NATIONAL RESEARCH UNIVERSITY

p(B|A)yesgroup.ru

UNIVERSITY OF SUSSEX

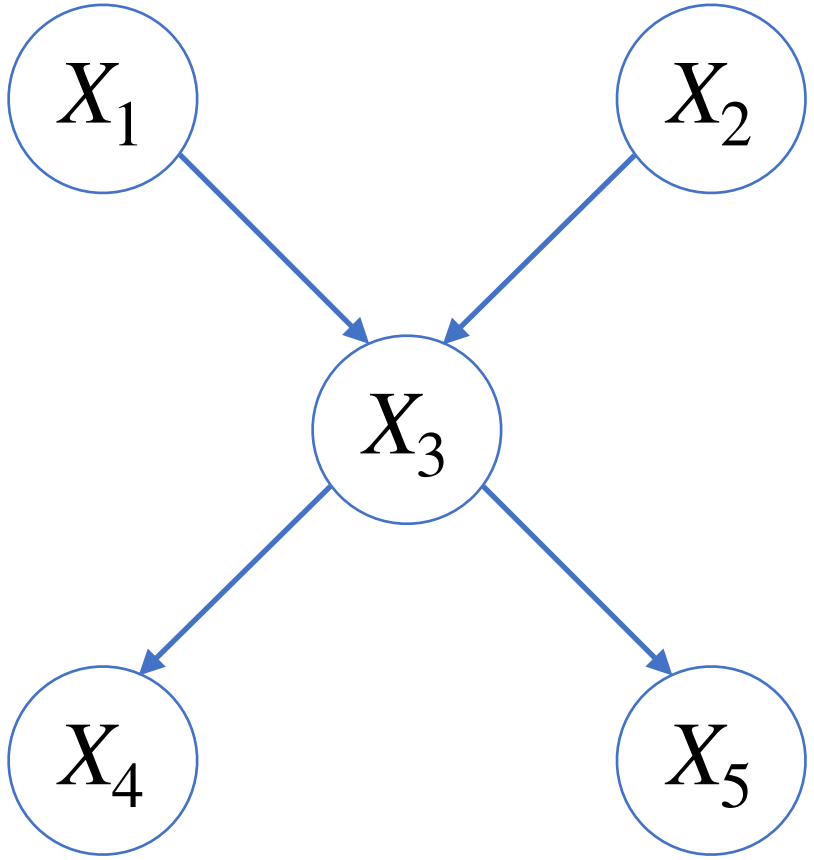
Low-variance Black-box Gradient Estimates for the Plackett-Luce Distribution

Artyom Gadetsky* Kirill Struminsky* Christopher Robinson Novi Quadrianto Dmitry Vetrov

Motivation: Causal Structure Learning

- $\mathbf{X} \in \mathbb{R}^{n \times k}$ - data matrix
- W - **DAG** adjacency matrix

$$\min_{W} \frac{1}{2n} \|\mathbf{X} - W^T\mathbf{X}\|_F^2 + \lambda \|vec(W)\|_1$$



Motivation: Order-based Causal Structure Learning

- **DAG constraint is hard**
- Use **topological order parametrization**

$$W \text{ - DAG} \iff W = PAP^T$$

- P - permutation matrix
- A - strictly upper triangular matrix

Motivation: Order-based Causal Structure Learning

- **DAG constraint is hard**
- Use **topological order parametrization**

$$\min_{P} \left[\min_{A} \frac{1}{2n} \|\mathbf{X} - PAP^T\mathbf{X}\|_F^2 + \lambda \|vec(A)\|_1 \right]$$

Motivation: Order-based Causal Structure Learning

- **DAG constraint is hard**
- Use **topological order parametrization**

$$\min_{P} \left[\min_{A} \frac{1}{2n} \|\mathbf{X} - PAP^T\mathbf{X}\|_F^2 + \lambda \|vec(A)\|_1 \right]$$

Motivation: Order-based Causal Structure Learning

- **DAG constraint is hard**
- Use **topological order parametrization**

$$\min_{P} \left[\min_{A} \frac{1}{2n} \|\mathbf{X} - PAP^T\mathbf{X}\|_F^2 + \lambda \|vec(A)\|_1 \right] = \min_{P} Q(P, \mathbf{X})$$

- Function $Q(P, \mathbf{X})$ involves optimization - consider as **black-box**

Optimization w.r.t permutations

$$\min_{b} f(b)$$

- $b = (b_1, \ldots, b_k) \in S_k$ - permutation
- $f(b)$ - function, can be black-box

- Potential use cases:
  - ranking
  - travelling salesman problem
  - order-based causal structure learning

From combinatorial to variational optimization

$$\min_b f(b) \leq \min_\theta \mathbb{E}_{p(b|\theta)} f(b)$$

- $b = (b_1, \ldots, b_k) \in S_k$ - permutation
- $f(b)$ - function, can be black-box
- $p(b \mid \theta)$ - distribution over permutations parametrized by θ

- Related work:
  - Relaxation to doubly-stochastic matrix (Mena et al. 2018)
  - Relaxation to unimodal row-stochastic matrix (Grover et al. 2019)

From combinatorial to variational optimization

$$\min_b f(b) \leq \min_\theta \mathbb{E}_{p(b|\theta)} f(b)$$

- $b = (b_1, \ldots, b_k) \in S_k$ - permutation
- $f(b)$ - function, can be black-box
- $p(b \mid \theta)$ - distribution over permutations parametrized by θ

- Related work:
  - Relaxation to doubly-stochastic matrix (Mena et al. 2018)
  - Relaxation to unimodal row-stochastic matrix (Grover et al. 2019)

Plackett-Luce distribution

$$p(b \mid \theta) = \prod_{j=1}^{k} \frac{\exp \theta_{b_j}}{\sum_{u=j}^{k} \exp \theta_{b_u}}$$

- $b = (b_1, \ldots, b_k) \in S_k$ - permutation
- $\theta = (\theta_1, \ldots, \theta_k)$ - scores

- Distribution can be seen as sequential generation without replacement from the categorical distribution

Stochastic estimators of the gradient

$$\nabla_\theta \mathbb{E}_{p(b|\theta)} f(b) - ?$$

Stochastic estimators of the gradient

$$\nabla_\theta \mathbb{E}_{p(b|\theta)} f(b) - ?$$

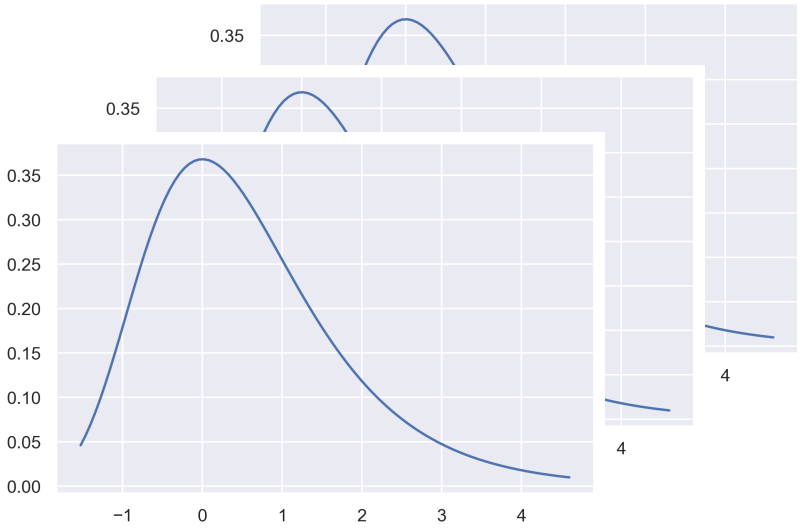
- REINFORCE (Williams 1992)
  - works for any type of variables & arbitrary f
  - high variance

Stochastic estimators of the gradient

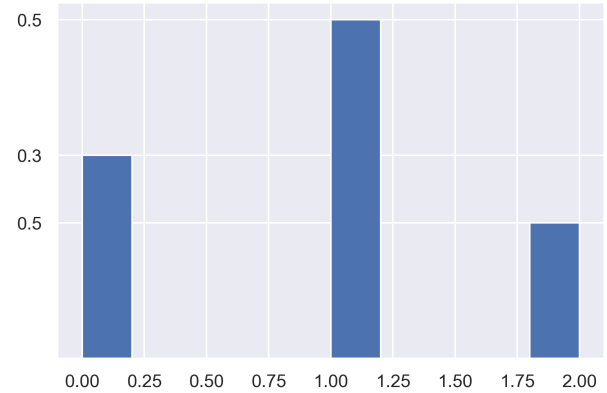
$$\nabla_\theta \mathbb{E}_{p(b|\theta)} f(b) - ?$$

- REINFORCE (Williams 1992)
  - works for any type of variables & arbitrary f
  - high variance
- Reparametrization trick (Kingma et al. 2013, Rezende et al. 2014)
  - works only for continuous distributions & differentiable f
  - low-variance

Gumbel-max trick



$$+\ H(z) = \arg\max_i z \ \Rightarrow$$



$$z_i \sim Gumbel(\theta_i, 1)$$

$$b \sim Categorical(\theta)$$

RELAX for the categorical random variable

Start with the initial expectation

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(z|\theta)} f(H(z)) = \mathbb{E}_{p(z|\theta)}[f(H(z)) - c_\phi(z)] + \mathbb{E}_{p(z|\theta)}[c_\phi(z)]$$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX for the categorical random variable

Use Gumbel-max trick for the reparametrization

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(z|\theta)} f(H(z)) = \mathbb{E}_{p(z|\theta)}[f(H(z)) - c_\phi(z)] + \mathbb{E}_{p(z|\theta)}[c_\phi(z)]$$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX for the categorical random variable

Add parametrized control variate c_ϕ to reduce variance

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(z|\theta)} f(H(z)) = \mathbb{E}_{p(z|\theta)}[f(H(z)) - c_\phi(z)] + \mathbb{E}_{p(z|\theta)}[c_\phi(z)]$$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX for the categorical random variable

Add parametrized control variate c_ϕ to reduce variance

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(z|\theta)} f(H(z)) = \mathbb{E}_{p(z|\theta)}[f(H(z)) - c_\phi(z)] + \mathbb{E}_{p(z|\theta)}[c_\phi(z)]$$

- Use REINFORCE
- Use reparametrization trick

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX for the categorical random variable

To further reduce variance use conditional reparametrization

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(b|\theta)} f(b) - \mathbb{E}_{p(b|\theta)} \mathbb{E}_{p(z|b,\theta)} c_\phi(z) + \mathbb{E}_{p(z|\theta)} c_\phi(z)$$

- Use REINFORCE
- Use reparametrization trick

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX requirements

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(b|\theta)} f(b) - \mathbb{E}_{p(b|\theta)} \mathbb{E}_{p(z|b,\theta)} c_\phi(z) + \mathbb{E}_{p(z|\theta)} c_\phi(z)$$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX requirements

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(b|\theta)} f(b) - \mathbb{E}_{p(b|\theta)} \mathbb{E}_{p(z|b,\theta)} c_\phi(z) + \mathbb{E}_{p(z|\theta)} c_\phi(z)$$

Requires reparametrization trick for $p(z \mid \theta)$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX requirements

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(b|\theta)} f(b) - \mathbb{E}_{p(b|\theta)} \mathbb{E}_{p(z|b,\theta)} c_\phi(z) + \mathbb{E}_{p(z|\theta)} c_\phi(z)$$

Requires deterministic mapping $b = H(z)$ and density $p(b \mid \theta)$

Requires reparametrization trick for $p(z \mid \theta)$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

RELAX requirements

$$\mathbb{E}_{p(b|\theta)} f(b) = \mathbb{E}_{p(b|\theta)} f(b) - \mathbb{E}_{p(b|\theta)} \mathbb{E}_{p(z|b,\theta)} c_\phi(z) + \mathbb{E}_{p(z|\theta)} c_\phi(z)$$

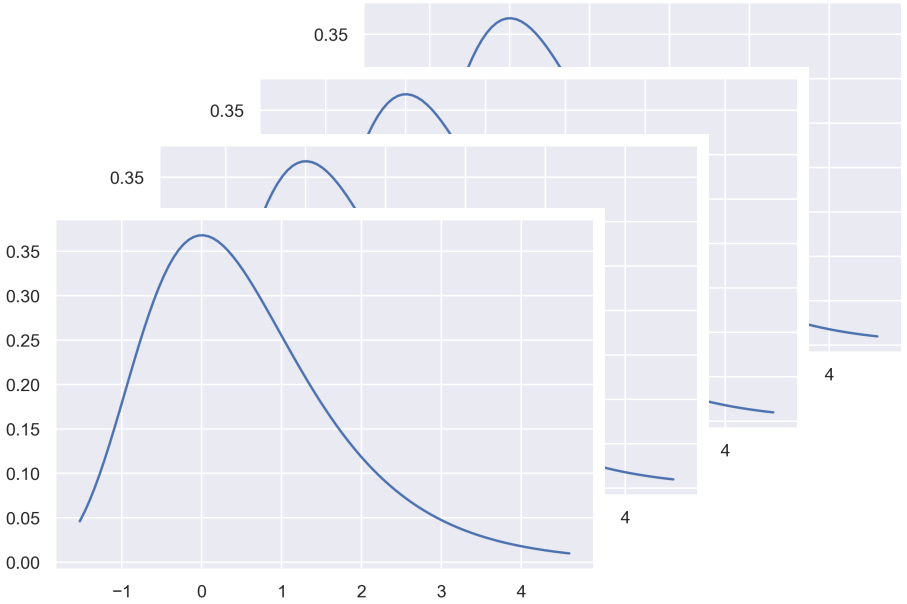
Requires deterministic mapping $b = H(z)$ and density $p(b \mid \theta)$

Requires reparametrization trick for $p(z \mid b, \theta)$

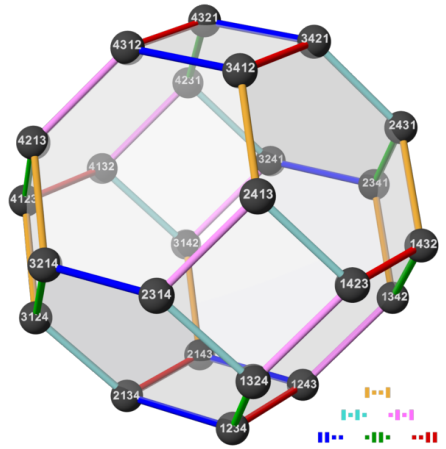
Requires reparametrization trick for $p(z \mid \theta)$

Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. Grathwohl et al. 2018

Gumbel-max trick for the Plackett-Luce



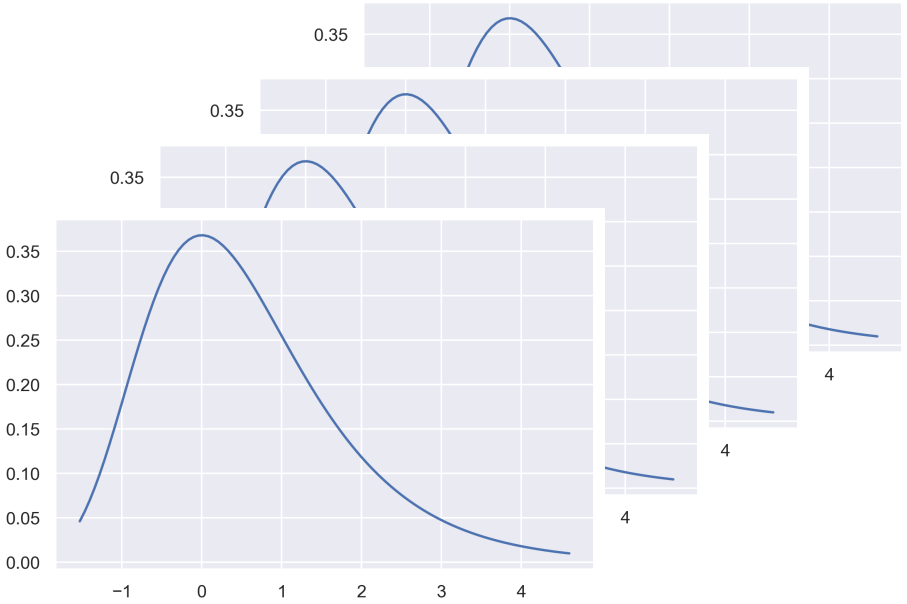
$$+\ H(z) = \arg\operatorname{sort}(z) \Rightarrow$$



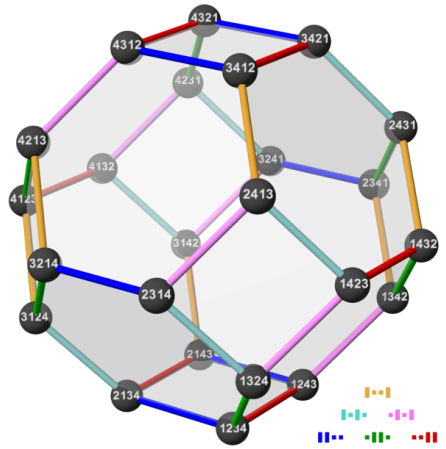
$$z_i \sim Gumbel(\theta_i, 1)$$

$$b \sim PlackettLuce(\theta)$$

Gumbel-max trick for the Plackett-Luce



$$+ \; H(z) = \arg\operatorname{sort}(z) \;\Rightarrow$$



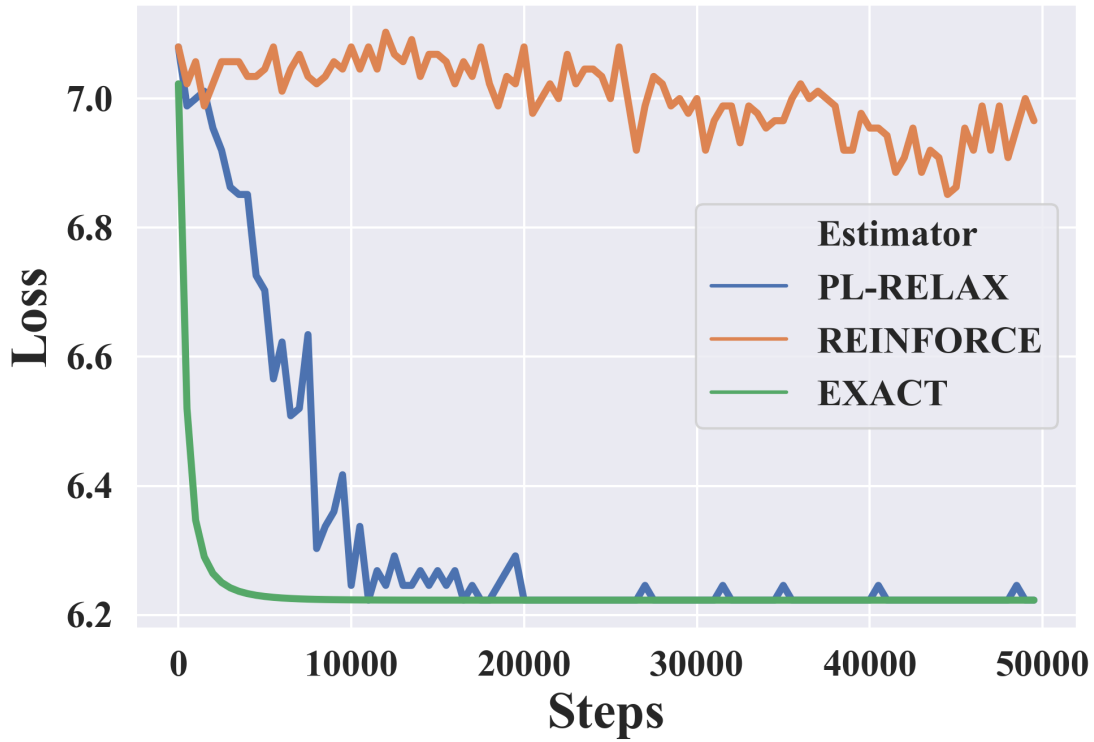
$$z_i \sim Gumbel(\theta_i, 1)$$

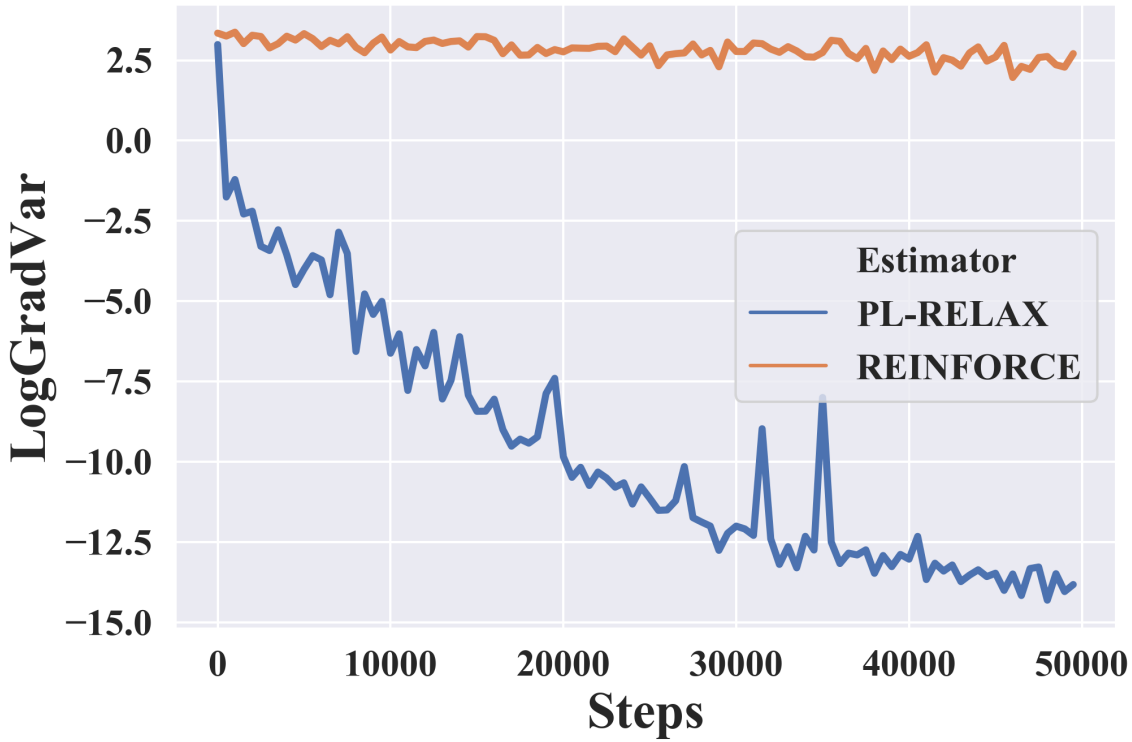
$$b \sim PlackettLuce(\theta)$$

Reparametrization trick for $p(z \mid b, \theta)$ derived in the paper

Toy Experiment

$$\min_{\theta} \mathbb{E}_{p(b|\theta)} \|P_b - P_t\|_F^2$$





Order-based Causal Structure Learning

$$\min_{\theta} \mathbb{E}_{p(b|\theta)} \hat{Q}(b, \mathrm{X})$$

	ER1			
	Val $\widehat{Q} - \widehat{Q}^*$	SHD	SHD-CPDAG	SID
PL-RELAX	-1.8±1.3	19.2±6.9	20.6±7.8	103.2±55.5
SINKHORN$_{ECP}$	5.5±7.0	30.0±6.3	30.8±5.8	151.8±35.1
URS$_{ECP}$	10.3±4.7	41.0±2.4	40.0±2.7	177.6±17.1
SINKHORN	90.3±35.8	49.6±4.3	49.6±4.3	275.0±42.5
URS	90.3±35.8	49.6±4.3	49.6±4.3	275.0±42.5
GREEDY-SP	N/A	38.2±21.6	38.2±24.6	151.6±84.3
RANDOM	271.0±71.6	99.4±9.3	99.8±9.5	301.2±60.4

Outcomes of the talk

Tonight
Poster #RU8893

- We derived stochastic gradient estimator which
  - allows optimization w.r.t. permutations for arbitrary function f
  - has low variance

- More details at arXiv: http://bit.ly/arx1v
- Code is available: http://bit.ly/g17hu8