Statistical Mechanics and Information Theory in Approximate Robust Inference

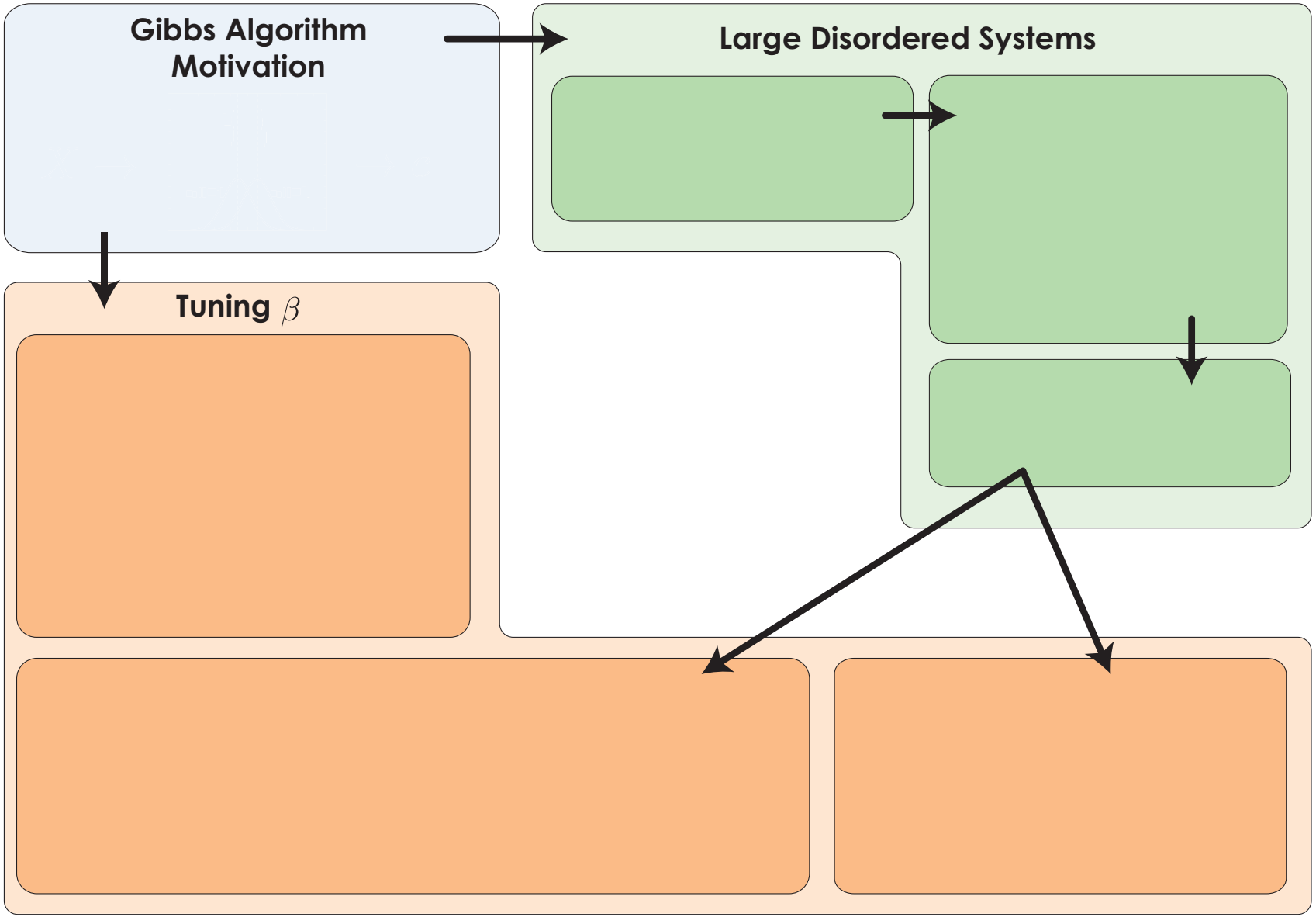
Doctoral Examination
Candidate: Alexey Gronskiy

Examination Committee:
Prof. Dr. Joachim M. Buhmann
Prof. Dr. Peter Widmayer
Prof. Dr. Wojciech Szpankowski

Chair: Prof. Dr. Onur Mutlu

ETH Zurich, D-INFK 6th March 2018

Roadmap


Gibbs Algorithm
Motivation
Large Disordered Systems
Tuning β

Regularizing by Gibbs Algorithm

Gibbs Algorithm
Motivation

Now Here

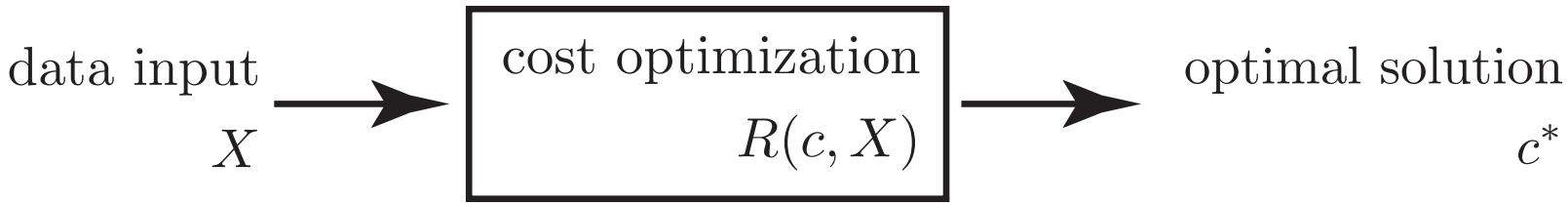
Generic Optimization Problem

As a motivation, we bring up a generic randomized optimization problem, where

- $X \in \mathcal{X}$ random data instance X as an element of set $\mathcal{X}$
- $c \in \mathcal{C}$ is solution c as element of solution set $\mathcal{C}$
- $R(c, X)$ is cost function of solution c given in a data instance X

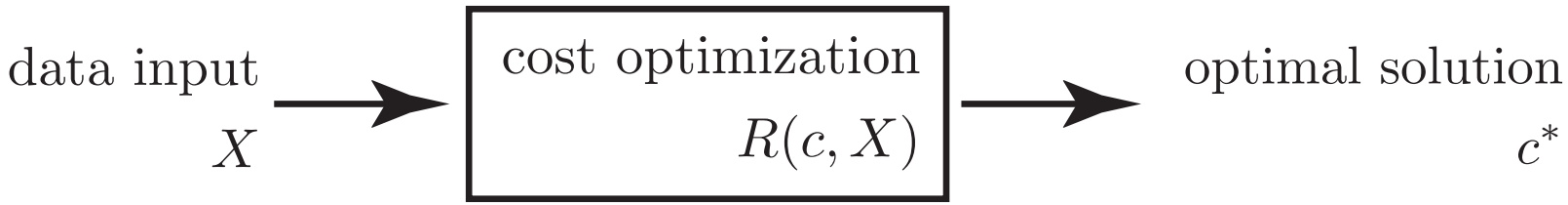
Regularizing by Gibbs Algorithm

- Recall possible standard approach (ERM):

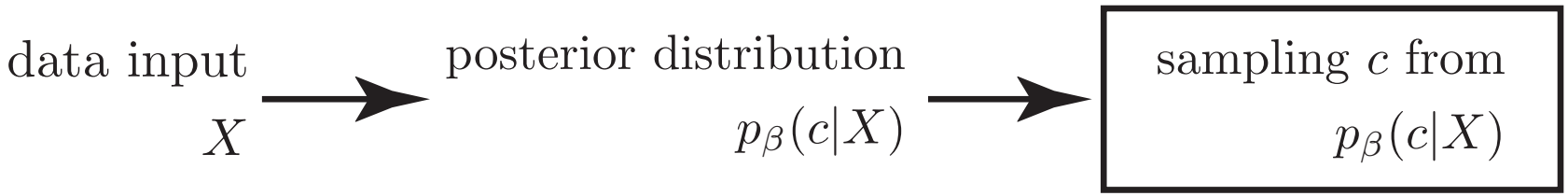


Regularizing by Gibbs Algorithm

- Recall possible standard approach (ERM):



- The approach we focus on:



Regularizer is inverse temperature β — controls "width"

Regularizing by Gibbs Algorithm

- We search for stochastic approximation:

$$X \longmapsto c \sim p(c|X)$$

Regularizing by Gibbs Algorithm

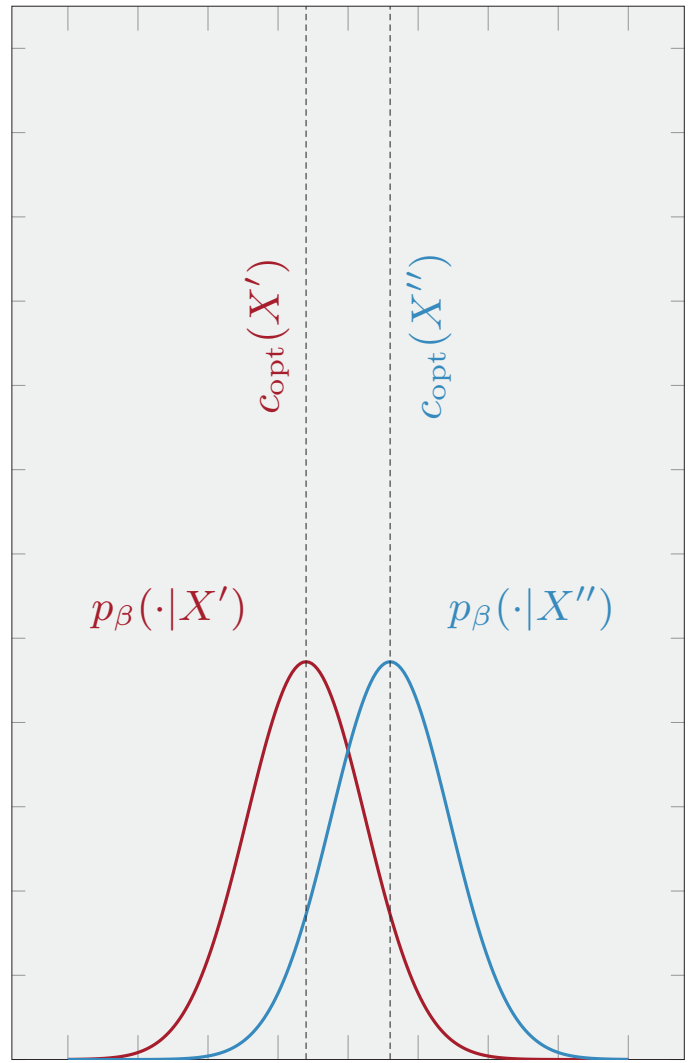
- We search for stochastic approximation:

$$X \longmapsto c \sim p(c|X)$$

- Define a **Gibbs posterior** over solutions:

$$\Downarrow$$

$$p_\beta(c|X) \propto \exp(-\beta \cdot R(c, X))$$



Motivation for Gibbs Distribution

- Represents the family of **maximum entropy** (1) distributions for the fixed expected costs:

$$p_\beta(c|X) \in \arg \max_{\substack{p(c|X):\\ \mathbb{E}[R(c,X)]=r}} H(p)$$

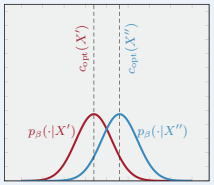
- Minimizes the expected risk, **regularized** by the input-output mutual information (2):

$$p_\beta(c|X) \in \arg \min_{p(c|X)} \left(\mathbb{E}[R(c,X)] + \frac{1}{\beta} I(c,X) \right)$$

(1) Jaynes (1957), *Information Theory and Statistical Mechanics*, Phys. Review, Ser. II
(2) Xu and Raginsky (2017), *Information-theoretic analysis of generalization capability of learning algorithms*, NIPS 2017

Large Disordered Systems

Gibbs Algorithm Motivation

$X \rightarrow$  $\rightarrow c$

Large Disordered Systems

Helmholtz Free Energy Motivation

Now Here

'In Vitro' Problem: sMBP

Main Result

Free Energy

Now

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

where the following is **Helmholtz free energy**:

$$\mathcal{F}(\beta, X) = \log Z(\beta, X),$$

here $Z(\beta, X)$ is **partition function**:

$$Z(\beta, X) = \sum_{c' \in \mathcal{C}} \exp(-\beta R(c', X))$$

Free Energy

Now

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

where the following is **Helmholtz free energy**:

$$\mathcal{F}(\beta, X) = \log Z(\beta, X),$$

here $Z(\beta, X)$ is **partition function**:

$$Z(\beta, X) = \sum_{c' \in \mathcal{C}} \exp(-\beta R(c', X))$$

Issue: Understanding the stochastic behavior of $\log Z(\beta, X)$ is known to be **hard**.

Statistical Mechanics of Free Energy

- Empirical quantity: $\log Z(\beta, X)$
- **Annealed** average (easy but "wrong"):

$$\mathcal{F}(\beta, X) = \log \mathbb{E}_X [Z(\beta, X)]$$

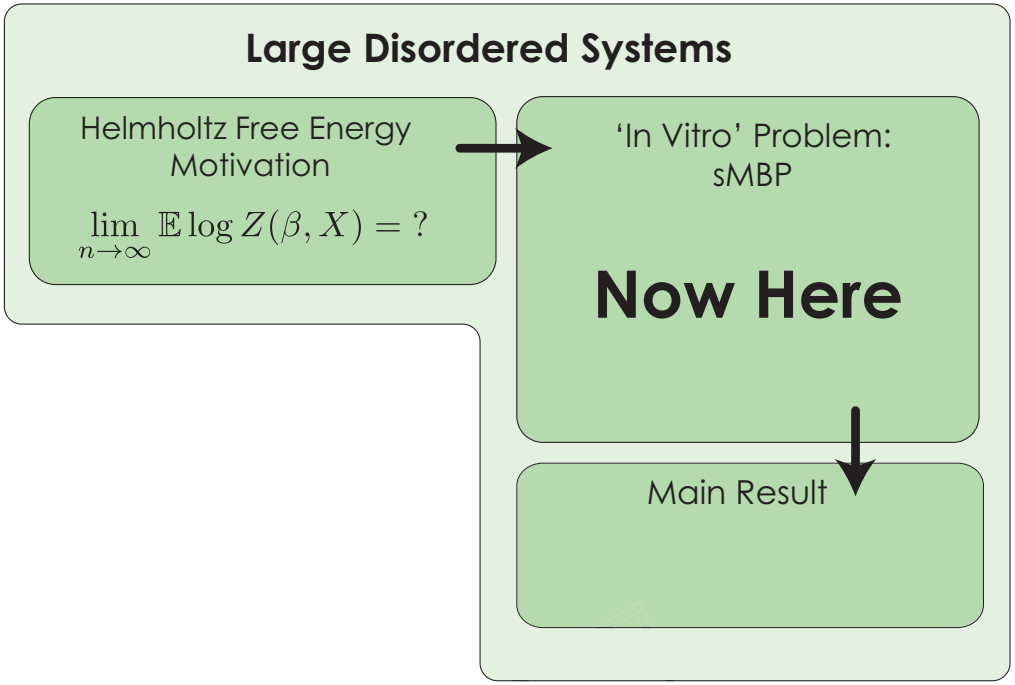
- **Quenched** average (hard but "correct"):

$$\mathcal{F}(\beta, X) = \mathbb{E}_X \left[\log Z(\beta, X)\right]$$

Our goal is to **asymptotically** study its quenched approximation.

$$\lim_{\text{size} \to \infty} \mathbb{E}_{\boldsymbol{X}} \log \boldsymbol{Z}(\boldsymbol{\beta}, \boldsymbol{X}) = \boldsymbol{?}$$

“In Vitro” Combinatorial Problem



"In Vitro" Combinatorial Problem: "Sparse" Minimum Bisection – I

- **Given:** complete graph with random edge weights

$$G = (V, E, X), \quad X = \{W_i\}_{i \in E}$$

"In Vitro" Combinatorial Problem: "Sparse" Minimum Bisection – I

- **Given:** complete graph with random edge weights

$$G = (V, E, X), \quad X = \{W_i\}_{i \in E}$$

- **Find:** two subgraphs

$$c = (U_1, U_2), \quad U_1 \sqcup U_2 \subsetneq V$$

of a small size d ("sparsity")

"In Vitro" Combinatorial Problem: "Sparse" Minimum Bisection – I

- **Given:** complete graph with random edge weights

$$G = (V, E, X), \quad X = \{W_i\}_{i \in E}$$

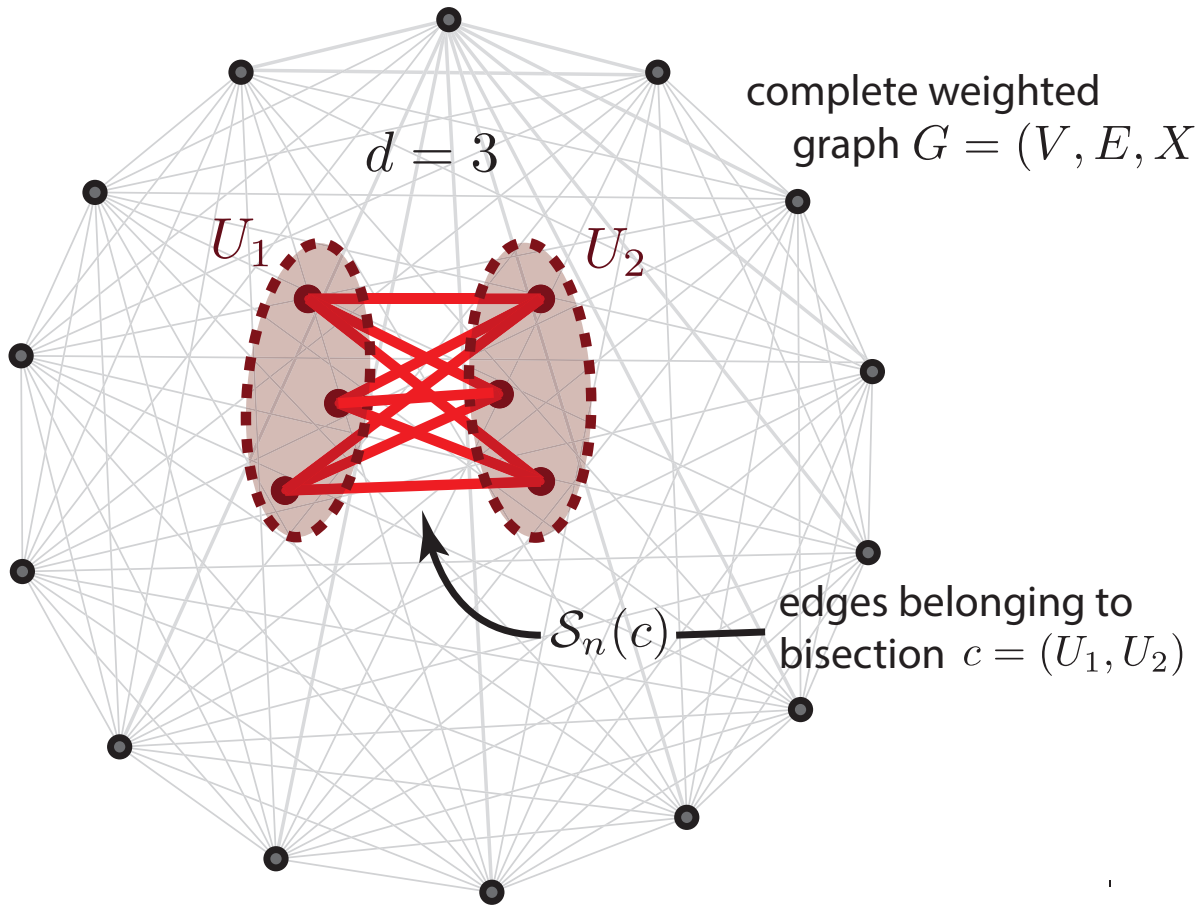
- **Find:** two subgraphs

$$c = (U_1, U_2), \quad U_1 \sqcup U_2 \subsetneq V$$

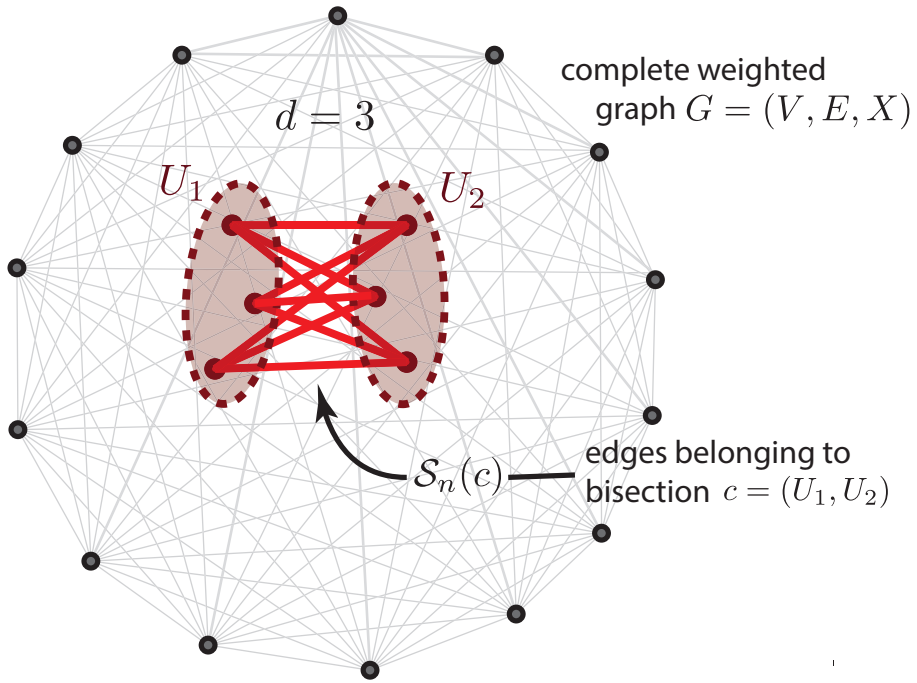
  of a small size d ("sparsity")
- Such that the **total edge cost** between them is minimal

$$\arg \min_{c=(U_1,U_2)} R(c, X) = \arg \min_{c=(U_1,U_2)} \sum_{i \in \text{edges}(c)} W_i$$

"In Vitro" Combinatorial Problem: "Sparse" Minimum Bisection – II



"In Vitro" Combinatorial Problem: "Sparse" Minimum Bisection – II



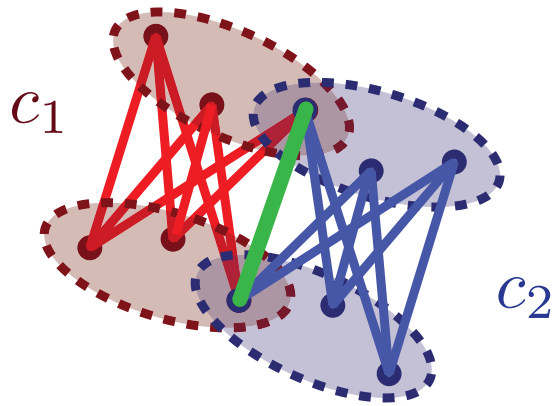
$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{X}} \log \boldsymbol{Z}(\boldsymbol{\beta}, \boldsymbol{X}) = ?$$

Free Energy: Short Overview – I

- Notorious difficulty of computing $\mathbb{E} \log Z$ is related (1)(2) to dependencies in solutions:

$$\mathrm{Cov}(R(c', X), R(c'', X)) \neq 0.$$

- In our case, the dependencies **exist** but are **small**.



- Large correlations require additional work (conjecture in appendix).

(1) Bovier (2012), *Statistical Mechanics of Disordered Systems: A Mathematical Perspective*, Cambridge
(2) Talagrand (2003), *Spin Glasses: A Challenge for Mathematicians*, Springer

Free Energy: Short Overview – II

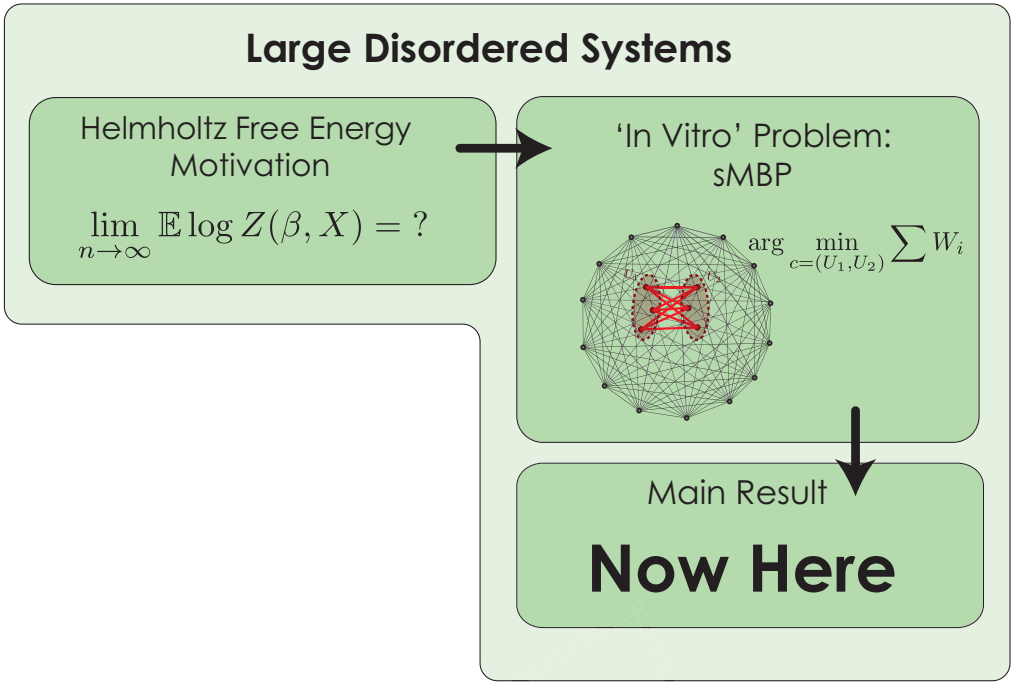
- Derrida (1) established Random Energy Model — a model where solution costs are independent;

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

- Vannimenus and Mézard (2) solved free energy for Traveling Salesman Problem: **small dependencies**.

(1) Derrida (1981), *Random-energy model: An exactly solvable model of disordered systems*, Phys. Rev. B 24
(2) Vannimenus and Mézard (1984), *On the Statistical Mechanics of Optimization Problems of the traveling Salesman Type*, J. de Physique Lettres, 45, 1984

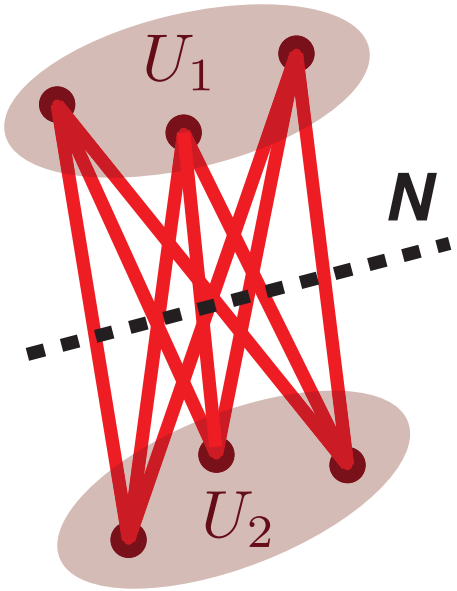
Main Result

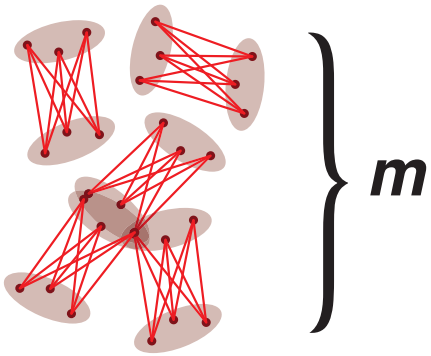

Large Disordered Systems
Helmholtz Free Energy Motivation
$\lim_{n\to\infty} \mathbb{E} \log Z(\beta, X) = ?$
'In Vitro' Problem: sMBP
$\arg \min_{c=(U_1,U_2)} \sum W_i$
Main Result
Now Here

Main Result: Setting

- N is # of solution parameters:

$$N := |\text{edges}(U_1, U_2)|$$





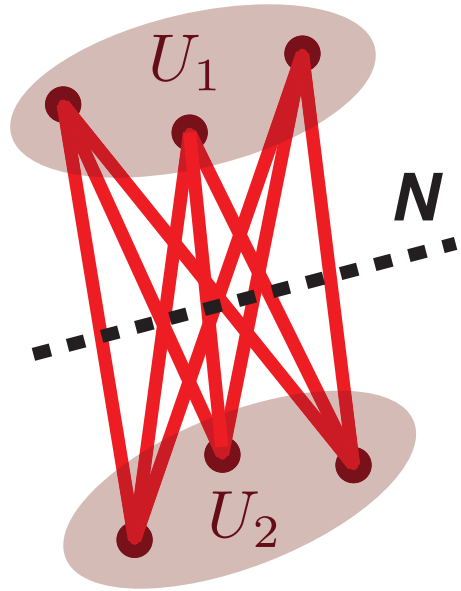
Main Result: Setting

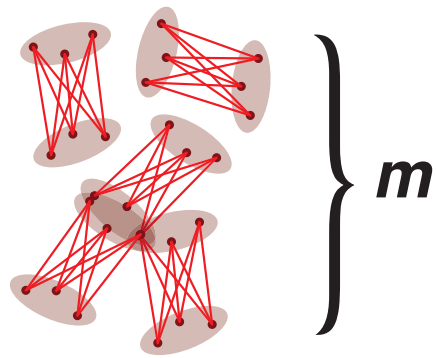
- N is $\#$ of solution parameters:

$$N := |\mathrm{edges}(U_1, U_2)|$$

- m is number of solutions:

$$m := |\mathcal{C}|$$





Main Result: Setting

- N is $\#$ of solution parameters:

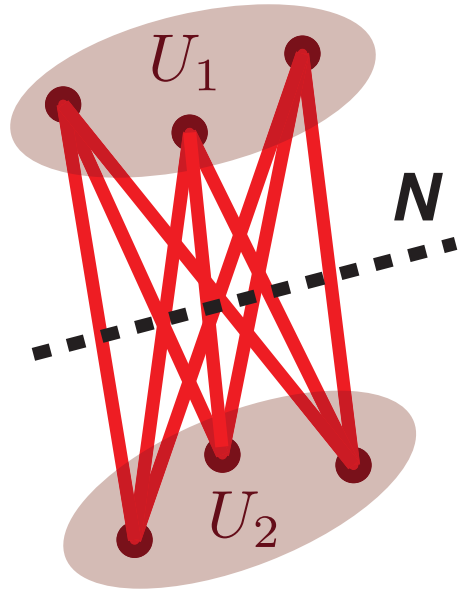
$$N := |\mathrm{edges}(U_1, U_2)|$$

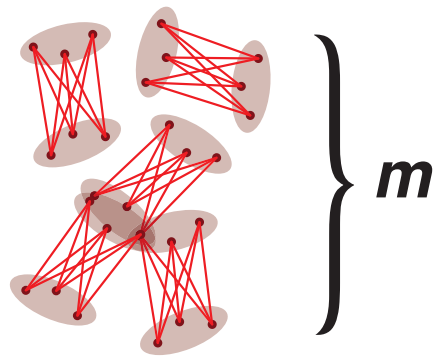
- m is number of solutions:

$$m := |\mathcal{C}|$$

- require to be "parameter rich":

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





Main Result

Main theorem: free energy asymptotics

Assume:

- Sparse MBP on a complete graph
- Edge weights mutually independent within any given c
- Weights mean μ and variance σ^2, MGF $< \infty$

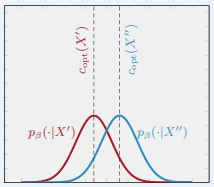
Then:

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

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

Tuning β for Approximation

Gibbs Algorithm Motivation

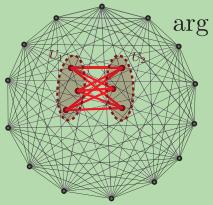
$X \rightarrow$  $\rightarrow c$

Large Disordered Systems

Helmholtz Free Energy Motivation

$$\lim_{n\to\infty} \mathbb{E} \log Z(\beta, X) = ?$$

'In Vitro' Problem: sMBP



$$\arg \min_{c=(U_1,U_2)} \sum W_i$$

Main Result

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

Tuning β

Motivation

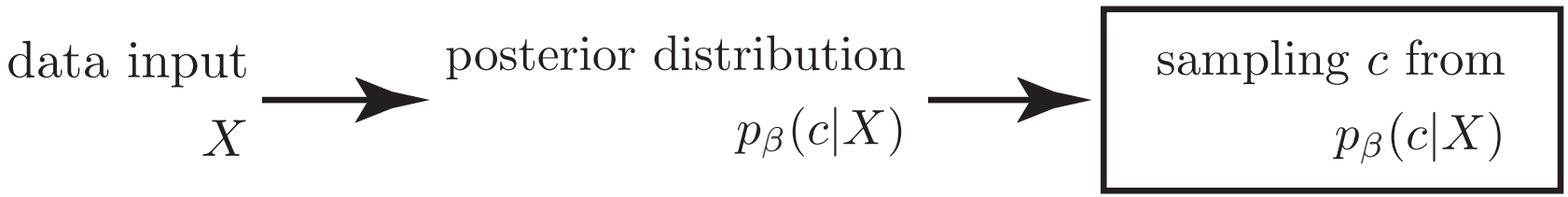
Now Here

Maximize Expected Log-Posterior Agreement (eLPA)

Minimize Average Gibbs Risk

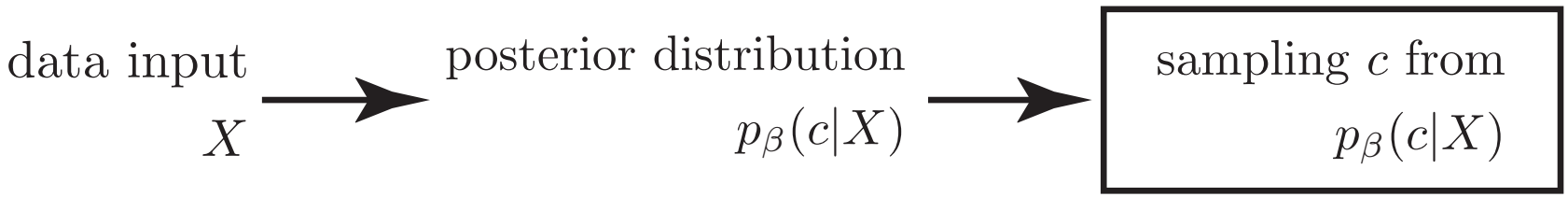
Role of β in Posterior Distribution

- Recall our approach:

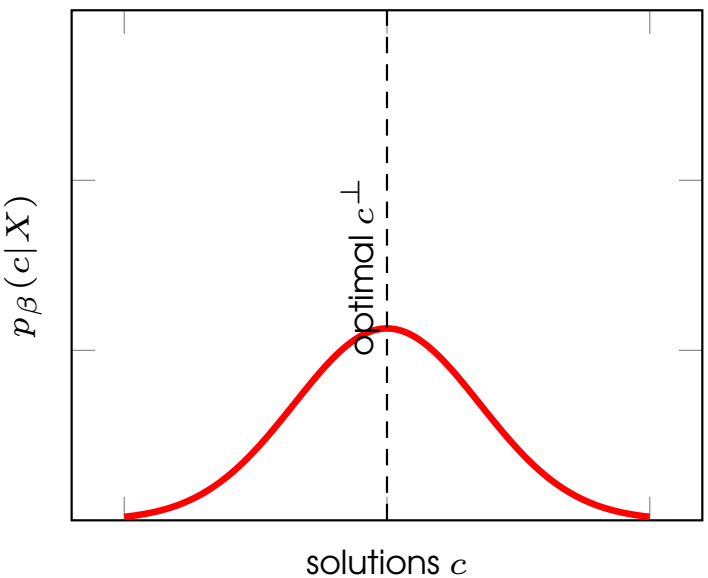


Role of β in Posterior Distribution

- Recall our approach:



Example (Gibbs Posterior)

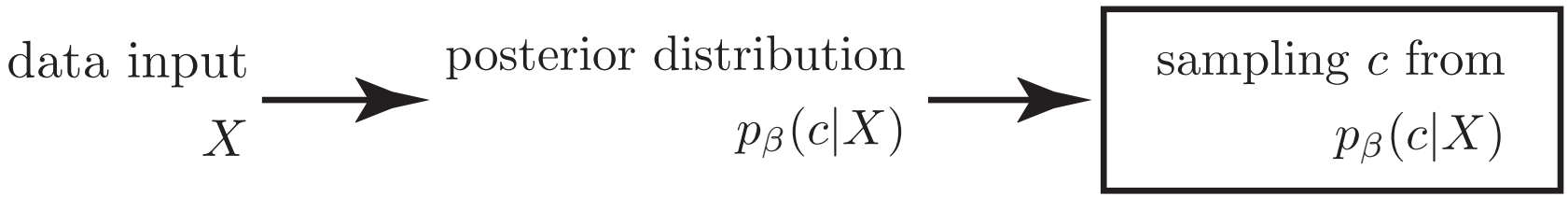


$$p_\beta(c|X) = \frac{\exp\big(-\beta R(c,X)\big)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\big(-\beta R(\tilde{c},X)\big)}$$

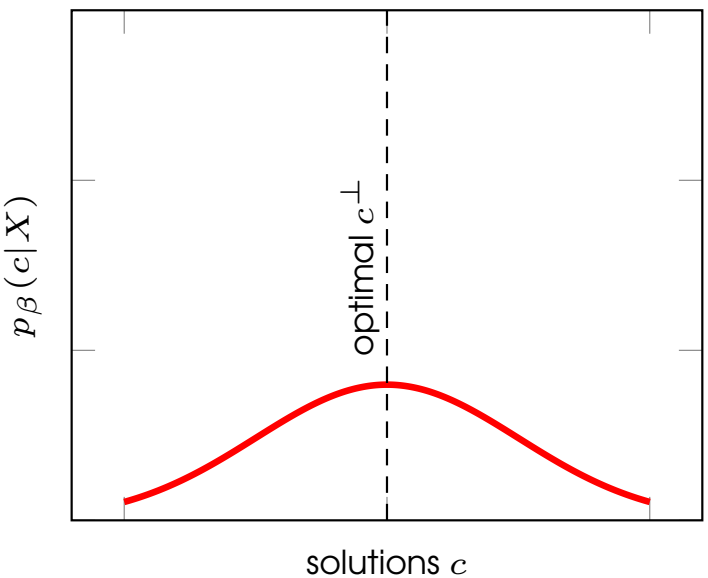
Picture: β is **decreasing**.

Role of β in Posterior Distribution

- Recall our approach:



Example (Gibbs Posterior)

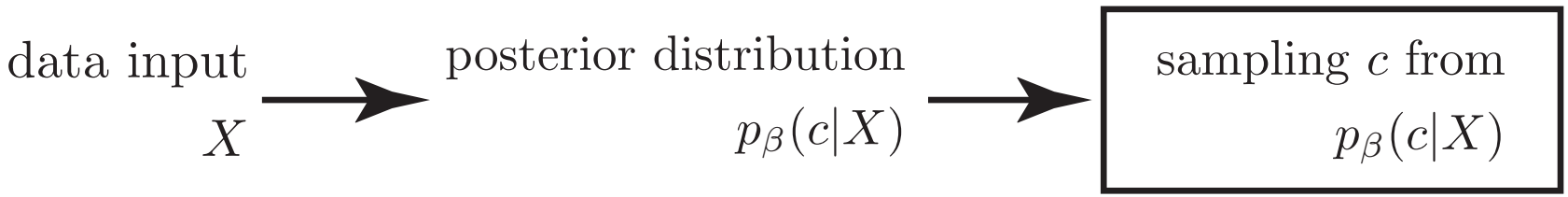


$$p_\beta(c|X) = \frac{\exp\big(-\beta R(c,X)\big)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\big(-\beta R(\tilde{c},X)\big)}$$

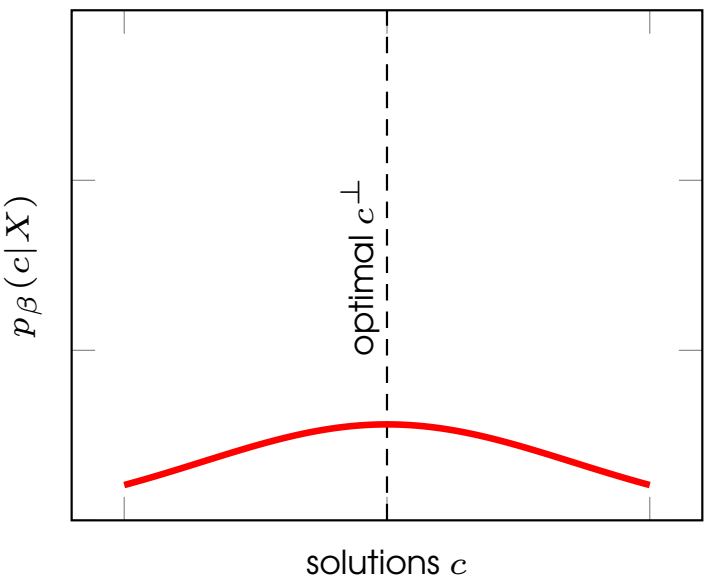
Picture: β is **decreasing**.

Role of β in Posterior Distribution

- Recall our approach:



Example (Gibbs Posterior)

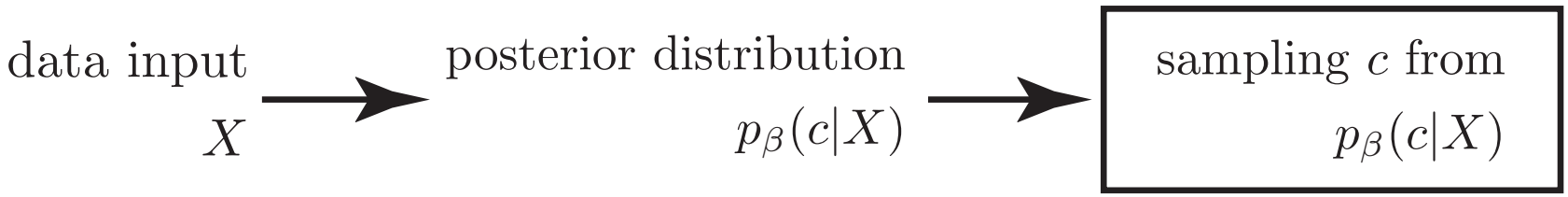


$$p_\beta(c|X) = \frac{\exp\big(-\beta R(c,X)\big)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\big(-\beta R(\tilde{c},X)\big)}$$

Picture: β is **decreasing**.

Tuning the β-Regularization

- Recall our approach:



- **Q:** how to tune β (the **regularizer**)?

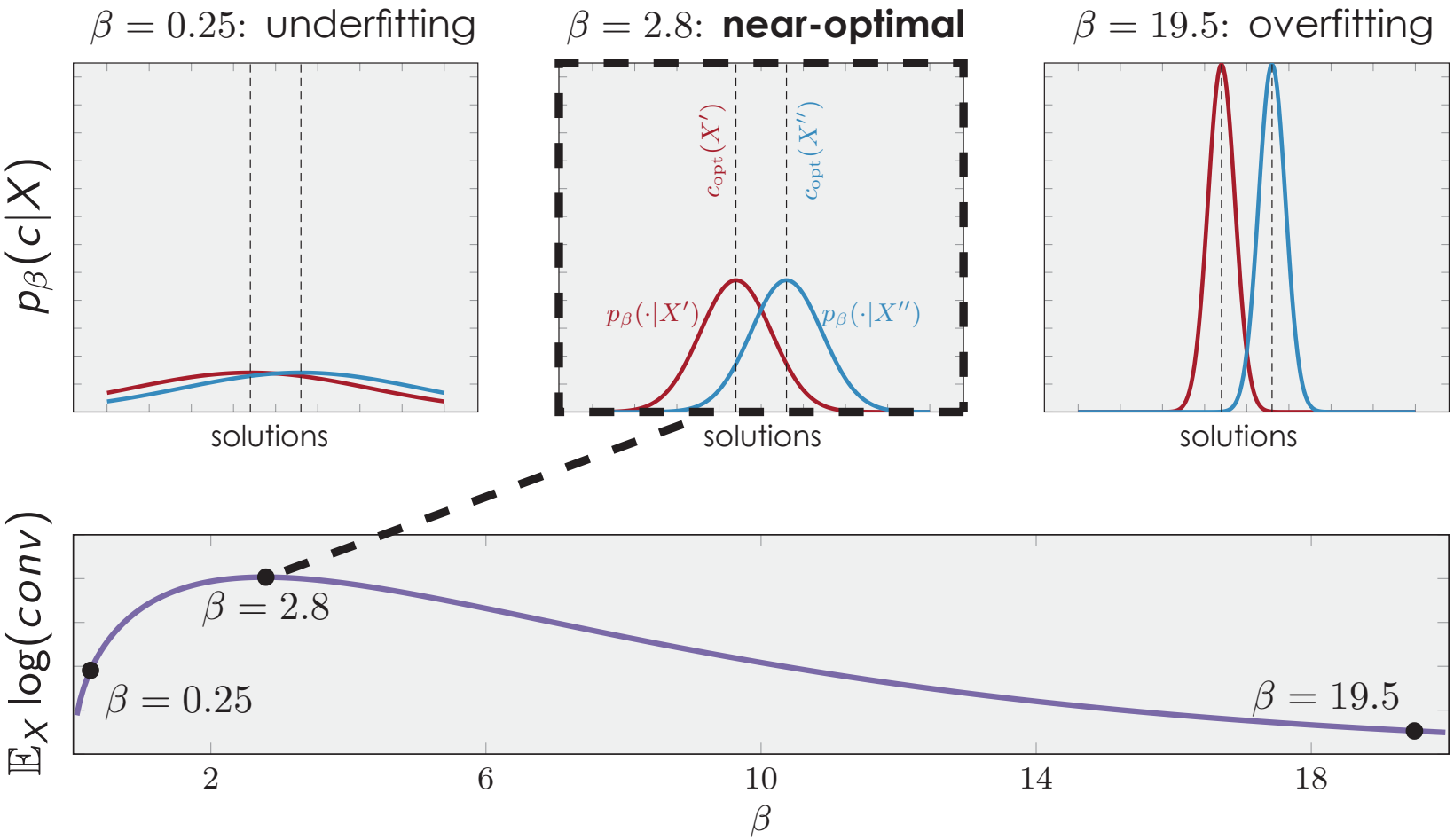
  A: maximize **expected log-posterior agreement** (eLPA) (1):

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

(1) Buhmann et al. (under review), *Posterior Agreement for Parameter-Rich Optimization Problems in the Asymptotic Limit*, J. Theor. Comp. Sc.

eLPA(β): Intuition

- It stabilizes solution output:



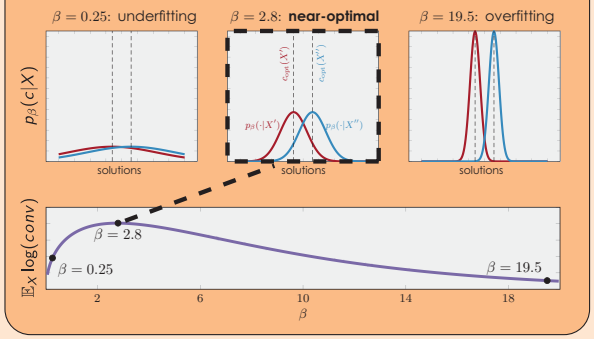
- eLPA $\approx$ cross entropy;

Computing and Maximizing eLPA(β)

Large Disordered Systems

Tuning β

Motivation



Main Result

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

Maximize Expected Log-Posterior Agreement (eLPA)

Now Here

Minimize Average Gibbs Risk

Theorem: eLPA(β) Asymptotics

- eLPA can be rewritten:

$$\mathbb{E}_{X',X''} \log \sum_c p_\beta(c|X') p_\beta(c|X'') =$$

$$\underbrace{\mathbb{E}_{X',X''} \log Z(\beta, X' + X'')}_{\text{Thm from above}}$$

$$- \underbrace{\mathbb{E}_{X'} \log Z(\beta, X')}_{\text{Thm from above}}$$

$$- \underbrace{\mathbb{E}_{X''} \log Z(\beta, X'')}_{\text{Thm from above}}$$

Theorem: eLPA(β) Asymptotics

Theorem: eLPA

- Previous setting and additive $\tilde{\sigma}$-noise on edges;
- Let $\gamma = \tilde{\sigma}/\sigma$ be **noise-to-signal ratio**.

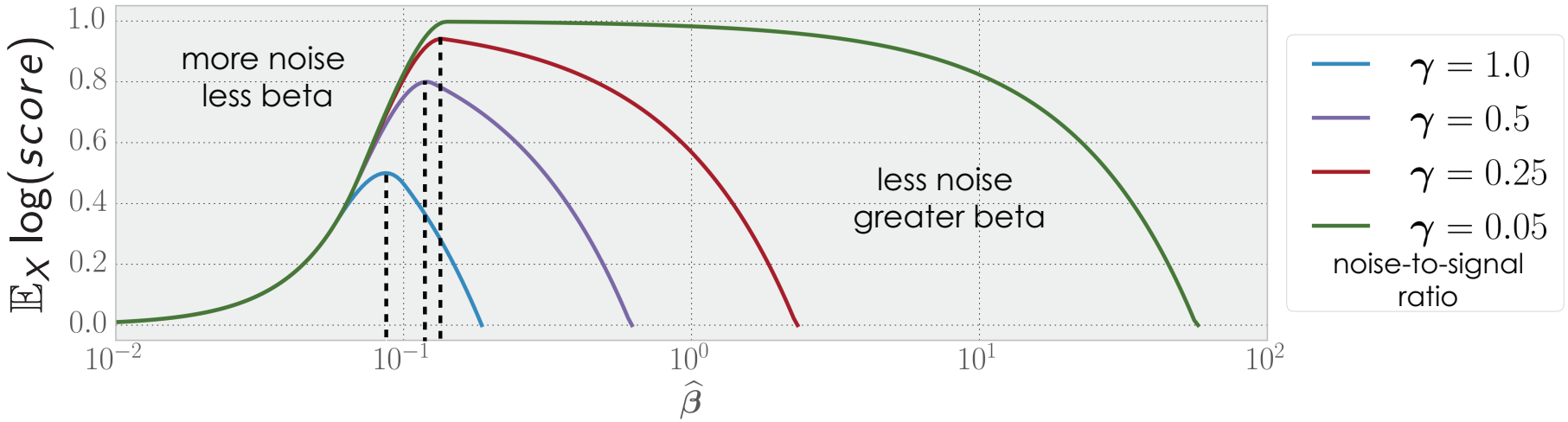
Then, eLPA satisfies

$$\lim_{n\to\infty} \frac{\mathrm{eLPA}(X', X'')}{\log m} = \eta(\widehat{\beta})$$

$$\eta(\widehat{\beta}) = \begin{cases} (\widehat{\beta}\sigma)^2, & \widehat{\beta}\sigma < \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \\ \widehat{\beta}\sigma\sqrt{2}\sqrt{4+2\gamma^2} - (\widehat{\beta}\sigma)^2(1+\gamma^2) - 1, & \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \le \widehat{\beta}\sigma < \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \\ \widehat{\beta}\sigma\sqrt{2}\left(\sqrt{4+2\gamma^2} - 2\sqrt{1+\gamma^2}\right) + 1, & \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \le \widehat{\beta}\sigma \end{cases}$$

Theorem: eLPA(β) Asymptotics

- Plot: ePLA has a clear maximum:



- Theorem: Optimal eLPA-temperature

Optimal eLPA-temperature is provided by:

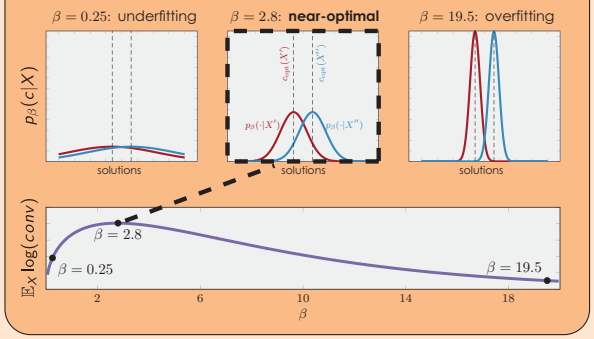
$$\hat{\beta}^*_{\text{eLPA}} := \arg\max_{\hat{\beta}} \text{eLPA}(\beta) = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}.$$

Expected Gibbs Risk Minimization

Large Disordered Systems

Tuning β

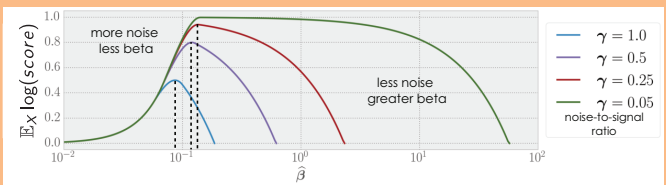
Motivation



Main Result

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

Maximize Expected Log-Posterior Agreement (eLPA)



$$\hat{\beta}^*_{\text{eLPA}} = \arg\max_{\hat{\beta}} \text{eLPA}(\beta) = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}.$$

Minimize Average Gibbs Risk

Now Here

Minimizing Expected Gibbs Risk

- Since the expected Gibbs risk equals

$$\mathbb{E}_{p_\beta(c|X),X}[R(c,X)] = -\frac{\partial}{\partial\beta}\mathbb{E}_X \log Z(\beta, X)$$

Minimizing Expected Gibbs Risk

- Since the expected Gibbs risk equals

$$\mathbb{E}_{p_\beta(c|X),X}[R(c,X)] = -\frac{\partial}{\partial\beta}\mathbb{E}_X \log Z(\beta, X)$$

- We can apply the same theorem to directly minimize it.

Theorem: Optimal Gibbs Risk-Temperature

$$\hat{\beta}^*_{\mathrm{GR}} := \arg\min_{\hat{\beta}} \mathbb{E}_{p_\beta(c|X)}[R(c,X)] = \frac{\sqrt{2+2\gamma^2}}{\sigma(1+\gamma^2)}.$$

Discrepancy: $\hat{\beta}^*_{\mathrm{eLPA}}$ and $\hat{\beta}^*_{\mathrm{GR}}$

- The eLPA gives

$$\hat{\beta}^*_{\mathrm{eLPA}} = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}.$$

- The expected Gibbs risk is minimized at

$$\hat{\beta}^*_{\mathrm{GR}} = \frac{\sqrt{2+\mathbf{2}\gamma^2}}{\sigma(1+\gamma^2)}.$$

- They related only via signal-to-noise ratio:

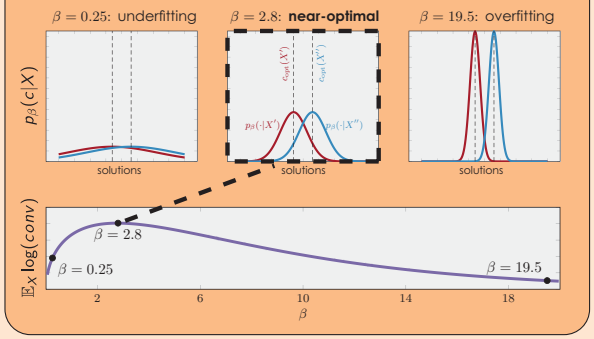
$$\frac{\hat{\beta}^*_{\mathrm{GR}}}{\hat{\beta}^*_{\mathrm{eLPA}}} = \sqrt{1+\frac{\gamma^2}{1+\gamma^2}}.$$

Finally

Large Disordered Systems

Tuning β

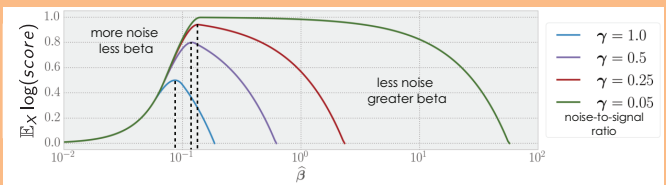
Motivation



Main Result

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

Maximize Expected Log-Posterior Agreement (eLPA)



$$\hat{\beta}^*_{\text{eLPA}} = \arg\max_{\hat{\beta}} \text{eLPA}(\beta) = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}.$$

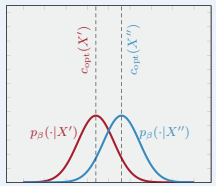
Minimize Average Gibbs Risk

$$\mathbb{E}_{p_\beta(c|X),X}[R(c,X)] = -\frac{\partial}{\partial\beta}\mathbb{E}_X \log Z(\beta, X)$$

$$\hat{\beta}^*_{\text{GR}} = \arg\min_{\hat{\beta}} \mathbb{E}_{p_\beta(c|X)}[R(c,X)] = \frac{\sqrt{2+2\gamma^2}}{\sigma(1+\gamma^2)}.$$

Overview and Outlook

Gibbs Algorithm Motivation

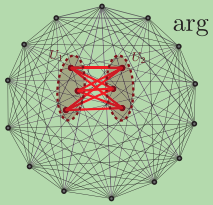
$X \rightarrow$  $\rightarrow c$

Large Disordered Systems

Helmholtz Free Energy Motivation

$$\lim_{n\to\infty} \mathbb{E} \log Z(\beta, X) = ?$$

'In Vitro' Problem: sMBP



$$\arg \min_{c=(U_1,U_2)} \sum W_i$$

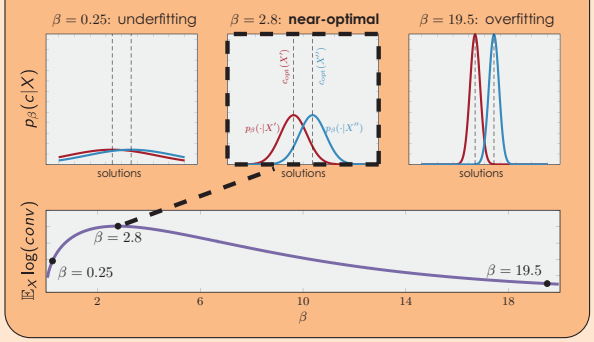
Main Result

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

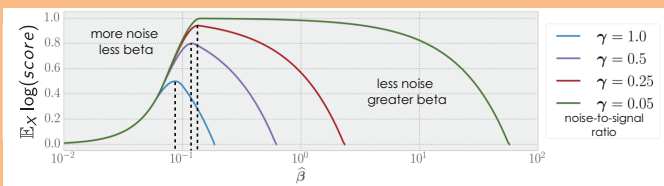
Contributions

Tuning β

Motivation



Maximize Expected Log-Posterior Agreement (eLPA)



$$\hat{\beta}^*_{\text{eLPA}} = \arg\max_{\hat{\beta}} \text{eLPA}(\beta) = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}.$$

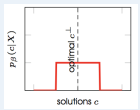
Minimize Average Gibbs Risk

$$\mathbb{E}_{p_\beta(c|X),X}[R(c,X)] = -\frac{\partial}{\partial\beta}\mathbb{E}_X \log Z(\beta, X)$$

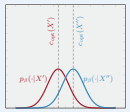
$$\hat{\beta}^*_{\text{GR}} = \arg\min_{\hat{\beta}} \mathbb{E}_{p_\beta(c|X)}[R(c,X)] = \frac{\sqrt{2+2\gamma^2}}{\sigma(1+\gamma^2)}.$$

More for Discussion: See Appendix

Uniform γ-Approximation Motivation

$X \to$  $\to c$

Gibbs Algorithm Motivation

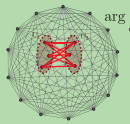
$X \to$  $\to c$

Large Disordered Systems

Helmholtz Free Energy Motivation

$$\lim_{n\to\infty} \mathbb{E} \log Z(\beta, X) = ?$$

'In Vitro' Problem: sMBP



$$\arg \min_{c=(U_1,U_2)} \sum W_i$$

'In Vitro' Problem: Lawler QAP

$$\arg\min_{\pi\in S_n} = \sum_{i,j=1}^{n} Q_{i,j,\pi(i),\pi(j)}$$

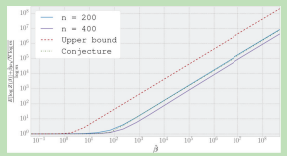
Main Result

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

Ad-Hoc Correction Conjecture

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

$$\alpha = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_{\mathcal{D}} R(c, X)}{\mathbb{E}_{\mathcal{D}} \mathrm{Var}_X R(c, X)}} = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_c R(c, X)}{N\sigma^2}}$$

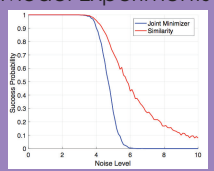


Tuning γ

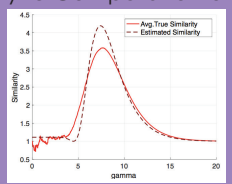
Similarity (ASC) Score

$$\arg\max_{\gamma} \hat{I}_\gamma(X', X'') = \log\left(\frac{|C| \, |\Delta\mathcal{C}_\gamma(X', X'')|}{|\mathcal{C}_\gamma(X')| \, |\mathcal{C}_\gamma(X'')|}\right),$$

Model Experiments



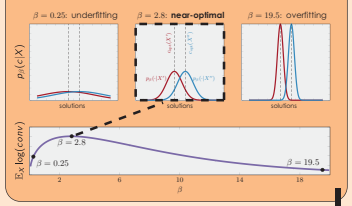
Analytic Computation of γ^*



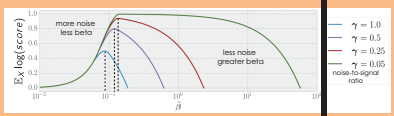
$$\gamma^* = \arg\max \left(\frac{\mathbb{E}[V]}{\mathbb{E}[W]} - \frac{\mathrm{Cov}(V, W)}{\mathbb{E}[W]^2} + \frac{\mathrm{Var}[W] \cdot \mathbb{E}[V]}{\mathbb{E}[W]^3} \right)$$

Tuning β

Motivation



Maximize Expected Log-Posterior Agreement (eLPA)



$$\hat{\beta}^*_{\mathrm{eLPA}} = \arg\max_{\hat{\beta}} \mathrm{eLPA}(\beta) = \frac{\sqrt{2+\gamma^2}}{\sigma(1+\gamma^2)}.$$

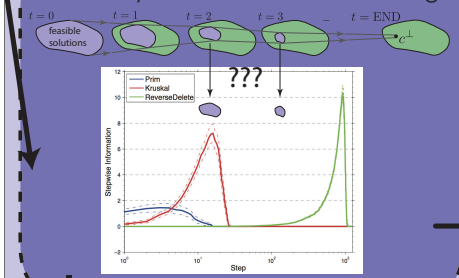
Minimize Average Gibbs Risk

$$\mathbb{E}_{p_\beta(c|X),X}[R(c, X)] = -\frac{\partial}{\partial\beta}\mathbb{E}_X \log Z(\beta, X)$$

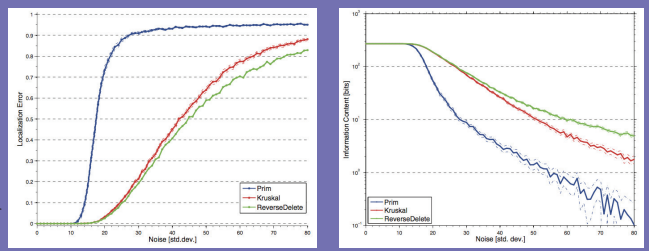
$$\hat{\beta}^*_{\mathrm{GR}} = \arg\min_{\hat{\beta}} \mathbb{E}_{p_\beta(c|X)}[R(c, X)] = \frac{\sqrt{2+2\gamma^2}}{\sigma(1+\gamma^2)}.$$

t-Step Uniform Approximation for Contractove Algorithms

eLPA Dynamics for Contractve Alg



Ranking of Localization Error



Efficient Computation of eLPA

Theorem (Kirchhoff's Matrix-Tree Theorem). *The total number of labeled trees on a graph $G = (V, E)$ with adjacency matrix M^G_{adj}, degree matrix M^G_{deg} is equal to (any) cofactor of the matrix $L = M^G_{\mathrm{deg}} - M^G_{\mathrm{adj}}$.*

List of Publications

— **Gronskiy, A.**, Buhmann, J. M., 2014. *How informative are minimum spanning tree algorithms?*
ISIT 2014

— Buhmann, J. M., **Gronskiy, A.**, Szpankowski, W., 2014. *Free energy rates for a class of very noisy optimization problems.*
AofA 2014

— Buhmann, J. M., Dumazert, J., **Gronskiy, A.**, Szpankowski, W., 2017. *Phase transitions in parameter rich optimization problems.*
SODA-ANALCO 2017

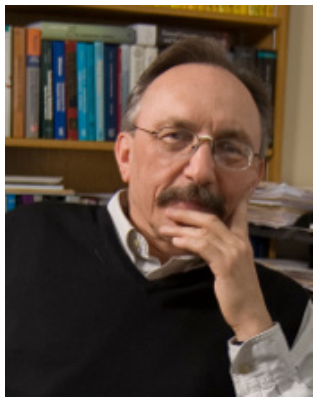
— Buhmann, J., **Gronskiy, A.**, Mihalák, M., Pröger, T., Šrámek, R., Widmayer, P., 2017. *Robust optimization in the presence of uncertainty: A generic approach.*
J. of Comp. Sci. and Systems

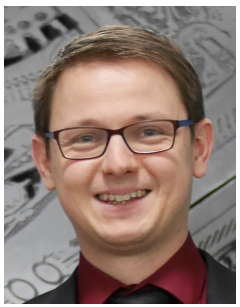
— (under rev.) **Gronskiy, A.**, Buhmann, J. M., Szpankowski, W., 2018. *Free Energy Asymptotics for Problems with Weak Solution Dependencies.*
ISIT 2018

— (under rev.) Buhmann, J. M., Dumazert, J., **Gronskiy, A.**, Szpankowski, W., 2018. *Posterior Agreement for Large Parameter-Rich Optimization Problems.*
J. of Theor Comp. Sci.

Thanks for your attention!









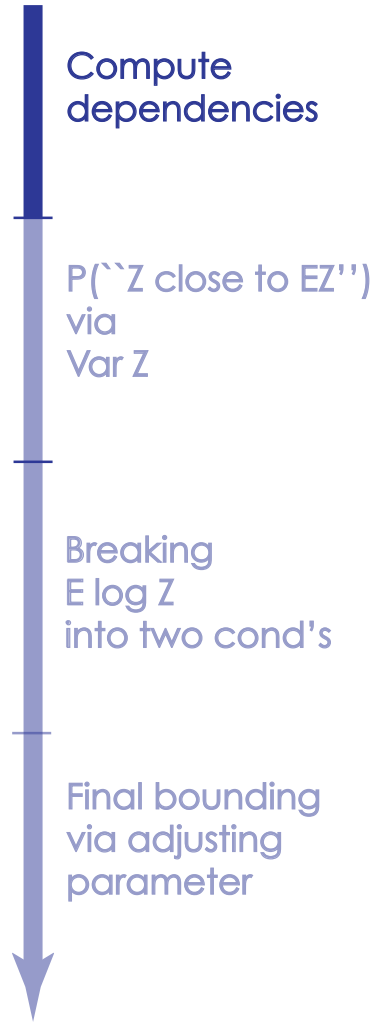


Appendix: Entropy and Free Energy

$$
\begin{aligned}
H(p_\beta) &= -\sum_{c \in \mathcal{C}} p_\beta(c) \log p_\beta(c) \\
&= \log Z(\beta) - \mathbb{E}_{p_\beta(c)}[R(c)] \\
&= \log Z(\beta) - \sum_{c \in \mathcal{C}} \frac{-\beta R(c) e^{-\beta R(c)}}{Z(\beta)} \\
&= \log Z(\beta) - \beta \frac{\partial}{\partial \beta} \log Z(\beta).
\end{aligned}
$$

Appendix: 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

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

Appendix: 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 $\mathbb{P}(A)$

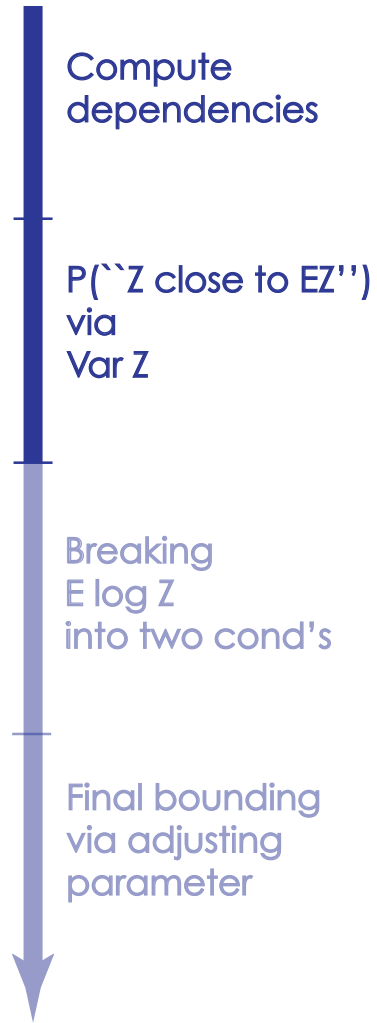
Fact 2

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

Lemma 3 (Buhmann et al., 2014)

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

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



Appendix: Proof Outline – III

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

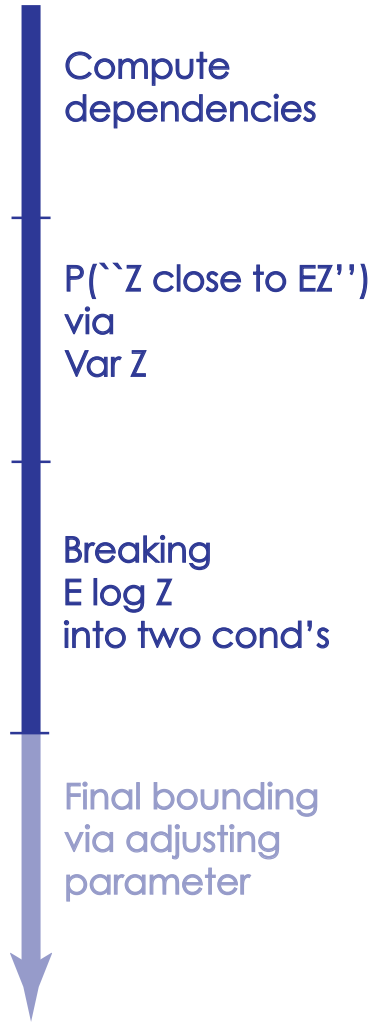
$$\mathbb{E}\log Z = \mathbb{E}[\log Z \mid A]\cdot\mathbb{P}(A) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})]$$
$$\geq (\log\mathbb{E}Z + \log\epsilon)\mathbb{P}(A) + \mathbb{E}[\log Z\mathbb{1}(\bar{A})]$$

Fact 4

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

Fact 5

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

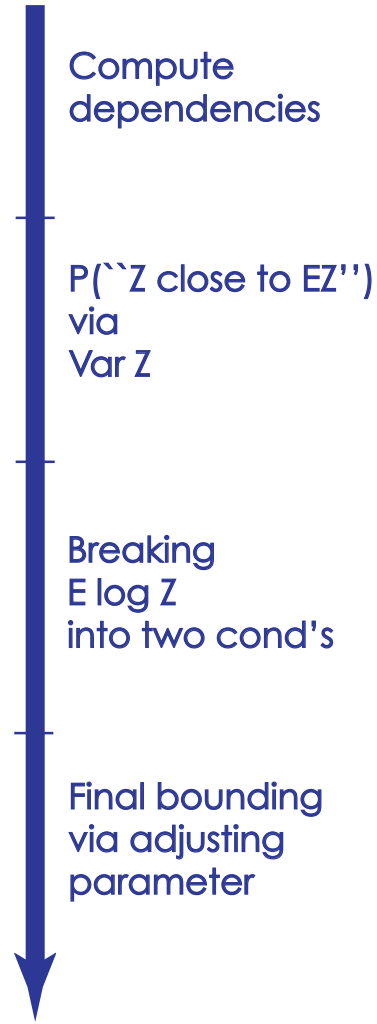


Appendix: Proof Outline – IV

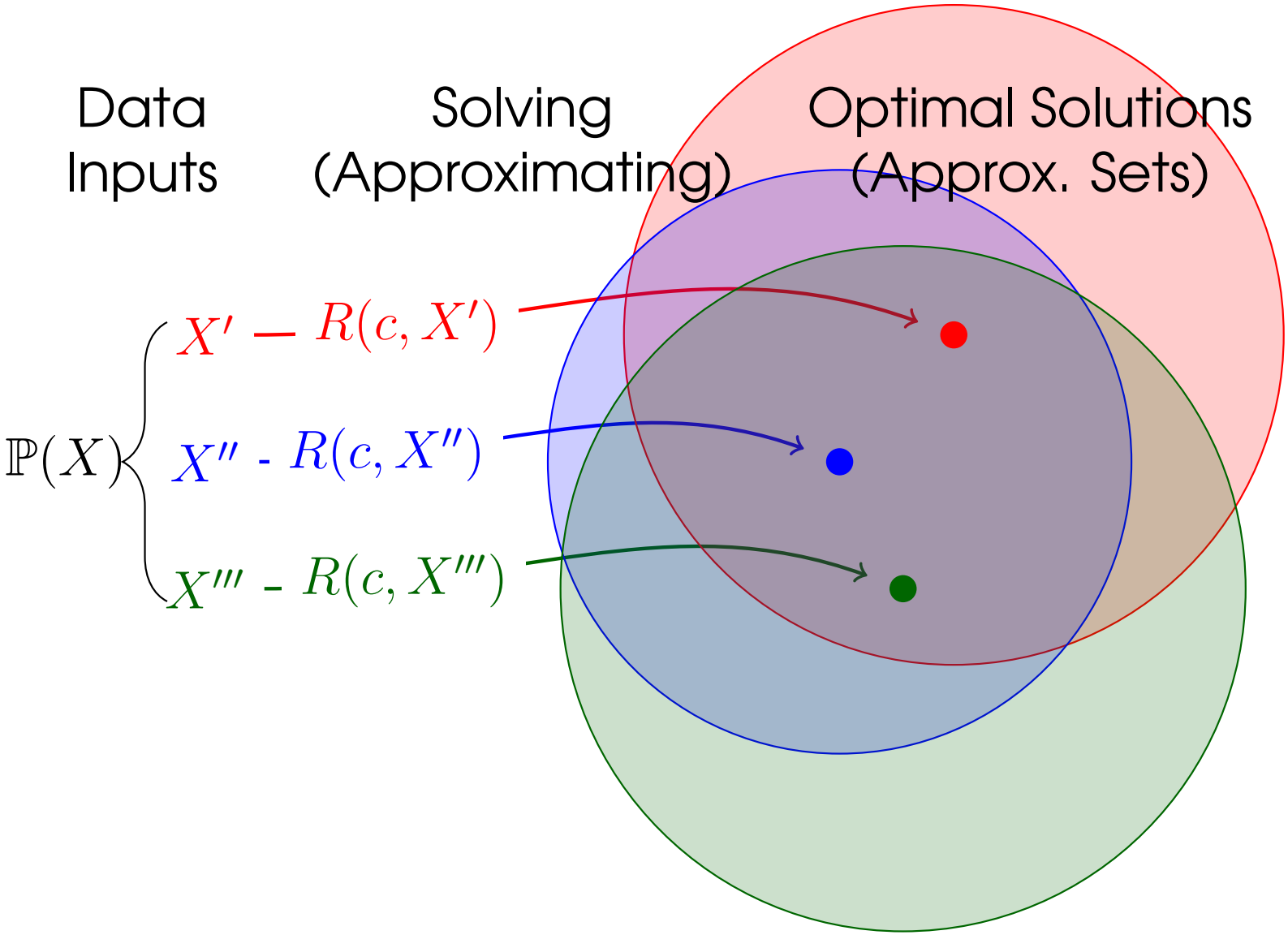
- Finally, the right choice of ϵ for two regimes of β gives the phase transition in **lower bound**

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

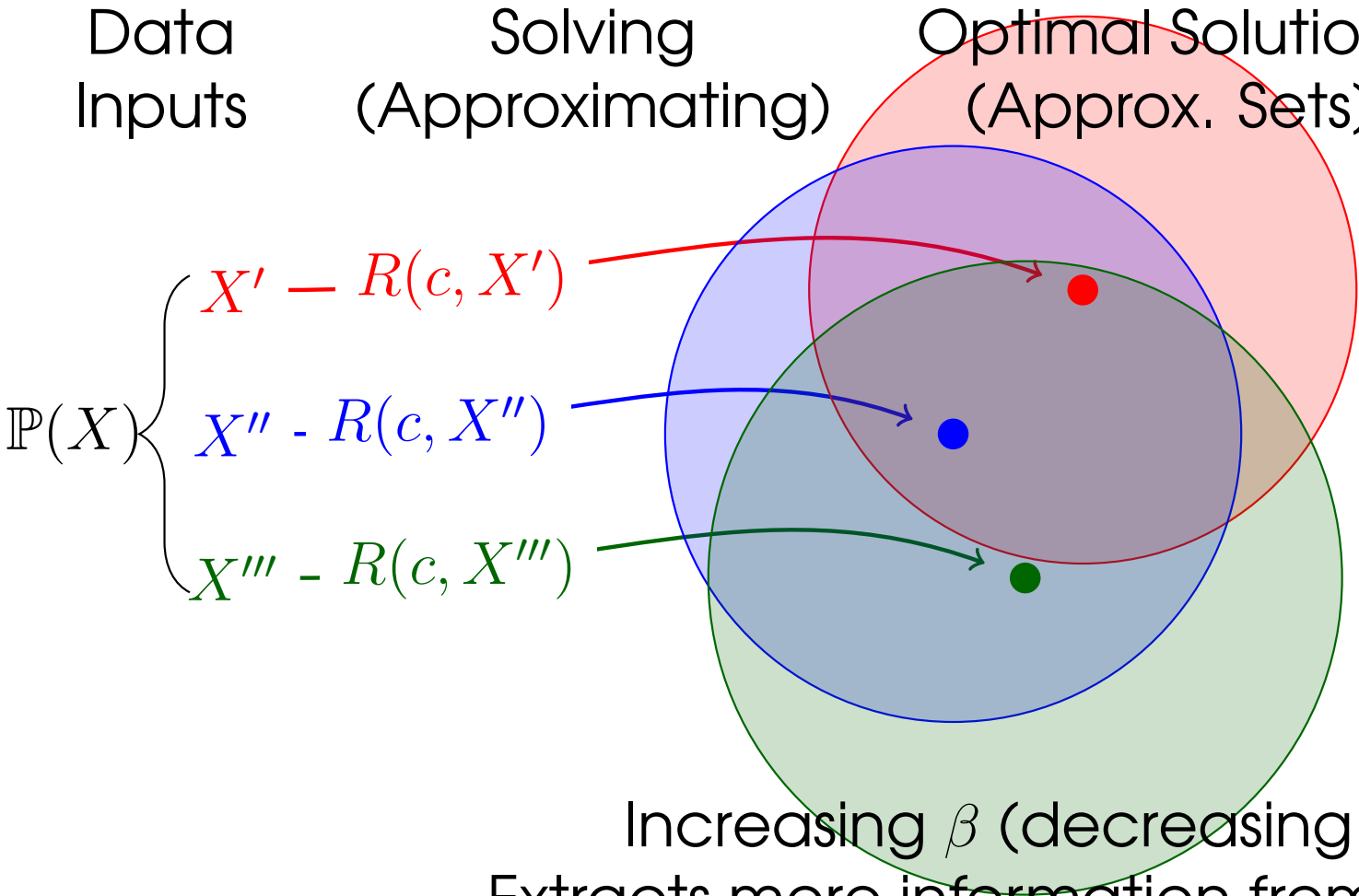
- The **same** phase transition happens for **upper bound**, — easier to prove (no need to compute dependencies)



Intuition: Informativeness vs Robustness

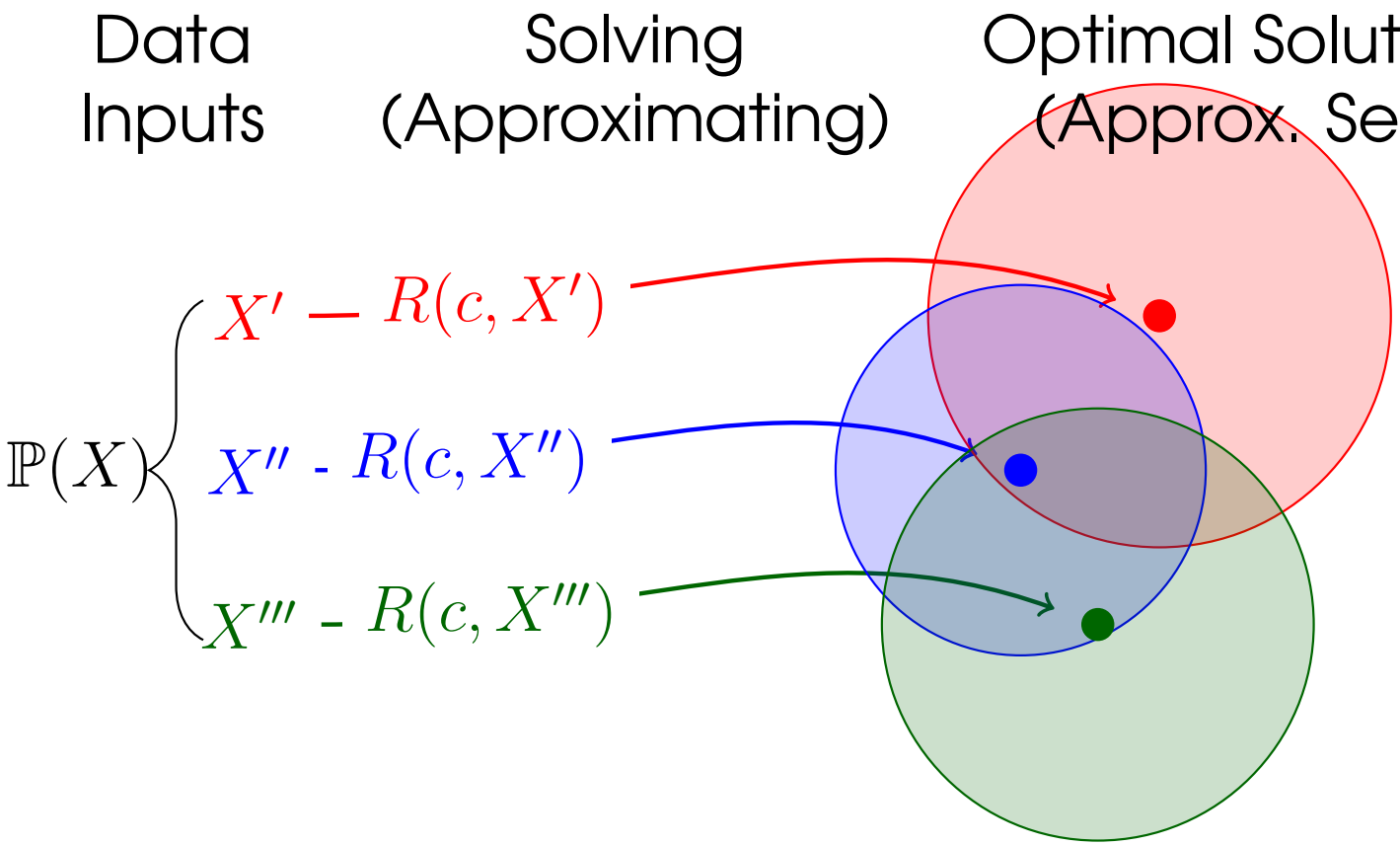


Intuition: Informativeness vs Robustness



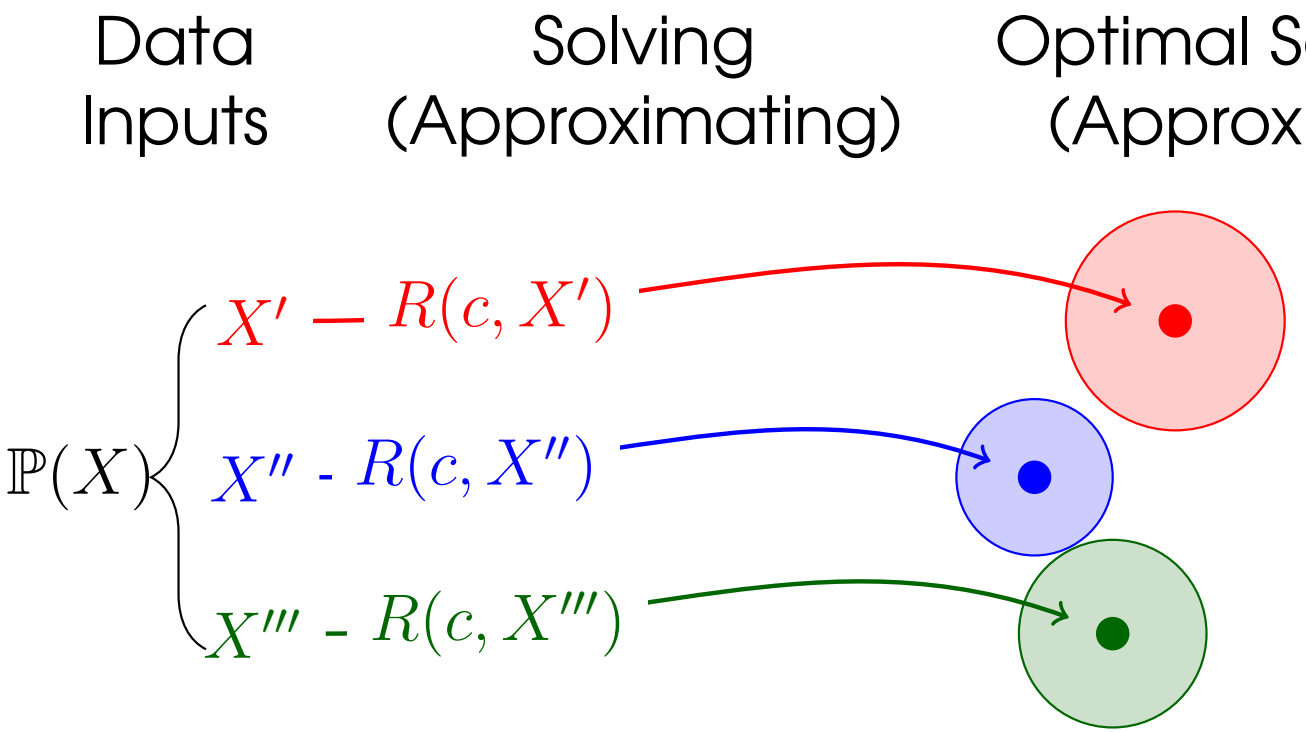
Increasing β (decreasing γ)
Extracts more information from input.

Intuition: Informativeness vs Robustness



Increasing β (decreasing γ)
Extracts more information from input.

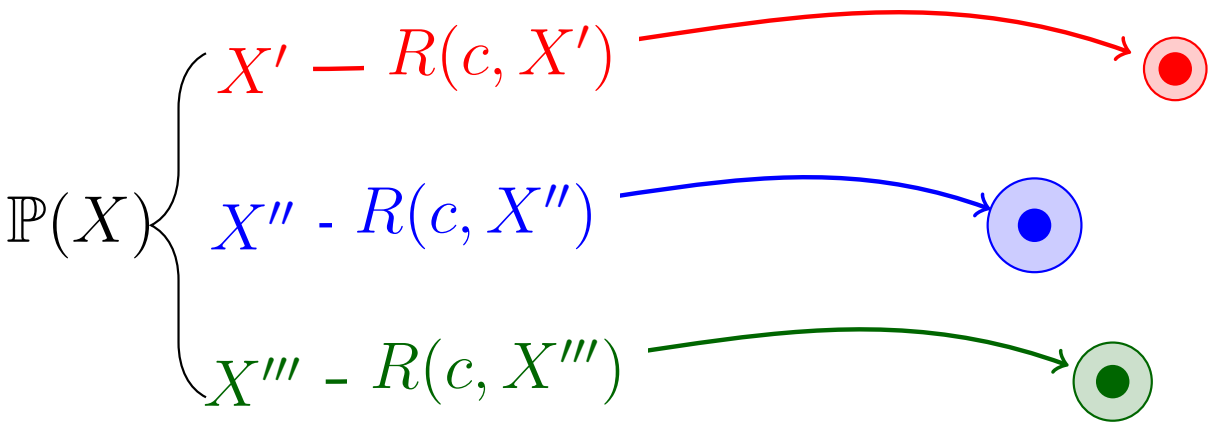
Intuition: Informativeness vs Robustness



Increasing β (decreasing γ)
Extracts more information from input.

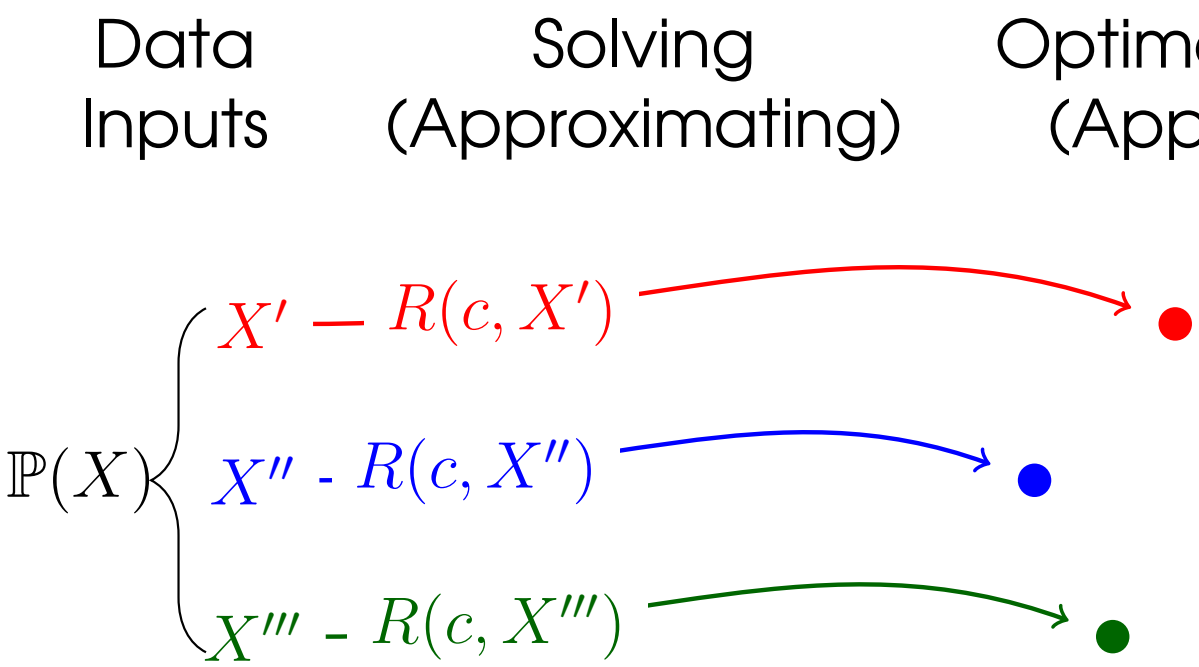
Intuition: Informativeness vs Robustness

Data Inputs | Solving (Approximating) | Optimal Solutions (Approx. Sets)



Increasing β (decreasing γ)
Extracts more information from input.

Intuition: Informativeness vs Robustness



Increasing β (decreasing γ)
Extracts more information from input.

Appendix: Free Energy Conjecture

Conjecture 5.1. *Consider a class of combinatorial optimization problems complying with Common Theorem Setting, weights W_i having mean μ and variance σ^2. Then the free energy satisfies*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\log Z(\beta, X)] + \widehat{\beta}\mu\sqrt{N \log m}}{\log m} = \begin{cases} 1 + \alpha^2 \frac{\widehat{\beta}^2\sigma^2}{2}, & \widehat{\beta} < \frac{\sqrt{2}}{\alpha\sigma} \\ \alpha\widehat{\beta}\sigma\sqrt{2}, & \widehat{\beta} \geq \frac{\sqrt{2}}{\alpha\sigma} \end{cases} \tag{5.144}$$

$$\alpha = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_{\mathcal{D}} R(c, X)}{\mathbb{E}_{\mathcal{D}} \mathrm{Var}_X R(c, X)}} = \sqrt{\frac{\mathbb{E}_X \mathrm{Var}_c R(c, X)}{N\sigma^2}}$$

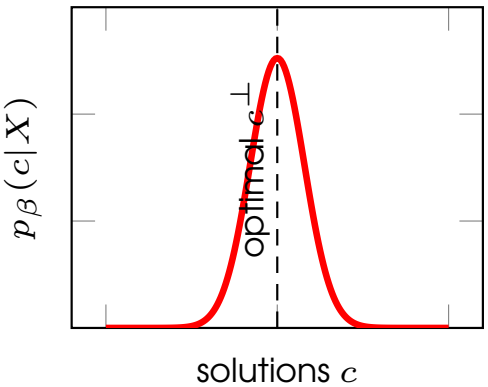
Statement 5.1. *For sMBP and Lawler QAP, Conjecture 5.1 turns into the proven asymptotics, since $\alpha = 1$.*

Appendix: Replica Trick

$$\ln Z = \lim_{n \to 0} \frac{Z^n - 1}{n}$$

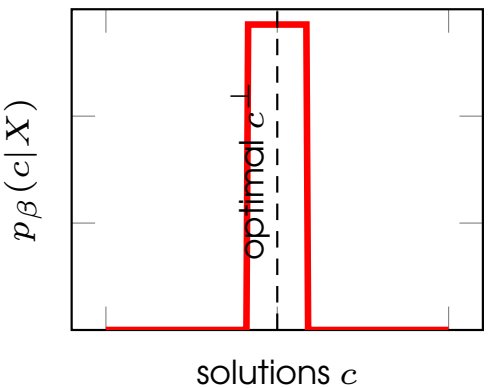
Appendix: Various Posteriors

Example (Gibbs Posterior)



$$p_\beta(c|X) = \frac{\exp\big(-\beta R(c,X)\big)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\big(-\beta R(\tilde{c},X)\big)}$$

Example (Bounded-Support Uniform)



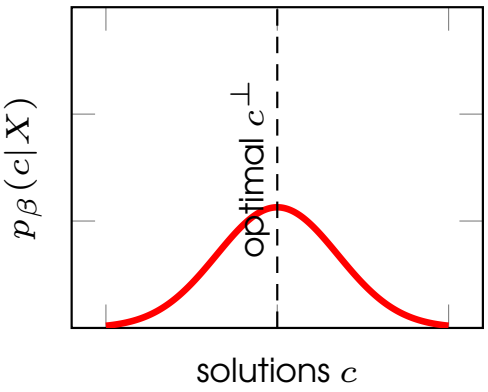
$$p_\beta(c|X) = \mathrm{Uniform}\big(C_\gamma(X)\big)$$

where C_γ is γ-**approximation set**:

$$C_\gamma(X) := \big\{c \in \mathcal{C} \;\big|\; R(c,X) - R(c^\perp,X) \le \gamma\big\}.$$

Appendix: Various Posteriors

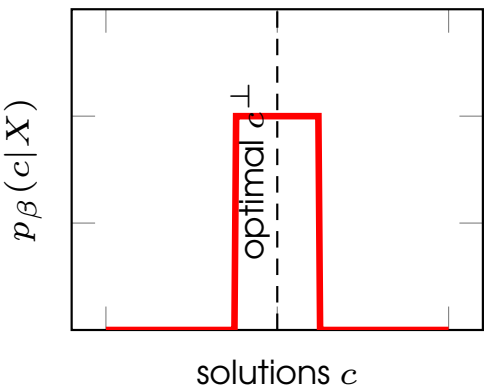
Example (Gibbs Posterior)



$$p_\beta(c|X) = \frac{\exp\bigl(-\beta R(c,X)\bigr)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\bigl(-\beta R(\tilde{c},X)\bigr)}$$

Picture: β is **decreasing**.

Example (Bounded-Support Uniform)



$$p_\beta(c|X) = \mathrm{Uniform}\bigl(C_\gamma(X)\bigr)$$

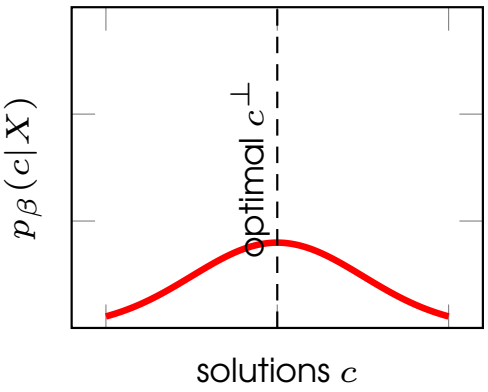
where C_γ is γ**-approximation set**:

$$C_\gamma(X) := \bigl\{c \in \mathcal{C} \;\big|\; R(c,X) - R(c^\perp,X) \le \gamma\bigr\}.$$

Picture: γ is **increasing**.

Appendix: Various Posteriors

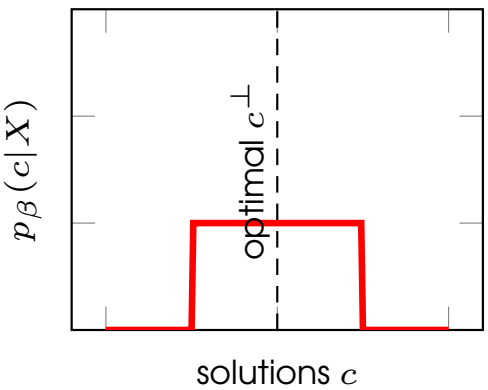
Example (Gibbs Posterior)



$$p_\beta(c|X) = \frac{\exp\bigl(-\beta R(c,X)\bigr)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\bigl(-\beta R(\tilde{c},X)\bigr)}$$

Picture: β is **decreasing**.

Example (Bounded-Support Uniform)



$$p_\beta(c|X) = \mathrm{Uniform}\bigl(C_\gamma(X)\bigr)$$

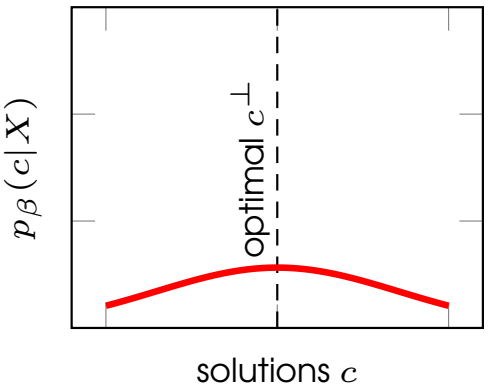
where C_γ is γ**-approximation set**:

$$C_\gamma(X) := \bigl\{c \in \mathcal{C} \;\big|\; R(c,X) - R(c^\perp,X) \le \gamma\bigr\}.$$

Picture: γ is **increasing**.

Appendix: Various Posteriors

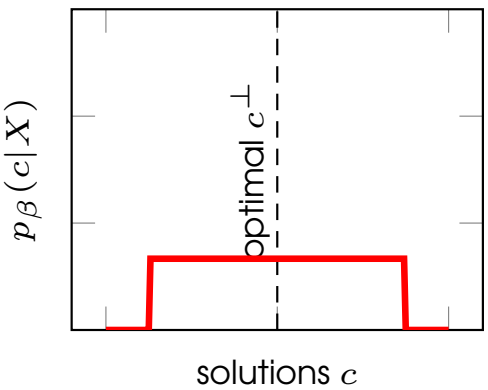
Example (Gibbs Posterior)



$$p_\beta(c|X) = \frac{\exp\bigl(-\beta R(c,X)\bigr)}{\sum_{\tilde{c}\in\mathcal{C}} \exp\bigl(-\beta R(\tilde{c},X)\bigr)}$$

Picture: β is **decreasing**.

Example (Bounded-Support Uniform)



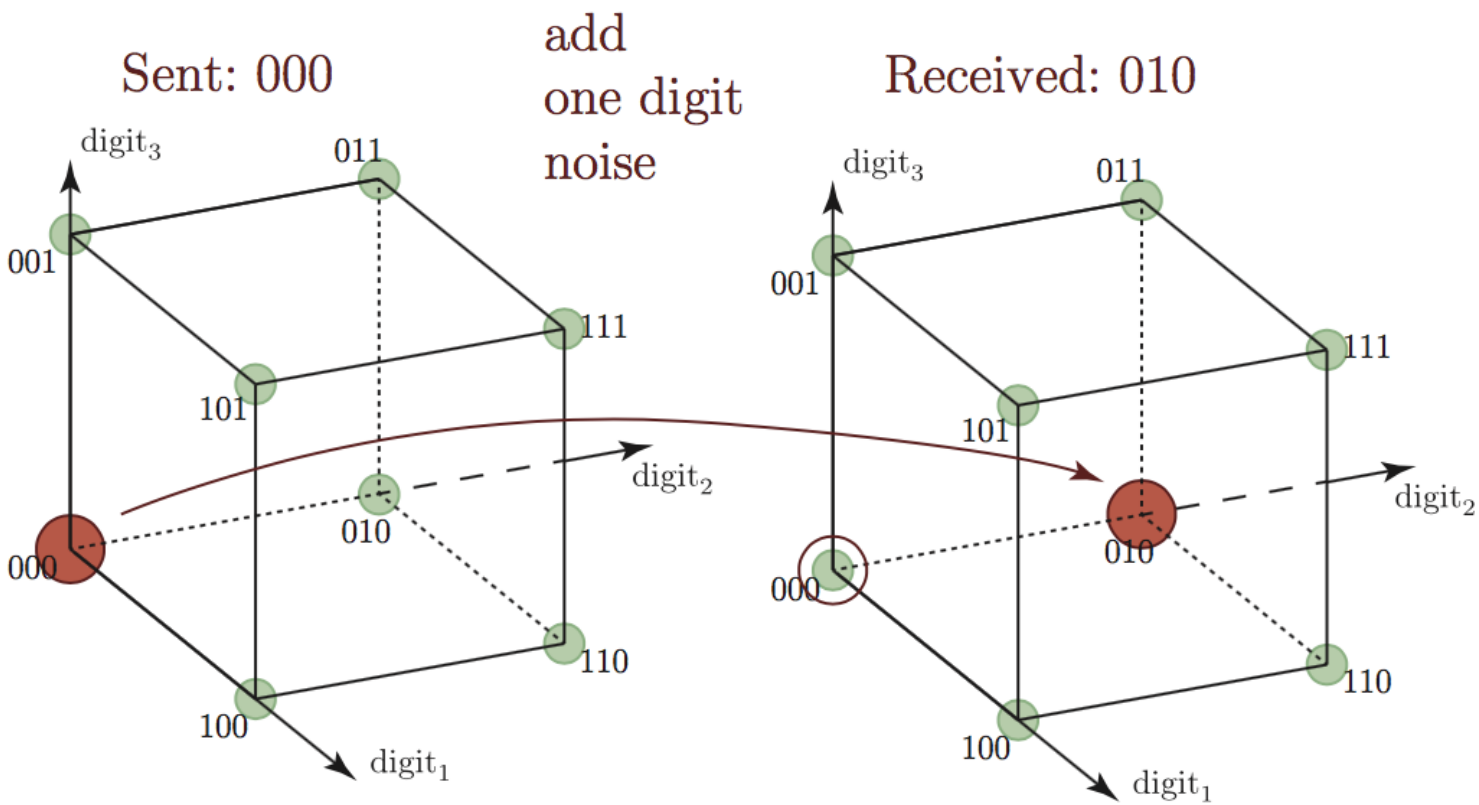
$$p_\beta(c|X) = \mathrm{Uniform}\bigl(C_\gamma(X)\bigr)$$

where C_γ is γ**-approximation set**:

$$C_\gamma(X) := \bigl\{c \in \mathcal{C} \;\big|\; R(c,X) - R(c^\perp,X) \le \gamma\bigr\}.$$

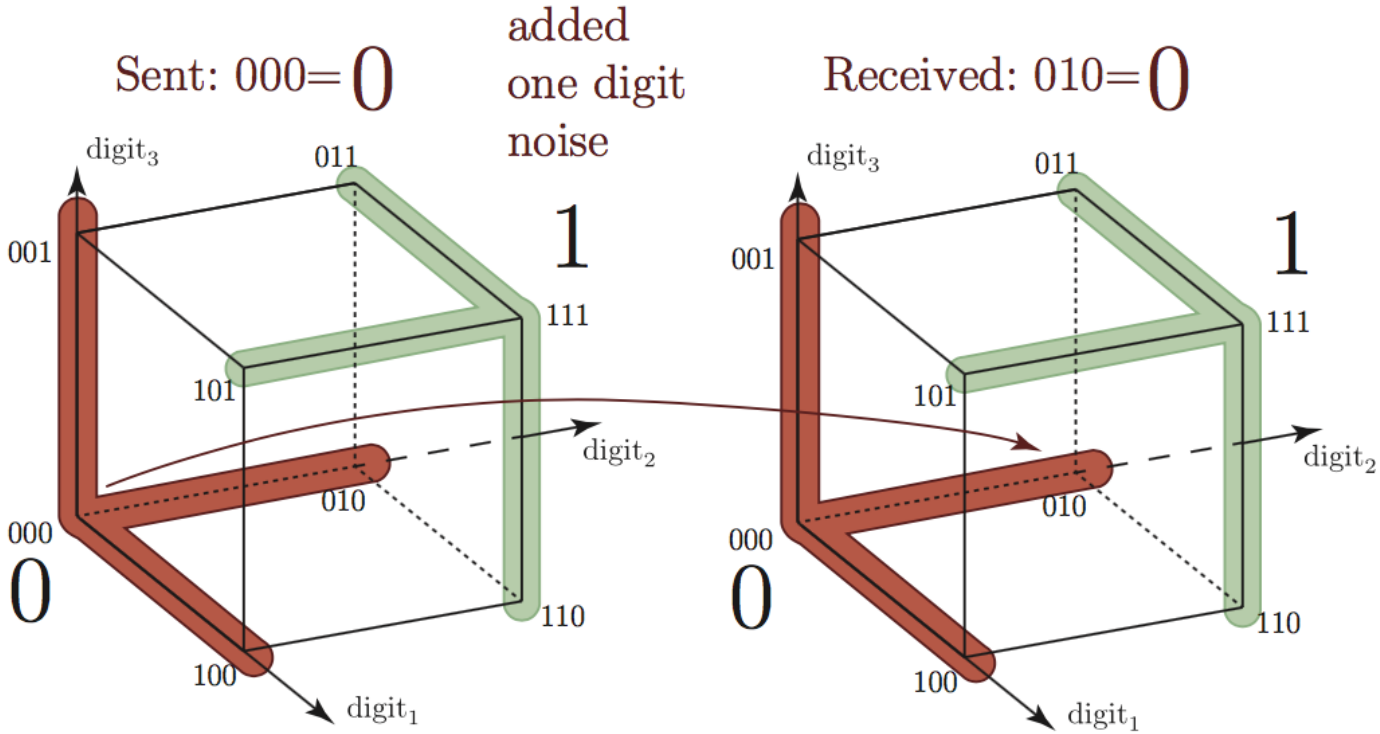
Picture: γ is **increasing**.

Appendix: Shannon Coding – I



(a) High rate ($R_{\text{code}} = 1$), but no way to correct the error (red: sent and received codes).

Appendix: Shannon Coding – II



(b) Lower rate ($R_{\text{code}} = 1/3$), correcting one digit error (red: sent and received codes).

Figure 3.3 Dealing with one digit error. Case **(a)** is high rate option with eight codebook vectors, leading to a low error-correcting capacity (in fact, no error can be tolerated). Case **(b)** is lower rate option with two codebook vectors leading to a higher error-correcting capacity (one digit error can be tolerated, two digits not).

Appendix: Coding Capacity – I

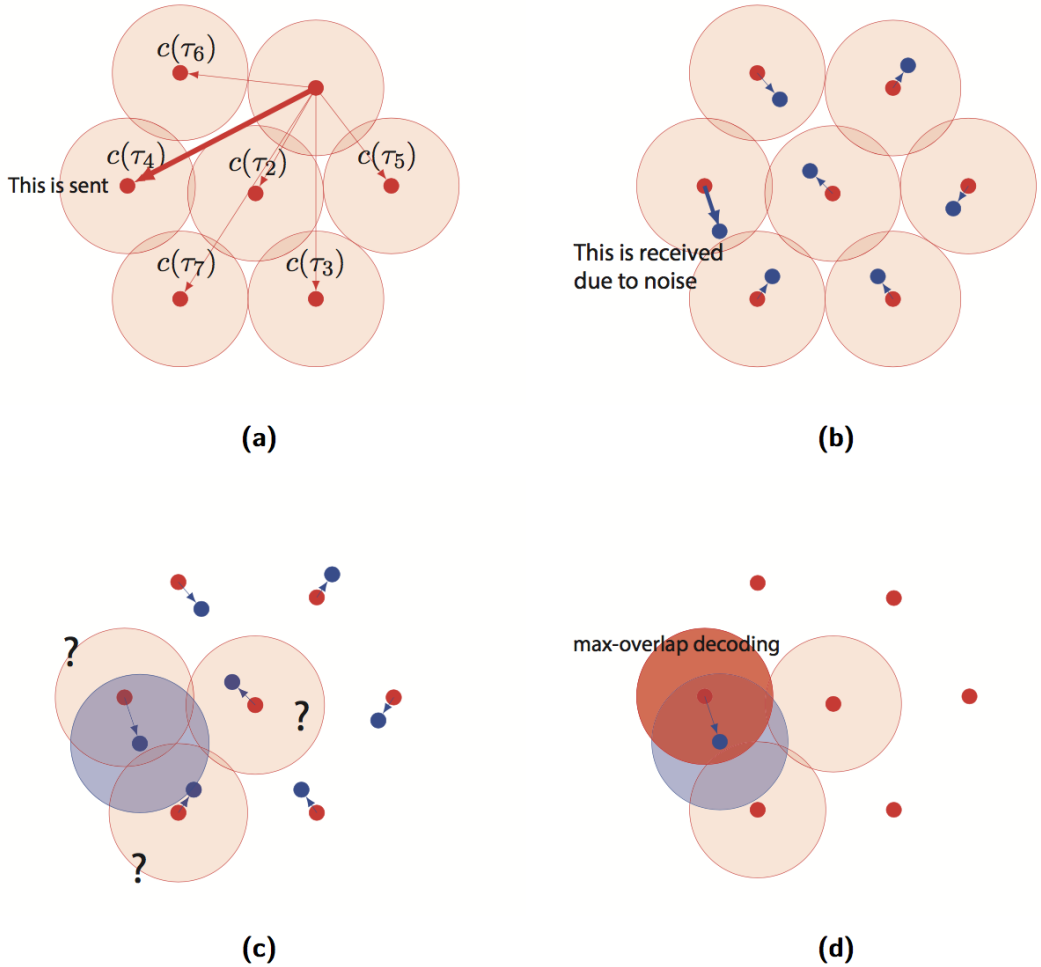


Figure 3.4 Process of correct decoding by approximation sets in the solution space: **(a)** X' is set and sender sends τ_4; **(b)** due to noise which replaces X' by X'', all the minimizers move around (red to blue) in the solution space; **(c)** the received solution is surrounded by its approximation set (blue) and overlaps are considered; **(d)** decoded solution (dark red) happens to be τ_4 which was initially sent (correct decoding).

Appendix: Coding Capacity – II

(b) (c)

Figure 3.5 Decreased γ and increased code rate leads to incorrect decoding: **(a)** same setting (i.e. same noise) as in Figure 3.4, but added more codebook vectors; **(b)** due to noise which replaces X' by X'', all the minimizers move around (red to blue) in the solution space, **(c)** decoded solution (dark red) happens to wrong (incorrect decoding).

Appendix: ASC Derivation – I

Before we proceed, we will denote the intersection (3.15) as follows:

$$\Delta\mathcal{C}_\gamma^\tau := \mathcal{C}_\gamma(\tau \circ X') \cap \mathcal{C}_\gamma(\tau_{\text{send}} \circ X''). \tag{3.18}$$

Due to the union bound, it holds that

$$\mathbb{P}(\widehat{\tau} \neq \tau_{\text{send}} | \tau_{\text{send}}) \leq \sum_{\tau \in \mathbb{T}} \mathbb{P}\big(|\Delta\mathcal{C}_\gamma^\tau| \geq |\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}| \;\big|\; \tau_{\text{send}}\big), \tag{3.19}$$

i.e. for decoding error to occur, one has to encounter an approximation set which is yielded by a wrong transformation, but happens to be closer to the received approximation set (this is illustrated in Figure 3.5(c)). The last bound can be rewritten via the indicator function:

$$\mathbb{P}(\widehat{\tau} \neq \tau_{\text{send}} | \tau_{\text{send}}) \leq \sum_{\tau \in \mathbb{T}} \mathbb{E}_{PG}\big[\mathbb{1}\{|\Delta\mathcal{C}_\gamma^\tau| \geq |\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}|\} \;\big|\; \tau_{\text{send}}\big], \tag{3.20}$$

where the expectation is taken w.r.t. the problem generation process $X', X'' \sim PG(\cdot|X^0)$. We further utilize the monotonicity of log function:

$$\mathbb{1}\{|\Delta\mathcal{C}_\gamma^\tau| \geq |\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}|\} = \mathbb{1}\{\log|\Delta\mathcal{C}_\gamma^\tau| \geq \log|\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}|\} \tag{3.21}$$

Appendix: ASC Derivation – II

and the fact that $\mathbb{1}\{x \geq 0\} \leq \exp(x)$ to come to the following:

$$\mathbb{E}_{PG}\Big(\mathbb{1}\{|\Delta\mathcal{C}_\gamma^\tau| \geq |\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}|\} \;\Big|\; \tau_{\text{send}}\Big) \leq \frac{|\mathcal{C}_\gamma(X')|\;|\mathcal{C}_\gamma(X'')|}{|\mathbb{T}|\;|\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}|}, \tag{3.22}$$

where the product in the nominator comes from the fact that, under our generation process, the data instances X' and X'' are independent given X^0, see (3.8).

In the spirit of Shannon (1948), we use the random coding argument here: all the τ are identically distributed and independent, hence the above can be rewritten:

$$\mathbb{P}(\widehat{\tau} \neq \tau_{\text{send}} | \tau_{\text{send}}) \leq (|\mathbb{T}| - 1)\exp(-I_\gamma(\tau_{\text{send}}, \widehat{\tau})), \tag{3.23}$$

where

$$I_\gamma(\tau_{\text{send}}, \widehat{\tau}) \coloneqq \mathbb{E}\log\Big(\frac{|\mathbb{T}|\;|\Delta\mathcal{C}_\gamma^{\tau_{\text{send}}}|}{|\mathcal{C}_\gamma(X')|\;|\mathcal{C}_\gamma(X'')|}\Big). \tag{3.24}$$

Appendix: Theorem eLPA Full

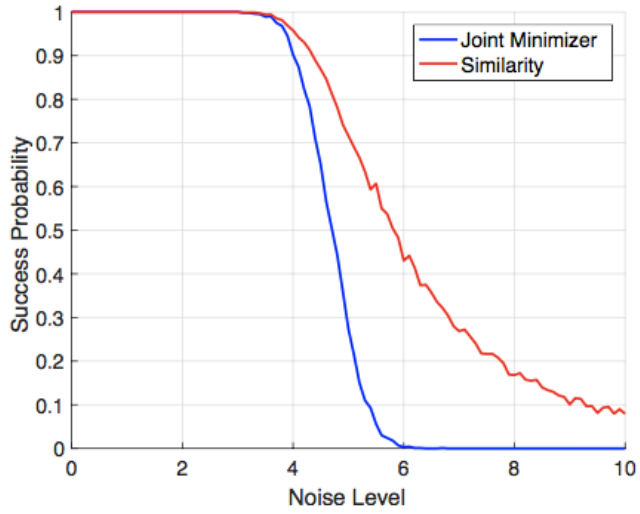
Theorem: eLPA

As previously, edge weights mutually independent within any given solution; have mean μ and variance σ^2. Assume then additive noise $\delta X'$, $\delta X''$ with mean 0, and variance $\tilde{\sigma}^2$, all the sets of the same size. Let $\gamma = \tilde{\sigma}/\sigma$ be noise-to-signal ratio. Then, eLPA satisfies

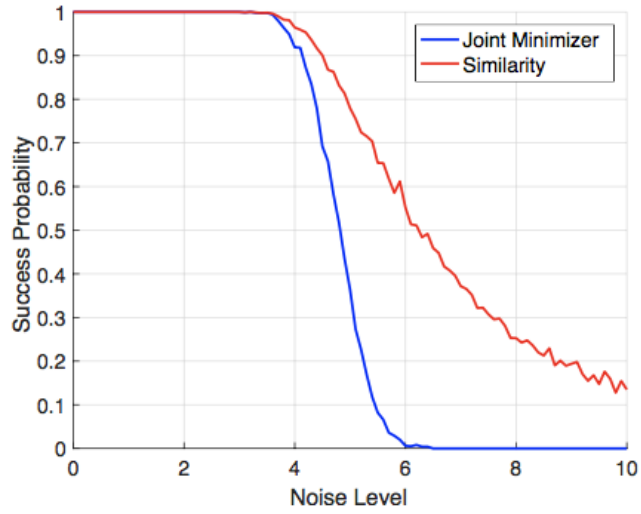
$$\lim_{n\to\infty} \frac{\mathrm{eLPA}(X', X'')}{\log m} = \eta(\widehat{\beta})$$

$$\eta(\widehat{\beta}) = \begin{cases} (\widehat{\beta}\sigma)^2, & \widehat{\beta}\sigma < \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \\ \widehat{\beta}\sigma\sqrt{2}\sqrt{4+2\gamma^2} - (\widehat{\beta}\sigma)^2(1+\gamma^2) - 1, & \frac{\sqrt{2}}{\sqrt{4+2\gamma^2}} \le \widehat{\beta}\sigma < \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \\ \widehat{\beta}\sigma\sqrt{2}\big(\sqrt{4+2\gamma^2} - 2\sqrt{1+\gamma^2}\big) + 1, & \frac{\sqrt{2}}{\sqrt{1+\gamma^2}} \le \widehat{\beta}\sigma \end{cases}$$

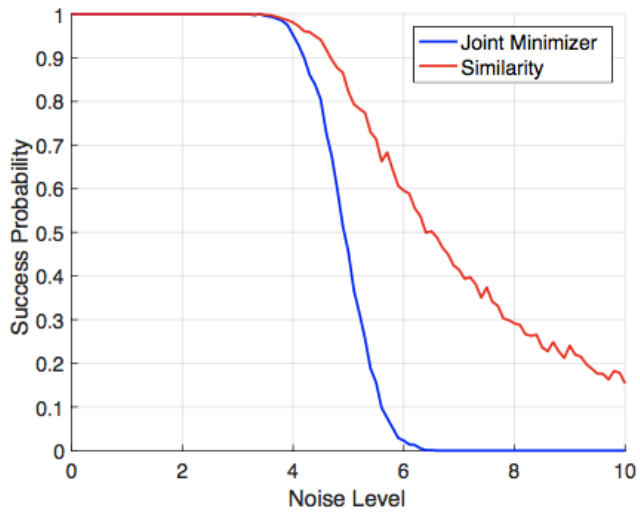
Appendix: γ-Similarity Approach



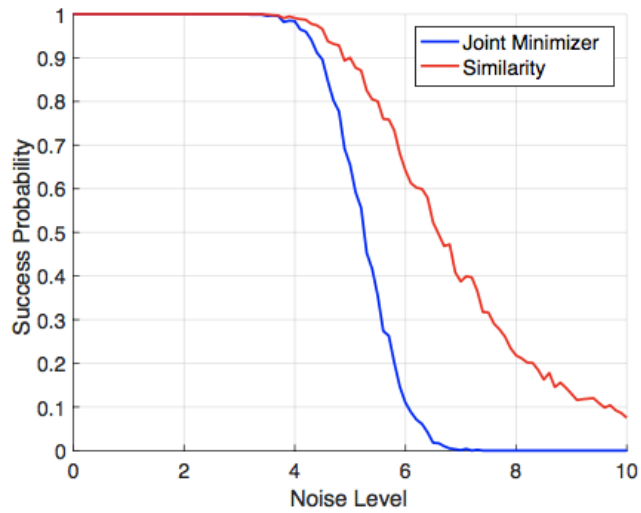
(a) 5% of solutions are stable.



(b) 10% of solutions are stable.

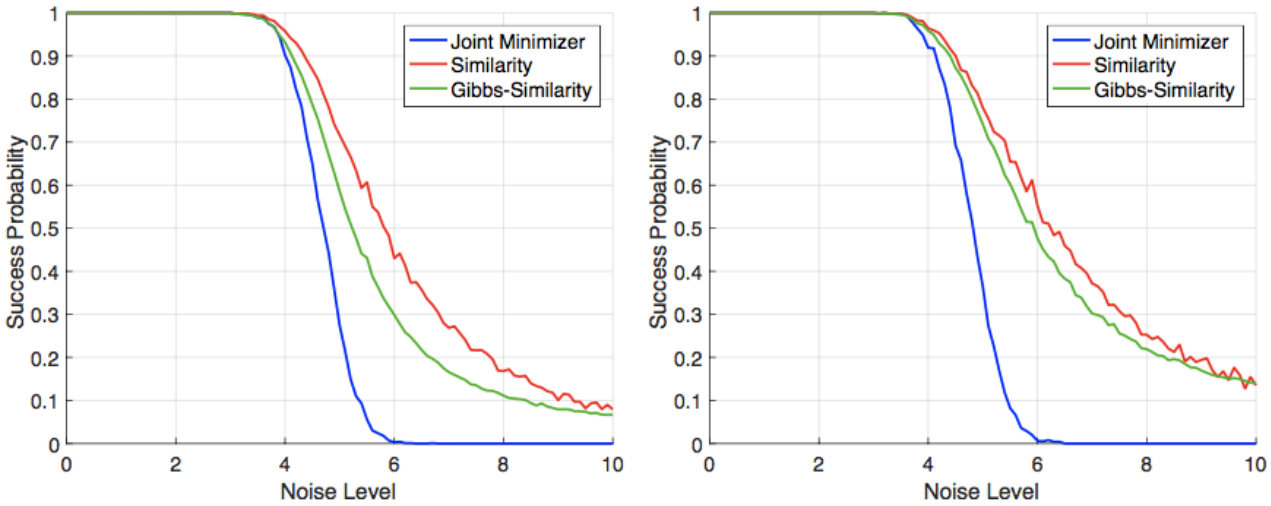


(c) 20% of solutions are stable.



(d) 50% of solutions are stable.

Appendix: Gibbs Similarity Approach



(a) 5% of solutions are stable. **(b)** 10% of solutions are stable.

Figure 3.11 Gibbs relaxation shows almost the same performance. Experimental results where 5% **(a)** and 10% **(b)**. Model and setting are the same as in Section 3.6.

Appendix: Estimate of Similarity – I

Theorem 3.5. *Let* $\gamma > 0$, $V = |\mathcal{C}_\gamma(X') \cap \mathcal{C}_\gamma(X'')|$, $W = |\mathcal{C}_\gamma(X')| \cdot |\mathcal{C}_\gamma(X'')|$, m *be the minimum cost of a solution in both* X' *and* X'' *(i.e., the calibrating assumption is satisfied), and* F_s *and* F_u *denote the cumulative density functions of the stable and the unstable solutions, respectively, evaluated at* $m + \gamma$. *Then, the expected similarity* (3.47) *can be approximated by the estimated similarity*

$$S_\gamma^{\mathrm{EXP}} \sim \widehat{S}_\gamma := |\mathcal{C}| \left(\frac{\mathbb{E}[V]}{\mathbb{E}[W]} - \frac{\mathrm{Cov}(V, W)}{\mathbb{E}[W]^2} + \frac{\mathrm{Var}[W] \cdot \mathbb{E}[V]}{\mathbb{E}[W]^3} \right) \tag{3.50}$$

Appendix: Estimate of Similarity – II

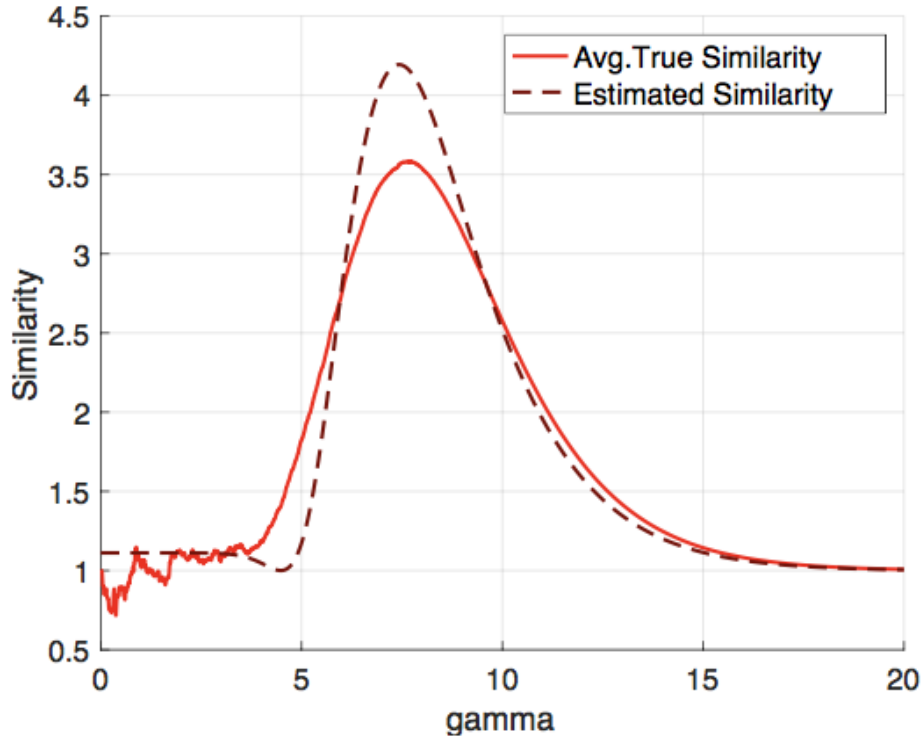
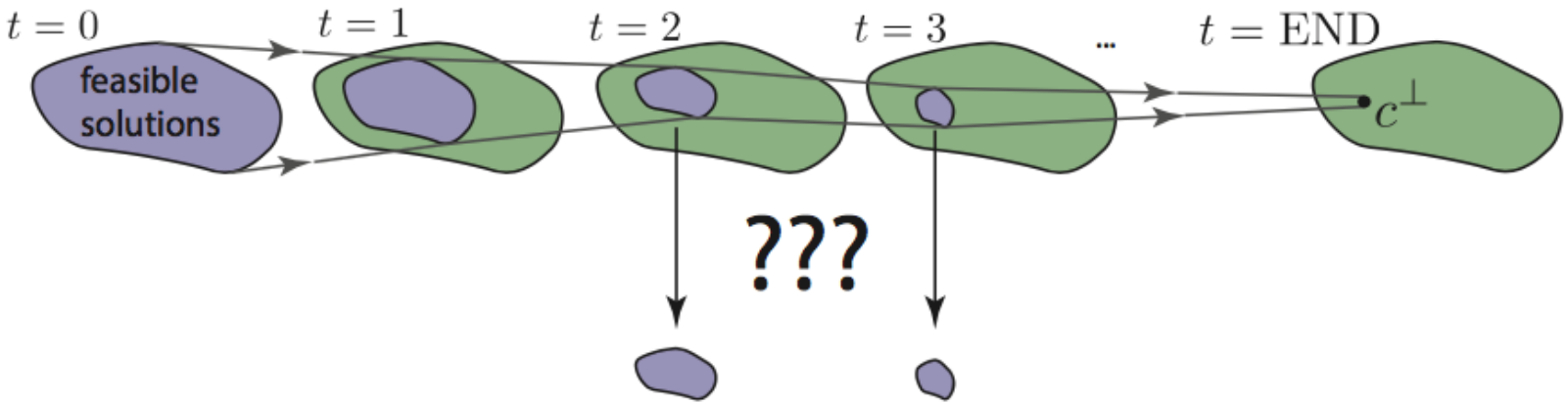


Figure 3.10 Average vs. estimated similarity for $\sigma_u = 1$ **(a)**, and for $\sigma_u = 5$ **(b)**.

relatively well, especially for larger values of γ. Although the discrepancy grows with the noise (which is natural due to the Taylor expansion used in the proof), probably the most important thing to note is that the positions of the γ^* computed based on $\widehat{S}_\gamma$ and $\bar{S}_\gamma$ remain the same.

Appendix: Algorithmic t-eLPA

4.4.3 Algorithmic ASC Score and Optimal Stopping



Appendix: MST – Maximum t-eLPA

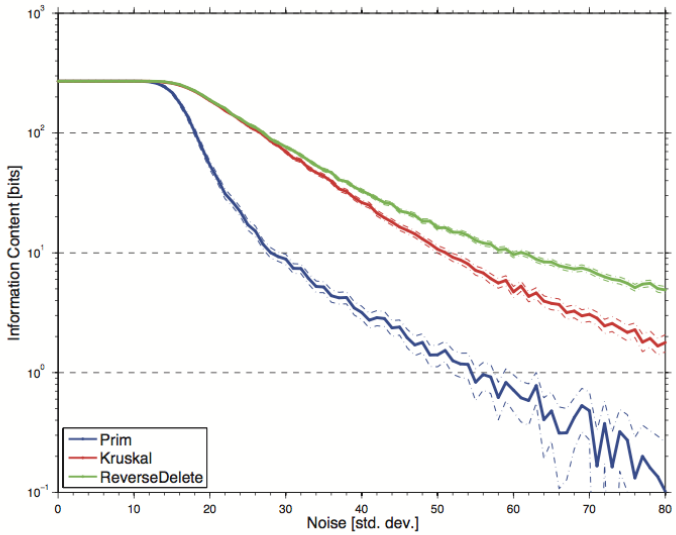


Figure 4.4 Gaussian noise model: information content

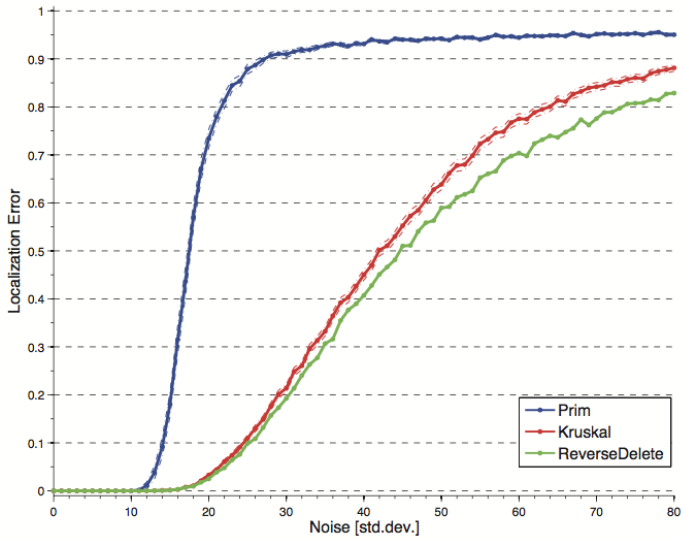


Figure 4.5 Gaussian noise model: localization error

Appendix: Algorithmic t-eLPA

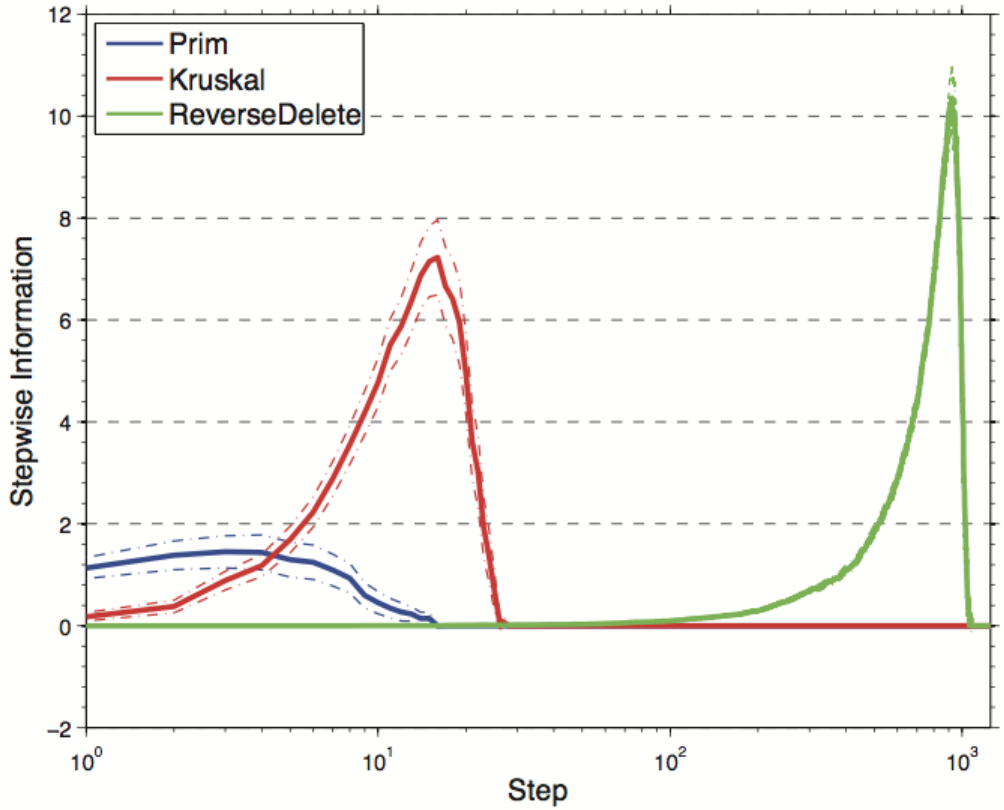


Figure 4.7 Gaussian noise model: stepwise algorithmic information defined in (4.6) ($\sigma = 48$)