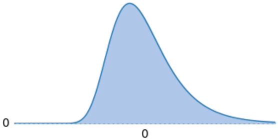
Lost Relatives of the Gumbel Trick

Matej Balog

Nilesh Tripuraneni

Zoubin Ghahramani

Adrian Weller



ICML 2017, Sydney, Australia

Slides: http://matejbalog.eu/en/research/

Code: https://github.com/matejbalog/gumbel-relatives

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$

Gumbel trick

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$

Gumbel trick $\longrightarrow$ Estimate
$Z := \sum_{j=1}^{N} \tilde{p}_j$

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$

Gumbel trick $\longrightarrow$ Estimate
$Z := \sum_{j=1}^{N} \tilde{p}_j$

In discrete graphical models, bound
$\mathcal{L} \leq Z \leq \mathcal{U}$

Lost Relatives of the Gumbel Trick

Competing exponential clocks $\xrightarrow{-\ln t - c}$ **Gumbel trick**

Gumbel trick → Sample from $\mathbf{p} = (p_1, \ldots, p_N)$

Gumbel trick → Estimate $Z := \sum_{j=1}^{N} \tilde{p}_j$

Gumbel trick → In discrete graphical models, bound $\mathcal{L} \leq Z \leq \mathcal{U}$

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$

Competing exponential clocks

Fréchet tricks

Gumbel trick

Exponential trick

Weibull tricks

...

Estimate
$Z := \sum_{j=1}^{N} \tilde{p}_j$

In discrete graphical models, bound
$\mathcal{L} \leq Z \leq \mathcal{U}$

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$
tricks equivalent

Competing exponential clocks

Fréchet tricks

Gumbel trick

Exponential trick

Weibull tricks

...

Estimate
$Z := \sum_{j=1}^{N} \tilde{p}_j$

In discrete graphical models, bound
$\mathcal{L} \leq Z \leq \mathcal{U}$

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$
tricks equivalent

Competing exponential clocks

Fréchet tricks

Gumbel trick

Exponential trick

Weibull tricks

...

Estimate
$Z := \sum_{j=1}^{N} \tilde{p}_j$
family of estimators $\hat{Z}$

In discrete graphical models, bound
$\mathcal{L} \leq Z \leq \mathcal{U}$

Lost Relatives of the Gumbel Trick

Sample from
$\mathbf{p} = (p_1, \ldots, p_N)$
tricks equivalent

Competing exponential clocks

Fréchet tricks

Gumbel trick

Exponential trick

Weibull tricks

...

Estimate
$Z := \sum_{j=1}^{N} \tilde{p}_j$
family of estimators $\hat{Z}$

In discrete graphical models, bound
$\mathcal{L} \leq Z \leq \mathcal{U}$
family of bounds
$\mathcal{L}(\alpha)$, $\mathcal{U}(\alpha)$

Competing Exponential Clocks

N independent clocks started simultaneously:

$\sim \mathrm{Exp}(\lambda_1)$

$\sim \mathrm{Exp}(\lambda_2)$

$\cdots$

$\sim \mathrm{Exp}(\lambda_N)$

Competing Exponential Clocks

N independent clocks started simultaneously:

$\sim \text{Exp}(\lambda_1)$

$\sim \text{Exp}(\lambda_2)$

$\dots$

$\sim \text{Exp}(\lambda_N)$

1. Probability q_i that clock i is the first to ring:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$$

Competing Exponential Clocks

N independent clocks started simultaneously:

$\sim \text{Exp}(\lambda_1)$

$\sim \text{Exp}(\lambda_2)$

$\dots$

$\sim \text{Exp}(\lambda_N)$

1. Probability q_i that clock i is the first to ring:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$$

2. First time T when a clock rings:

$$T \sim \text{Exp}(\lambda_1 + \cdots + \lambda_N)$$

$$\mathbb{E}[T] = \frac{1}{\lambda_1 + \cdots + \lambda_N}$$

Competing Exponential Clocks

Unnormalized probability mass function $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_N)$:

$\sim \text{Exp}(\lambda_1)$

$\sim \text{Exp}(\lambda_2)$

$\ldots$

$\sim \text{Exp}(\lambda_N)$

1. Probability q_i that clock i is the first to ring:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j}$$

2. First time T when a clock rings:

$$T \sim \text{Exp}(\lambda_1 + \cdots + \lambda_N)$$
$$\mathbb{E}[T] = \frac{1}{\lambda_1 + \cdots + \lambda_N}$$

Competing Exponential Clocks

Unnormalized probability mass function $\tilde{\mathbf{p}} = (\tilde{p}_1, \dots, \tilde{p}_N)$:

$\sim \text{Exp}(\tilde{p}_1)$

$\sim \text{Exp}(\tilde{p}_2)$

$\dots$

$\sim \text{Exp}(\tilde{p}_N)$

1. Probability q_i that clock i is the first to ring:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^N \lambda_j} = \frac{\tilde{p}_i}{\sum_{j=1}^N \tilde{p}_j} = \frac{\tilde{p}_i}{Z} = p_i$$

2. First time T when a clock rings:

$$T \sim \text{Exp}(\lambda_1 + \dots + \lambda_N) = \text{Exp}(Z)$$

$$\mathbb{E}[T] = \frac{1}{\lambda_1 + \dots + \lambda_N} = \frac{1}{Z}$$

Competing Exponential Clocks

Unnormalized probability mass function $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_N) = (e^{\theta_1}, \ldots, e^{\theta_N})$:

$-ln($  $) - c$ $\sim \theta_1 + \text{Gumbel}(0)$

$-ln($  $) - c$ $\sim \theta_2 + \text{Gumbel}(0)$

$\ldots$

$-ln($  $) - c$ $\sim \theta_N + \text{Gumbel}(0)$

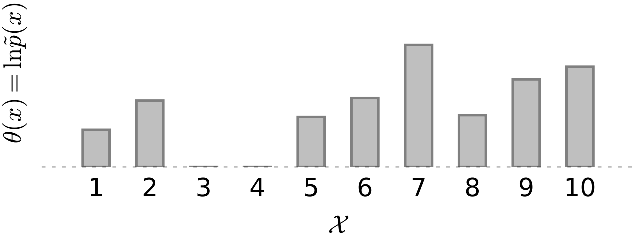
1. Probability q_i of $\tilde{\theta}_i$ being the maximal perturbed potential:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j} = \frac{\tilde{p}_i}{\sum_{j=1}^{N} \tilde{p}_j} = \frac{\tilde{p}_i}{Z} = p_i$$

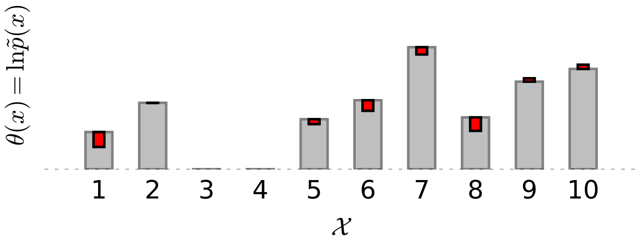
2. Maximal perturbed potential:

$$-\ln T - c \sim \ln Z + \text{Gumbel}(0)$$
$$\mathbb{E}[-lnT - c] = \ln Z$$

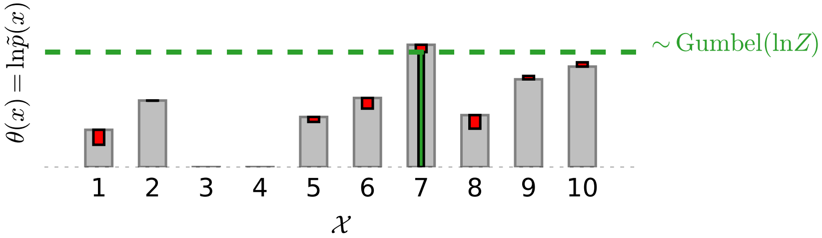
Classical Gumbel trick



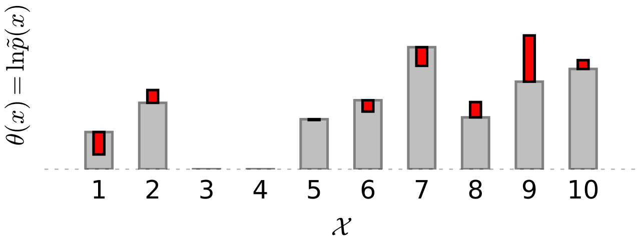
Classical Gumbel trick



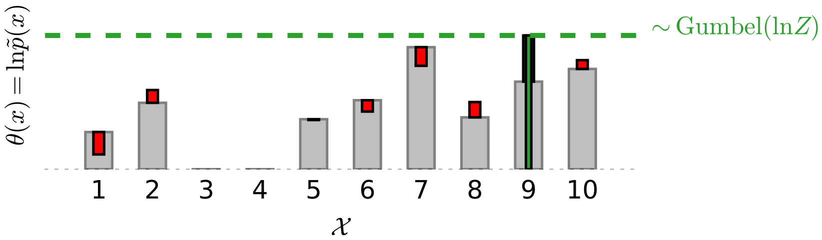
Classical Gumbel trick



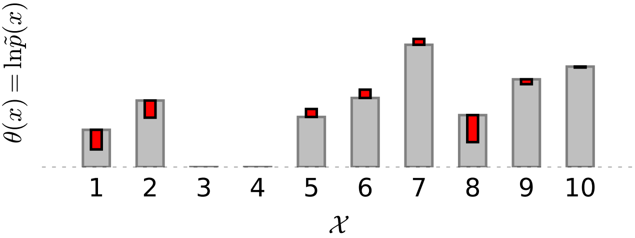
Classical Gumbel trick



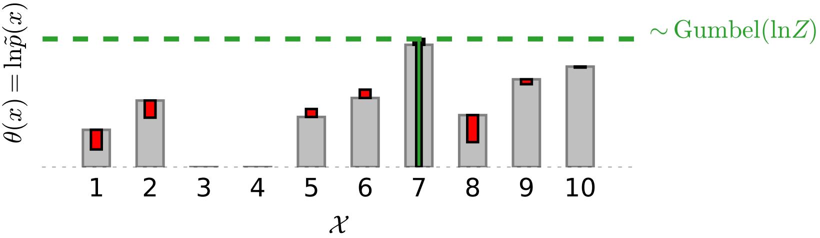
Classical Gumbel trick



Classical Gumbel trick



Classical Gumbel trick



Generalizing the Gumbel trick

Recall:

$$\text{Competing exponential clocks} \xrightarrow{-\ln x - c} \text{Gumbel trick}$$

$$\text{Exp}(Z) \text{ random var.} \xrightarrow{-\ln x - c} \ln Z + \text{Gumbel}(0) \text{ random var.}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{-\ln x - c} \text{Estimator of } \ln Z$$

Generalizing the Gumbel trick

Recall:

$$\text{Competing exponential clocks} \xrightarrow{-\ln x - c} \text{Gumbel trick}$$

$$\text{Exp}(Z) \text{ random var.} \xrightarrow{-\ln x - c} \ln Z + \text{Gumbel}(0) \text{ random var.}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{-\ln x - c} \text{Estimator of } \ln Z$$

Unbiased estimator of $\ln Z$:

$$\widehat{\ln Z} = \frac{1}{M} \sum_{m=1}^{M} G_m \qquad \text{where} \qquad G_1, \dots, G_M \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(\ln Z)$$

Generalizing the Gumbel trick

Recall:

$$\text{Competing exponential clocks} \xrightarrow{-\ln x - c} \text{Gumbel trick}$$

$$\text{Exp}(Z) \text{ random var.} \xrightarrow{-\ln x - c} \ln Z + \text{Gumbel}(0) \text{ random var.}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{-\ln x - c} \text{Estimator of } \ln Z$$

Unbiased estimator of $\ln Z$:

$$\widehat{\ln Z} = \frac{1}{M} \sum_{m=1}^{M} G_m \qquad \text{where} \qquad G_1, \ldots, G_M \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(\ln Z)$$

Biased estimator of Z: $\hat{Z} = \exp \widehat{\ln Z}$.

Generalizing the Gumbel trick

Recall:

$$\text{Competing exponential clocks} \xrightarrow{-\ln x - c} \text{Gumbel trick}$$

$$\text{Exp}(Z) \text{ random var.} \xrightarrow{-\ln x - c} \ln Z + \text{Gumbel}(0) \text{ random var.}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{-\ln x - c} \text{Estimator of } \ln Z$$

Unbiased estimator of $\ln Z$:

$$\widehat{\ln Z} = \frac{1}{M} \sum_{m=1}^{M} G_m \qquad \text{where} \qquad G_1, \dots, G_M \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(\ln Z)$$

Biased estimator of Z: $\hat{Z} = \exp \widehat{\ln Z}$.

How about applying different functions to the clocks?

Example: Weibull tricks

Unnormalized probability mass function $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_N)$:

$\sim \text{Exp}(\tilde{p}_1)$

$\sim \text{Exp}(\tilde{p}_2)$

$\ldots$

$\sim \text{Exp}(\tilde{p}_N)$

1. Probability q_i that clock i is the first to ring:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^N \lambda_j} = \frac{\tilde{p}_i}{\sum_{j=1}^N \tilde{p}_j} = \frac{\tilde{p}_i}{Z} = p_i$$

2. First time T when a clock rings:

$$T \sim \text{Exp}(\lambda_1 + \cdots + \lambda_N) = \text{Exp}(Z)$$
$$\mathbb{E}[T] = \frac{1}{\lambda_1 + \cdots + \lambda_N} = \frac{1}{Z}$$

Example: Weibull tricks

Unnormalized probability mass function $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_N)$:

$\Big($  $\Big)^{\alpha} \sim \frac{1}{\tilde{p}_1^{\alpha}}\,\mathrm{Weibull}(1, \frac{1}{\alpha})$

$\Big($  $\Big)^{\alpha} \sim \frac{1}{\tilde{p}_2^{\alpha}}\,\mathrm{Weibull}(1, \frac{1}{\alpha})$

$\ldots$

$\Big($  $\Big)^{\alpha} \sim \frac{1}{\tilde{p}_N^{\alpha}}\,\mathrm{Weibull}(1, \frac{1}{\alpha})$

1. Probability q_i that clock i is the first to ring:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^N \lambda_j} = \frac{\tilde{p}_i}{\sum_{j=1}^N \tilde{p}_j} = \frac{\tilde{p}_i}{Z} = p_i$$

2. First time T when a clock rings:

$$T \sim \mathrm{Exp}(\lambda_1 + \cdots + \lambda_N) = \mathrm{Exp}(Z)$$
$$\mathbb{E}[T] = \frac{1}{\lambda_1 + \cdots + \lambda_N} = \frac{1}{Z}$$

Example: Weibull tricks

Unnormalized probability mass function $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_N)$:

$\Big($  $\Big)^{\alpha}$ $\sim \frac{1}{\tilde{p}_1^{\alpha}} \text{Weibull}(1, \frac{1}{\alpha})$

$\Big($  $\Big)^{\alpha}$ $\sim \frac{1}{\tilde{p}_2^{\alpha}} \text{Weibull}(1, \frac{1}{\alpha})$

$\ldots$

$\Big($  $\Big)^{\alpha}$ $\sim \frac{1}{\tilde{p}_N^{\alpha}} \text{Weibull}(1, \frac{1}{\alpha})$

1. Probability q_i of the minimum being at configuration i:

$$q_i = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j} = \frac{\tilde{p}_i}{\sum_{j=1}^{N} \tilde{p}_j} = \frac{\tilde{p}_i}{Z} = p_i$$

2. Transformed first ringing time:

$$T^{\alpha} \sim \frac{1}{Z^{\alpha}} \text{Weibull}(1, \frac{1}{\alpha})$$
$$\mathbb{E}[T^{\alpha}] = Z^{-\alpha} \Gamma(1 + \alpha)$$

Generalizing the Gumbel trick

$$\text{Competing exponential clocks} \xrightarrow{x^\alpha} \text{Weibull trick}$$

$$\text{Exp}(Z) \text{ random variable} \xrightarrow{x^\alpha} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha}) \text{ rand. variable}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{x^\alpha} \text{Estimator of } Z^{-\alpha}\Gamma(1+\alpha)$$

Generalizing the Gumbel trick

$$\text{Competing exponential clocks} \xrightarrow{x^\alpha} \text{Weibull trick}$$

$$\text{Exp}(Z) \text{ random variable} \xrightarrow{x^\alpha} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha}) \text{ rand. variable}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{x^\alpha} \text{Estimator of } Z^{-\alpha}\Gamma(1+\alpha)$$

Unbiased estimator of $Z^{-\alpha}$:

$$\widehat{Z^{-\alpha}} = \frac{1}{\Gamma(1+\alpha)} \frac{1}{M} \sum_{m=1}^{M} W_m \quad \text{where} \quad \{W_m\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha})$$

Generalizing the Gumbel trick

$$\text{Competing exponential clocks} \xrightarrow{x^\alpha} \text{Weibull trick}$$

$$\text{Exp}(Z) \text{ random variable} \xrightarrow{x^\alpha} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha}) \text{ rand. variable}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{x^\alpha} \text{Estimator of } Z^{-\alpha}\Gamma(1+\alpha)$$

Unbiased estimator of $Z^{-\alpha}$:

$$\widehat{Z^{-\alpha}} = \frac{1}{\Gamma(1+\alpha)} \frac{1}{M} \sum_{m=1}^{M} W_m \quad \text{where} \quad \{W_m\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha})$$

$$= \frac{1}{\Gamma(1+\alpha)} \frac{1}{M} \sum_{m=1}^{M} e^{-\alpha(G_m + c)} \quad \text{where} \quad \{G_m\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(\ln Z)$$

Generalizing the Gumbel trick

$$\text{Competing exponential clocks} \xrightarrow{x^\alpha} \text{Weibull trick}$$

$$\text{Exp}(Z) \text{ random variable} \xrightarrow{x^\alpha} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha}) \text{ rand. variable}$$

$$\text{Estimator of } \frac{1}{Z} \xrightarrow{x^\alpha} \text{Estimator of } Z^{-\alpha}\Gamma(1+\alpha)$$

Unbiased estimator of $Z^{-\alpha}$:

$$\widehat{Z^{-\alpha}} = \frac{1}{\Gamma(1+\alpha)} \frac{1}{M} \sum_{m=1}^{M} W_m \quad \text{where} \quad \{W_m\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} \text{Weibull}(Z^{-\alpha}, \frac{1}{\alpha})$$

$$= \frac{1}{\Gamma(1+\alpha)} \frac{1}{M} \sum_{m=1}^{M} e^{-\alpha(G_m + c)} \quad \text{where} \quad \{G_m\}_{m=1}^{M} \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(\ln Z)$$

Biased estimator of Z: $\hat{Z} = \left(\widehat{Z^{-\alpha}}\right)^{-1/\alpha}$.

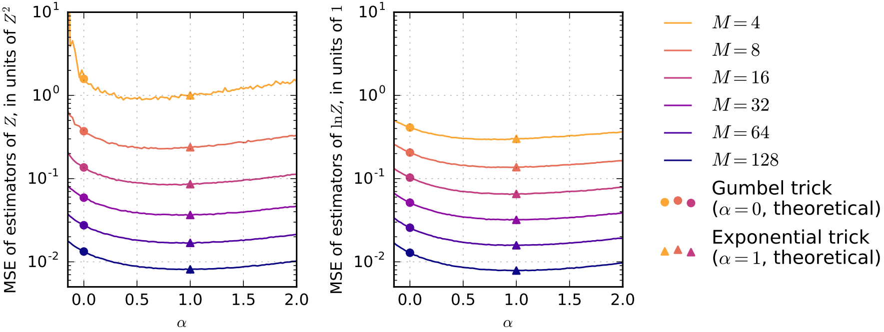
Generalizing the Gumbel trick

Trick	$g(x)$	Mean $f(Z)$
Gumbel	$-\ln x - c$	$\ln Z$
Exponential	x	$\frac{1}{Z}$
Weibull α	x^{α}, $\alpha > 0$	$Z^{-\alpha}\Gamma(1+\alpha)$
Fréchet α	x^{α}, $\alpha \in (-1, 0)$	$Z^{-\alpha}\Gamma(1+\alpha)$
Pareto	e^{x}	$\frac{Z}{Z-1}$ for $Z > 1$
Tail t	$\mathbb{1}_{\{x>t\}}$	e^{-tZ}

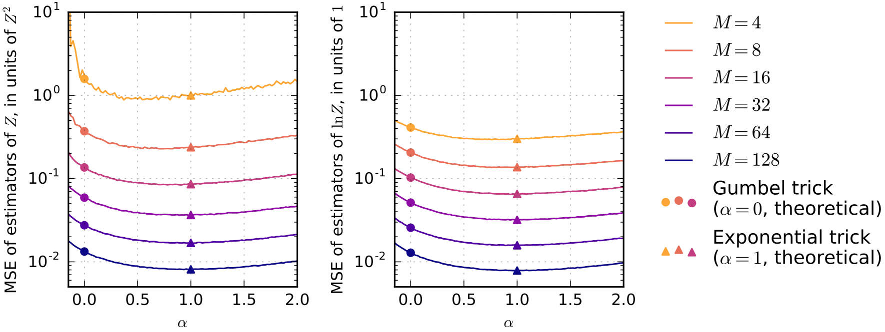
Generalizing the Gumbel trick

Trick	$g(x)$	Mean $f(Z)$	Asymptotic var. of $\hat{Z}$
Gumbel	$-\ln x - c$	$\ln Z$	$\frac{\pi^2}{6} Z^2$
Exponential	x	$\frac{1}{Z}$	Z^2
Weibull α	x^α, $\alpha > 0$	$Z^{-\alpha}\Gamma(1+\alpha)$	$\frac{1}{\alpha^2}\left(\frac{\Gamma(1+2\alpha)}{\Gamma(1+\alpha)^2} - 1\right) Z^2$
Fréchet α	x^α, $\alpha \in (-1, 0)$	$Z^{-\alpha}\Gamma(1+\alpha)$	$\frac{1}{\alpha^2}\left(\frac{\Gamma(1+2\alpha)}{\Gamma(1+\alpha)^2} - 1\right) Z^2$
Pareto	e^x	$\frac{Z}{Z-1}$ for $Z > 1$	$\frac{Z^2}{(Z-2)^2}$
Tail t	$\mathbb{1}_{\{x>t\}}$	e^{-tZ}	$\frac{(1-e^{-tZ})^2}{t^2}$

Comparing tricks



Comparing tricks



Takeaway:

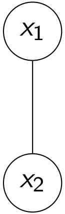
Save up to 40% samples by using the Exponential trick instead of the Gumbel trick. Applications:

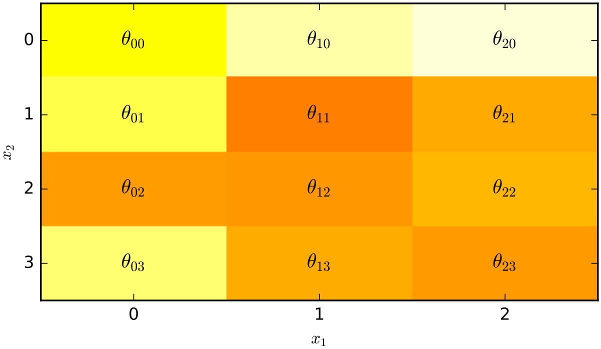
- *Large-scale discrete sampling* of [Chen and Ghahramani, 2016]
- *A* sampling* of [Maddison et al., 2014]

Low-rank perturbations

Discrete graphical model with n variables: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Example
$n = 2$

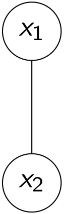


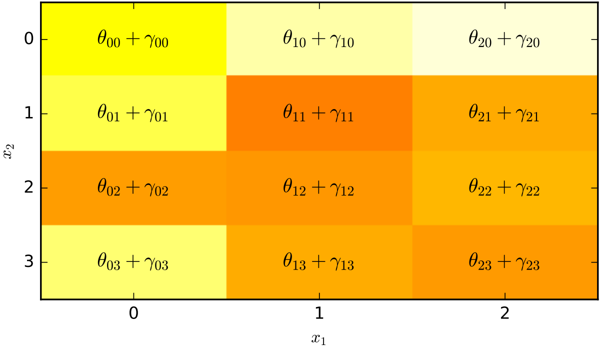


Low-rank perturbations

Discrete graphical model with n variables: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Example
$n = 2$

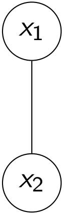


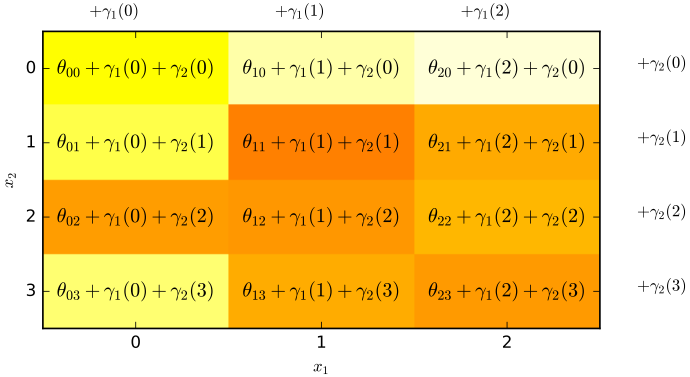


Low-rank perturbations

Discrete graphical model with n variables: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Example
$n = 2$

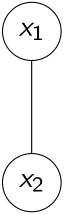


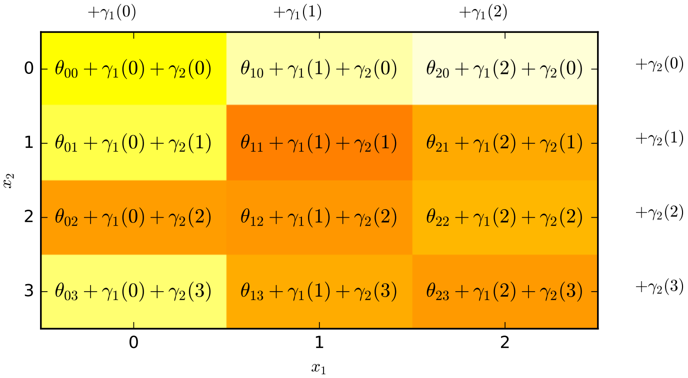


Low-rank perturbations

Discrete graphical model with n variables: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Example
$n = 2$



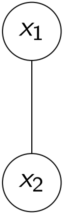


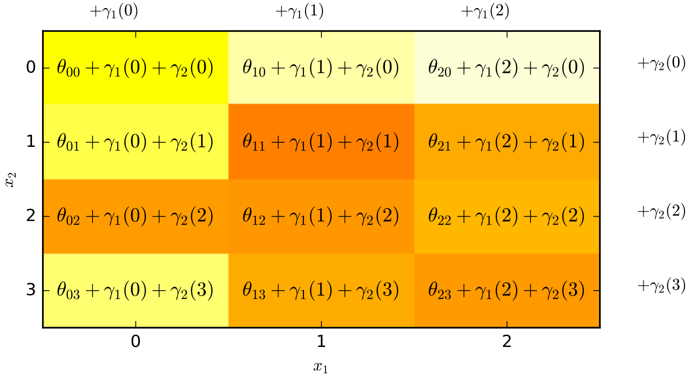
$$U := \max_{\mathbf{x} \in \mathcal{X}} \left\{ \theta(\mathbf{x}) + \sum_{i=1}^{n} \gamma_i(x_i) \right\}$$

Low-rank perturbations

Discrete graphical model with n variables: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Example
$n = 2$



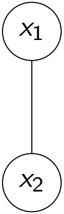


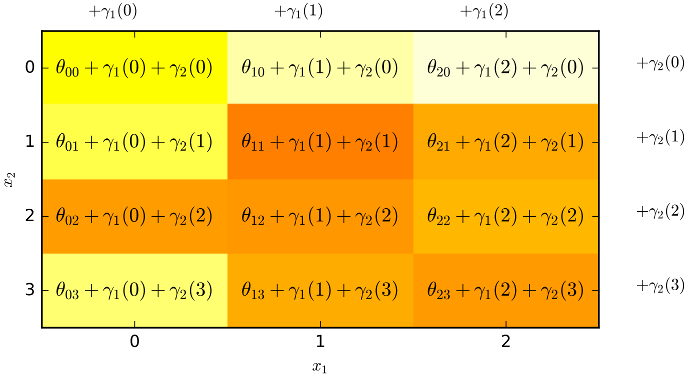
$$\ln Z \le \mathbb{E}[U]$$

Low-rank perturbations

Discrete graphical model with n variables: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.

Example
$n = 2$

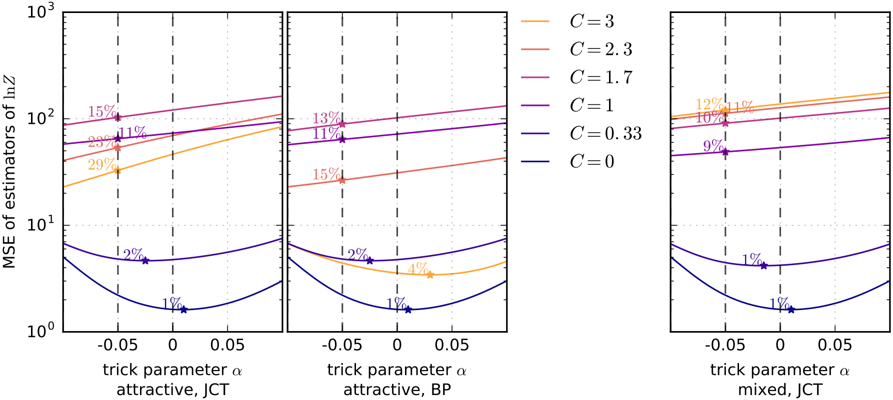




Gumbel trick: $\ln Z \leq \mathbb{E}[U]$

Fréchet and Weibull tricks: $\ln Z \leq C(\alpha, n) - \frac{1}{\alpha} \ln \mathbb{E}_\gamma \left[e^{-\alpha U} \right]$

Low-rank perturbations



Also in the paper

Low-rank perturbations with the Frechét, Weibull and Exponential tricks:

1. **Clamping** never hurts $\ln Z$ estimation using any of the Fréchet or Weibull upper bounds $\mathcal{U}(\alpha)$.
2. **Sequential samplers** for the Gibbs distribution p.
3. **Lower bounds** $\ln Z \geq \mathcal{L}(\alpha)$, where

$$\mathcal{L}(\alpha) := c + \frac{\ln \Gamma(1+\alpha)}{\alpha} - \frac{1}{n\alpha} \ln \mathbb{E}\left[\exp\left(-n\alpha L\right)\right].$$

4. **Advantages** of the Gumbel trick: theory such as

$$\underbrace{(\mathcal{U}(0) - \ln Z)}_{\text{error in } \ln Z \text{ bound}} + \underbrace{\mathrm{KL}(x^* \parallel p)}_{\text{sampling error}} = \underbrace{\mathbb{E}_{\gamma_i}\left[\gamma_i(x_i^*)\right] - H(x^*)}_{\text{error in entropy estimation}}.$$

Thank you

Matej Balog
University of Cambridge & MPI-IS Tübingen

Nilesh Tripuraneni
UC Berkeley

Zoubin Ghahramani
University of Cambridge & Uber AI Labs

Adrian Weller
University of Cambridge & Alan Turing Institute

Poster: Tuesday (today) 18:30–22:00 at Gallery #**117**

Slides: http://matejbalog.eu/en/research/
Code: https://github.com/matejbalog/gumbel-relatives