Free Energy Rates for Parameter Rich Optimization Problems

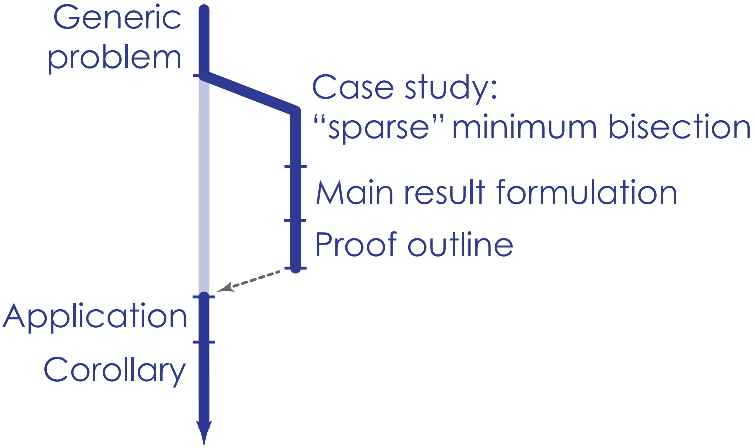
Joachim M. Buhmann[1], Julien Dumazert[1],
Alexey Gronskiy[1], Wojciech Szpankowski[2]

[1]ETH Zurich, {jbuhmann, alexeygr}@inf.ethz.ch,
julien.dumazert@gmail.com,

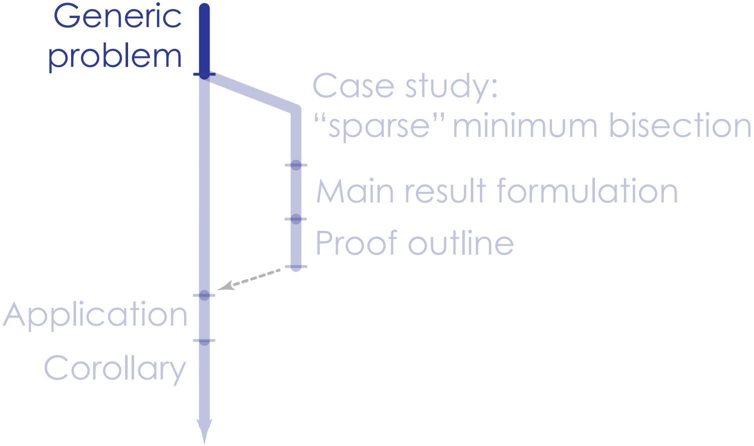
[2]Purdue University, spa@cs.purdue.edu

AofA Workshop
Princeton, 20th June 2017

Roadmap


Generic problem
Case study:
“sparse” minimum bisection
Main result formulation
Proof outline
Application
Corollary

Generic Problem



Generic Problem: Gibbs Sampling

- We search for solutions:
  $X \to c$

Generic Problem: Gibbs Sampling

- We search for solutions:
  $X \to c$

- Input X (e. g. a graph) is random!

Generic Problem: Gibbs Sampling

- We search for solutions:
  $X \rightarrow c$

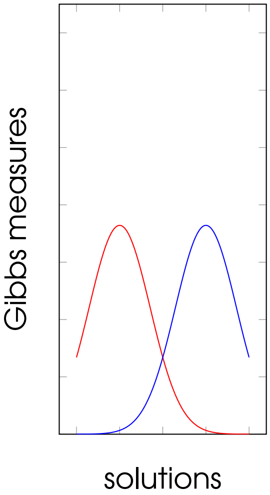
- Input X (e. g. a graph) is random!

- Want solution c still to be stable

  $$\Downarrow$$

  define a Maximum Entropy Gibbs measure over c:

  $$p(c|X) \propto \exp(-\beta \cdot \mathrm{cost}(c,X))$$



Generic Problem: "free energy"

Now

$$p(c|X) = \exp\big(-\beta \cdot \text{cost} - \mathcal{F}(X)\big),$$

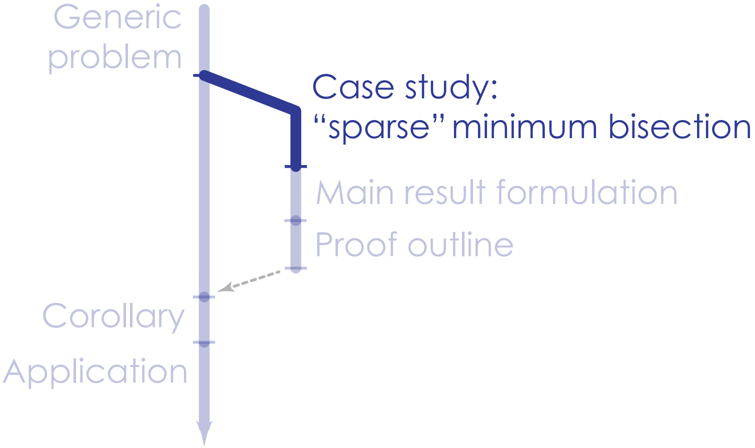
where the following is **free energy**:

$$\mathcal{F}(X) = \log Z(X) \quad \text{(here } Z(X) \text{ is } \textbf{partition function}\text{)}.$$

Goal: compute expected "free energy" — **hard**:

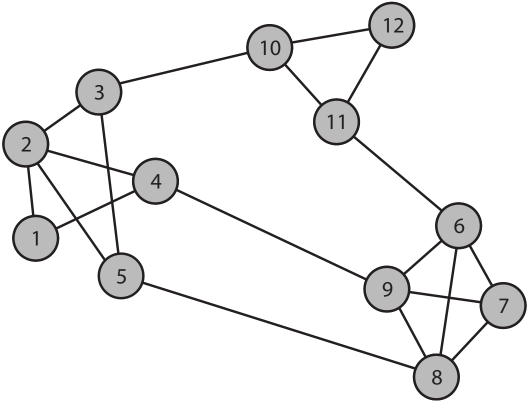
$$\mathbb{E}_X \mathcal{F}(X) = \mathbb{E}_X \log Z(X).$$

Case Study: “Sparse” Minimum Bisection



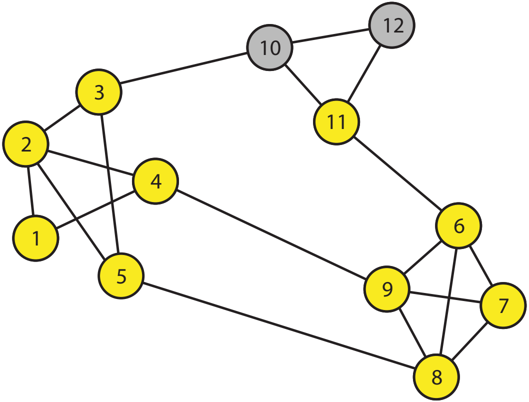
Case Study: “Sparse” Minimum Bisection

Find two subgraphs of some small size d with minimum edge cost between them.



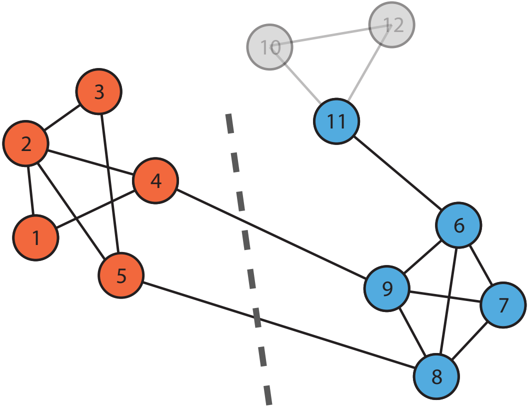
Case Study: "Sparse" Minimum Bisection

Find two subgraphs of some small size d with minimum edge cost between them.



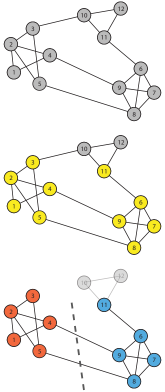
Case Study: “Sparse” Minimum Bisection

Find two subgraphs of some small size d with minimum edge cost between them.



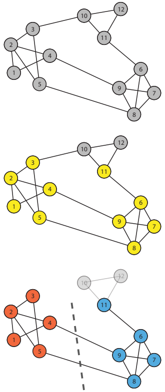
Case Study: "Sparse" Minimum Bisection

- flashback to "Generic Problem": in high dimensions may require solving via sampling!



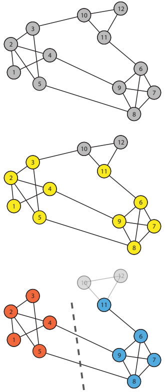
Case Study: "Sparse" Minimum Bisection

- flashback to "Generic Problem": in high dimensions may require solving via sampling!
- solutions **are dependent** on each other — not Random Energy Model (next slide)!



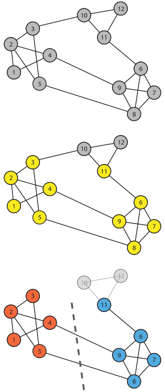
Case Study: "Sparse" Minimum Bisection

- flashback to "Generic Problem": in high dimensions may require solving via sampling!
- solutions **are dependent** on each other — not Random Energy Model (next slide)!
- **contribution**: free energy asymptotics solved!



Case Study: “Sparse“ Minimum Bisection

- flashback to “Generic Problem“: in high dimensions may require solving via sampling!
- solutions **are dependent** on each other — not Random Energy Model (next slide)!
- **contribution**: free energy asymptotics solved!
- another case study Lawler Quadratic Assignment Problem solved as well: **(see the paper)**.



Free Energy: History and Advances

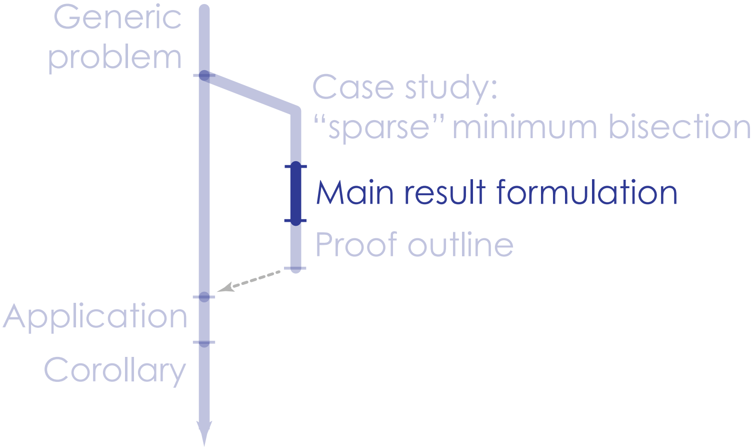
- **(Derrida'80)** established REM — no dependencies;

Free Energy: History and Advances

- **(Derrida'80)** established REM — no dependencies;
- **(Talagrand'02)**: systematization; simple proof of REM free energy phase transition and beyond:

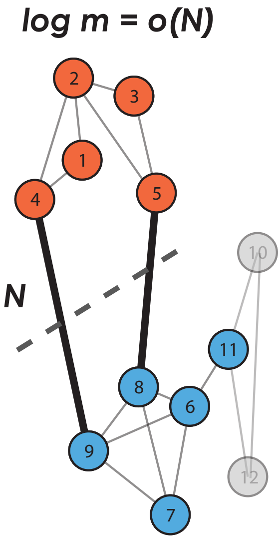
$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)]}{n} = \begin{cases} \frac{\beta^2}{4} + \log 2 & \beta < 2\sqrt{\log 2}, \\ \beta\sqrt{\log 2} & \beta \geq 2\sqrt{\log 2}. \end{cases}$$

Main Result


Generic problem
Case study: "sparse" minimum bisection
Main result formulation
Proof outline
Application
Corollary

Main Result: Setting

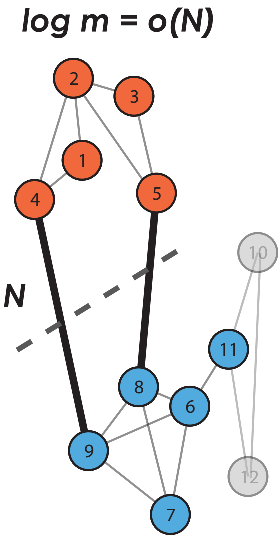
- m is number of solutions



Main Result: Setting

- m is number of solutions

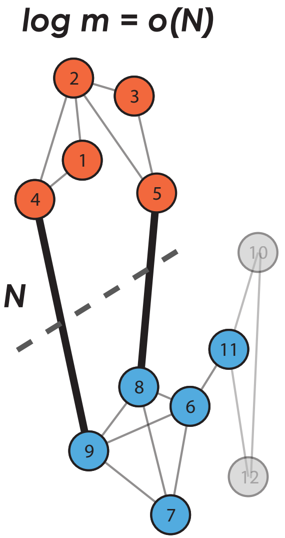
- N is size of one solution (number of parameters)



Main Result: Setting

- m is number of solutions
- N is size of one solution (number of parameters)
- require to be "parameter rich"

$$\log m = o(N)$$



Main Result: 2^{nd} Order Phase Transition

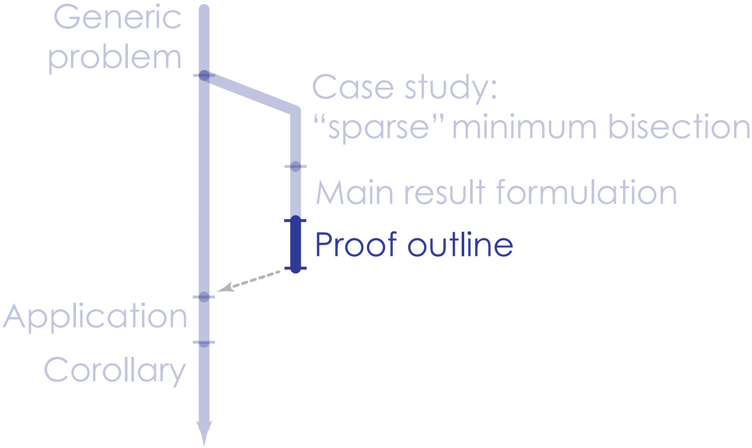
Main theorem

Consider sparse MBP on a complete graph; edge weights mutually independent within any given solution; have mean μ and variance σ^2. Then, the following holds:

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z] + \hat{\beta}\mu\sqrt{N\log m}}{\log m} = \begin{cases} 1 + \frac{\hat{\beta}^2\sigma^2}{2}, & \hat{\beta} < \frac{\sqrt{2}}{\sigma}, \\ \hat{\beta}\sigma\sqrt{2}, & \hat{\beta} \geq \frac{\sqrt{2}}{\sigma} \end{cases}$$

provided $\log n \ll d \ll \frac{n^{2/7}}{\sqrt{\log n}}$.

Proof Outline



Proof Outline – I

- Introduce "solution overlap" D: average intersection of two bisections
- D is key to understand **dependencies** (remember we are no REM!)

Lemma 1

The following holds

$$\mathbb{E}_{\text{rand choice}} D = \mathcal{O}(d^4/n).$$

Compute dependencies

P(``Z close to EZ'') via Var Z

Breaking E log Z into two cond's

Final bounding via adjusting parameter

Proof Outline – II

- Introduce event A: happens when Z is close to $\mathbb{E}Z$, i. e. $A := \{Z \geq \epsilon \mathbb{E}Z\}$
- Goal is to compute $\mathbb{P}(A)$

Fact 1

$\mathbb{P}(A)$ can be bounded by Var$Z$ via Chebychev.

Lemma 2 (Buhmann et al., 2014)

Var$Z$ can be asymptotically approximated via $\mathbb{E}_{\text{rand choice}}D$:

$$\mathrm{Var}Z \sim (\mathbb{E}Z)^2 \left(\sigma^2 \beta^2 \mathbb{E}_{\text{rand choice}} D\right).$$

Compute dependencies

P(``Z close to EZ'') via Var Z

Breaking E log Z into two cond's

Final bounding via adjusting parameter

Proof Outline – III

- Break $\mathbb{E}\log Z$ into

$$\begin{aligned}\mathbb{E}\log Z &= \mathbb{E}[\log Z \mid A]\cdot\mathbb{P}(A) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})]\\ &\geq (\log\mathbb{E}Z + \log\epsilon)\mathbb{P}(A) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})]\end{aligned}$$

Fact 3

Can expand $\log\mathbb{E}Z$ via Taylor expansion (used assumptions of Theorem) and bound $\mathbb{P}(A)$ from previous.

Fact 4

$\mathbb{E}[\log Z\mathbb{1}(\bar{A})]$: enough to bound loosely

Compute dependencies

P(``Z close to EZ'') via Var Z

Breaking E log Z into two cond's

Final bounding via adjusting parameter

Proof Outline – IV

- Finally, the right choice of ϵ for two regimes of β gives the phase transition in **lower bound**

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z] + \cdots}{\cdots} > \begin{cases} 1 + \frac{\hat\beta^2\sigma^2}{2}, & \hat\beta < \frac{\sqrt{2}}{\sigma}, \\ \hat\beta\sigma\sqrt{2}, & \hat\beta \geq \frac{\sqrt{2}}{\sigma}. \end{cases}$$

- The **same** phase transition happens for **upper bound**, — easier to prove (no computing dependencies)

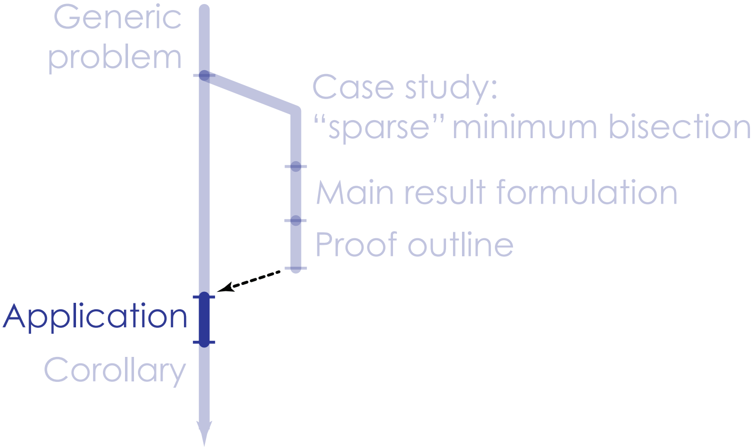
Compute dependencies

P(``Z close to EZ'') via Var Z

Breaking E log Z into two cond's

Final bounding via adjusting parameter

Possible Application



Application Setting: Gibbs Sampling Regularizer

- Remember our approach:

data input X $\longrightarrow$ posterior distribution $p_\beta(c|X)$ $\longrightarrow$ **sampling from $p_{\beta^*}(c|X)$**

Application Setting: Gibbs Sampling Regularizer

- Remember our approach:

data input X $\longrightarrow$ posterior distribution $p_\beta(c|X)$ $\longrightarrow$ $\boxed{\text{sampling from } p_{\beta^*}(c|X)}$

- **Q:** what is **regularizing** here?
  A: choice of β — essentially controls "width"

Application Setting: Gibbs Sampling Regularizer

- Remember our approach:

$$\begin{array}{c} \text{data input} \\ X \end{array} \longrightarrow \begin{array}{c} \text{posterior distribution} \\ p_\beta(c|X) \end{array} \longrightarrow \boxed{\begin{array}{c} \text{sampling from} \\ p_{\beta^*}(c|X) \end{array}}$$

- **Q:** what is **regularizing** here?
  A: choice of β — essentially controls "width"
- **Q:** how?
  A: maximize **expected log-convolution**:

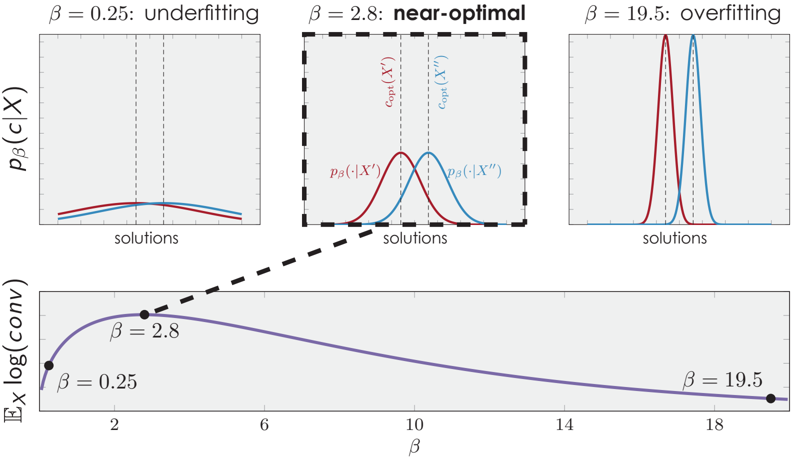
$$\beta^* = \arg\max_\beta \mathbb{E}_{X',X''}\left[\log \sum_c p_\beta(c|X')p_\beta(c|X'')\right]$$

Application Setting: Intuition

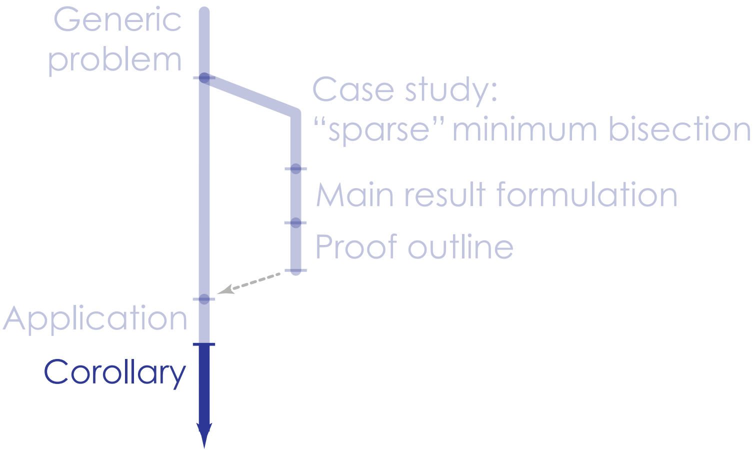
- Log-convolution is “almost” cross entropy;

Application Setting: Intuition

- Log-convolution is "almost" cross entropy;
- It stabilizes solution output:



Corollary and Application



Corollary: Log-Convolution

- The log-convolution score can be rewritten:

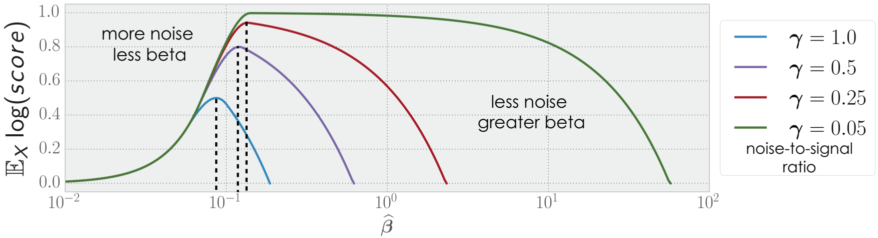
$$\mathbb{E} \log \sum_c p_\beta(c|X') p_\beta(c|X'') = \underbrace{\mathbb{E} \log Z(\beta, X' + X'')}_{\text{Thm}} - 2 \underbrace{\mathbb{E} \log Z(\beta, X)}_{\text{Thm}}$$

Corollary: Log-Convolution

- The log-convolution score can be rewritten:

$$\mathbb{E} \log \sum_c p_\beta(c|X')p_\beta(c|X'') = \underbrace{\mathbb{E} \log Z(\beta, X' + X'')}_{\text{Thm}} - 2\, \underbrace{\mathbb{E} \log Z(\beta, X)}_{\text{Thm}}$$

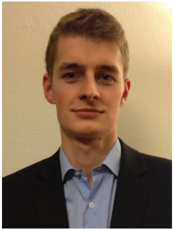
- Thus possible to analytically compute score:

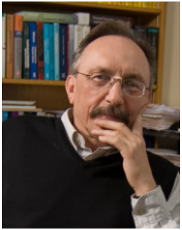


Conclusion

- We computed asymptotically precisely the value of $\mathbb{E} \log Z$ with small dependencies
- It has applications to model validation (submitted to J. of Theor. CS) in machine learning: method involves comparing two fluctuating Gibbs distributions
- Still not found a general way (and probably there is no such way)

Thanks for your attention!





References I

- E. T. Jaynes,
  Information Theory and Statistical Mechanics.
  Phys. Rev., 106, 1957, p. 620.
- W. Szpankowski,
  Combinatorial optimization problems for which almost every algorithm is asymptotically optimal.
  Optimization, 33, 1995, pp. 359–368.
- M. Talagrand.
  Spin Glasses: A Challenge for Mathematicians.
  Springer, NY, 2003.

References II

- J. Vannimenus and M. Mézard,
  On the Statistical Mechanics of Optimization Problems of the traveling Salesman Type.
  J. de Physique Lettres, 45, 1984, pp. 1145–1153.

- J. M. Buhmann, A. Gronskiy and W. Szpankowski
  Free Energy Rates for a Class of Combinatorial Optimization Problems
  AofA'14, Paris

- J. M. Buhmann, A. Gronskiy, J. Dumazert and W. Szpankowski
  Phase Transitions in Parameter Rich Optimization Problems
  SODA/ANALCO'17, Barcelona