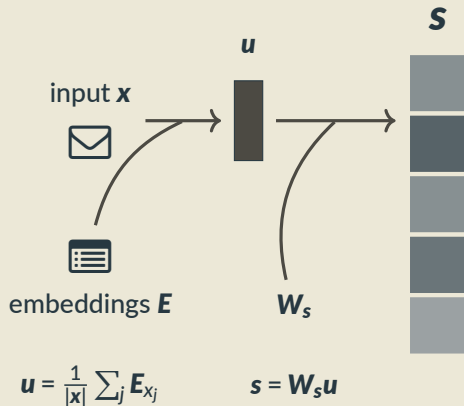
- Unstructured case: simplex Δ



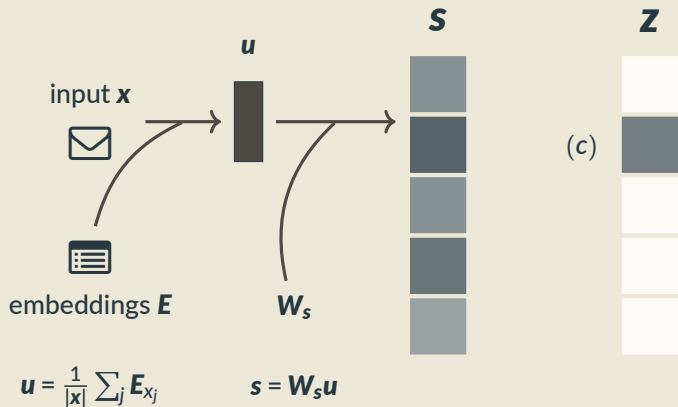
- Structured case: marginal polytope \mathcal{M}



Example: Regression with latent categorization

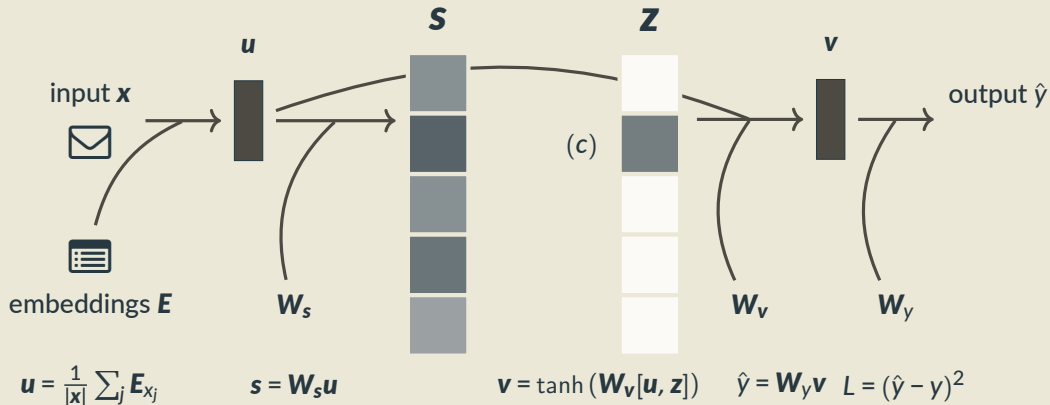


Example: Regression with latent categorization



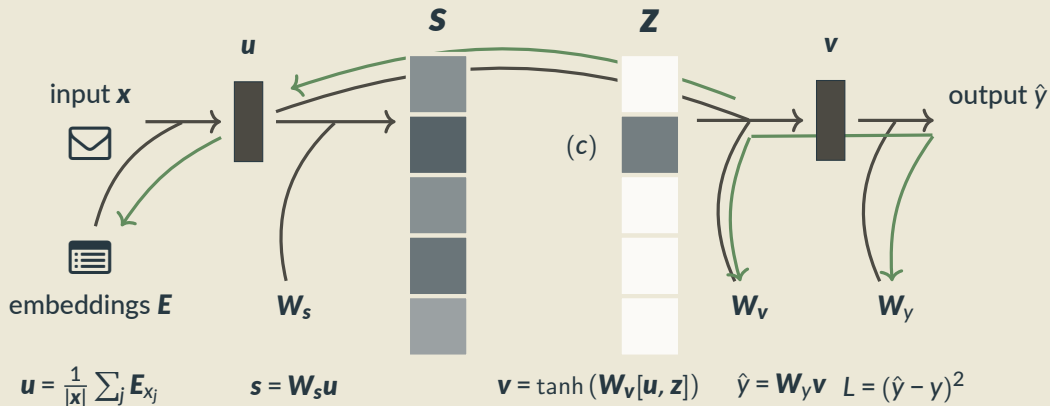
predict topic c ($z = e_c$)

Example: Regression with latent categorization

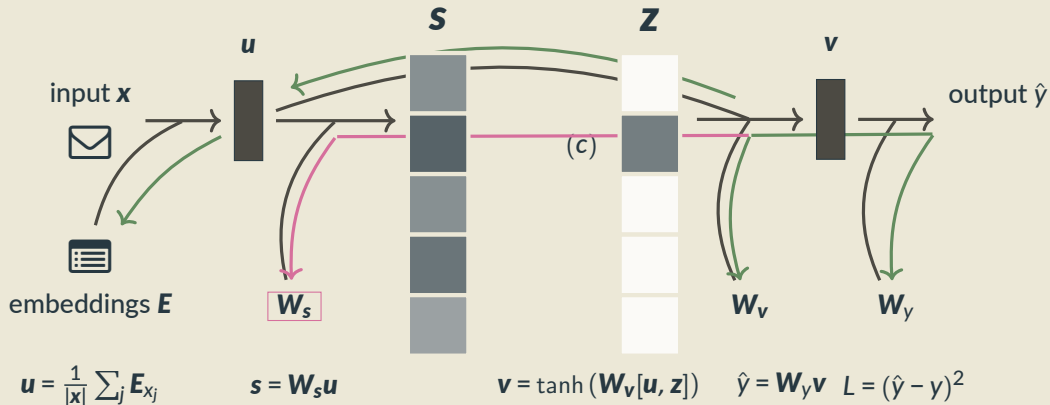


predict topic c ($z = e_c$)

Example: Regression with latent categorization

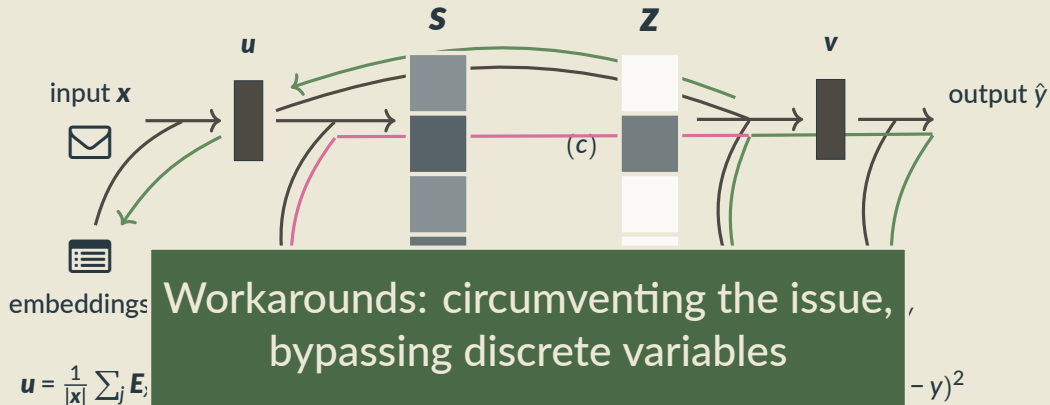


Example: Regression with latent categorization

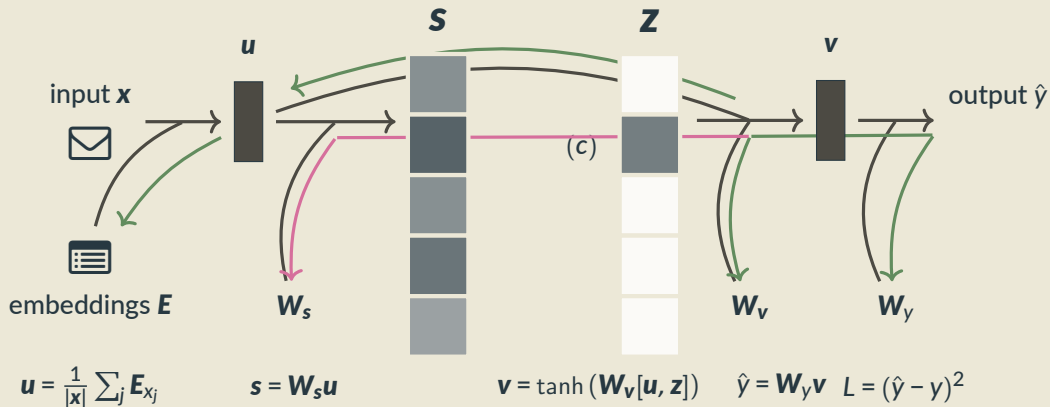


$$\frac{\partial L}{\partial W_s} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v} \frac{\partial v}{\partial z} \underbrace{\frac{\partial z}{\partial s}}_{\equiv 0} \frac{\partial s}{\partial W_s}$$

Example: Regression with latent categorization

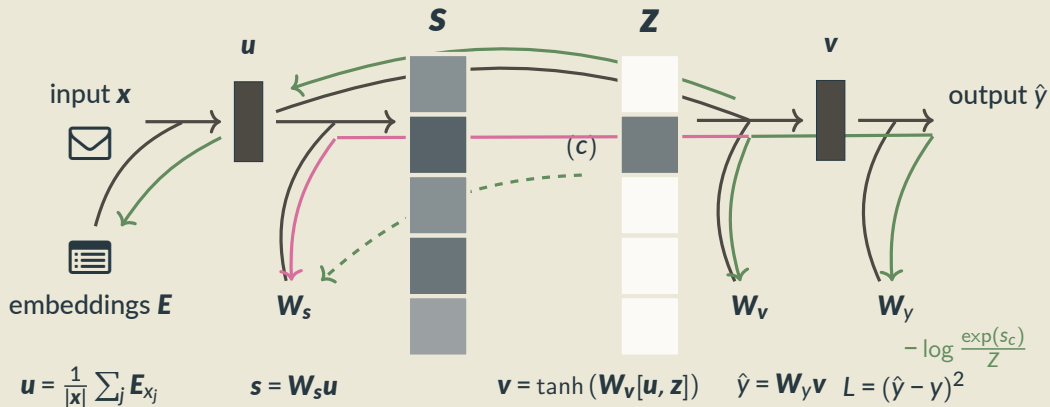


Example: Regression with latent categorization



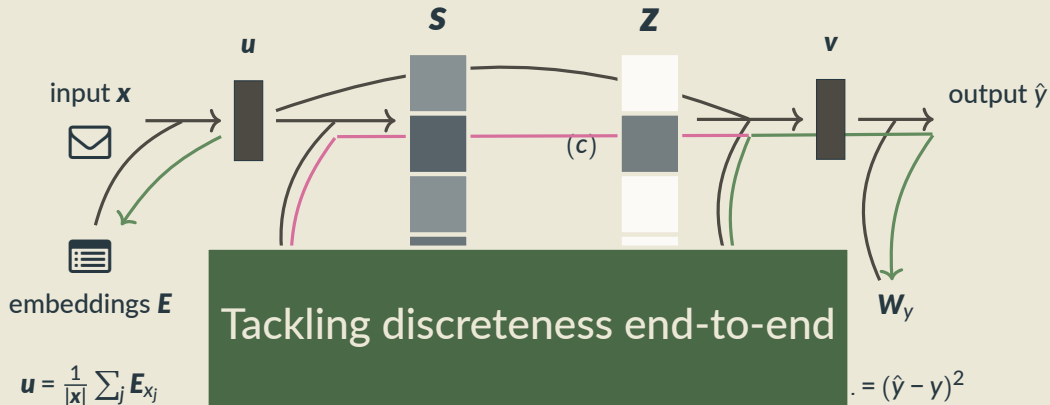
Option 1. Pretrain latent classifier W_s

Example: Regression with latent categorization

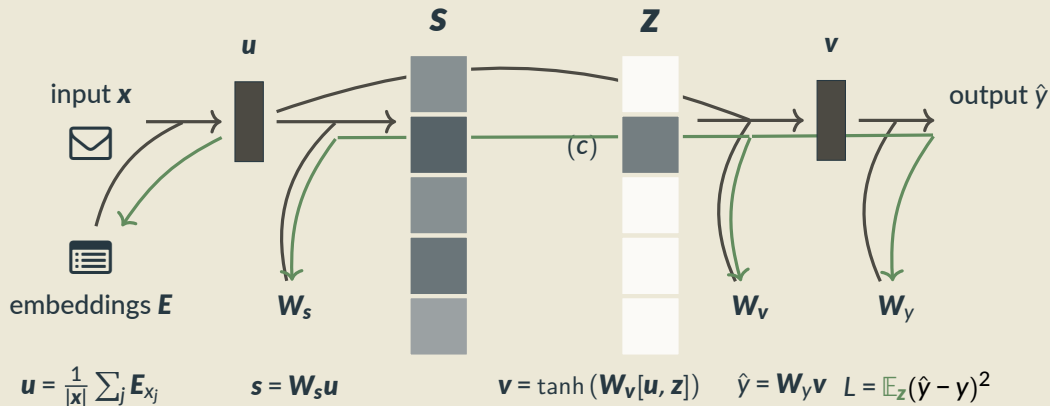


Option 2. Multi-task learning

Example: Regression with latent categorization

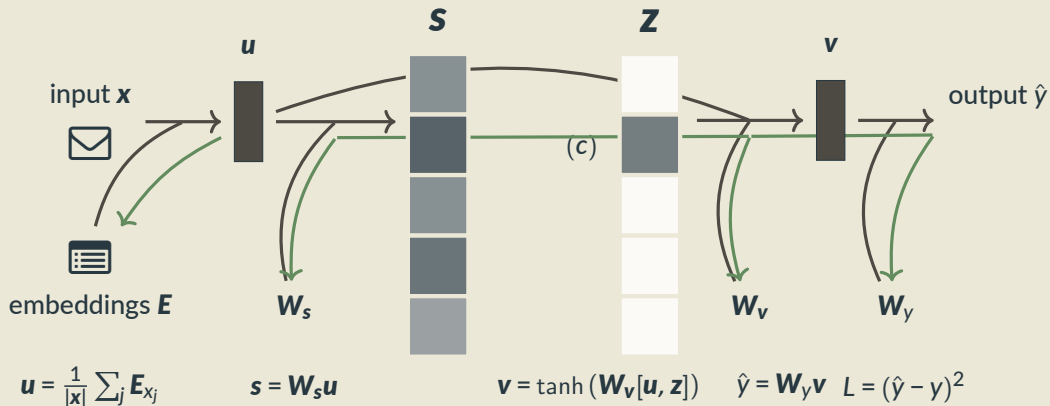


Example: Regression with latent categorization



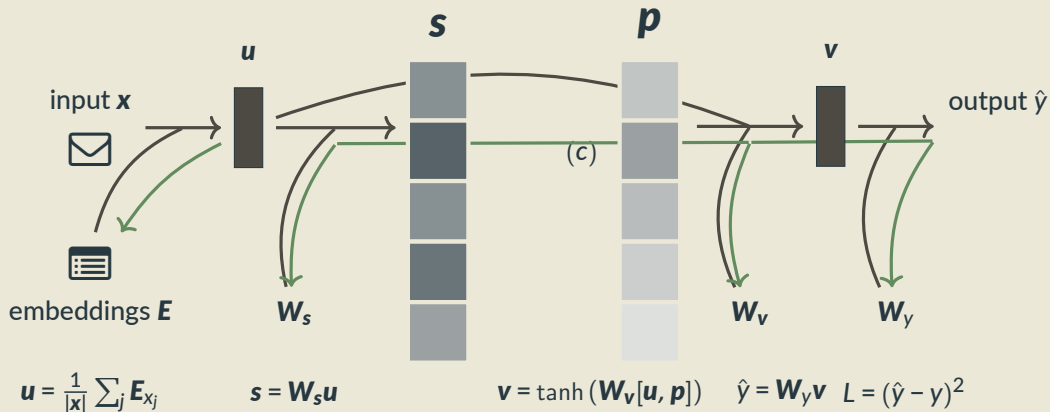
Option 3. Stochasticity! $\frac{\partial \mathbb{E}_z (\hat{y}(z) - y)^2}{\partial W_s} \neq 0$

Example: Regression with latent categorization



Option 4. Gradient surrogates (e.g. straight-through, $\frac{\partial z}{\partial s} \leftarrow I$)

Example: Regression with latent categorization



Option 5. Continuous relaxation (e.g. softmax)

Dealing with discrete latent variables

1. Pre-train external classifier
2. Multi-task learning
3. Stochastic latent variables
4. Gradient surrogates
5. Continuous relaxation

Dealing with discrete latent variables

1. Pre-train external classifier
2. Multi-task learning
3. Stochastic latent variables (Part 2)
4. Gradient surrogates (Part 3)
5. Continuous relaxation (Part 4)

Roadmap of the tutorial

- Part 1: Introduction ✓
- Part 2: Reinforcement learning

Coffee Break

- Part 3: Gradient surrogates
- Part 4: End-to-end differentiable models (1/2)

Coffee Break

- Part 4: End-to-end differentiable models (2/2)
- Part 5: Conclusions

II. Reinforcement Learning Methods

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree z ,

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree \mathbf{z} , we can make a prediction $\hat{y}(\mathbf{z}; x)$

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree \mathbf{z} , we can make a prediction $\hat{y}(\mathbf{z}; x)$ and incur a loss,

$$L(\hat{y}(\mathbf{z}; x), y)$$

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree \mathbf{z} , we can make a prediction $\hat{y}(\mathbf{z}; x)$ and incur a loss,

$$L(\hat{y}(\mathbf{z}; x), y) \text{ or simply } L(\mathbf{z})$$

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree \mathbf{z} , we can make a prediction $\hat{y}(\mathbf{z}; x)$ and incur a loss,

$$L(\hat{y}(\mathbf{z}; x), y) \text{ or simply } L(\mathbf{z})$$

- But we don't know \mathbf{z} !

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree \mathbf{z} , we can make a prediction $\hat{y}(\mathbf{z}; x)$ and incur a loss,

$$L(\hat{y}(\mathbf{z}; x), y) \text{ or simply } L(\mathbf{z})$$

- But we don't know \mathbf{z} !
- In this section:
we jointly learn a structured prediction model $\pi_{\theta}(\mathbf{z} \mid x)$

Latent structure via marginalization

- Given a sentence-label pair (x, y) and its **known** parse tree \mathbf{z} , we can make a prediction $\hat{y}(\mathbf{z}; x)$ and incur a loss,

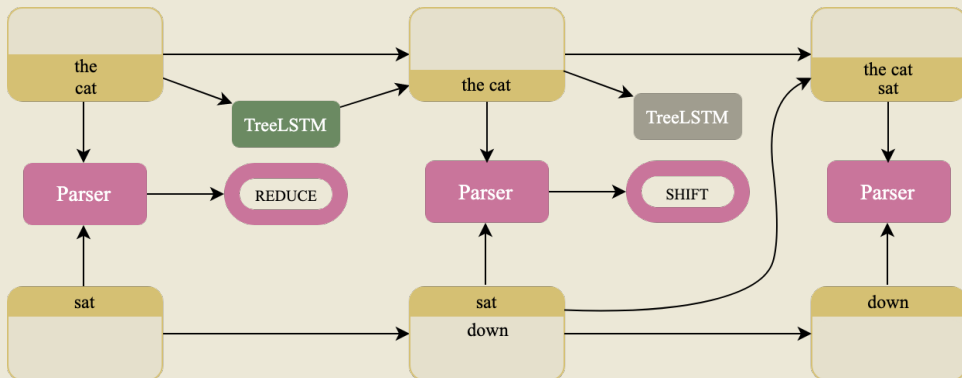
$$L(\hat{y}(\mathbf{z}; x), y) \text{ or simply } L(\mathbf{z})$$

- But we don't know \mathbf{z} !
- In this section:
we jointly learn a structured prediction model $\pi_{\theta}(\mathbf{z} \mid x)$ by optimizing the **expected loss**,

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z} \mid x)}[L(\mathbf{z})]$$

**But first, supervised
SPINN**

Stack-augmented Parser-Interpreter Neural-Network



Stack-augmented Parser-Interpreter Neural-Network

- Joint learning: Combines a constituency parser and a sentence representation model.

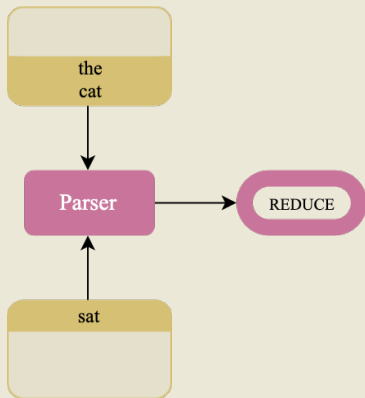
Stack-augmented Parser-Interpreter Neural-Network

- Joint learning: Combines a constituency parser and a sentence representation model.
- The parser, $f_{\theta}(x)$ is a transition-based **shift-reduce** parser. It looks at top two elements of stack and top element of the buffer.

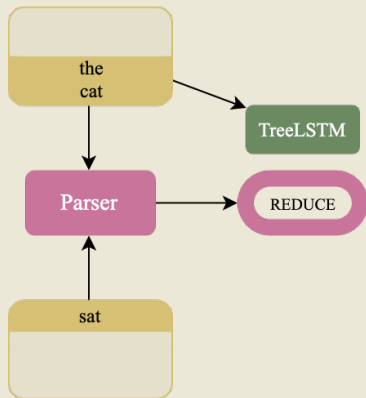
Stack-augmented Parser-Interpreter Neural-Network

- Joint learning: Combines a constituency parser and a sentence representation model.
- The parser, $f_{\theta}(x)$ is a transition-based **shift-reduce** parser. It looks at top two elements of stack and top element of the buffer.
- **TreeLSTM** combines top two elements of the stack when the parser chooses the REDUCE action.

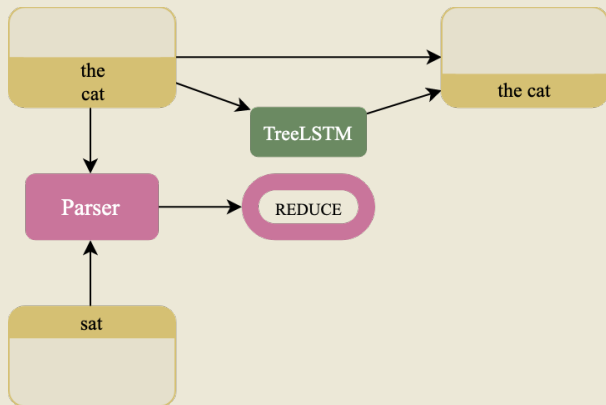
Stack-augmented Parser-Interpreter Neural-Network



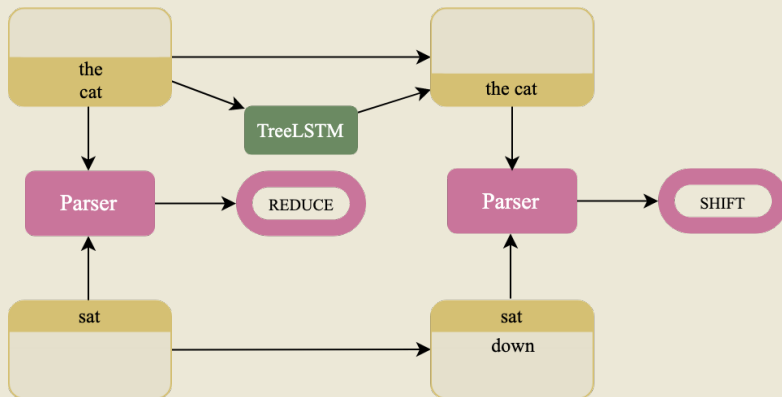
Stack-augmented Parser-Interpreter Neural-Network



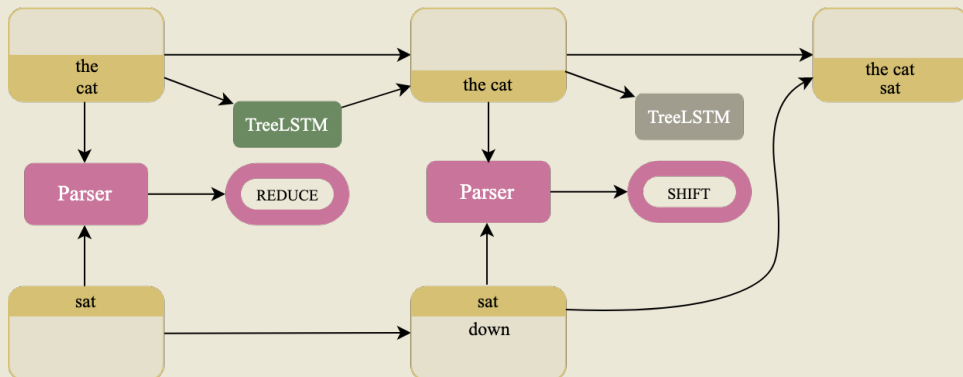
Stack-augmented Parser-Interpreter Neural-Network



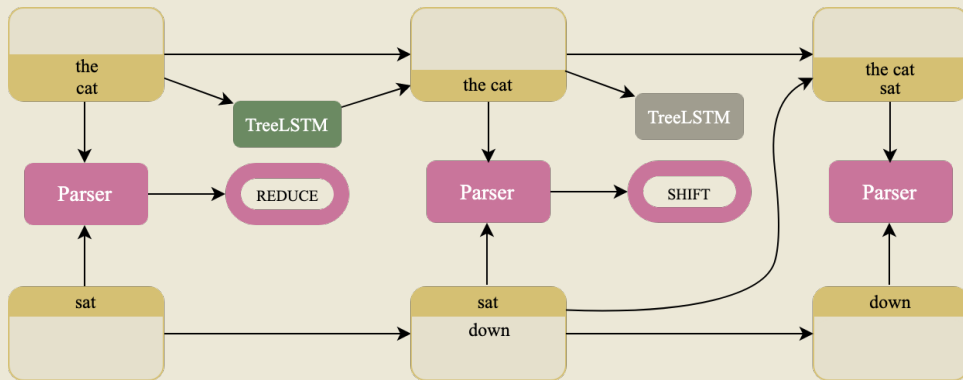
Stack-augmented Parser-Interpreter Neural-Network



Stack-augmented Parser-Interpreter Neural-Network



Stack-augmented Parser-Interpreter Neural-Network



Shift-Reduce parsing

We can write a shift-reduce style parse as a sequence of Bernoulli random variables,

$$\mathbf{z} = \{z_1, \dots, z_{2L-1}\}$$

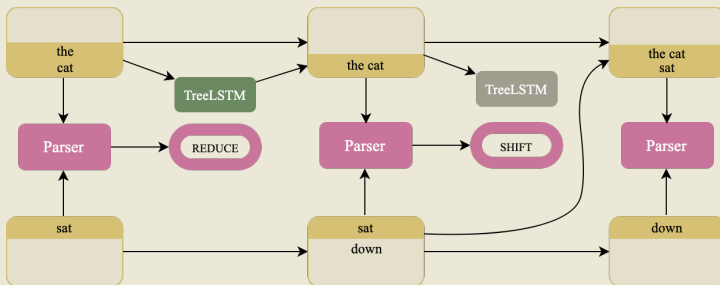
where, $z_j \in \{0, 1\} \ \forall j \in [1, 2L - 1]$

Shift-Reduce parsing

A sequence of Bernoulli trials but with conditional dependence,

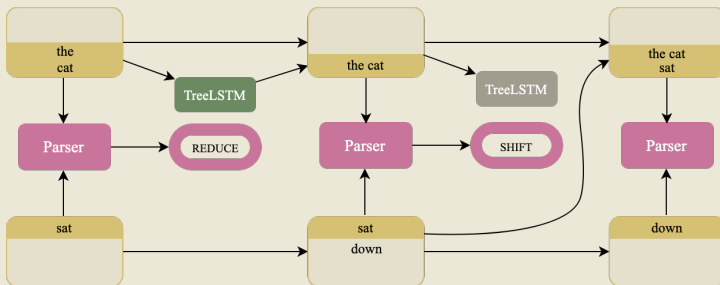
$$p(z_1, z_2, \dots, z_{2L-1}) = \prod_{j=1}^{2L-1} p(z_j \mid z_{<j})$$

Latent structure learning with SPINN



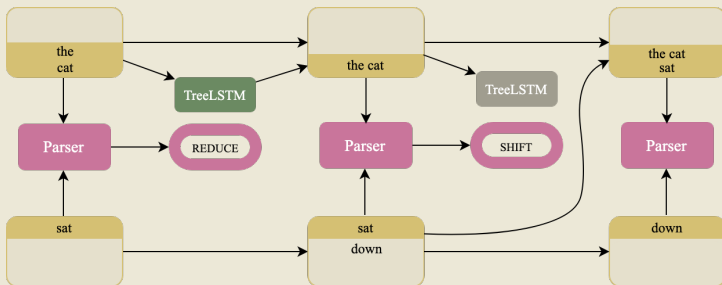
Latent structure learning with SPINN

- But now, remove syntactic supervision from SPINN.



Latent structure learning with SPINN

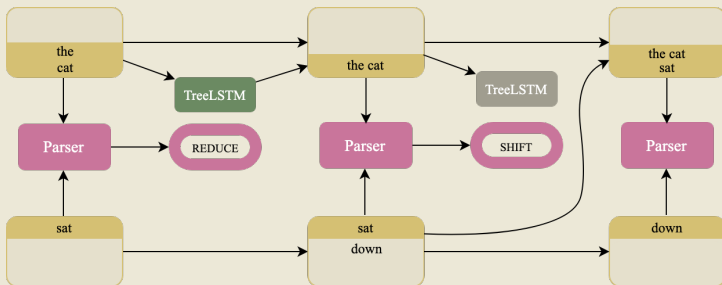
- But now, remove syntactic supervision from SPINN.



- We model the parse, \mathbf{z} , as a latent variable scored by $f_{\theta}(\mathbf{x})$.

Latent structure learning with SPINN

- But now, remove syntactic supervision from SPINN.



- We model the parse, \mathbf{z} , as a latent variable scored by $f_{\theta}(\mathbf{x})$.
- With shift-reduce parsing, we're making discrete decisions
 \Rightarrow REINFORCE as a “natural” solution.

Unsupervised SPINN

Unsupervised SPINN

No syntactic supervision.

Only reward is from the downstream task.

We only get this reward after parsing the full sentence.

SPINN with REINFORCE

Some basic terminology,

- The action space is $z_j \in \{\text{SHIFT}, \text{REDUCE}\}$, and \mathbf{z} is a sequence of actions.

SPINN with REINFORCE

Some basic terminology,

- The action space is $z_j \in \{\text{SHIFT}, \text{REDUCE}\}$, and \mathbf{z} is a sequence of actions.
- Training parser network parameters, θ with REINFORCE

SPINN with REINFORCE

Some basic terminology,

- The action space is $z_j \in \{\text{SHIFT}, \text{REDUCE}\}$, and \mathbf{z} is a sequence of actions.
- Training parser network parameters, θ with REINFORCE
- The state, \mathbf{h} , is the top two elements of the stack and the top element of the buffer.

SPINN with REINFORCE

Some basic terminology,

- The action space is $z_j \in \{\text{SHIFT}, \text{REDUCE}\}$, and \mathbf{z} is a sequence of actions.
- Training parser network parameters, $\boldsymbol{\theta}$ with REINFORCE
- The state, \mathbf{h} , is the top two elements of the stack and the top element of the buffer.
- Learning a policy network $\pi(\mathbf{z} \mid \mathbf{h}; \boldsymbol{\theta})$

SPINN with REINFORCE

Some basic terminology,

- The action space is $z_j \in \{\text{SHIFT}, \text{REDUCE}\}$, and \mathbf{z} is a sequence of actions.
- Training parser network parameters, $\boldsymbol{\theta}$ with REINFORCE
- The state, \mathbf{h} , is the top two elements of the stack and the top element of the buffer.
- Learning a policy network $\pi(\mathbf{z} \mid \mathbf{h}; \boldsymbol{\theta})$
- Maximize the reward, where \mathcal{R} is performance on the downstream task like sentence classification.

SPINN with REINFORCE

Some basic terminology,

- The action space is $z_j \in \{\text{SHIFT}, \text{REDUCE}\}$, and \mathbf{z} is a sequence of actions.
 - Training parser network parameters, θ with REINFORCE
 - The state \mathbf{h} is the top two elements of the stack and the top element of the buffer
 - Learn
 - Max
- NOTE: Only a single reward at the end of parsing.
- ke
sentence classification.

Through the looking glass of REINFORCE

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})} [L(\mathbf{z})]$$

Through the looking glass of REINFORCE

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)} [L(\mathbf{z})] = \nabla_{\boldsymbol{\theta}} \left[\sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \right]$$

(Need to turn it into $\mathbb{E}[\cdot]$ so we can MC-estimate)

Through the looking glass of REINFORCE

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)} [L(\mathbf{z})] = \nabla_{\boldsymbol{\theta}} \left[\sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \right]$$

(Need to turn it into $\mathbb{E}[\cdot]$ so we can MC-estimate)

$$= \sum_{\mathbf{z}} L(\mathbf{z}) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)$$

Through the looking glass of REINFORCE

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)} [L(\mathbf{z})] = \nabla_{\boldsymbol{\theta}} \left[\sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \right]$$

(Need to turn it into $\mathbb{E}[\cdot]$ so we can MC-estimate)

$$= \sum_{\mathbf{z}} L(\mathbf{z}) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)$$

$$\nabla \log f = \frac{\nabla f}{f}, \text{ so } \nabla f = f \nabla \log f.$$

Through the looking glass of REINFORCE

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)} [L(\mathbf{z})] = \nabla_{\boldsymbol{\theta}} \left[\sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \right]$$

(Need to turn it into $\mathbb{E}[\cdot]$ so we can MC-estimate)

$$= \sum_{\mathbf{z}} L(\mathbf{z}) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)$$

$$= \sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)$$

Through the looking glass of REINFORCE

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)} [L(\mathbf{z})] = \nabla_{\boldsymbol{\theta}} \left[\sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \right]$$

(Need to turn it into $\mathbb{E}[\cdot]$ so we can MC-estimate)

$$\begin{aligned} &= \sum_{\mathbf{z}} L(\mathbf{z}) \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \\ &= \sum_{\mathbf{z}} L(\mathbf{z}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{z} | x) \\ &= \mathbb{E}_{\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)} [L(\mathbf{z}) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)] \end{aligned}$$

SPINN with REINFORCE, aka RL-SPINN

Yogatama et al. [2017] uses REINFORCE to train SPINN!

SPINN with REINFORCE, aka RL-SPINN

Yogatama et al. [2017] uses REINFORCE to train SPINN!

However, this vanilla implementation isn't very effective at learning syntax.

SPINN with REINFORCE, aka RL-SPINN

Yogatama et al. [2017] uses REINFORCE to train SPINN!

However, this vanilla implementation isn't very effective at learning syntax.

This model fails to solve a simple toy problem.

Toy problem: ListOps



Toy problem: ListOps

Model	Accuracy
LSTM	74.4
RL-SPINN	64.8
TreeLSTM with ground-truth trees	98.7

Toy problem: ListOps

Model	Accuracy
LSTM	4.4
RL-SPIN	4.8
TreeLST	8.7

But why?

RL-SPINN's Troubles

This system faces at least two big problems,

RL-SPINN's Troubles

This system faces at least two big problems,

1. High variance of gradients
2. Coadaptation

High variance

- We have a single reward at the end of parsing.

High variance

- We have a single reward at the end of parsing.
- We are sampling parses from very large search space!
Catalan number of binary trees.

High variance

- We have a single reward at the end of parsing.
- We are sampling parses from very large search space!

Catalan number of binary trees.

3 tokens \Rightarrow 5 trees

5 tokens \Rightarrow 42 trees

10 tokens \Rightarrow 16796 trees

High variance

- We have a single reward at the end of parsing.
- We are sampling parses from very large search space!
Catalan number of binary trees.
- And the policy is stochastic.

High variance

So, sometimes the policy lands in a “rewarding state”:

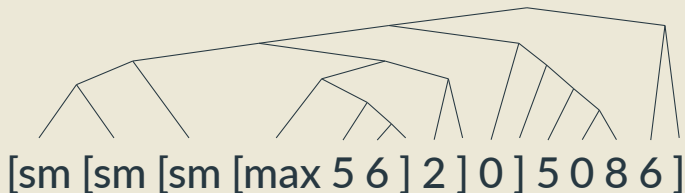


Figure: Truth: 7; Pred: 7

High variance

Sometimes it doesn't:



Figure: Truth: 6; Pred: 5

High variance

Catalan number of parses means we need many many samples to lower variance!

High variance

Catalan number of parses means we need many many samples to lower variance!

Possible solutions:

1. Gradient normalization
2. Control variates, aka baselines

Control variates

- A simple control variate: moving average of recent rewards

Control variates

- A simple control variate: moving average of recent rewards
- Parameters are updated using the advantage which is the difference between the reward, \mathcal{R} , and the baseline prediction.

Control variates

- A simple control variate: moving average of recent rewards
- Parameters are updated using the advantage which is the difference between the reward, \mathcal{R} , and the baseline prediction.

So,

$$\nabla \mathbb{E}_{\mathbf{z} \sim \pi(\mathbf{z})} = \mathbb{E}_{\mathbf{z} \sim \pi(\mathbf{z})} [(L(\mathbf{z}) - b(\mathbf{x})) \nabla \pi(\mathbf{z})]$$

Issues with SPINN with REINFORCE

This system faces two big problems,

1. High variance of gradients
2. Coadaptation

Coadaptation problem

Learning composition function parameters ϕ with backpropagation,
and parser parameters θ with REINFORCE.

Coadaptation problem

Learning composition function parameters ϕ with backpropagation, and parser parameters θ with REINFORCE.

Generally, ϕ will be learned more quickly than θ , making it harder to explore the parsing search space and optimize for θ .

Coadaptation problem

Learning composition function parameters ϕ with backpropagation, and parser parameters θ with REINFORCE.

Generally, ϕ will be learned more quickly than θ , making it harder to explore the parsing search space and optimize for θ .

Difference in variance of two gradient estimates.

Coadaptation problem

Learning composition function parameters ϕ with backpropagation, and parser parameters θ with REINFORCE.

Generally, ϕ will be learned more quickly than θ , making it harder to explore the parsing search space and optimize for θ .

Difference in variance of two gradient estimates.

Possible solution: Proximal Policy Optimization [Schulman et al., 2017].

Making REINFORCE+SPINN work

Havrylov et al. [2019] use,

1. Input dependent control variate
2. Gradient normalization
3. Proximal Policy Optimization

Making REINFORCE+SPINN work

Havrylov et al. [2019] use,

1. Input dependent control variate
2. Gradient normalization
3. Proximal Policy Optimization

They solve ListOps!

Making REINFORCE+SPINN work

Havrylov et al. [2019] use,

1. Input dependent control variate
2. Gradient normalization
3. Proximal Policy Optimization

They solve ListOps!

However, does not learn English grammars.

Should I? Shouldn't I?

- Unbiased!

Should I? Shouldn't I?

- Unbiased!
- High variance 🙄

Should I? Shouldn't I?

- Unbiased!
- In a simple setting, with enough tricks, it can work! 😊
- High variance 😞

Should I? Shouldn't I?

- Unbiased!
- In a simple setting, with enough tricks, it can work! 😊
- High variance 😞
- Has not yet been very effective at learning English syntax.

Roadmap of the tutorial

- Part 1: Introduction ✓
- Part 2: Reinforcement learning ✓

Coffee Break

- Part 3: Gradient surrogates
- Part 4: End-to-end differentiable models (1/2)

Coffee Break

- Part 4: End-to-end differentiable models (2/2)
- Part 5: Conclusions

III. Gradient Surrogates

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.
- Struggled with variance & sampling.

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.
- Struggled with variance & sampling.

In this section:

- Consider the **deterministic alternative**:

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.
- Struggled with variance & sampling.

In this section:

- Consider the **deterministic alternative**:

pick highest-score structure $\hat{\mathbf{z}}(x) := \arg \max_{\mathbf{z} \in \mathcal{M}} \pi_{\theta}(\mathbf{z} | x)$

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.
- Struggled with variance & sampling.

In this section:

- Consider the **deterministic alternative**:

pick highest-score structure
incur loss

$$\hat{\mathbf{z}}(x) := \arg \max_{\mathbf{z} \in \mathcal{M}} \pi_{\theta}(\mathbf{z} | x) \\ L(\hat{\mathbf{z}}(x))$$

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.
- Struggled with variance & sampling.

In this section:

- Consider the **deterministic alternative**:

pick highest-score structure

incur loss

$$\hat{\mathbf{z}}(x) := \arg \max_{\mathbf{z} \in \mathcal{M}} \pi_{\theta}(\mathbf{z} | x) \\ L(\hat{\mathbf{z}}(x))$$

- 3A: try to optimize the deterministic loss directly

So far:

- Tackled **expected loss** in a **stochastic computation graph**

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- Optimized with the **REINFORCE** estimator.
- Struggled with variance & sampling.

In this section:

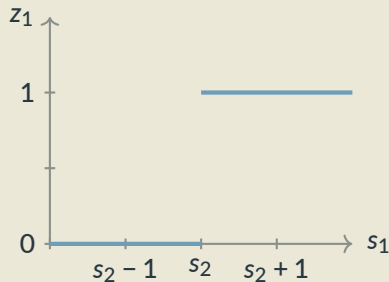
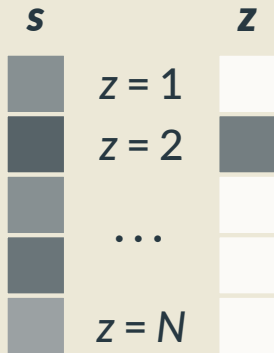
- Consider the **deterministic alternative**:

pick highest-score structure $\hat{\mathbf{z}}(x) := \arg \max_{\mathbf{z} \in \mathcal{M}} \pi_{\theta}(\mathbf{z} | x)$
incur loss $L(\hat{\mathbf{z}}(x))$

- 3A: try to optimize the deterministic loss directly
- 3B: use this strategy to reduce variance in the stochastic model.

Recap: The argmax problem

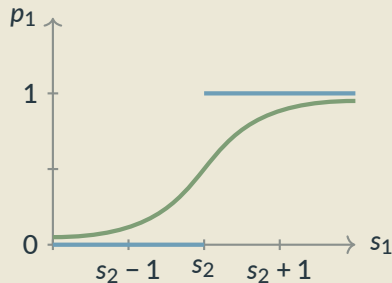
$$\mathbf{z} = \arg \max(\mathbf{s})$$



$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = \mathbf{0}$$

Softmax

$$p_j = \exp(s_j)/Z$$



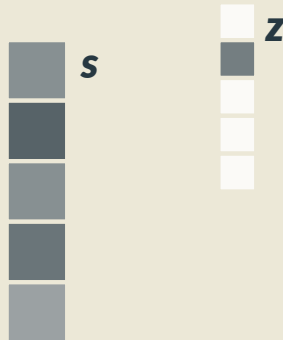
$$\frac{\partial \mathbf{p}}{\partial \mathbf{s}} = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$$

Straight-Through Estimator



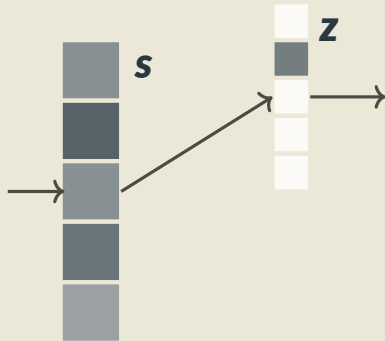
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$



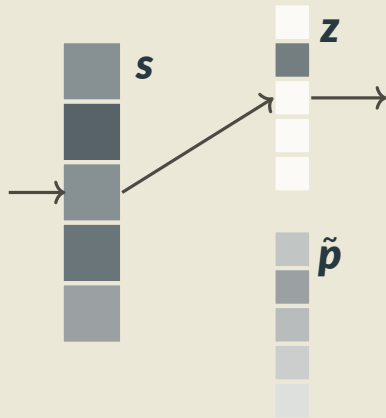
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$



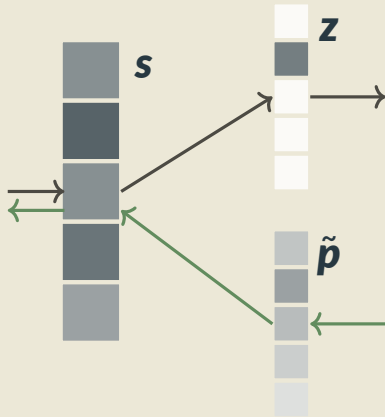
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$
- Backward: pretend \mathbf{z} was some continuous $\tilde{\mathbf{p}}$; $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} \neq \mathbf{0}$



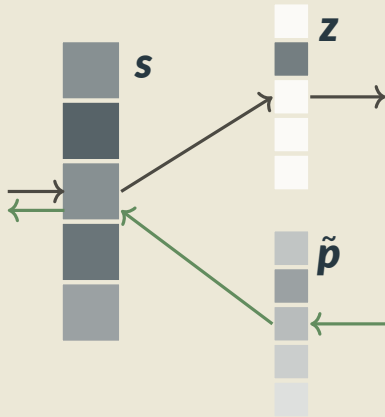
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$
- Backward: pretend \mathbf{z} was some continuous $\tilde{\mathbf{p}}$; $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} \neq \mathbf{0}$



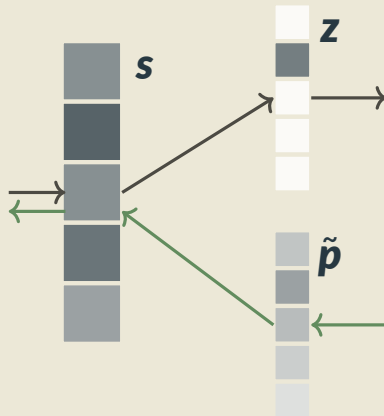
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$
- Backward: pretend \mathbf{z} was some continuous $\tilde{\mathbf{p}}$; $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} \neq \mathbf{0}$
 - simplest: identity, $\tilde{\mathbf{p}}(\mathbf{s}) = \mathbf{s}$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \mathbf{I}$



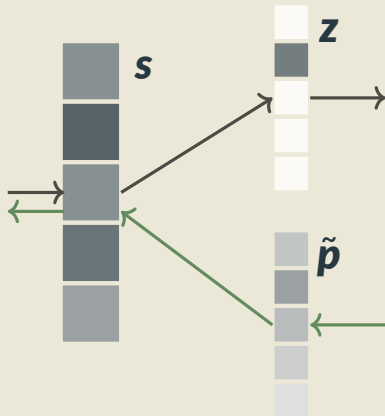
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$
- Backward: pretend \mathbf{z} was some continuous $\tilde{\mathbf{p}}$; $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} \neq \mathbf{0}$
 - simplest: identity, $\tilde{\mathbf{p}}(\mathbf{s}) = \mathbf{s}$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \mathbf{I}$
 - others, e.g. softmax $\tilde{\mathbf{p}}(\mathbf{s}) = \text{softmax}(\mathbf{s})$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \text{diag}(\tilde{\mathbf{p}}) - \tilde{\mathbf{p}}\tilde{\mathbf{p}}^T$



Straight-Through Estimator

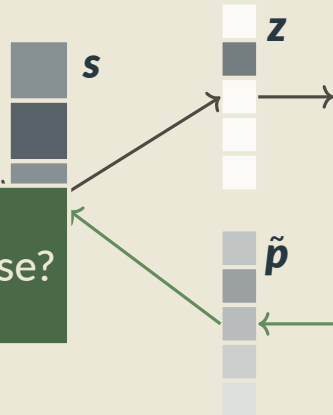
- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$
- Backward: pretend \mathbf{z} was some continuous $\tilde{\mathbf{p}}$; $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} \neq \mathbf{0}$
 - simplest: identity, $\tilde{\mathbf{p}}(\mathbf{s}) = \mathbf{s}$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \mathbf{I}$
 - others, e.g. softmax $\tilde{\mathbf{p}}(\mathbf{s}) = \text{softmax}(\mathbf{s})$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \text{diag}(\tilde{\mathbf{p}}) - \tilde{\mathbf{p}}\tilde{\mathbf{p}}^T$
- More explanation in a while



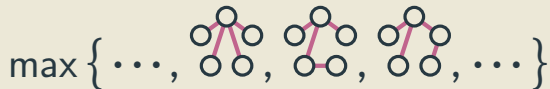
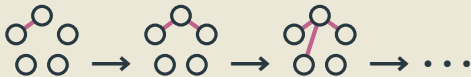
Straight-Through Estimator

- Forward: $\mathbf{z} = \arg \max(\mathbf{s})$
- Backward: pretend \mathbf{z} was some continuous $\tilde{\mathbf{p}}$; $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} \neq \mathbf{0}$
 - simplest: identity, $\tilde{\mathbf{p}}(\mathbf{s}) = \mathbf{s}$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \mathbf{I}$
 - others, e.g. softmax $\tilde{\mathbf{p}}(\mathbf{s}) = \text{softmax}(\mathbf{s})$, $\frac{\partial \tilde{\mathbf{p}}}{\partial \mathbf{s}} = \text{diag}(\tilde{\mathbf{p}}) - \tilde{\mathbf{p}}\tilde{\mathbf{p}}^T$
- More explanation

What about the structured case?



Dealing with the combinatorial explosion



1. Incremental structures

- Build structure **greedily**, as sequence of discrete choices (e.g., shift-reduce).
- Scores (partial structure, action) tuples.
- **Advantages:** flexible, rich histories.
- **Disadvantages:** greedy, local decisions are suboptimal, error propagation.

2. Factorization into parts

- Optimizes **globally** (e.g. Viterbi, Chu-Liu-Edmonds, Kuhn-Munkres).
- Scores smaller parts.
- **Advantages:** optimal, elegant, can handle hard & global constraints.
- **Disadvantages:** strong assumptions.

STE for incremental structures

STE for incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)

STE for incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.

STE for incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- In this case, we just apply the straight-through estimator for each step.

STE for incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- In this case, we just apply the straight-through estimator for each step.
- Forward: the **highest scoring action** for each step

STE for incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- In this case, we just apply the straight-through estimator for each step.
- Forward: the **highest scoring action** for each step
- Backward: pretend that we had used a **differentiable surrogate function**

STE for incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- In this case, we just apply the straight-through estimator for each step.
- Forward: the **highest scoring action** for each step
- Backward: pretend that we had used a **differentiable surrogate function**

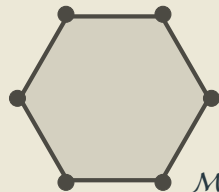
Example: Latent Tree Learning with Differentiable Parsers: Shift-Reduce Parsing and Chart Parsing [Maillard and Clark, 2018] (STE through beam search).

STE for the factorized approach

Requires a bit more work:

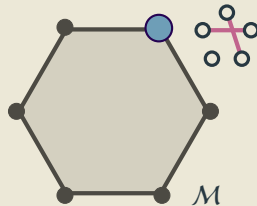
- Recap: marginal polytope
- Predicting structures globally: Maximum A Posteriori (MAP)
- Deriving Straight-Through and SPIGOT

The structured case: Marginal polytope



The structured case: Marginal polytope

- Each vertex corresponds to one such *bit* vector \mathbf{z}



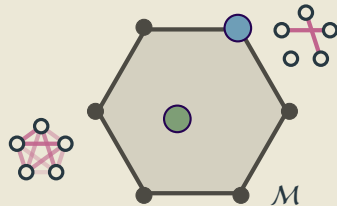
The structured case: Marginal polytope

- Each vertex corresponds to one such *bit vector* \mathbf{z}
- Points inside correspond to *marginal distributions*: convex combinations of structured objects

$$\boldsymbol{\mu} = \underbrace{p_1 \mathbf{z}_1 + \dots + p_N \mathbf{z}_N}_{\text{exponentially many terms}}, \quad \mathbf{p} \in \Delta.$$

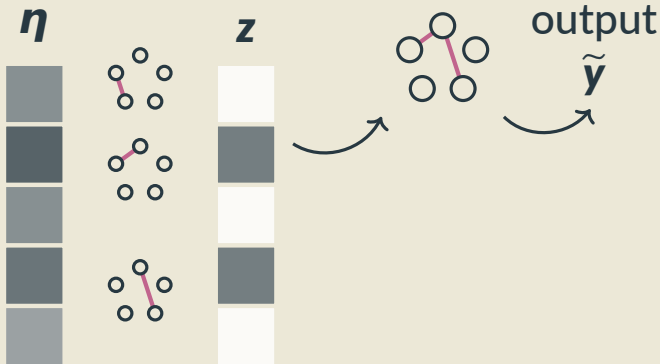
$$\begin{aligned} p_1 &= 0.2, & \mathbf{z}_1 &= [1, 0, 0, 0, 1, 0, 0, 0, 1] \\ p_2 &= 0.7, & \mathbf{z}_2 &= [0, 0, 1, 0, 0, 1, 1, 0, 0] \\ p_3 &= 0.1, & \mathbf{z}_3 &= [1, 0, 0, 0, 1, 0, 0, 1, 0] \end{aligned}$$

$$\Rightarrow \boldsymbol{\mu} = [.3, 0, .7, 0, .3, .7, .7, .1, .2].$$



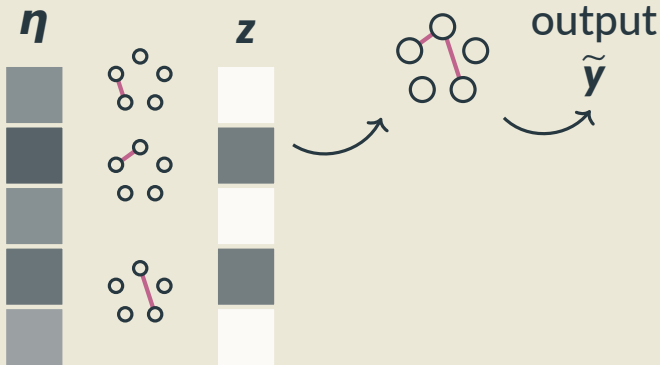
Predicting structures from scores of parts

- $\eta(i \rightarrow j)$: score of arc $i \rightarrow j$
- $z(i \rightarrow j)$: is arc $i \rightarrow j$ selected?



Predicting structures from scores of parts

- $\eta(i \rightarrow j)$: score of arc $i \rightarrow j$
- $z(i \rightarrow j)$: is arc $i \rightarrow j$ selected?
- Task-specific algorithm for the highest-scoring structure.



Algorithms for specific structures

Best structure (MAP)

Sequence tagging

Viterbi
[Rabiner, 1989]

Constituent trees

CKY
[Kasami, 1966, Younger, 1967]
[Cocke and Schwartz, 1970]

Temporal alignments

DTW
[Sakoe and Chiba, 1978]

Dependency trees

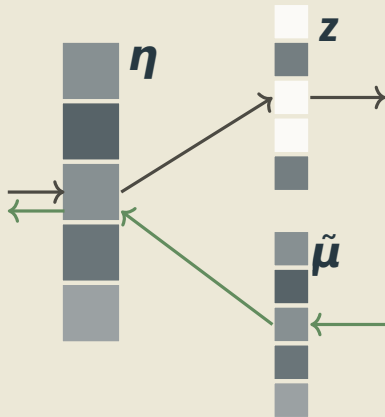
Max. Spanning Arborescence
[Chu and Liu, 1965, Edmonds, 1967]

Assignments

Kuhn-Munkres
[Kuhn, 1955, Jonker and Volgenant, 1987]

Structured Straight-Through

- Forward pass:
Find highest-scoring structure:
 $\mathbf{z} = \arg \max_{\mathbf{z} \in \mathcal{Z}} \boldsymbol{\eta}^\top \mathbf{z}$
- Backward pass:
pretend we used $\tilde{\boldsymbol{\mu}} = \boldsymbol{\eta}$.



Straight-Through Estimator

Revisited

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.
- if we had labels (multi-task learning), $L_{\text{MTL}} = L(\hat{y}(\mathbf{z}), y) + L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}})$

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.
- if we had labels (multi-task learning), $L_{\text{MTL}} = L(\hat{y}(\mathbf{z}), y) + L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}})$
- One choice: perceptron loss $L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}}) = \mathbf{s}^T \mathbf{z} - \mathbf{s}^T \mathbf{z}^{\text{true}}$; $\frac{\partial L_{\text{hid}}}{\partial \mathbf{s}} = \mathbf{z} - \mathbf{z}^{\text{true}}$.

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.
- if we had labels (multi-task learning), $L_{\text{MTL}} = L(\hat{y}(\mathbf{z}), y) + L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}})$
- One choice: perceptron loss $L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}}) = \mathbf{s}^T \mathbf{z} - \mathbf{s}^T \mathbf{z}^{\text{true}}$; $\frac{\partial L_{\text{hid}}}{\partial \mathbf{s}} = \mathbf{z} - \mathbf{z}^{\text{true}}$.
- We don't have labels! Induce labels by “pulling back” the downstream target:
the “best” (unconstrained) latent value would be: $\arg \min_{\tilde{\mathbf{z}} \in \mathbb{R}^D} L(\hat{y}(\tilde{\mathbf{z}}), y)$

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.
- if we had labels (multi-task learning), $L_{\text{MTL}} = L(\hat{\mathbf{y}}(\mathbf{z}), y) + L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}})$
- One choice: perceptron loss $L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}}) = \mathbf{s}^T \mathbf{z} - \mathbf{s}^T \mathbf{z}^{\text{true}}$; $\frac{\partial L_{\text{hid}}}{\partial \mathbf{s}} = \mathbf{z} - \mathbf{z}^{\text{true}}$.
- We don't have labels! Induce labels by “pulling back” the downstream target:
the “best” (unconstrained) latent value would be: $\arg \min_{\tilde{\mathbf{z}} \in \mathbb{R}^D} L(\hat{\mathbf{y}}(\tilde{\mathbf{z}}), y)$
- One gradient descent step starting from \mathbf{z} : $\mathbf{z}^{\text{true}} \leftarrow \mathbf{z} - \frac{\partial L}{\partial \mathbf{z}}$

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.
- if we had labels (multi-task learning), $L_{\text{MTL}} = L(\hat{y}(\mathbf{z}), y) + L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}})$
- One choice: perceptron loss $L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}}) = \mathbf{s}^T \mathbf{z} - \mathbf{s}^T \mathbf{z}^{\text{true}}$; $\frac{\partial L_{\text{hid}}}{\partial \mathbf{s}} = \mathbf{z} - \mathbf{z}^{\text{true}}$.
- We don't have labels! Induce labels by “pulling back” the downstream target:
the “best” (unconstrained) latent value would be: $\arg \min_{\tilde{\mathbf{z}} \in \mathbb{R}^D} L(\hat{y}(\tilde{\mathbf{z}}), y)$
- One gradient descent step starting from \mathbf{z} : $\mathbf{z}^{\text{true}} \leftarrow \mathbf{z} - \frac{\partial L}{\partial \mathbf{z}}$

$$\frac{\partial L_{\text{MTL}}}{\partial \mathbf{s}} = \underbrace{\frac{\partial L}{\partial \mathbf{s}}}_{=0} + \frac{\partial L_{\text{hid}}}{\partial \mathbf{s}}$$

Straight-Through Estimator

Revisited

- In the forward pass, $\mathbf{z} = \arg \max(\mathbf{s})$.
- if we had labels (multi-task learning), $L_{\text{MTL}} = L(\hat{y}(\mathbf{z}), y) + L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}})$
- One choice: perceptron loss $L_{\text{hid}}(\mathbf{s}, \mathbf{z}^{\text{true}}) = \mathbf{s}^T \mathbf{z} - \mathbf{s}^T \mathbf{z}^{\text{true}}$; $\frac{\partial L_{\text{hid}}}{\partial \mathbf{s}} = \mathbf{z} - \mathbf{z}^{\text{true}}$.
- We don't have labels! Induce labels by “pulling back” the downstream target:
the “best” (unconstrained) latent value would be: $\arg \min_{\tilde{\mathbf{z}} \in \mathbb{R}^D} L(\hat{y}(\tilde{\mathbf{z}}), y)$
- One gradient descent step starting from \mathbf{z} : $\mathbf{z}^{\text{true}} \leftarrow \mathbf{z} - \frac{\partial L}{\partial \mathbf{z}}$

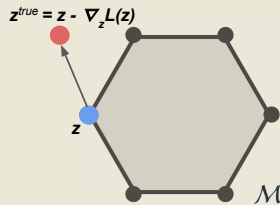
$$\frac{\partial L_{\text{MTL}}}{\partial \mathbf{s}} = \underbrace{\frac{\partial L}{\partial \mathbf{s}}}_{=0} + \frac{\partial L_{\text{hid}}}{\partial \mathbf{s}} = \mathbf{z} - \left(\mathbf{z} - \frac{\partial L}{\partial \mathbf{z}} \right) = \frac{\partial L}{\partial \mathbf{z}}$$

Straight-Through in the structured case

- Structured STE: perceptron update with induced annotation

$$\arg \min_{\boldsymbol{\mu} \in \mathbb{R}^D} L(\hat{y}(\boldsymbol{\mu}), y) \approx \mathbf{z} - \nabla_{\mathbf{z}} L(\mathbf{z}) \rightarrow \mathbf{z}^{\text{true}}$$

(one step of gradient descent)



Straight-Through in the structured case

- Structured STE: perceptron update with induced annotation

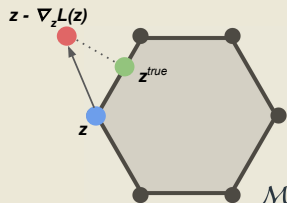
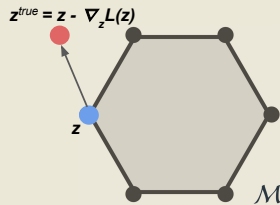
$$\arg \min_{\boldsymbol{\mu} \in \mathbb{R}^D} L(\hat{y}(\boldsymbol{\mu}), y) \approx \mathbf{z} - \nabla_{\mathbf{z}} L(\mathbf{z}) \rightarrow \mathbf{z}^{\text{true}}$$

(one step of gradient descent)

- SPIGOT takes into account the constraints; uses the induced annotation

$$\arg \min_{\boldsymbol{\mu} \in \mathcal{M}} L(\hat{y}(\boldsymbol{\mu}), y) \approx \text{Proj}_{\mathcal{M}}(\mathbf{z} - \nabla_{\mathbf{z}} L(\mathbf{z})) \rightarrow \mathbf{z}^{\text{true}}$$

(one step of *projected* gradient descent!)



Straight-Through in the structured case

- Structured STE: perceptron update with induced annotation

$$\arg \min_{\boldsymbol{\mu} \in \mathbb{R}^D} L(\hat{y}(\boldsymbol{\mu}), y) \approx \mathbf{z} - \nabla_{\mathbf{z}} L(\mathbf{z}) \rightarrow \mathbf{z}^{\text{true}}$$

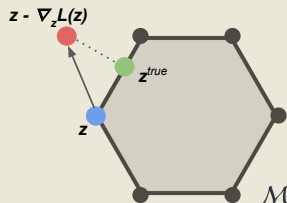
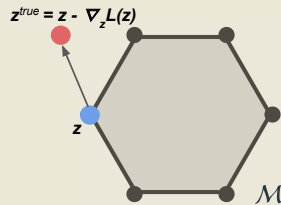
(one step of gradient descent)

- SPIGOT takes into account the constraints; uses the induced annotation

$$\arg \min_{\boldsymbol{\mu} \in \mathcal{M}} L(\hat{y}(\boldsymbol{\mu}), y) \approx \text{Proj}_{\mathcal{M}}(\mathbf{z} - \nabla_{\mathbf{z}} L(\mathbf{z})) \rightarrow \mathbf{z}^{\text{true}}$$

(one step of *projected* gradient descent!)

- We discuss a generic way to compute the projection in part 4.



Summary: Straight-Through Estimator

Summary: Straight-Through Estimator

We saw how to use the *Straight-Through Estimator* to allow learning models with *argmax* in the middle of the computation graph.

Summary: Straight-Through Estimator

We saw how to use the *Straight-Through Estimator* to allow learning models with *argmax* in the middle of the computation graph.

We were optimizing $L(\hat{\mathbf{z}}(x))$

Summary: Straight-Through Estimator

We saw how to use the *Straight-Through Estimator* to allow learning models with *argmax* in the middle of the computation graph.

We were optimizing $L(\hat{\mathbf{z}}(x))$

Now we will see how to apply STE for stochastic graphs, as an alternative approach of REINFORCE.

Stochastic node in the computation graph

Stochastic node in the computation graph

Recall the stochastic objective:

$$\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}[L(\mathbf{z})]$$

Stochastic node in the computation graph

Recall the stochastic objective:

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- REINFORCE (previous section).

Stochastic node in the computation graph

Recall the stochastic objective:

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- REINFORCE (previous section). High variance. 🙄

Stochastic node in the computation graph

Recall the stochastic objective:

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- REINFORCE (previous section). High variance. 😞
- An alternative is using the *reparameterization trick* [Kingma and Welling, 2014].

Categorical reparameterization

Categorical reparameterization

- Sampling from a categorical value in the middle of the computation graph.

$$\mathbf{z} \sim \pi_{\theta}(\mathbf{z} \mid \mathbf{x}) \propto \exp \mathbf{s}_{\theta}(\mathbf{z} \mid \mathbf{x})$$

s



z



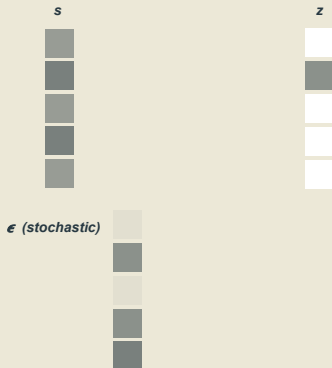
Categorical reparameterization

- Sampling from a categorical value in the middle of the computation graph.
 $\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \propto \exp \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$
- What is the gradient of a sample $\frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}}$?!



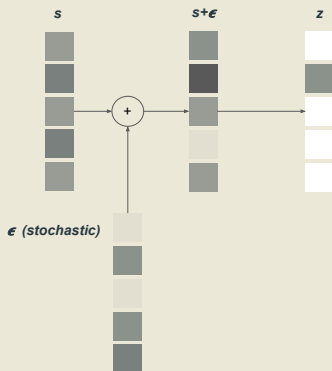
Categorical reparameterization

- Sampling from a categorical value in the middle of the computation graph.
 $\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \propto \exp \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$
- What is the gradient of a sample $\frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}}$?!
- Reparameterization: Move the stochasticity out of the gradient path.



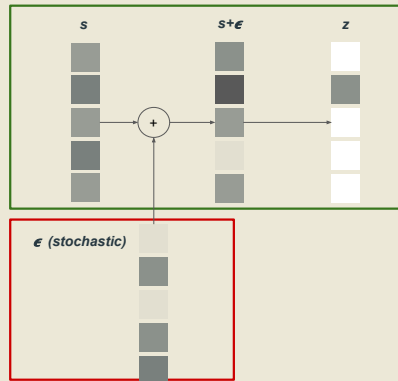
Categorical reparameterization

- Sampling from a categorical value in the middle of the computation graph.
 $\mathbf{z} \sim \pi_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x}) \propto \exp \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$
- What is the gradient of a sample $\frac{\partial \mathbf{z}}{\partial \boldsymbol{\theta}}$?!
- Reparameterization: Move the stochasticity out of the gradient path.
- Makes \mathbf{z} deterministic w.r.t. \mathbf{s} !



Categorical reparameterization

- Sampling from a categorical value in the middle of the computation graph.
 $\mathbf{z} \sim \pi_{\theta}(\mathbf{z} \mid \mathbf{x}) \propto \exp \mathbf{s}_{\theta}(\mathbf{z} \mid \mathbf{x})$
- What is the gradient of a sample $\frac{\partial \mathbf{z}}{\partial \theta}$?!
- Reparameterization: Move the stochasticity out of the gradient path.
- Makes \mathbf{z} deterministic w.r.t. \mathbf{s} !



Categorical reparameterization

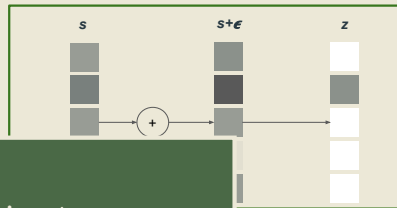
- Sampling from a categorical value in the middle of the computation graph.

$$\mathbf{z} \sim \pi_{\theta}(\mathbf{z} | \mathbf{x}) \propto \exp \mathbf{s}_{\theta}(\mathbf{z} | \mathbf{x})$$

- What is the gradient of the loss with respect to \mathbf{z} ?

- Reparameterizing stochasticity

- Makes \mathbf{z} deterministic



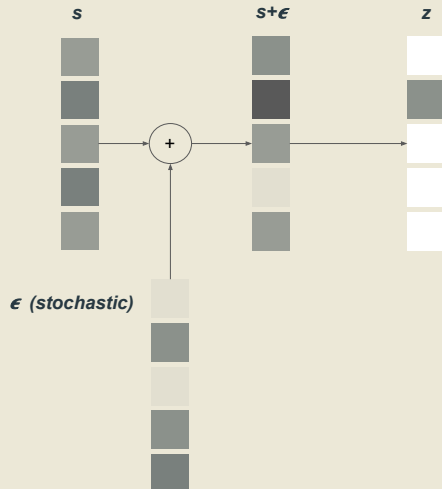
As a result:

Stochasticity is moved as an input.

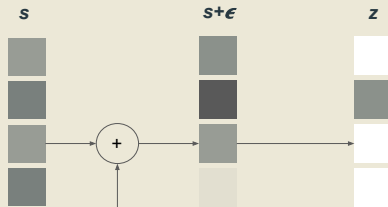
We can backpropagate through the deterministic input to \mathbf{z} .



Categorical reparameterization



Categorical reparameterization



How do we sample from a categorical variable?



Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling: 2. The Gumbel-Max trick:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

2. The Gumbel-Max trick:

- $u_i \sim \text{Uniform}(0, 1)$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

2. The Gumbel-Max trick:

- $u_i \sim \text{Uniform}(0, 1)$
- $\epsilon_i = -\log(-\log(u_i))$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

2. The Gumbel-Max trick:

- $u_i \sim \text{Uniform}(0, 1)$
- $\epsilon_i = -\log(-\log(u_i))$
- $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

2. The Gumbel-Max trick:

- $u_i \sim \text{Uniform}(0, 1)$
- $\epsilon_i = -\log(-\log(u_i))$
- $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$

The two methods are equivalent. (*Not obvious, but we will not prove it now.*)

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

2. The Gumbel-Max trick:

- $u_i \sim \text{Uniform}(0, 1)$
- $\epsilon_i = -\log(-\log(u_i))$
- $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$

The two methods are equivalent. (*Not obvious, but we will not prove it now.*)
Requires sampling from the Standard Gumbel Distribution $G(0,1)$.

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$ s.t. $c_t \leq u < c_{t+1}$

2. The Gumbel-Max trick:

- $u_i \sim \text{Uniform}(0, 1)$
- $\epsilon_i = -\log(-\log(u_i))$
- $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$

The two methods are equivalent. *(Not obvious, but we will not prove it now.)*

Requires sampling from the Standard Gumbel Distribution $G(0,1)$.

Derivation & more info: [Adams, 2013, Vieira, 2014]

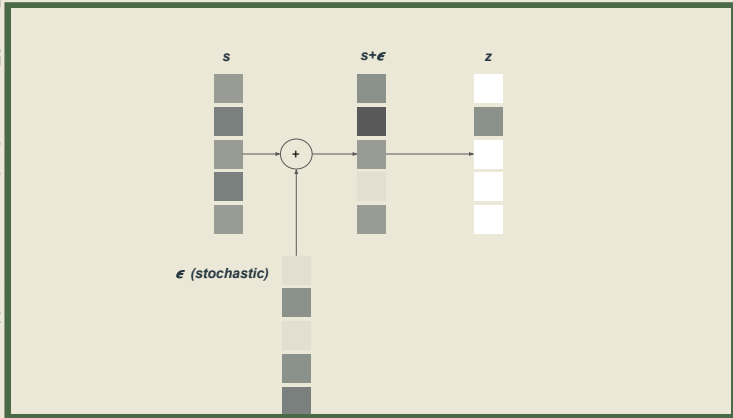
Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling • 2 The Gumbel-Max trick:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$

The t
f



now.)
).

Sampling from a categorical variable

We want to sample from a categorical variable with scores \mathbf{s} (class i has a score s_i)

1. Inverse transform sampling 2 The Gumbel-Max trick:

- $\mathbf{p} = \text{softmax}(\mathbf{s})$
- $c_i = \sum_{j \leq i} p_j$
- $u \sim \text{Uniform}(0, 1)$
- return $\mathbf{z} = \mathbf{e}_t$

The t
f



now.)
).

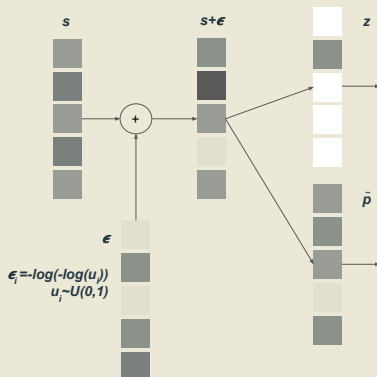
Straight-Through Gumbel Estimator

Apply a variant of the Straight-Through Estimator to Gumbel-Max!

Straight-Through Gumbel Estimator

Apply a variant of the Straight-Through Estimator to Gumbel-Max!

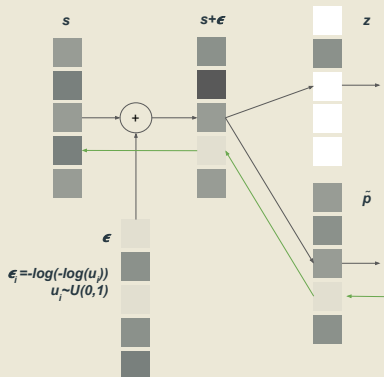
- Forward: $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$



Straight-Through Gumbel Estimator

Apply a variant of the Straight-Through Estimator to Gumbel-Max!

- Forward: $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$
- Backward: pretend we had done
 $\tilde{\mathbf{p}} = \text{softmax}(\mathbf{s} + \boldsymbol{\epsilon})$



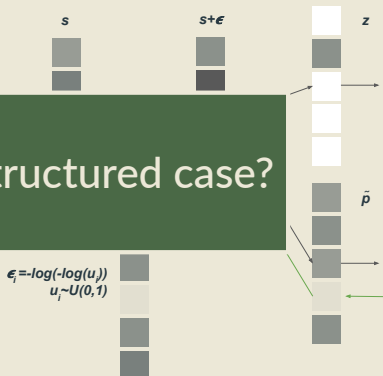
Straight-Through Gumbel Estimator

Apply a variant of the Straight-Through Estimator to Gumbel-Max!

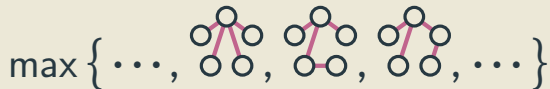
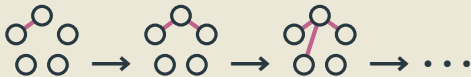
- Forward: $\mathbf{z} = \arg \max(\mathbf{s} + \boldsymbol{\epsilon})$
- Backward: pretend we had done

$$\tilde{\mathbf{p}} = \text{softmax}(\mathbf{s} + \boldsymbol{\epsilon})$$

What about the structured case?



Dealing with the combinatorial explosion



1. Incremental structures

- Build structure **greedily**, as sequence of discrete choices (e.g., shift-reduce).
- Scores (partial structure, action) tuples.
- **Advantages:** flexible, rich histories.
- **Disadvantages:** greedy, local decisions are suboptimal, error propagation.

2. Factorization into parts

- Optimizes **globally** (e.g. Viterbi, Chu-Liu-Edmonds, Kuhn-Munkres).
- Scores smaller parts.
- **Advantages:** optimal, elegant, can handle hard & global constraints.
- **Disadvantages:** strong assumptions.

Sampling from incremental structures

Sampling from incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)

Sampling from incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.

Sampling from incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- Reparameterize the scores with Gumbel-Max - now we have a deterministic node.

Sampling from incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- Reparameterize the scores with Gumbel-Max - now we have a deterministic node.
- Forward: the **argmax** from the reparameterized scores for each step

Sampling from incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- Reparameterize the scores with Gumbel-Max - now we have a deterministic node.
- Forward: the **argmax** from the reparameterized scores for each step
- Backward: pretend we had used a **differentiable surrogate function**

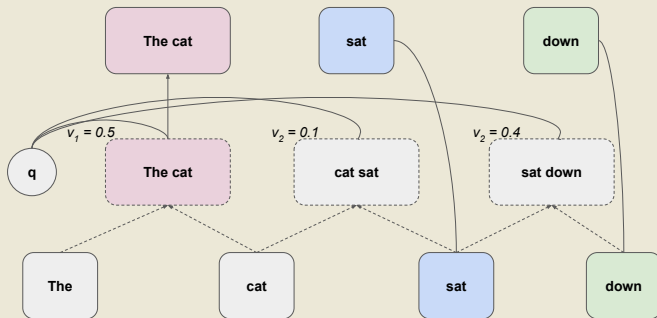
Sampling from incremental structures

- Build a structure as a sequence of discrete choices (e.g., shift-reduce)
- Assigns a score to any (partial structure, action) tuple.
- Reparameterize the scores with Gumbel-Max - now we have a deterministic node.
- Forward: the **argmax** from the reparameterized scores for each step
- Backward: pretend we had used a **differentiable surrogate function**

Example: Gumbel Tree-LSTM [Choi et al., 2018].

Example: Gumbel Tree-LSTM

- Building task-specific tree structures.
- Straight-Through Gumbel-Softmax at each step to select one arc.



Sampling from factorized models

Perturb-and-MAP

Reparameterize by **perturbing the arc scores**. (inexact!)

Sampling from factorized models

Perturb-and-MAP

Reparameterize by **perturbing the arc scores**. (inexact!)

- Sample from the standard Gumbel distribution.
- $\epsilon \sim G(0, 1)$

Sampling from factorized models

Perturb-and-MAP

Reparameterize by **perturbing the arc scores**. (inexact!)

- Sample from the standard Gumbel distribution.
- $\epsilon \sim G(0, 1)$
- Perturb the arc scores with the Gumbel noise.
- $\tilde{\eta} = \eta + \epsilon$

Sampling from factorized models

Perturb-and-MAP

Reparameterize by **perturbing the arc scores**. (inexact!)

- Sample from the standard Gumbel distribution.
- Perturb the arc scores with the Gumbel noise.
- Compute MAP (task-specific algorithm).
- $\epsilon \sim G(0, 1)$
- $\tilde{\eta} = \eta + \epsilon$
- $\arg \max_{\mathbf{z} \in \mathcal{Z}} \tilde{\eta}^T \mathbf{z}$

Sampling from factorized models

Perturb-and-MAP

Reparameterize by **perturbing the arc scores**. (inexact!)

- Sample from the standard Gumbel distribution.
- Perturb the arc scores with the Gumbel noise.
- Compute MAP (task-specific algorithm).
- Backward: we could use Straight-Through with Identity.
- $\epsilon \sim G(0, 1)$
- $\tilde{\eta} = \eta + \epsilon$
- $\arg \max_{\mathbf{z} \in \mathcal{Z}} \tilde{\eta}^T \mathbf{z}$

Summary: Gradient surrogates

- Based on the **Straight-Through Estimator**.
- Can be used for stochastic or deterministic computation graphs.
- **Forward pass**: Get an argmax (might be structured).
- **Backpropagation**: use a function, which we hope is close to argmax.
- Examples:
 - Argmax for iterative structures and factorization into parts
 - Sampling from iterative structures and factorization into parts

Gradient surrogates: Pros and cons

Pros

- Do not suffer from the high variance problem of REINFORCE.
- Allow for flexibility to select or sample a latent structured in the middle of the computation graph.
- Efficient computation.

Cons

- The Gumbel sampling with Perturb-and-MAP is an approximation.
- Bias, due to function mismatch on the backpropagation
(next section will address this problem.)

Overview

$$\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}|x)}[L(\mathbf{z})]$$

$$L(\arg \max_{\mathbf{z}} \pi_{\boldsymbol{\theta}}(\mathbf{z} | x))$$

- REINFORCE
- Straight-Through Gumbel (Perturb & MAP)
- Straight-Through
- SPIGOT

Overview

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

$$L(\arg \max_{\mathbf{z}} \pi_{\theta}(\mathbf{z} | x))$$

$$L(\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[\mathbf{z}])$$

- REINFORCE
- Straight-Through Gumbel (Perturb & MAP)
- Straight-Through
- SPIGOT
- Structured Attn. Nets
- SparseMAP

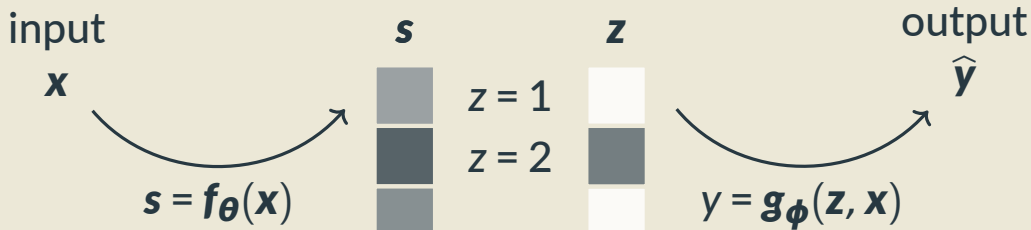
And more, in the next section!

IV. End-to-end Differentiable Relaxations

End-to-end differentiable relaxations

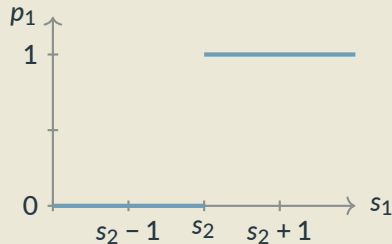
1. Digging into softmax
2. Alternatives to softmax
3. Generalizing to structured prediction
4. Stochasticity and global structures

Recall: Discrete choices & differentiability

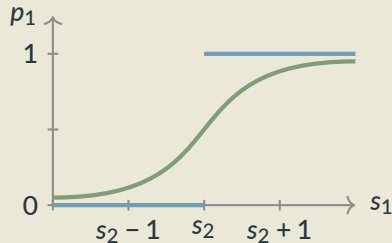
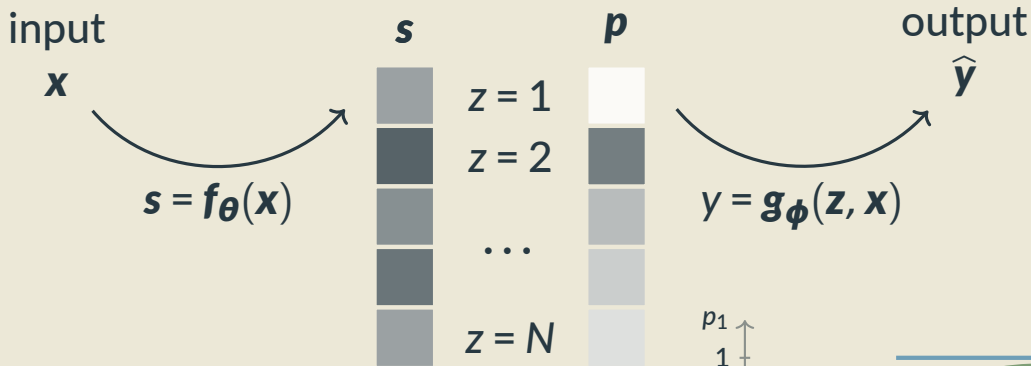


$$\frac{\partial \mathbf{z}}{\partial \mathbf{s}} = \mathbf{0} \text{ or n/a}$$

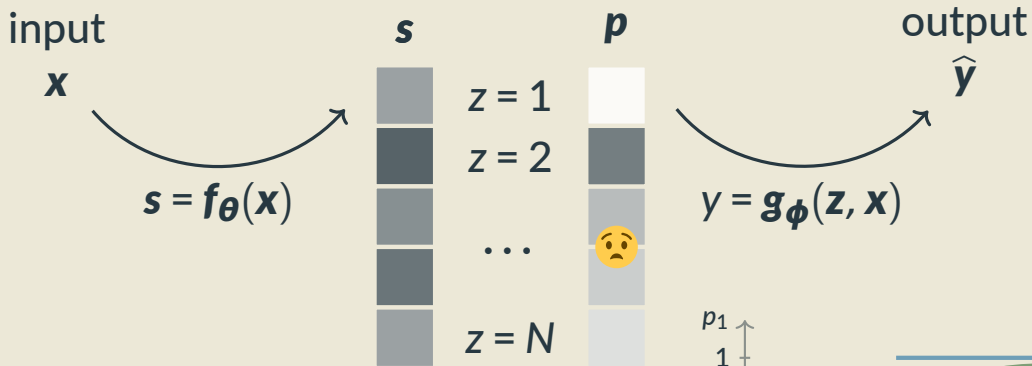
(argmax)



One solution: smooth relaxation



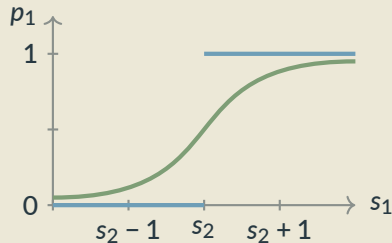
One solution: smooth relaxation



$\mathbf{p} = \text{softmax}(\mathbf{s}) = \mathbb{E}[\mathbf{z}]$, i.e.
replace $\mathbb{E}[f(\mathbf{z})]$ with $f(\mathbb{E}[\mathbf{z}])$

$$\frac{\partial \mathbf{p}}{\partial \mathbf{s}} = \text{😊}$$

(softmax)



Overview

$$\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}|x)}[L(\mathbf{z})]$$

$$L\left(\arg \max_{\mathbf{z}} \pi_{\boldsymbol{\theta}}(\mathbf{z} | x)\right)$$

$$L\left(\mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{z}|x)}[\mathbf{z}]\right)$$

- REINFORCE
- Straight-Through Gumbel (Perturb & MAP)
- Straight-Through
- SPIGOT

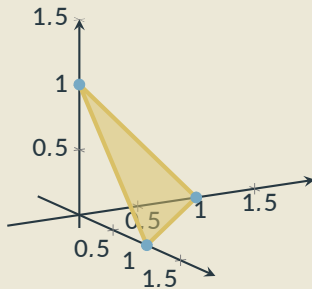
What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

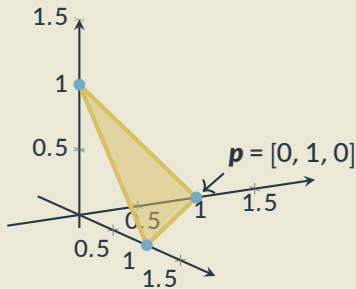
$\mathbf{p} \in \Delta$: probability distribution over choices



What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

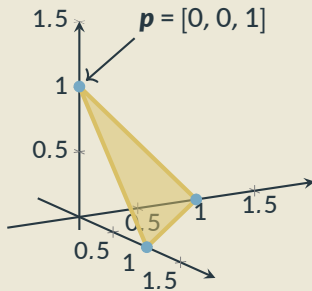
$\mathbf{p} \in \Delta$: probability distribution over choices



What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

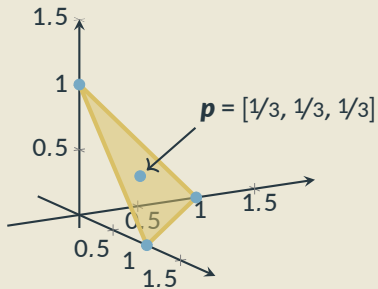
$\mathbf{p} \in \Delta$: probability distribution over choices



What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

$\mathbf{p} \in \Delta$: probability distribution over choices

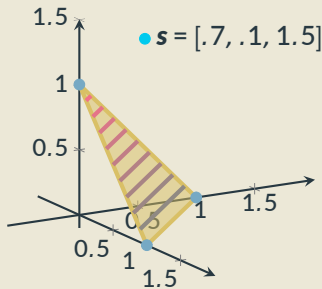


What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

$\mathbf{p} \in \Delta$: probability distribution over choices

Expected score under \mathbf{p} : $\mathbb{E}_{i \sim \mathbf{p}} s_i = \mathbf{p}^\top \mathbf{s}$



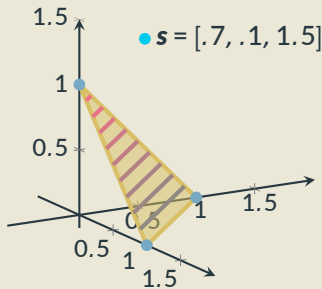
What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

$\mathbf{p} \in \Delta$: probability distribution over choices

Expected score under \mathbf{p} : $\mathbb{E}_{i \sim \mathbf{p}} s_i = \mathbf{p}^\top \mathbf{s}$

argmax



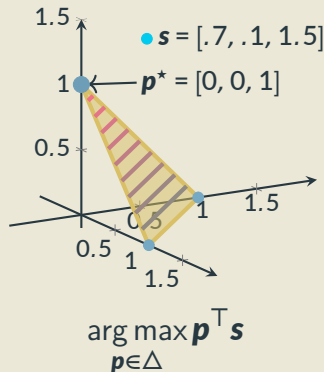
What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

$\mathbf{p} \in \Delta$: probability distribution over choices

Expected score under \mathbf{p} : $\mathbb{E}_{i \sim \mathbf{p}} s_i = \mathbf{p}^\top \mathbf{s}$

argmax maximizes **expected score**



What is softmax?

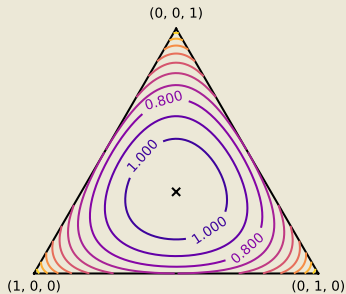
Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

$\mathbf{p} \in \Delta$: probability distribution over choices

Expected score under \mathbf{p} : $\mathbb{E}_{i \sim \mathbf{p}} s_i = \mathbf{p}^\top \mathbf{s}$

argmax maximizes **expected score**

Shannon entropy of \mathbf{p} : $H(\mathbf{p}) = -\sum_i p_i \log p_i$



What is softmax?

Often defined via $p_i = \frac{\exp s_i}{\sum_j \exp s_j}$, but where does it come from?

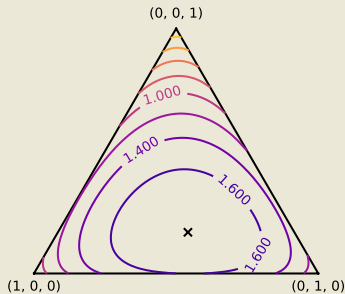
$\mathbf{p} \in \Delta$: probability distribution over choices

Expected score under \mathbf{p} : $\mathbb{E}_{i \sim \mathbf{p}} s_i = \mathbf{p}^\top \mathbf{s}$

argmax maximizes **expected score**

Shannon entropy of \mathbf{p} : $H(\mathbf{p}) = -\sum_i p_i \log p_i$

softmax maximizes **expected score + entropy**:



$$\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} &\text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ &\text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = - \sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$0 = \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau$$

$$\mathbf{p}^\top \boldsymbol{\nu} = 0$$

$$\mathbf{p} \in \Delta$$

$$\boldsymbol{\nu} \geq 0$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\log p_i = s_i + \nu_i - (\tau + 1)$$

$$\text{maximize} \quad \sum_j p_j s_j - p_j \log p_j$$

$$\text{subject to} \quad \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$0 = \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau$$

$$\mathbf{p}^\top \boldsymbol{\nu} = 0$$

$$\mathbf{p} \in \Delta$$

$$\boldsymbol{\nu} \geq 0$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

$$\begin{aligned} \log p_i &= s_i + \nu_i - (\tau + 1) \\ \text{if } p_i &= 0, \text{ r.h.s. must be } -\infty, \\ \text{thus } p_i &> 0, \text{ so } \nu_i = 0. \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$0 = \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau$$

$$\mathbf{p}^\top \boldsymbol{\nu} = 0$$

$$\mathbf{p} \in \Delta$$

$$\boldsymbol{\nu} \geq 0$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

$$\begin{aligned} \log p_i &= s_i + \nu_i - (\tau + 1) \\ \text{if } p_i &= 0, \text{ r.h.s. must be } -\infty, \\ \text{thus } p_i &> 0, \text{ so } \nu_i = 0. \end{aligned}$$

$$p_i = \exp(s_i) / \exp(\tau + 1) = \exp(s_i) / Z$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$0 = \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau$$

$$\mathbf{p}^\top \boldsymbol{\nu} = 0$$

$$\mathbf{p} \in \Delta$$

$$\boldsymbol{\nu} \geq 0$$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$\begin{aligned} 0 &= \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau \\ \mathbf{p}^\top \boldsymbol{\nu} &= 0 \\ \mathbf{p} &\in \Delta \\ \boldsymbol{\nu} &\geq 0 \end{aligned}$$

$$\begin{aligned} \log p_i &= s_i + \nu_i - (\tau + 1) \\ \text{if } p_i &= 0, \text{ r.h.s. must be } -\infty, \\ \text{thus } p_i &> 0, \text{ so } \nu_i = 0. \end{aligned}$$

$$p_i = \exp(s_i) / \exp(\tau + 1) = \exp(s_i) / Z$$

Must find Z such that $\sum_j p_j = 1$.

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$\begin{aligned} 0 &= \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau \\ \mathbf{p}^\top \boldsymbol{\nu} &= 0 \\ \mathbf{p} &\in \Delta \\ \boldsymbol{\nu} &\geq 0 \end{aligned}$$

$$\begin{aligned} \log p_i &= s_i + \nu_i - (\tau + 1) \\ \text{if } p_i &= 0, \text{ r.h.s. must be } -\infty, \\ \text{thus } p_i &> 0, \text{ so } \nu_i = 0. \end{aligned}$$

$$p_i = \exp(s_i) / \exp(\tau + 1) = \exp(s_i) / Z$$

Must find Z such that $\sum_j p_j = 1$.

Answer: $Z = \sum_j \exp(s_j)$

Variational form of softmax

Proposition. The unique solution to $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^\top \mathbf{s} + H(\mathbf{p})$ is given by $p_j = \frac{\exp s_j}{\sum_i \exp s_i}$.

Explicit form of the optimization problem:

$$\begin{aligned} & \text{maximize} && \sum_j p_j s_j - p_j \log p_j \\ & \text{subject to} && \mathbf{p} \geq 0, \mathbf{p}^\top \mathbf{1} = 1 \end{aligned}$$

Lagrangian:

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -\sum_j p_j s_j - p_j \log p_j - \mathbf{p}^\top \boldsymbol{\nu} + \tau(\mathbf{p}^\top \mathbf{1} - 1)$$

Optimality conditions (KKT):

$$0 = \nabla_{p_i} \mathcal{L}(\mathbf{p}, \boldsymbol{\nu}, \tau) = -s_i + \log p_i + 1 - \nu_i + \tau$$

$$\mathbf{p}^\top \boldsymbol{\nu} = 0$$

$$\mathbf{p} \in \Delta$$

$$\boldsymbol{\nu} \geq 0$$

$$\begin{aligned} \log p_i &= s_i + \nu_i - (\tau + 1) \\ \text{if } p_i &= 0, \text{ r.h.s. must be } -\infty, \\ \text{thus } p_i &> 0, \text{ so } \nu_i = 0. \end{aligned}$$

$$p_i = \exp(s_i) / \exp(\tau + 1) = \exp(s_i) / Z$$

Must find Z such that $\sum_j p_j = 1$.

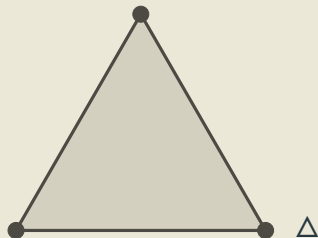
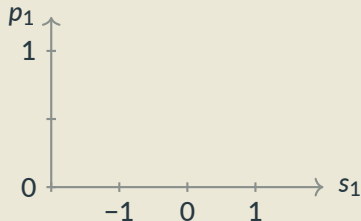
Answer: $Z = \sum_j \exp(s_j)$

$$\text{So, } p_i = \frac{\exp(s_i)}{\sum_j \exp(s_j)}.$$

Classic result, e.g., [Boyd and Vandenberghe, 2004, Wainwright and Jordan, 2008]

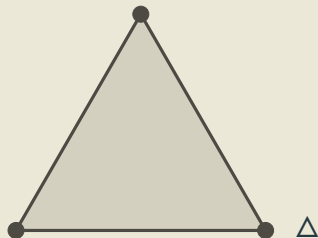
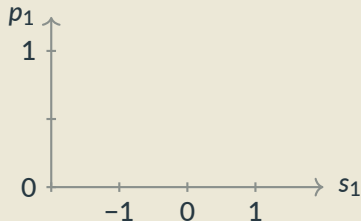
Generalizing softmax: Smoothed argmaxes

$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$



Generalizing softmax: Smoothed argmaxes

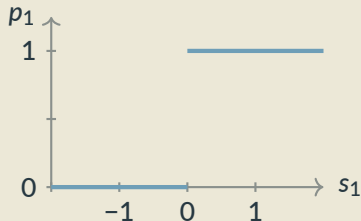
$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$



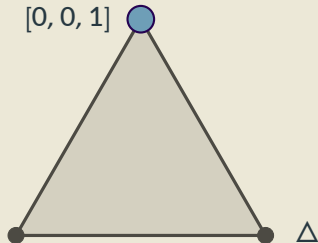
Generalizing softmax: Smoothed argmaxes

$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$

- argmax: $\Omega(\mathbf{p}) = 0$



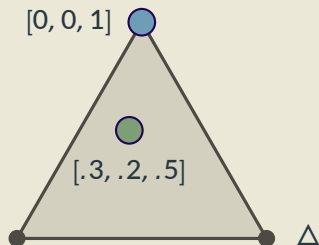
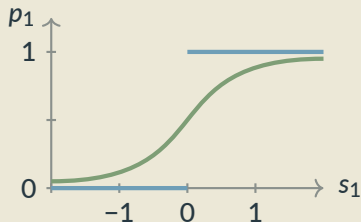
$[0, 0, 1]$



Generalizing softmax: Smoothed argmaxes

$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$

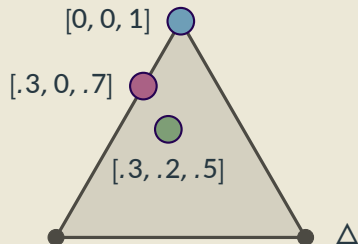
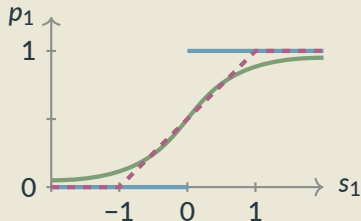
- argmax: $\Omega(\mathbf{p}) = 0$
- softmax: $\Omega(\mathbf{p}) = \sum_j p_j \log p_j$



Generalizing softmax: Smoothed argmaxes

$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$

- argmax: $\Omega(\mathbf{p}) = 0$
- softmax: $\Omega(\mathbf{p}) = \sum_j p_j \log p_j$
- sparsemax: $\Omega(\mathbf{p}) = 1/2 \|\mathbf{p}\|_2^2$

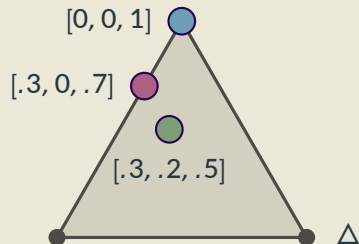
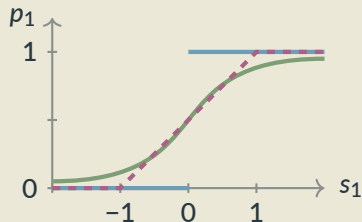


Generalizing softmax: Smoothed argmaxes

$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$

- argmax: $\Omega(\mathbf{p}) = 0$
- softmax: $\Omega(\mathbf{p}) = \sum_j p_j \log p_j$
- sparsemax: $\Omega(\mathbf{p}) = 1/2 \|\mathbf{p}\|_2^2$
- α -entmax: $\Omega(\mathbf{p}) = 1/\alpha(\alpha-1) \sum_j p_j^{\alpha}$

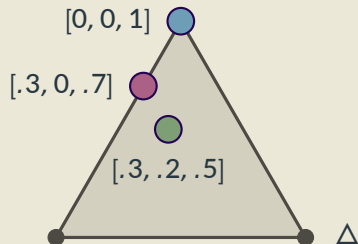
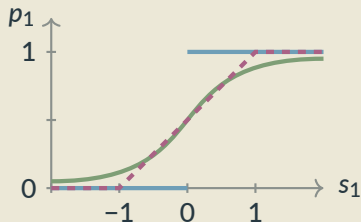
Generalized entropy interpolates in between [Tsallis, 1988]
 Used in Sparse Seq2Seq: [Peters et al., 2019] and Adaptively
 Sparse Transformers [Correia et al., 2019]



Generalizing softmax: Smoothed argmaxes

$$\hat{\mathbf{p}}_{\Omega}(\mathbf{s}) = \arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^{\top} \mathbf{s} - \Omega(\mathbf{p})$$

- argmax: $\Omega(\mathbf{p}) = 0$
- softmax: $\Omega(\mathbf{p}) = \sum_j p_j \log p_j$
- sparsemax: $\Omega(\mathbf{p}) = 1/2 \|\mathbf{p}\|_2^2$
- α -entmax: $\Omega(\mathbf{p}) = 1/\alpha(\alpha-1) \sum_j p_j^{\alpha}$
- fusedmax: $\Omega(\mathbf{p}) = 1/2 \|\mathbf{p}\|_2^2 + \sum_j |p_j - p_{j-1}|$
- csparsesmax: $\Omega(\mathbf{p}) = 1/2 \|\mathbf{p}\|_2^2 + \iota(\mathbf{a} \leq \mathbf{p} \leq \mathbf{b})$
- csoftmax: $\Omega(\mathbf{p}) = \sum_j p_j \log p_j + \iota(\mathbf{a} \leq \mathbf{p} \leq \mathbf{b})$



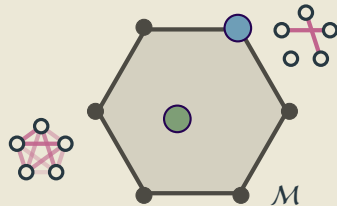
The structured case: Marginal polytope

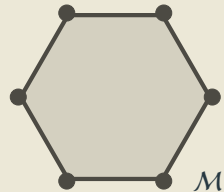
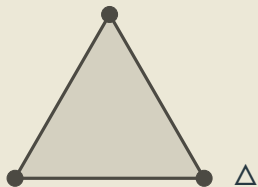
- Each vertex corresponds to one such *bit vector* \mathbf{z}
- Points inside correspond to *marginal distributions*: convex combinations of structured objects

$$\boldsymbol{\mu} = \underbrace{p_1 \mathbf{z}_1 + \dots + p_N \mathbf{z}_N}_{\text{exponentially many terms}}, \quad \mathbf{p} \in \Delta.$$

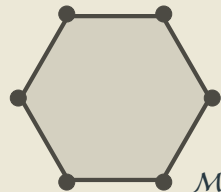
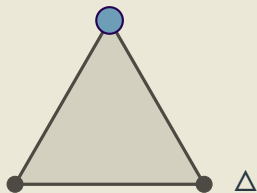
$$\begin{aligned} p_1 &= 0.2, & \mathbf{z}_1 &= [1, 0, 0, 0, 1, 0, 0, 0, 1] \\ p_2 &= 0.7, & \mathbf{z}_2 &= [0, 0, 1, 0, 0, 1, 1, 0, 0] \\ p_3 &= 0.1, & \mathbf{z}_3 &= [1, 0, 0, 0, 1, 0, 0, 1, 0] \end{aligned}$$

$$\Rightarrow \boldsymbol{\mu} = [.3, 0, .7, 0, .3, .7, .7, .1, .2].$$

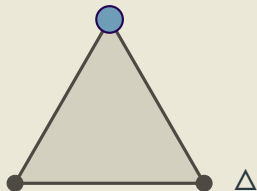




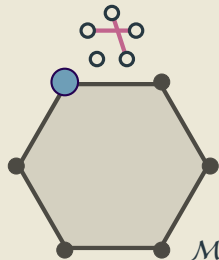
- $\operatorname{argmax}_{\mathbf{p} \in \Delta} \arg \max \mathbf{p}^T \mathbf{s}$



• $\text{argmax}_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$

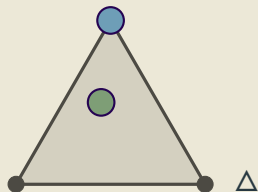


• $\text{MAP} \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$

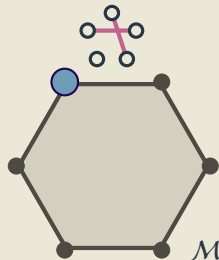


- **argmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$

- **softmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} + H(\mathbf{p})$

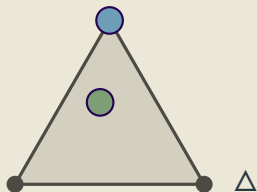


- **MAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$



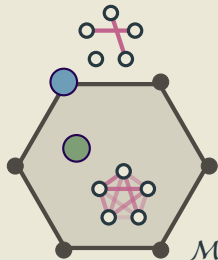
- **argmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$

- **softmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} + H(\mathbf{p})$



- **MAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$

- **marginals** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta} + \tilde{H}(\boldsymbol{\mu})$



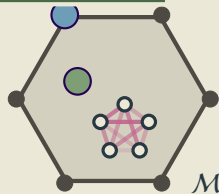
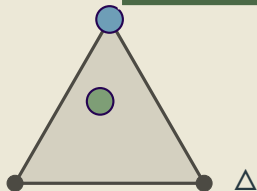
- **argmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$

- **softmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} + H(\mathbf{p})$

- **MAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$

- **marginals** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta} + \tilde{H}(\boldsymbol{\mu})$

Just like softmax relaxes argmax,
marginals relax MAP **differentiably!**



- **argmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$

- **softmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} + H(\mathbf{p})$

- **MAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$

- **marginals** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta} + \tilde{H}(\boldsymbol{\mu})$

Just like softmax relaxes argmax,
marginals relax MAP **differentiably!**

Unlike argmax/softmax, computation is not obvious!



Algorithms for specific structures

	Best structure (MAP)	Marginals
Sequence tagging	Viterbi [Rabiner, 1989]	Forward-Backward [Rabiner, 1989]
Constituent trees	CKY [Kasami, 1966, Younger, 1967] [Cocke and Schwartz, 1970]	Inside-Outside [Baker, 1979]
Temporal alignments	DTW [Sakoe and Chiba, 1978]	Soft-DTW [Cuturi and Blondel, 2017]
Dependency trees	Max. Spanning Arborescence [Chu and Liu, 1965, Edmonds, 1967]	Matrix-Tree [Kirchhoff, 1847]
Assignments	Kuhn-Munkres [Kuhn, 1955, Jonker and Volgenant, 1987]	#P-complete [Valiant, 1979, Taskar, 2004]

Algorithms for specific structures

		Best structure (MAP)	Marginals
dyn. prog.	Sequence tagging	Viterbi [Rabiner, 1989]	Forward-Backward [Rabiner, 1989]
	Constituent trees	CKY [Kasami, 1966, Younger, 1967] [Cocke and Schwartz, 1970]	Inside-Outside [Baker, 1979]
	Temporal alignments	DTW [Sakoe and Chiba, 1978]	Soft-DTW [Cuturi and Blondel, 2017]
	Dependency trees	Max. Spanning Arborescence [Chu and Liu, 1965, Edmonds, 1967]	Matrix-Tree [Kirchhoff, 1847]
	Assignments	Kuhn-Munkres [Kuhn, 1955, Jonker and Volgenant, 1987]	#P-complete [Valiant, 1979, Taskar, 2004]

Derivatives of marginals 1: DP

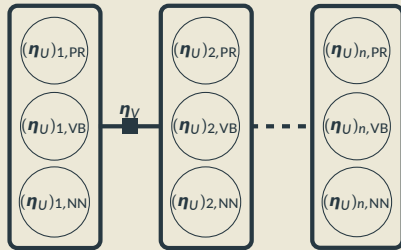
Dynamic programming: marginals by **Forward-Backward**, **Inside-Outside**, etc.

Derivatives of marginals 1: DP

Dynamic programming: marginals by **Forward-Backward, Inside-Outside**, etc.

Marginals in a sequence tagging model.

```
1 input:  $d$  tags,  $n$  tokens,  $\boldsymbol{\eta}_U \in \mathbb{R}^{n \times d}$ ,  $\boldsymbol{\eta}_V \in \mathbb{R}^{d \times d}$ 
2 initialize  $\boldsymbol{\alpha}_1 = \mathbf{0}$ ,  $\boldsymbol{\beta}_n = \mathbf{0}$ 
3 for  $i \in 2, \dots, n$  do                                # forward log-probabilities
4    $\alpha_{i,k} = \log \sum_{k'} \exp(\alpha_{i-1,k'} + (\boldsymbol{\eta}_U)_{i,k} + (\boldsymbol{\eta}_V)_{k',k})$  for all  $k$ 
5 for  $i \in n-1, \dots, 1$  do                                # backward log-probabilities
6    $\beta_{i,k} = \log \sum_{k'} \exp(\beta_{i+1,k'} + (\boldsymbol{\eta}_U)_{i+1,k'} + (\boldsymbol{\eta}_V)_{k,k'})$  for all  $k$ 
7  $Z = \sum_k \exp \alpha_{n,k}$                                     # partition function
8 return  $\boldsymbol{\mu} = \exp(\boldsymbol{\alpha} + \boldsymbol{\beta} - \log Z)$                 # marginals
```



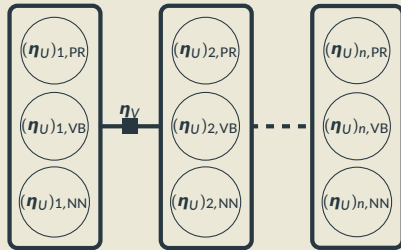
Derivatives of marginals 1: DP

Dynamic programming: marginals by **Forward-Backward, Inside-Outside**, etc.

- Alg. consists of differentiable ops: PyTorch autograd can handle it! (v. bad idea)

Marginals in a sequence tagging model.

```
1 input:  $d$  tags,  $n$  tokens,  $\boldsymbol{\eta}_U \in \mathbb{R}^{n \times d}$ ,  $\boldsymbol{\eta}_V \in \mathbb{R}^{d \times d}$ 
2 initialize  $\boldsymbol{\alpha}_1 = \mathbf{0}$ ,  $\boldsymbol{\beta}_n = \mathbf{0}$ 
3 for  $i \in 2, \dots, n$  do                                # forward log-probabilities
4    $\alpha_{i,k} = \log \sum_{k'} \exp(\alpha_{i-1,k'} + (\boldsymbol{\eta}_U)_{i,k} + (\boldsymbol{\eta}_V)_{k',k})$  for all  $k$ 
5 for  $i \in n-1, \dots, 1$  do                                # backward log-probabilities
6    $\beta_{i,k} = \log \sum_{k'} \exp(\beta_{i+1,k'} + (\boldsymbol{\eta}_U)_{i+1,k'} + (\boldsymbol{\eta}_V)_{k,k'})$  for all  $k$ 
7  $Z = \sum_k \exp \alpha_{n,k}$                                     # partition function
8 return  $\boldsymbol{\mu} = \exp(\boldsymbol{\alpha} + \boldsymbol{\beta} - \log Z)$                 # marginals
```



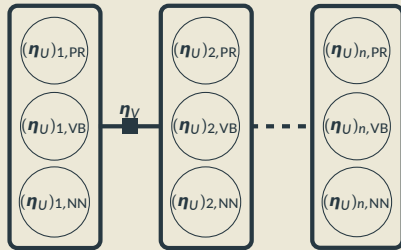
Derivatives of marginals 1: DP

Dynamic programming: marginals by **Forward-Backward, Inside-Outside**, etc.

- Alg. consists of differentiable ops: PyTorch autograd can handle it! (v. bad idea)
- Better book-keeping: Li and Eisner [2009], Mensch and Blondel [2018]

Marginals in a sequence tagging model.

```
1 input:  $d$  tags,  $n$  tokens,  $\boldsymbol{\eta}_U \in \mathbb{R}^{n \times d}$ ,  $\boldsymbol{\eta}_V \in \mathbb{R}^{d \times d}$ 
2 initialize  $\boldsymbol{\alpha}_1 = \mathbf{0}$ ,  $\boldsymbol{\beta}_n = \mathbf{0}$ 
3 for  $i \in 2, \dots, n$  do                                # forward log-probabilities
4    $\alpha_{i,k} = \log \sum_{k'} \exp(\alpha_{i-1,k'} + (\boldsymbol{\eta}_U)_{i,k} + (\boldsymbol{\eta}_V)_{k',k})$  for all  $k$ 
5 for  $i \in n-1, \dots, 1$  do                                # backward log-probabilities
6    $\beta_{i,k} = \log \sum_{k'} \exp(\beta_{i+1,k'} + (\boldsymbol{\eta}_U)_{i+1,k'} + (\boldsymbol{\eta}_V)_{k,k'})$  for all  $k$ 
7  $Z = \sum_k \exp \alpha_{n,k}$                                 # partition function
8 return  $\boldsymbol{\mu} = \exp(\boldsymbol{\alpha} + \boldsymbol{\beta} - \log Z)$                 # marginals
```



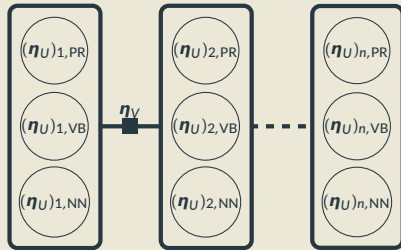
Derivatives of marginals 1: DP

Dynamic programming: marginals by **Forward-Backward, Inside-Outside**, etc.

- Alg. consists of differentiable ops: PyTorch autograd can handle it! (v. bad idea)
- Better book-keeping: Li and Eisner [2009], Mensch and Blondel [2018]
- With circular dependencies, this breaks! Can get an approximation [Stoyanov et al., 2011]

Marginals in a sequence tagging model.

```
1 input:  $d$  tags,  $n$  tokens,  $\boldsymbol{\eta}_U \in \mathbb{R}^{n \times d}$ ,  $\boldsymbol{\eta}_V \in \mathbb{R}^{d \times d}$ 
2 initialize  $\boldsymbol{\alpha}_1 = \mathbf{0}$ ,  $\boldsymbol{\beta}_n = \mathbf{0}$ 
3 for  $i \in 2, \dots, n$  do                                # forward log-probabilities
4    $\alpha_{i,k} = \log \sum_{k'} \exp(\alpha_{i-1,k'} + (\boldsymbol{\eta}_U)_{i,k} + (\boldsymbol{\eta}_V)_{k',k})$    for all  $k$ 
5 for  $i \in n-1, \dots, 1$  do                                # backward log-probabilities
6    $\beta_{i,k} = \log \sum_{k'} \exp(\beta_{i+1,k'} + (\boldsymbol{\eta}_U)_{i+1,k'} + (\boldsymbol{\eta}_V)_{k,k'})$    for all  $k$ 
7  $Z = \sum_k \exp \alpha_{n,k}$                                 # partition function
8 return  $\boldsymbol{\mu} = \exp(\boldsymbol{\alpha} + \boldsymbol{\beta} - \log Z)$                 # marginals
```



Derivatives of marginals 2: Matrix-Tree

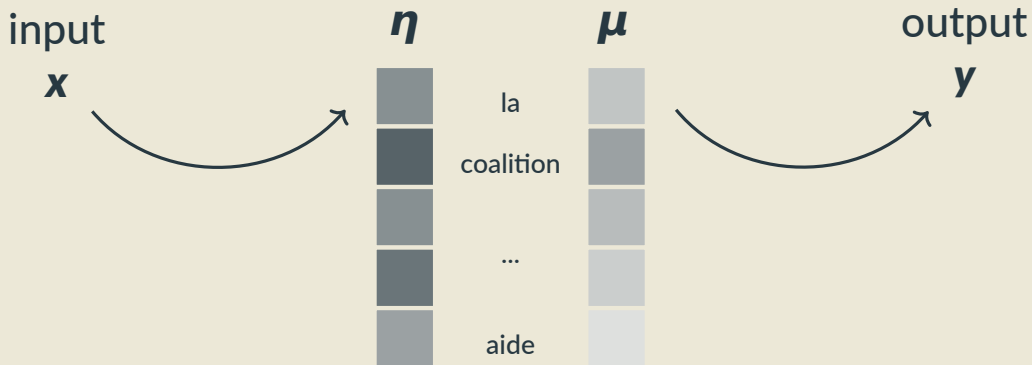
$\mathbf{L}(\mathbf{s})$: Laplacian of the edge score graph

$$Z = \det \mathbf{L}(\mathbf{s})$$

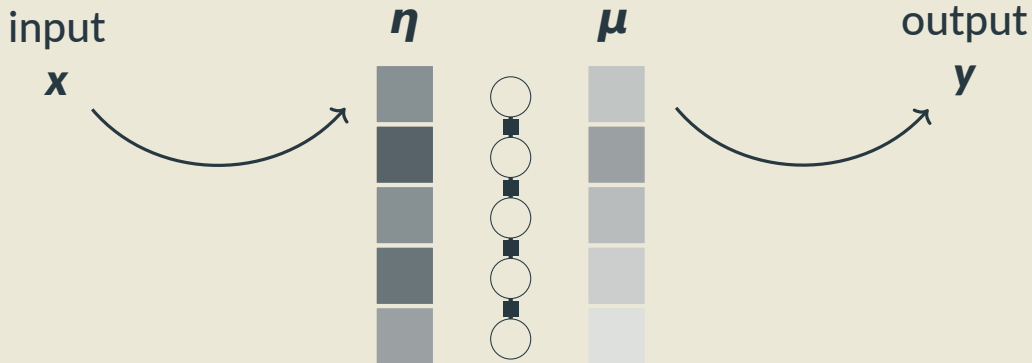
$$\boldsymbol{\mu} = \mathbf{L}(\mathbf{s})^{-1}$$

$$\nabla \boldsymbol{\mu} = \nabla \mathbf{L}^{-1} = \mathbf{L}^{-1} \left(\frac{\partial \mathbf{L}}{\partial \boldsymbol{\eta}} \right) \mathbf{L}^{-1}$$

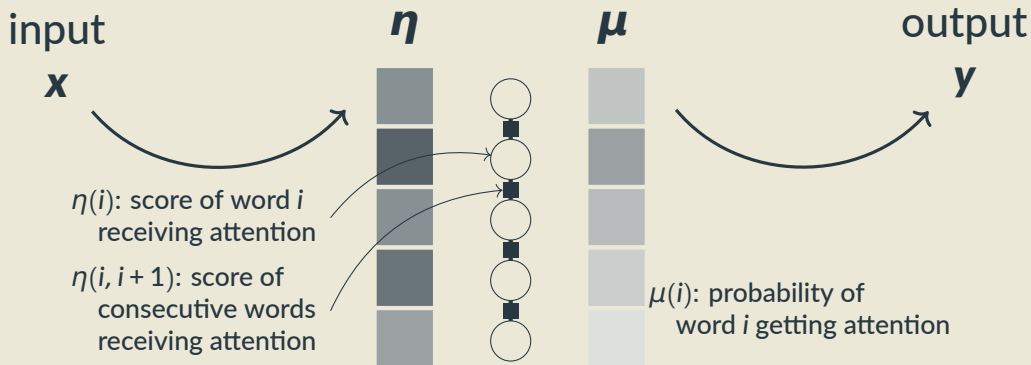
Structured Attention Networks



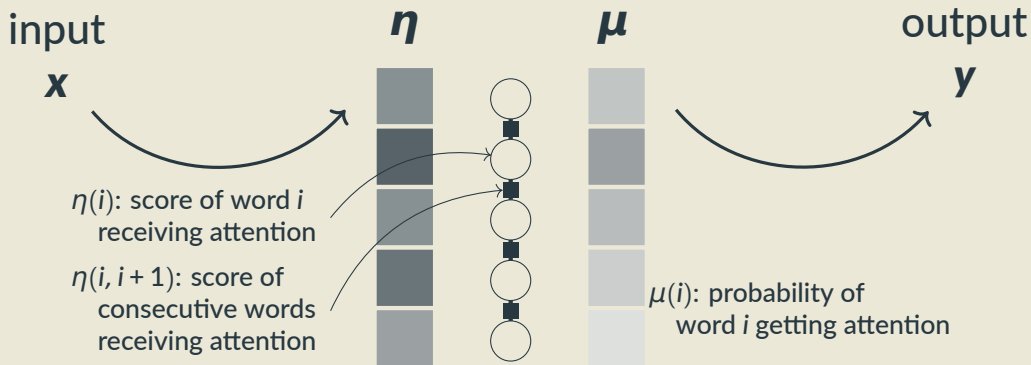
Structured Attention Networks



Structured Attention Networks

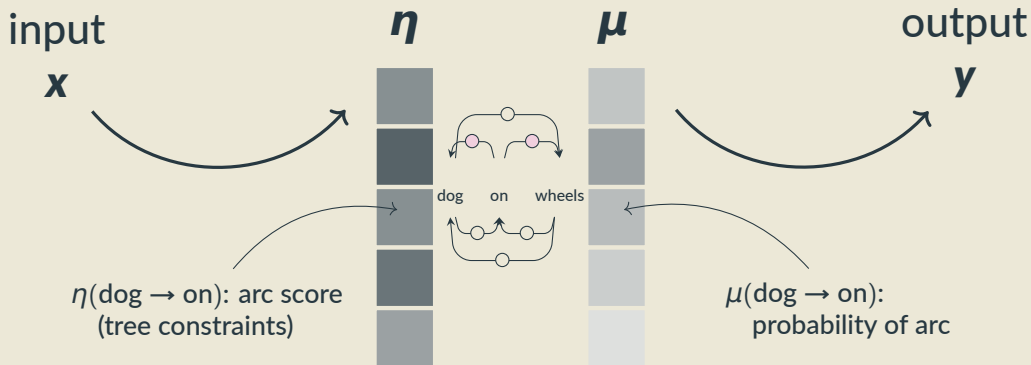


Structured Attention Networks



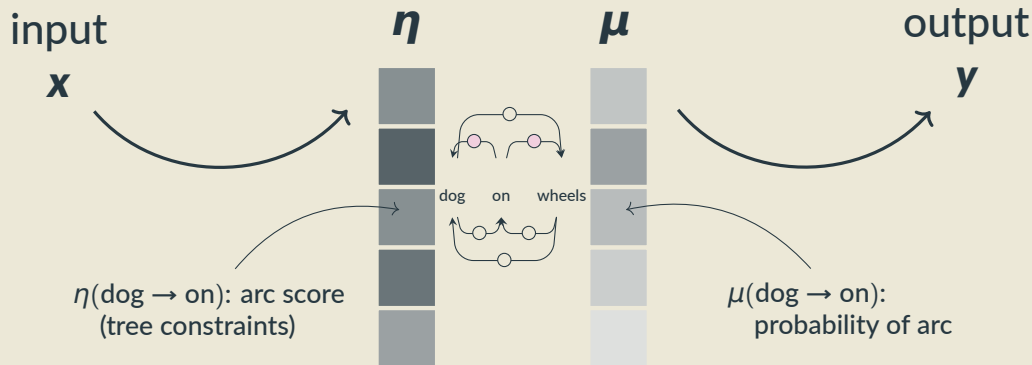
CRF marginals (from *forward-backward*) give attention weights $\in (0, 1)$

Structured Attention Networks



CRF marginals (from *forward-backward*) give attention weights $\in (0, 1)$
 Similar idea for projective dependency trees with *inside-outside*

Structured Attention Networks



CRF marginals (from *forward-backward*) give attention weights $\in (0, 1)$

Similar idea for projective dependency trees with *inside-outside*

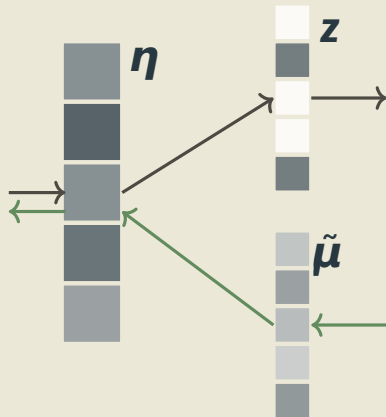
and non-projective with the Matrix-Tree theorem [Liu and Lapata, 2018].

Differentiable Perturb & Parse

Extending Gumbel-Softmax to structured stochastic models

- Forward pass:
sample structure \mathbf{z} (approximately)
$$\mathbf{z} = \arg \max_{\mathbf{z} \in \mathcal{Z}} (\boldsymbol{\eta} + \boldsymbol{\epsilon})^\top \mathbf{z}$$
- Backward pass:
pretend we did marginal inference
$$\tilde{\boldsymbol{\mu}} = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} (\boldsymbol{\eta} + \boldsymbol{\epsilon})^\top \mathbf{z} + \tilde{H}(\boldsymbol{\mu})$$

(or some similar relaxation)



Back-propagating through marginals

Pros:

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Networks:) All computations exact.

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Networks:) All computations exact.

Cons:

- (Structured Attention Networks:) forward pass marginals are dense;
(fixed by Perturb & MAP, at cost of rough approximation)

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Networks:) All computations exact.

Cons:

- (Structured Attention Networks:) forward pass marginals are dense;
(fixed by Perturb & MAP, at cost of rough approximation)
- Efficient & numerically stable back-propagation through DPs is tricky;
(somewhat alleviated by Mensch and Blondel [2018])

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Networks:) All computations exact.

Cons:

- (Structured Attention Networks:) forward pass marginals are dense;
(fixed by Perturb & MAP, at cost of rough approximation)
- Efficient & numerically stable back-propagation through DPs is tricky;
(somewhat alleviated by Mensch and Blondel [2018])
- Not applicable when marginals are unavailable.

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Networks:) All computations exact.

Cons:

- (Structured Attention Networks:) forward pass marginals are dense;
(fixed by Perturb & MAP, at cost of rough approximation)
- Efficient & numerically stable back-propagation through DPs is tricky;
(somewhat alleviated by Mensch and Blondel [2018])
- Not applicable when marginals are unavailable.
- Case-by-case algorithms required, can get tedious.

Back-propagating through marginals

Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Net

Cons:

- (Structured Attention Net (fixed by Perturb & MA
- Efficient & numerically stable (somewhat alleviated b
- Not applicable when mar
- Case-by-case algorithms

```

procedure BACKPROPINSIDEOUTSIDE( $\theta, p, \nabla_p^T$ )
  for  $s, t = 1, \dots, n, s \neq t$  do
     $d[s, t] \leftarrow \log p[s, t] \otimes \log \nabla_p^T[s, t]$ 
     $\nabla_p^T[s, t] \leftarrow \log \nabla_p^T[s, t] \leftarrow -\infty$ 
    for  $s = 1, \dots, n - 1$  do
      for  $t = s + 1, \dots, n$  do
         $\nabla_p^T[s, t, R, 0], \nabla_p^T[s, t, R, 1] \leftarrow d[s, t]$ 
         $\nabla_p^T[1, n, R, 1] \leftarrow -d[s, t]$ 
        if  $s > 1$  then
           $\nabla_p^T[s, t, L, 0], \nabla_p^T[s, t, L, 1] \leftarrow -d[s, t]$ 
           $\nabla_p^T[1, n, R, 1] \leftarrow -d[s, t]$ 
    for  $k = 1, \dots, n$  do
      for  $s = 1, \dots, n - k$  do
         $t \leftarrow s + k$ 
         $v \leftarrow \nabla_p^T[s, t, R, 0] \otimes \beta[s, t, R, 0]$ 
        for  $u = t, \dots, n$  do
           $\nabla_p^T[s, u, R, 0], \nabla_p^T[s, u, R, 1] \leftarrow v \otimes \beta[s, u, R, 1] \otimes \alpha[t, u, R, 1]$ 
        if  $s > 1$  then
           $v \leftarrow \nabla_p^T[s, t, L, 0] \otimes \beta[s, t, L, 0]$ 
          for  $u = t, \dots, n$  do
             $\nabla_p^T[s, u, L, 1], \nabla_p^T[s, u, L, 1] \leftarrow v \otimes \beta[s, u, L, 1] \otimes \alpha[t, u, L, 1]$ 
           $v \leftarrow \nabla_p^T[s, t, L, 1] \otimes \beta[s, t, L, 1]$ 
          for  $u = t, \dots, n$  do
             $\nabla_p^T[s, u, L, 1], \nabla_p^T[s, u, L, 0] \leftarrow v \otimes \beta[s, u, L, 1] \otimes \alpha[t, u, L, 1]$ 
        for  $u = 1, \dots, s - 1$  do
           $\gamma \leftarrow \beta[u, t, R, 0] \otimes \alpha[u, s - 1, R, 1] \otimes \theta_{s,t}$ 
           $\nabla_p^T[s, t, R, 0], \nabla_p^T[s, t, R, 1] \leftarrow \gamma \otimes \log \nabla_p^T[s, u] \leftarrow v \otimes \gamma$ 
           $\gamma \leftarrow \beta[u, t, L, 0] \otimes \alpha[u, s - 1, R, 1] \otimes \theta_{s,t}$ 
           $\nabla_p^T[s, t, L, 0], \nabla_p^T[s, t, L, 1] \leftarrow \gamma \otimes \log \nabla_p^T[s, u] \leftarrow v \otimes \gamma$ 
           $v \leftarrow \nabla_p^T[s, t, R, 1] \otimes \beta[s, t, R, 1]$ 
          for  $u = 1, \dots, s$  do
             $\nabla_p^T[s, u, R, 1], \nabla_p^T[s, u, R, 0] \leftarrow v \otimes \beta[s, u, R, 1] \otimes \alpha[u, s, R, 0]$ 
        for  $u = t + 1, \dots, n$  do
           $v \leftarrow \beta[s, u, R, 0] \otimes \alpha[t + 1, u, L, 1] \otimes \theta_{s,t}$ 
           $\nabla_p^T[s, u, R, 0], \nabla_p^T[s, u, L, 1], \log \nabla_p^T[s, u] \leftarrow v \otimes \gamma$ 
           $\gamma \leftarrow \beta[s, u, L, 0] \otimes \alpha[t + 1, u, L, 1] \otimes \theta_{s,t}$ 
           $\nabla_p^T[s, u, L, 0], \nabla_p^T[s, u, L, 1], \log \nabla_p^T[s, u] \leftarrow v \otimes \gamma$ 
    for  $k = n, \dots, 1$  do
      for  $s = 1, \dots, n - k$  do
         $t \leftarrow s + k$ 
         $v \leftarrow \nabla_p^T[s, t, R, 1] \otimes \alpha[s, t, R, 1]$ 
        for  $u = s + 1, \dots, t$  do
           $\nabla_p^T[s, u, R, 0], \nabla_p^T[s, u, R, 1] \leftarrow v \otimes \alpha[s, u, R, 0] \otimes \alpha[u, t, R, 1]$ 
        if  $s > 1$  then
           $v \leftarrow \nabla_p^T[s, t, L, 1] \otimes \alpha[s, t, L, 1]$ 
          for  $u = s, \dots, t - 1$  do
             $\nabla_p^T[s, u, L, 1], \nabla_p^T[s, u, L, 0] \leftarrow v \otimes \alpha[s, u, L, 1] \otimes \alpha[u, t, L, 0]$ 
           $v \leftarrow \nabla_p^T[s, t, L, 0] \otimes \alpha[s, t, L, 0]$ 
          for  $u = s, \dots, t - 1$  do
             $\gamma \leftarrow \alpha[s, u, R, 1] \otimes \alpha[u + 1, t, L, 1] \otimes \theta_{s,t}$ 
             $\nabla_p^T[s, u, R, 1], \nabla_p^T[s, u, L, 1], \log \nabla_p^T[s, u] \leftarrow v \otimes \gamma$ 
           $v \leftarrow \nabla_p^T[s, t, R, 0] \otimes \alpha[s, t, R, 0]$ 
          for  $u = s, \dots, t - 1$  do
             $\gamma \leftarrow \alpha[s, u, R, 1] \otimes \alpha[u + 1, t, L, 1] \otimes \theta_{s,t}$ 
             $\nabla_p^T[s, u, R, 1], \nabla_p^T[s, u, L, 1], \log \nabla_p^T[s, u] \leftarrow v \otimes \gamma$ 
      return sign exp log  $\nabla_p^T$ 
  
```

Figure 7: Backpropagation through the inside-outside algorithm to calculate the gradient with respect to the input potentials. ∇_p^T denotes the Jacobian of α with respect to θ (∇_p^T is the gradient with respect to θ). $\alpha, \beta \leftarrow \infty$ means $\alpha = \alpha \otimes \infty$ and $\beta = \beta \otimes \infty$.

xact.

inals are dense;
nation)

ugh DPs is tricky;
[8])

Back-propagating through marginals

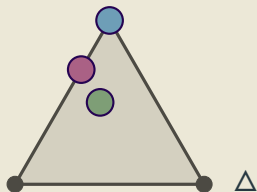
Pros:

- Familiar algorithms for NLPers,
- (Structured Attention Networks:) All computations exact.

Cons:

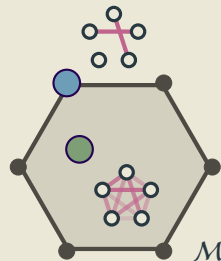
- (Structured Attention Networks:) forward pass marginals are dense;
(fixed by Perturb & MAP, at cost of rough approximation)
- Efficient & numerically stable back-propagation through DPs is tricky;
(somewhat alleviated by Mensch and Blondel [2018])
- Not applicable when marginals are unavailable.
- Case-by-case algorithms required, can get tedious.

- **argmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$
- **softmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} + H(\mathbf{p})$
- **sparsemax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} - 1/2 \|\mathbf{p}\|^2$

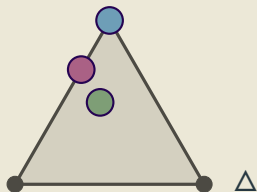


● **MAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$

● **marginals** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta} + \tilde{H}(\boldsymbol{\mu})$



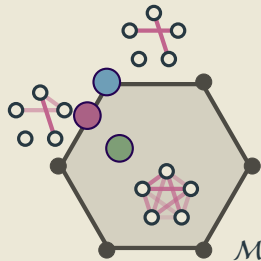
- **argmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s}$
- **softmax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} + H(\mathbf{p})$
- **sparsemax** $\arg \max_{\mathbf{p} \in \Delta} \mathbf{p}^T \mathbf{s} - 1/2 \|\mathbf{p}\|^2$



- **MAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta}$

- **marginals** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta} + \tilde{H}(\boldsymbol{\mu})$

- **SparseMAP** $\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^T \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$



SparseMAP solution

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

$$= \begin{array}{c} \circ \quad \circ \\ \diagup \quad \diagdown \\ \circ \quad \circ \end{array} = .6 \begin{array}{c} \circ \quad \circ \\ \diagup \quad \diagdown \\ \circ \quad \circ \end{array} + .4 \begin{array}{c} \circ \quad \circ \\ \diagup \quad \diagdown \\ \circ \quad \circ \end{array}$$

($\boldsymbol{\mu}^*$ is unique, but may have multiple decompositions \mathbf{p} . Active Set recovers a sparse one.)

Algorithms for SparseMAP

$$\boldsymbol{\mu}^{\star} = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^{\top} \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

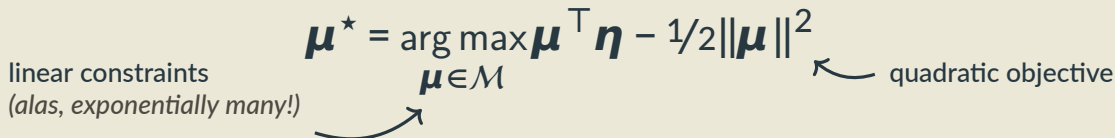
This is also $\text{proj}_{\mathcal{M}}$ required by SPIGOT!

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective

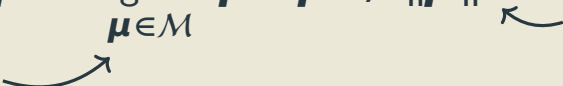


Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective



Conditional Gradient

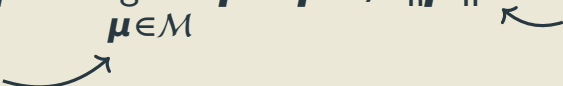
[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective



Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner of \mathcal{M}

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective

Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner of \mathcal{M}

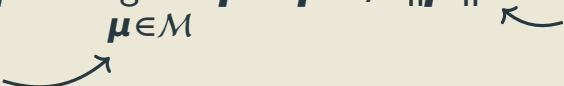
$$\arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \underbrace{(\boldsymbol{\eta} - \boldsymbol{\mu}^{(t-1)})}_{\tilde{\boldsymbol{\eta}}}$$

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective



Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner of \mathcal{M}
- update the (sparse) coefficients of \boldsymbol{p}
 - Update rules: vanilla, away-step, pairwise

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective

Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner of \mathcal{M}
- update the (sparse) coefficients of \boldsymbol{p}
 - Update rules: vanilla, away-step, pairwise
 - Quadratic objective: **Active Set**

a.k.a. Min-Norm Point, [Wolfe, 1976]

[Martins et al., 2015, Nocedal and Wright, 1999]

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective

Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner
- update the (sparse)
 - Update rules: vanilla
 - Quadratic objective:

Active Set achieves
finite & linear convergence!

a.k.a. Min-Norm Point, [Wolfe, 1976]

[Martins et al., 2015, Nocedal and Wright, 1999]

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective

Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner of \mathcal{M}
- update the (sparse) coefficients of \mathbf{p}
 - Update rules: vanilla, away-step, pairwise
 - Quadratic objective: **Active Set**
a.k.a. Min-Norm Point, [Wolfe, 1976]
[Martins et al., 2015, Nocedal and Wright, 1999]

Backward pass

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \text{ is sparse}$$

Algorithms for SparseMAP

linear constraints
(*alas, exponentially many!*)

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

quadratic objective

Conditional Gradient

[Frank and Wolfe, 1956, Lacoste-Julien and Jaggi, 2015]

- select a new corner of \mathcal{M}
- update the (sparse) coefficients of \mathbf{p}
 - Update rules: vanilla, away-step, pairwise
 - Quadratic objective: **Active Set**
a.k.a. Min-Norm Point, [Wolfe, 1976]
[Martins et al., 2015, Nocedal and Wright, 1999]

Backward pass

$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}$ is sparse
computing $\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)^\top d\mathbf{y}$
takes $\mathcal{O}(\dim(\boldsymbol{\mu}) \text{nnz}(\mathbf{p}^*))$

Algorithms for SparseMAP

$$\boldsymbol{\mu}^* = \arg \max_{\boldsymbol{\mu} \in \mathcal{M}} \boldsymbol{\mu}^\top \boldsymbol{\eta} - 1/2 \|\boldsymbol{\mu}\|^2$$

linear constraints
(*alas, exponentially many!*)

quadratic objective

Condition

pass

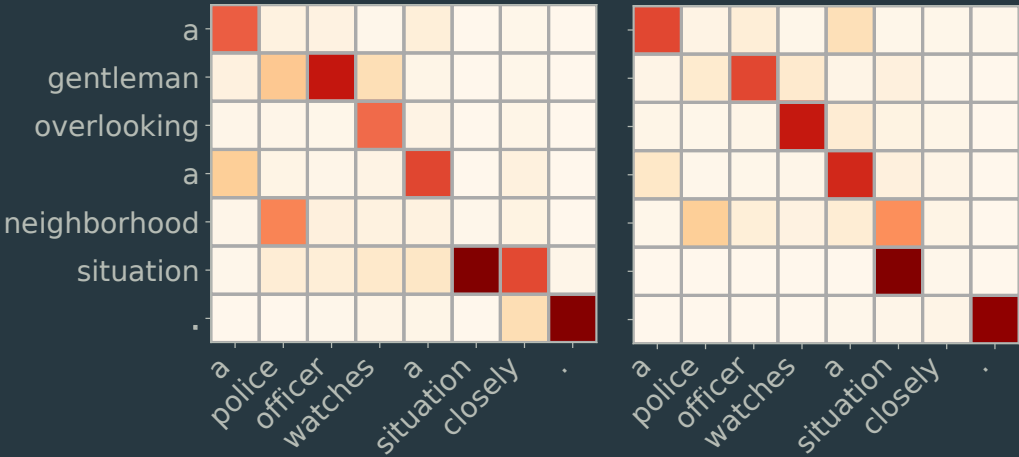
Completely modular: just add MAP

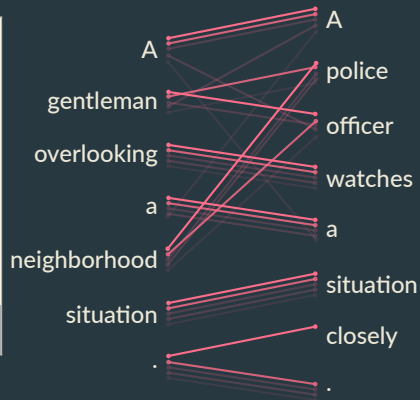
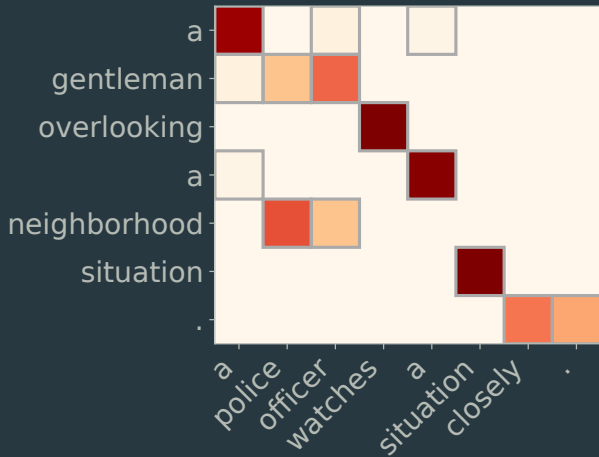
[Frank and Wolfe, 1956]

- select a new c
- update the (sparse) coefficients of \mathbf{p}
 - Update rules: vanilla, away-step, pairwise
 - Quadratic objective: **Active Set**
a.k.a. Min-Norm Point, [Wolfe, 1976]

[Martins et al., 2015, Nocedal and Wright, 1999]

computing $\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)^\top d\boldsymbol{\eta}$
takes $\mathcal{O}(\dim(\boldsymbol{\mu}) \text{nnz}(\mathbf{p}^*))$





Overview

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- REINFORCE
- Straight-Through Gumbel (Perturb & MAP)

$$L(\arg \max_{\mathbf{z}} \pi_{\theta}(\mathbf{z} | x))$$

- Straight-Through
- SPIGOT

$$L(\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[\mathbf{z}])$$

- Structured Attn. Nets
- SparseMAP

Structured latent variables without sampling

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}(\mathbf{z})) \pi(\mathbf{z} | \mathbf{x})$$

Structured latent variables without sampling

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\boldsymbol{\phi}}(\mathbf{z})) \pi_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{x})$$

Structured latent variables without sampling

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\boldsymbol{\phi}}(\mathbf{z})) \pi_{\boldsymbol{\theta}}(\mathbf{z} | \mathbf{x})$$

e.g., a TreeLSTM defined by \mathbf{z}

Structured latent variables without sampling

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\boldsymbol{\phi}}(\mathbf{z})) \pi_{\boldsymbol{\theta}}(\mathbf{z} | \mathbf{x})$$

e.g., a TreeLSTM defined by \mathbf{z}

parsing model,
using some scorer $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}; \mathbf{x})$

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\boldsymbol{\phi}(\mathbf{z})}) \pi_{\boldsymbol{\theta}}(\mathbf{z} | \mathbf{x})$$

parsing model,
using some scorer $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{z}; \mathbf{x})$

Exponentially large sum!

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | \mathbf{x})$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; \mathbf{x})$

idea 1

idea 2

idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\boldsymbol{\phi}}(\mathbf{z})) \pi_{\boldsymbol{\theta}}(\mathbf{z} | \mathbf{x})$$

How to define $\pi_{\boldsymbol{\theta}}$?

parsing model,
using some scorer $f_{\boldsymbol{\theta}}(\mathbf{z}; \mathbf{x})$

$$\sum_{h \in \mathcal{H}}$$

idea 1

idea 2

idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

idea 1

idea 2

idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

softmax

idea 2

idea 3

Structured latent variables without sampling

sum over
all possible trees


e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

softmax 

idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

idea 2

idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$



idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

idea 2

idea 3

softmax

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | \mathbf{x})$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; \mathbf{x})$

All methods we've seen require sampling; hard in general.

idea 2

idea 3

Structured latent variables without sampling

sum over
all possible trees



e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

softmax

idea 2 $\pi_{\theta}(\mathbf{z}) = 1$ if $\mathbf{z} = \text{MAP}(f_{\theta}(\cdot))$ else 0

argmax

idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

softmax



idea 2 $\pi_{\theta}(\mathbf{z}) = 1$ if $\mathbf{z} = \text{MAP}(f_{\theta}(\cdot))$ else 0

argmax



idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

softmax



idea 2 $\pi_{\theta}(\mathbf{z}) = 1$ if $\mathbf{z} = \text{MAP}(f_{\theta}(\cdot))$ else 0

argmax



idea 3

Structured latent variables without sampling

sum over
all possible trees

e.g., a TreeLSTM defined by \mathbf{z}

$$\mathbb{E}_{\mathbf{z}}[L(\mathbf{z})] = \sum_{\mathbf{z} \in \mathcal{Z}} L(\hat{\mathbf{y}}_{\phi}(\mathbf{z})) \pi_{\theta}(\mathbf{z} | x)$$

How to define π_{θ} ?

parsing model,
using some scorer $f_{\theta}(\mathbf{z}; x)$

$$\sum_{h \in \mathcal{H}} \frac{\partial \mathbb{E}[L(\mathbf{z})]}{\partial \theta}$$

idea 1 $\pi_{\theta}(\mathbf{z}) \propto \exp(f_{\theta}(\mathbf{z}))$

idea 2 $\pi_{\theta}(\mathbf{z}) = 1$ if $\mathbf{z} = \text{MAP}(f_{\theta}(\cdot))$ else 0

idea 3

softmax

argmax

SparseMAP



Structured latent variables without sampling

$$\begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} = .7x \quad \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + .3x \quad \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array}$$

Structured latent variables without sampling

$$\begin{array}{c} \text{•} \quad \text{•} \quad \text{•} \\ \text{↖} \quad \text{↗} \\ \text{•} \end{array} = .7x \quad \begin{array}{c} \text{•} \quad \text{•} \quad \text{•} \\ \text{↖} \quad \text{↗} \\ \text{•} \end{array} + .3x \quad \begin{array}{c} \text{•} \quad \text{•} \quad \text{•} \\ \text{↖} \quad \text{↗} \\ \text{•} \end{array} + 0x \quad \begin{array}{c} \text{•} \quad \text{•} \quad \text{•} \\ \text{↖} \quad \text{↗} \\ \text{•} \end{array} + \dots$$

Structured latent variables without sampling

$$\begin{aligned} \text{Diagram} &= .7 \times \text{Diagram} + .3 \times \text{Diagram} + 0 \times \text{Diagram} + \dots \\ \mathbb{E}[L(\mathbf{z})] &= .7 \times L(\text{Diagram}) + .3 \times L(\text{Diagram}) \end{aligned}$$

recall our shorthand $L(\mathbf{z}) = L(\hat{y}_{\phi}(\mathbf{z}), y)$

V. Conclusions

Stanford Natural Language Inference (Accuracy)

[Kim et al., 2017]

Simple Attention 86.2

Structured Attention 86.8

[Liu and Lapata, 2018]

100D Structured Attention 86.8

[Yogatama et al., 2017]

100D RL-SPINN 80.5

[Choi et al., 2018]

100D ST Gumbel-Tree 82.6

300D - 85.6

600D - 86.0

[Corro and Titov, 2019b]

Latent Tree + 1 GCN - 85.2

Latent Tree + 2 GCN - 86.2

[Havrylov et al., 2019]

100D TreeLSTM + tricks 84.3

Stanford Sentiment (Accuracy)

[Socher et al., 2013]

Bigram Naive Bayes 83.1

[Niculae et al., 2018b]

DepTreeLSTM w/ CoreNLP 83.2

DepTreeLSTM w/ SparseMAP 84.7

[Corro and Titov, 2019b]

GCN w/ CoreNLP 83.8

GCN w/ Perturb-and-MAP 84.6

[Choi et al., 2018]

ST Gumbel-Tree 90.7

[Havrylov et al., 2019]

TreeLSTM + tricks 90.2

Is it syntax?!

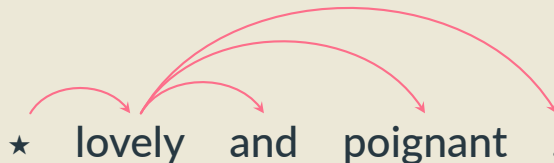
- Unlike e.g. unsupervised parsing, the structures we learn are guided by a **downstream objective** (typically discriminative).
- They don't typically resemble grammatical structure (yet) [Williams et al., 2018]
(future work: more inductive biases and constraints?)

Is it syntax?!

- Unlike e.g. unsupervised parsing, the structures we learn are guided by a **downstream objective** (typically discriminative).
- They don't typically resemble grammatical structure (yet) [Williams et al., 2018]
(future work: more inductive biases and constraints?)
- Common to compare latent structures with parser outputs.
But is this always a meaningful comparison?

Syntax vs. Composition Order

CoreNLP parse, $p = 21.4\%$

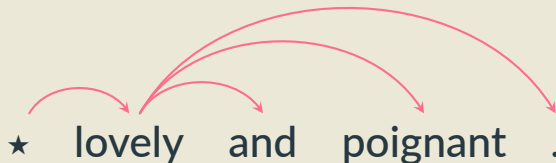


Syntax vs. Composition Order

$p = 22.6\%$

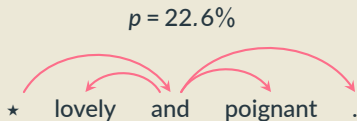


CoreNLP parse, $p = 21.4\%$

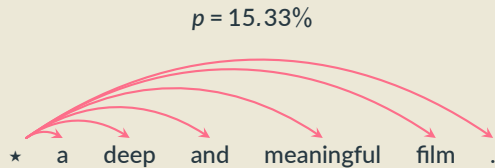
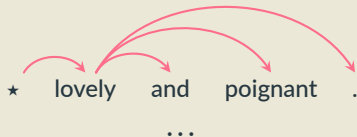


...

Syntax vs. Composition Order



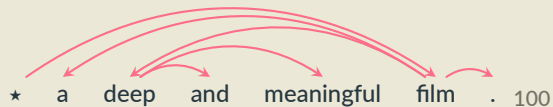
CoreNLP parse, $p = 21.4\%$



$p = 15.27\%$



...
CoreNLP parse, $p = 0\%$



Conclusions

- Latent structure models are desirable for interpretability, structural bias, and higher predictive power with fewer parameters.
- Stochastic latent variables can be dealt with RL or straight-through gradients.
- Deterministic argmax requires surrogate gradients (e.g. SPIGOT).
- Continuous relaxations of argmax include SANs and SparseMAP.
- Intuitively, some of these different methods are trying to do similar things or require the same building blocks (e.g. SPIGOT and SparseMAP).
- ... we didn't even get into deep *generative* models! These tools apply, but there are new challenges. [Corro and Titov, 2019a, Kim et al., 2019a,b, Kawakami et al., 2019]

Overview

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

- REINFORCE
- Straight-Through Gumbel (Perturb & MAP)
- SparseMAP

$$L(\arg \max_{\mathbf{z}} \pi_{\theta}(\mathbf{z} | x))$$

- Straight-Through
- SPIGOT

$$L(\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[\mathbf{z}])$$

- Structured Attn. Nets
- SparseMAP

Overview

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

$$L(\arg \max_{\mathbf{z}} \pi_{\theta}(\mathbf{z} | x))$$

$$L(\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[\mathbf{z}])$$

- REINFORCE^{SPL}
- Straight-Through Gumbel
(Perturb & MAP)^{SPL,MRG}
- SparseMAP^{MAP+}

- Straight-Through^{MAP,MRG}
- SPIGOT^{MAP+}

- Structured Attn. Nets^{MRG}
- SparseMAP^{MAP+}

Computation:

^{SPL}: Sampling. (Simple in incremental/unstructured, hard for most global structures.)

^{MAP}: Finding the highest-scoring structure.

^{MRG}: Marginal inference.

Overview

Thank you!

$$\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[L(\mathbf{z})]$$

$$L(\arg \max_{\mathbf{z}} \pi_{\theta}(\mathbf{z} | x))$$

$$L(\mathbb{E}_{\pi_{\theta}(\mathbf{z}|x)}[\mathbf{z}])$$

- REINFORCE^{SPL}
- Straight-Through Gumbel
(Perturb & MAP)^{SPL,MRG}
- SparseMAP^{MAP+}

- Straight-Through^{MAP,MRG}
- SPIGOT^{MAP+}

- Structured Attn. Nets^{MRG}
- SparseMAP^{MAP+}

Computation:

^{SPL}: Sampling. (Simple in incremental/unstructured, hard for most global structures.)

^{MAP}: Finding the highest-scoring structure.

^{MRG}: Marginal inference.

References I

- Ryan Adams. The gumbel-max trick for discrete distributions, 2013. URL <https://lips.cs.princeton.edu/the-gumbel-max-trick-for-discrete-distributions/>. Blog post.
- James K Baker. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132, 1979.
- Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
- Mathieu Blondel, André FT Martins, and Vlad Niculae. Learning classifiers with Fenchel-Young losses: Generalized entropies, margins, and algorithms. In *Proc. of AISTATS*, 2019.
- Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In *Proc. of ACL*, 2016. doi: 10.18653/v1/P16-1139.
- Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced LSTM for natural language inference. In *Proc. of ACL*, 2017.
- Jihun Choi, Kang Min Yoo, and Sang-goo Lee. Learning to compose task-specific tree structures. In *Proc. of AAAI*, 2018.

References II

- Yoeng-Jin Chu and Tseng-Hong Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965.
- William John Cocke and Jacob T Schwartz. *Programming languages and their compilers*. Courant Institute of Mathematical Sciences., 1970.
- Shay B Cohen, Karl Stratos, Michael Collins, Dean P Foster, and Lyle Ungar. Spectral learning of latent-variable PCFGs. In *Proc. of ACL*, 2012.
- Calo M Correia, Gon Vlad Niculae, and André FT Martins. Adaptively sparse transformers. In *Proc. of EMNLP-IJCNLP (to appear)*, 2019.
- Caio Corro and Ivan Titov. Differentiable Perturb-and-Parse: Semi-Supervised Parsing with a Structured Variational Autoencoder. In *Proc. of ICLR*, 2019a.
- Caio Corro and Ivan Titov. Learning latent trees with stochastic perturbations and differentiable dynamic programming. In *Proc. of ACL*, 2019b.
- Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *Proc. of ICML*, 2017.
- Jack Edmonds. Optimum branchings. *J. Res. Nat. Bur. Stand.*, 71B:233–240, 1967.
- Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Nav. Res. Log.*, 3(1-2):95–110, 1956.
- Serhii Havrylov, Germán Kruszewski, and Armand Joulin. Cooperative Learning of Disjoint Syntax and Semantics. In *Proc. NAACL-HLT*, 2019.

References III

- Geoffrey Hinton. Neural networks for machine learning. In *Coursera video lectures*, 2012.
- Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *Proc. of ICLR*, 2017.
- Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.
- Tadao Kasami. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*, 1966.
- Kazuya Kawakami, Chris Dyer, and Phil Blunsom. Learning to discover, ground and use words with segmental neural language models. In *Proc. of ACL*, 2019.
- Yoon Kim, Carl Denton, Loung Hoang, and Alexander M Rush. Structured attention networks. In *Proc. of ICLR*, 2017.
- Yoon Kim, Chris Dyer, and Alexander Rush. Compound probabilistic context-free grammars for grammar induction. In *Proc. of ACL*, 2019a.
- Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised recurrent neural network grammars. In *Proc. of NAACL-HLT*, 2019b.
- Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. 2014.
- Gustav Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497–508, 1847.

References IV

- Harold W Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Log.*, 2(1-2):83–97, 1955.
- Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Proc. of NeurIPS*, 2015.
- Zhifei Li and Jason Eisner. First-and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proc. of EMNLP*, 2009.
- Yang Liu and Mirella Lapata. Learning structured text representations. *TACL*, 6:63–75, 2018.
- Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. of ICLR*, 2016.
- Jean Maillard and Stephen Clark. Latent tree learning with differentiable parsers: Shift-Reduce parsing and chart parsing. *arXiv preprint arXiv:1806.00840*, 2018.
- Chaitanya Malaviya, Pedro Ferreira, and André FT Martins. Sparse and constrained attention for neural machine translation. In *Proc. of ACL*, 2018.
- André FT Martins and Ramón Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proc. of ICML*, 2016.
- André FT Martins and Julia Kreutzer. Learning what’s easy: Fully differentiable neural easy-first taggers. In *Proc. of EMNLP*, 2017.
- André FT Martins and Vlad Niculae. Notes on latent structure models and SPIGOT. *preprint arXiv:1907.10348*, 2019.

References V

- André FT Martins, Mário AT Figueiredo, Pedro MQ Aguiar, Noah A Smith, and Eric P Xing. AD3: Alternating directions dual decomposition for MAP inference in graphical models. *JMLR*, 16(1):495–545, 2015.
- Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proc. of ICML*, 2018.
- Nikita Nangia and Samuel Bowman. ListOps: A diagnostic dataset for latent tree learning. In *Proc. of NAACL SRW*, 2018.
- Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. In *Proc. of NeurIPS*, 2017.
- Vlad Niculae, André FT Martins, Mathieu Blondel, and Claire Cardie. SparseMAP: Differentiable sparse structured inference. In *Proc. of ICML*, 2018a.
- Vlad Niculae, André FT Martins, and Claire Cardie. Towards dynamic computation graphs via sparse latent structure. In *Proc. of EMNLP*, 2018b.
- Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer New York, 1999.
- George Papandreou and Alan L Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *Proc. of ICCV*, 2011.
- Hao Peng, Sam Thomson, and Noah A Smith. Backpropagating through structured argmax using a SPIGOT. In *Proc. of ACL*, 2018.
- Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proc. of ACL*, 2019.

References VI

- Slav Petrov and Dan Klein. Discriminative log-linear grammars with latent variables. In *Advances in neural information processing systems*, pages 1153–1160, 2008.
- Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 29(10):1848–1852, 2007.
- Lawrence R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *P. IEEE*, 77(2):257–286, 1989.
- Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, 26:43–49, 1978.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. EMNLP*, 2013.
- Veselin Stoyanov, Alexander Ropson, and Jason Eisner. Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proc. of AISTATS*, 2011.
- Ben Taskar. *Learning structured prediction models: A large margin approach*. PhD thesis, Stanford University, 2004.
- Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.
- Leslie G Valiant. The complexity of computing the permanent. *Theor. Comput. Sci.*, 8(2):189–201, 1979.

References VII

Tim Vieira. Gumbel-max trick, 2014. URL

<https://timvieira.github.io/blog/post/2014/07/31/gumbel-max-trick/>. Blog post.

Martin J Wainwright and Michael I Jordan. *Graphical models, exponential families, and variational inference.*, volume 1. Now Publishers, Inc., 2008.

Adina Williams, Andrew Drozdov, and Samuel R Bowman. Do latent tree learning models identify meaningful structure in sentences? *TACL*, 2018.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8, 1992.

Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. Learning to compose words into sentences with reinforcement learning. In *Proc. of ICLR*, 2017.

Daniel H Younger. Recognition and parsing of context-free languages in time n^3 . *Information and Control*, 10(2):189–208, 1967.