HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

p(B|A)yesgroup.ru

AIRI

Leveraging Recursive Gumbel-Max Trick for Approximate Inference in Combinatorial Spaces





Kirill Struminsky*   Artyom Gadetsky*   Denis Rakitin*   Danil Karpushkin   Dmitry Vetrov

Latent Combinatorial Structures in Beer Reviews

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

Appearance: 3.0 **Aroma: 4.0** Palate: 4.5 Taste: 4.0

Julian McAuley, Jure Leskovec. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews

Latent Combinatorial Structures in Beer Reviews

Predictor (Pours a slight tangerine orange and straw yellow. The head is nice and bubbly but fades very quickly with a little lacing. Smells like Wheat and European hops, a little yeast in there too. There is some fruit in there too, but you have to take a good whiff to get it. The taste is of wheat, a bit of malt, and a little fruit flavour in there too. Almost feels like drinking Champagne, medium mouthful otherwise. Easy to drink, but not something I'd be trying every night. **)** → **Aroma: 4.0**

Julian McAuley, Jure Leskovec. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews

Latent Combinatorial Structures in Beer Reviews

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

Appearance: 3.0 **Aroma: 4.0** Palate: 4.5 Taste: 4.0

Julian McAuley, Jure Leskovec. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews

Jianbo Chen, Le Song, Martin Wainwright, Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation.

Latent Combinatorial Structures in Beer Reviews

Explainer (Pours a slight tangerine orange and straw yellow. The head is nice and bubbly but fades very quickly with a little lacing. Smells like Wheat and European hops, a little yeast in there too. There is some fruit in there too, but you have to take a good whiff to get it. The taste is of wheat, a bit of malt, and a little fruit flavour in there too. Almost feels like drinking Champagne, medium mouthful otherwise. Easy to drink, but not something I'd be trying every night. **)** → **Predictor (** Smells like Wheat and European hops, a little fruit flavour **)** → **Aroma: 4.0**

Utilizing latent structure and prior knowledge gives us:

- Interpretability
- Faster inference for the predictor

Julian McAuley, Jure Leskovec. From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews

Jianbo Chen, Le Song, Martin Wainwright, Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation.

Sequence Modeling with Non-monotonic Orders

a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .

$$\prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$$

Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based decoding with automatically inferred generation order.
Dmitrii Emelianenko, Elena Voita, Pavel Serdyukov. Sequence Modeling with Unconstrained Generation Order

Sequence Modeling with Non-monotonic Orders

a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .

$$\prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$$

a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .
a cat sat on a mat .

$$\prod_{i=1}^{n} p(x_{\sigma_i} | x_{\sigma_1}, \ldots, x_{\sigma_{i-1}})$$

Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based decoding with automatically inferred generation order.
Dmitrii Emelianenko, Elena Voita, Pavel Serdyukov. Sequence Modeling with Unconstrained Generation Order

Outline of contributions

- Extension of Gumbel-Max Trick for Structured Latent Variables
  - How to sample discrete structure ✓
  - How to compute its probability ✓

- Construction of Gradient Estimators for Stochastic Optimization

$$\min_{\theta} \mathbb{E}_{p(X|\theta)} \mathcal{L}(X)$$

Gumbel-Max Trick

- $X \sim \text{Cat}(\text{softmax}(\theta)), \theta \in \mathbb{R}^n$

Gumbel-Max Trick

- $X \sim \text{Cat(softmax}(\theta)), \theta \in \mathbb{R}^n$

- Gumbel-Max Trick defines transformation
  - $g_i = -\log(-\log(u_i))$ for $u_i \sim U[0,1]^d$

  - Let $\phi(g, \theta) = \arg\max_{i=1\ldots d} (\theta_i + g_i)$ then $X = \phi(g, \theta)$

Gumbel-Max Trick properties

- $X = \arg\max_{i=1\ldots d} (\theta_i + g_i),\ g_i \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(g_i \,|\, 0)$

Gumbel-Max Trick properties

- $X = \arg\max_{i=1\ldots d} (\theta_i + g_i),\ g_i \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(g_i \mid 0)$
- $p(X, g_1, \ldots, g_d) = p(X) p(g_X) \prod_{i \neq X} p(g_i \mid g_X) =$

$$= \text{Cat}(X \mid \text{soft}\max(\theta)) \cdot \text{Gumbel}(g_X \mid \log \sum_i \exp \theta_i) \cdot \prod_{i \neq X} \text{TruncGumbel}(g_i \mid \theta_i, g_X)$$

Categorical r.v.

Maximal Gumbel

Remaining Gumbels are independent

Gumbel-Max Trick properties

- $X = \arg\max_{i=1\ldots d} \theta_i + g_i,\ g_i \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(g_i \mid 0)$
- $p(X, g_1, \ldots, g_d) = p(X)p(g_X)\prod_{i \neq X} p(g_i \mid g_X) =$

$$= \underbrace{\text{Cat}(X \mid \text{soft}\max(\theta))}_{\text{Categorical r.v.}} \cdot \underbrace{\text{Gumbel}(g_X \mid \log \sum_i \exp \theta_i)}_{\text{Maximal Gumbel}} \cdot \underbrace{\prod_{i \neq X} \text{TruncGumbel}(g_i \mid \theta_i, g_X)}_{\text{Remaining Gumbels are independent}}$$

Gumbel-Max for Permutations

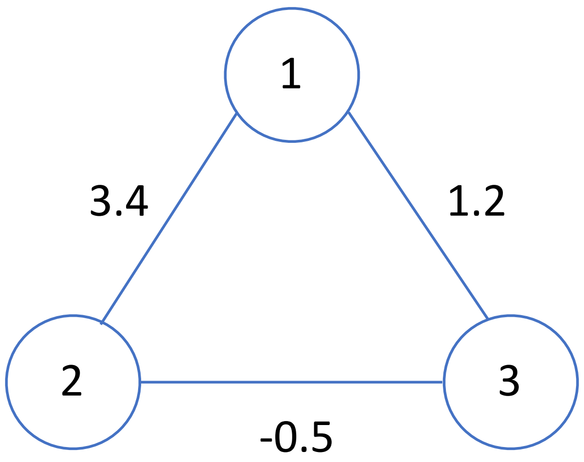
1.3	2.8	-0.5	3.4	-0.8	$\theta + g$
4	2	1	3	5	$X = \arg\operatorname{sort}(\theta + g)$

$$p(X|\theta) = \prod_{j=1}^{k} \frac{\exp \theta_{X_j}}{\sum_{u=j}^{k} \exp \theta_{X_u}}$$

Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, Dmitry Vetrov. Low-variance Black-box Gradient Estimates for the Plackett-Luce Distribution

Luce, R. D. Individual Choice Behavior: A Theoretical Analysis. Courier Corporation.

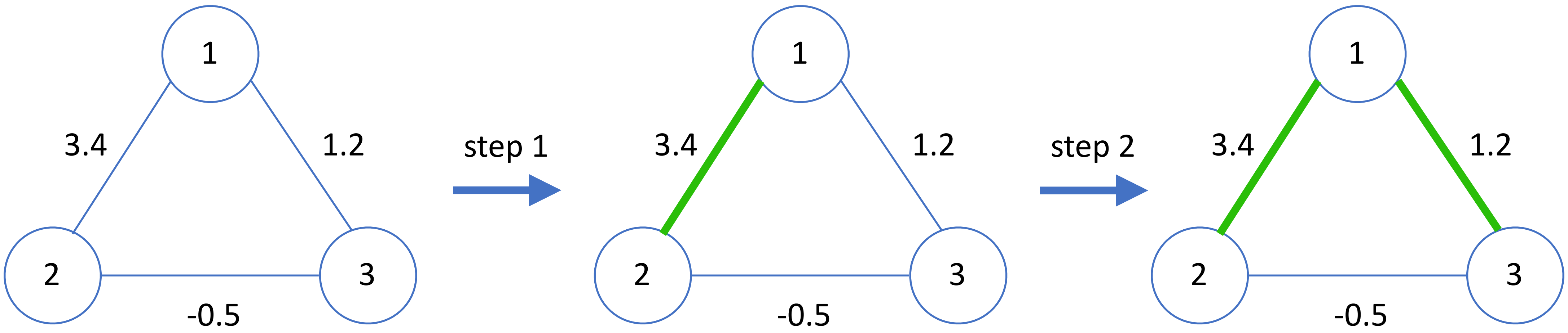
Gumbel-Max for Spanning Trees: Kruskal’s



$\theta_i + g_i$ - weights of the undirected graph

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, Chris J. Maddison. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables

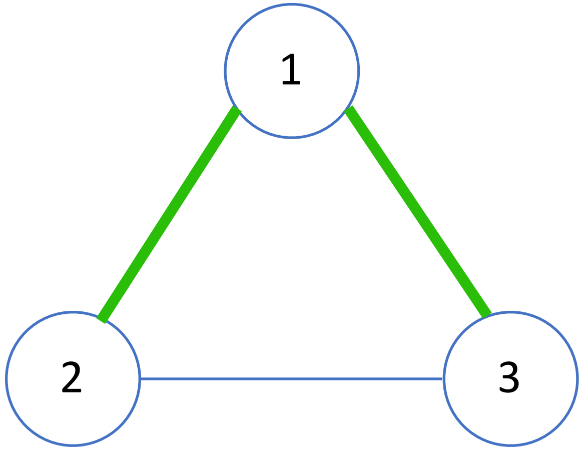
Gumbel-Max for Spanning Trees: Kruskal's



$\theta_i + g_i$ - weights of the undirected graph

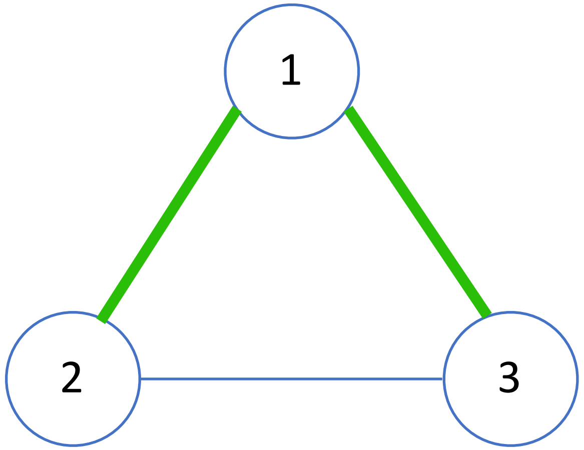
Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, Chris J. Maddison. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables

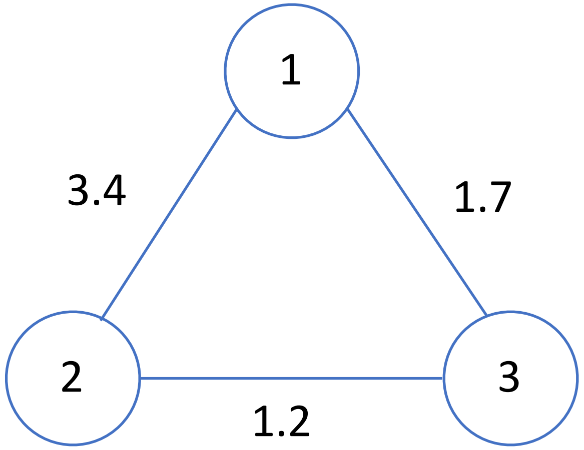
Trace of the algorithm



$p(X | \theta)$ - ?

Trace of the algorithm





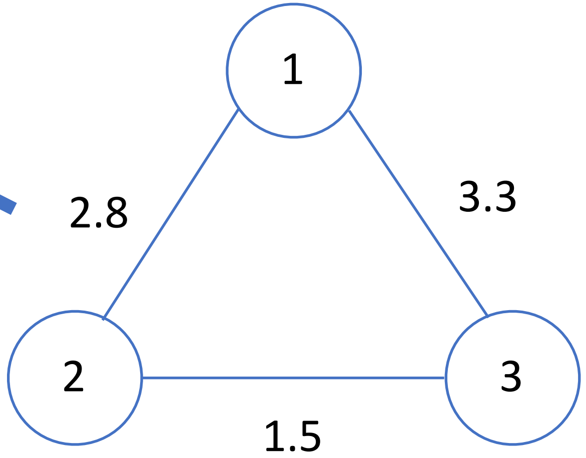
$$\theta_i + g_i$$

$$T = (\{1,2\}, \{1,3\})$$

$$\pi = (1,2)$$

$$p(X \mid \theta) = \sum_{\pi} p(X, \pi \mid \theta)$$

$$p(T \mid \theta) \stackrel{\text{def.}}{=} p(X, \pi \mid \theta)$$

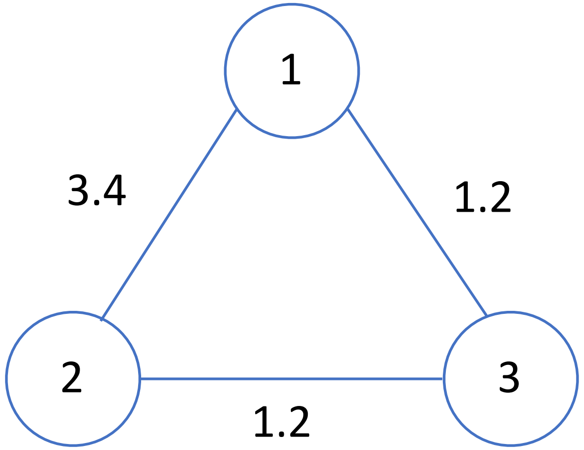


$$\theta_i + \tilde{g}_i$$

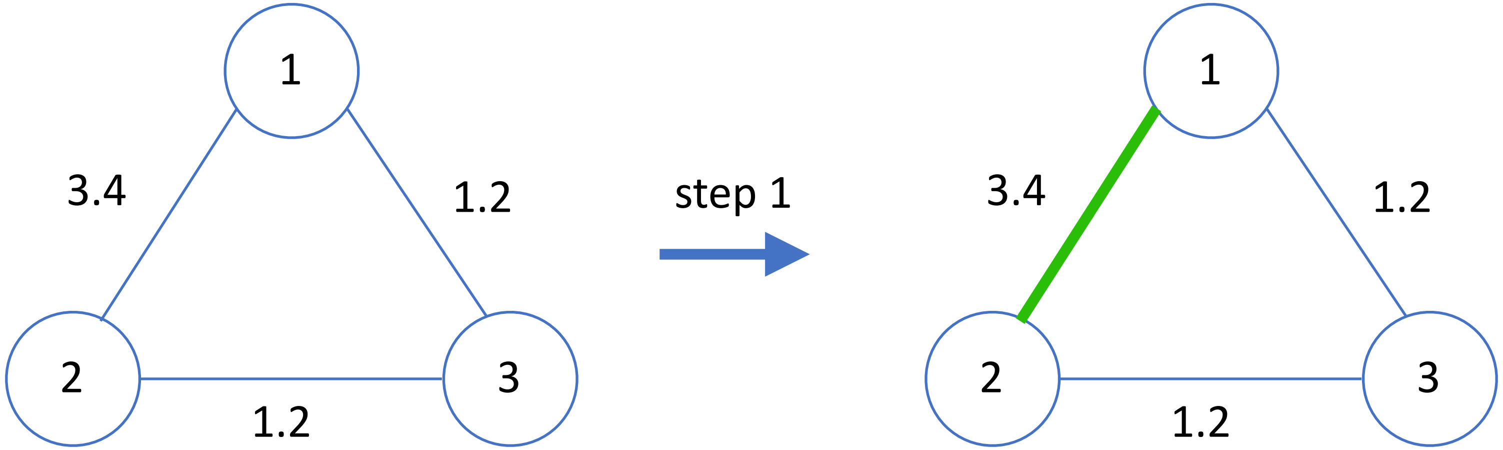
$$\tilde{T} = (\{1,3\}, \{1,2\})$$

$$\tilde{\pi} = (2,1)$$

When property does not hold: Prim’s



When property does not hold: Prim's



$$E_{12} \sim \text{Gumbel}(E_{12} \mid \log(\exp\theta_{12} + \exp\theta_{13}))$$
$$E_{13} \sim \text{TruncGumbel}(E_{13} \mid \theta_{13}, E_{12})$$
$$E_{23} \sim \text{Gumbel}(E_{23} \mid \theta_{23})$$

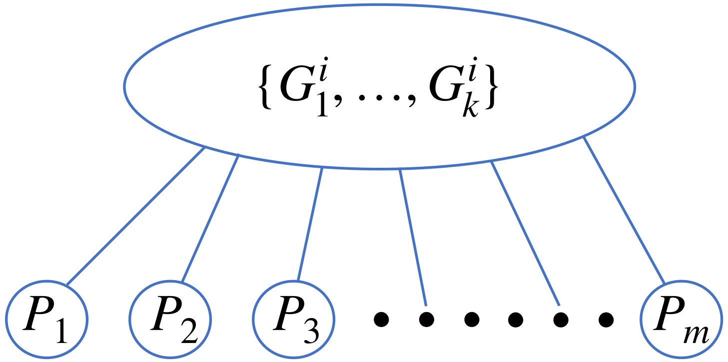
Stochastic Invariance in General

Recursion level i

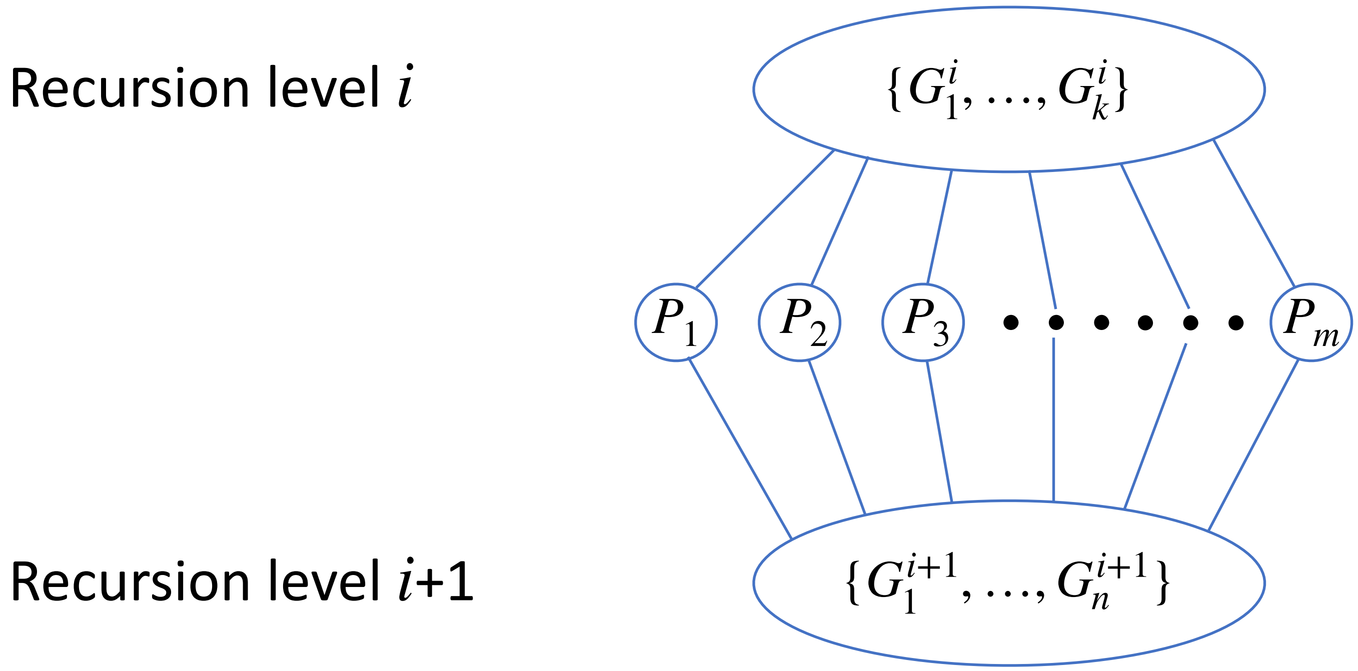
$\{G_1^i, \ldots, G_k^i\}$

Stochastic Invariance in General

Recursion level i



Stochastic Invariance in General



Stochastic Invariance

$\Downarrow$

- $X \sim p(X \mid \theta)$
- $p(T \mid \theta)$
- $\tilde{G} \sim p(G \mid T, \theta)$

Gradient Estimation

$$\frac{d}{d\theta}\mathbb{E}_{p(X|\theta)}\mathcal{L}(X)$$

Gradient Estimation

$$\frac{d}{d\theta}\mathbb{E}_{p(X|\theta)}\mathcal{L}(X)$$

$$\frac{d}{d\theta}\mathcal{L}(\phi(\epsilon,\theta))$$

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, Chris J. Maddison. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables

Gradient Estimation

$$\frac{d}{d\theta}\mathbb{E}_{p(X|\theta)}\mathcal{L}(X)$$

$$\frac{d}{d\theta}\mathcal{L}(\phi(\epsilon, \theta)) \qquad \mathcal{L}(\tilde{X})\frac{d}{d\theta}\log p(\tilde{X}|\theta)$$

Williams R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, Chris J. Maddison. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables

Gradient Estimation with Trace Variable

- Compute Score Function in Gumbel Space

$$g_\theta^G = \mathscr{L}(X)\,\nabla_\theta \log p(G\,|\,\theta)$$

Gradient Estimation with Trace Variable

- Compute Score Function in Gumbel Space

$$g_\theta^G = \mathscr{L}(X)\,\nabla_\theta \log p(G \mid \theta)$$

- But we can do better

use $g_\theta^T = \mathscr{L}(X)\,\nabla_\theta \log p(T \mid \theta)$

instead of $g_\theta^X = \mathscr{L}(X)\,\nabla_\theta \log p(X \mid \theta)$

Gradient Estimation with Trace Variable

- Compute Score Function in Gumbel Space

$$g_\theta^G = \mathscr{L}(X)\,\nabla_\theta \log p(G\,|\,\theta)$$

- But we can do better

use $g_\theta^T = \mathscr{L}(X)\,\nabla_\theta \log p(T\,|\,\theta)$

instead of $g_\theta^X = \mathscr{L}(X)\,\nabla_\theta \log p(X\,|\,\theta)$

- We can show that $g_\theta^T = \mathbb{E}_{G|T}\,[g_\theta^G]$ and $g_\theta^X = \mathbb{E}_{T|X}\,[g_\theta^T]$

$$\implies \text{Var}[g_\theta^X] \leq \text{Var}[g_\theta^T] \leq \text{Var}[g_\theta^G]$$

Conclusion

- Extension of Gumbel-Max Trick for Structured Latent Variables
  - How to sample discrete structure ✓
  - How to compute its probability ✓

- Construction of Gradient Estimators for Stochastic Optimization

$$\min_{\theta} \mathbb{E}_{p(X|\theta)} \mathcal{L}(X)$$