Perturbation, Optimization & Statistics workshop

Probabilistic inference by randomly perturbing max-solvers

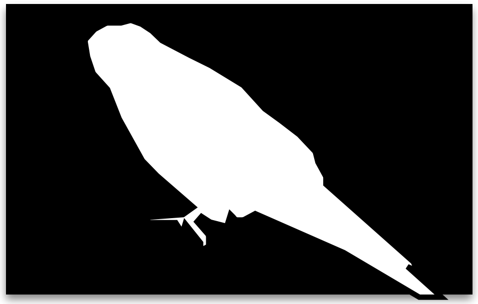
Tamir Hazan
University of Haifa

Inference in machine learning

- Complex structures dominate machine learning applications:

Inference in machine learning

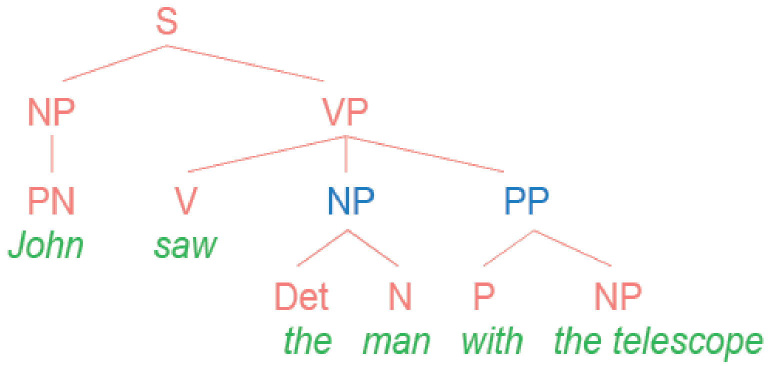
- Complex structures dominate machine learning applications:
  - Computer vision



Inference in machine learning

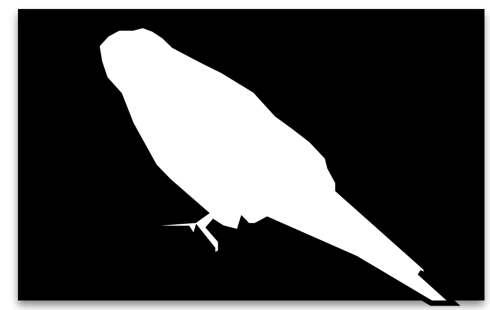
- Complex structures dominate machine learning applications:
  - Computer vision

  - Natural language processing

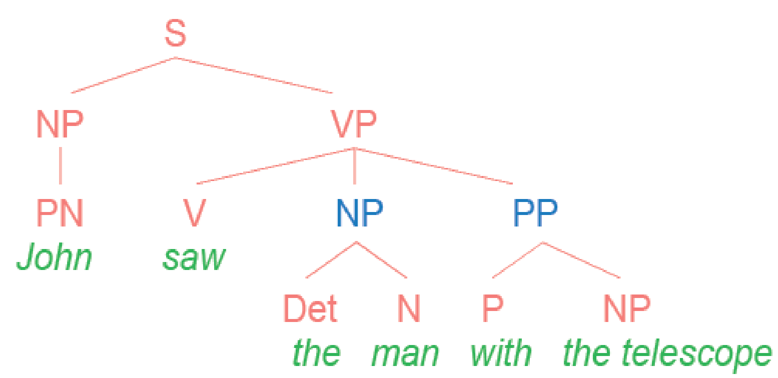


Inference in machine learning

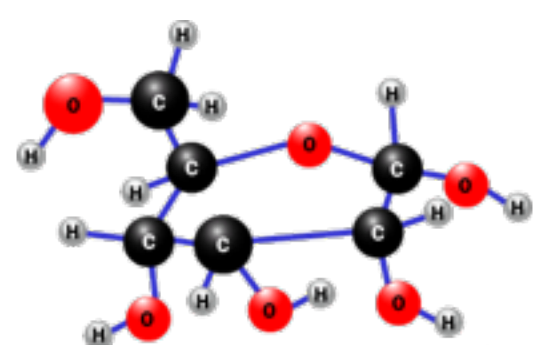
- Complex structures dominate machine learning applications:
  - Computer vision

  - Natural language processing



  - Computational biology



  - and more..

Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.

Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs with perturb-max

Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs with perturb-max
- Application: interactive annotation.
  - New entropy bounds for perturb-max models.

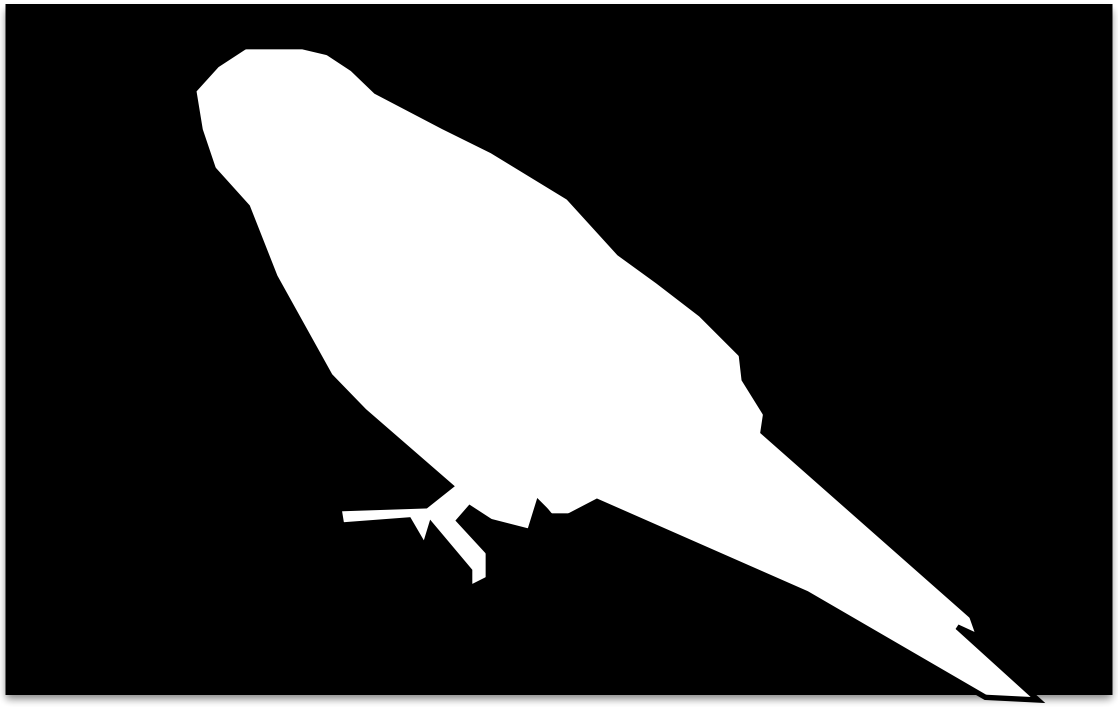
Inference in machine learning

- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$

Inference in machine learning

- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$

$$y \in \{0, 1\}^n$$



Inference in machine learning

- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$

Inference in machine learning

- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$
  - potential function that scores these structures

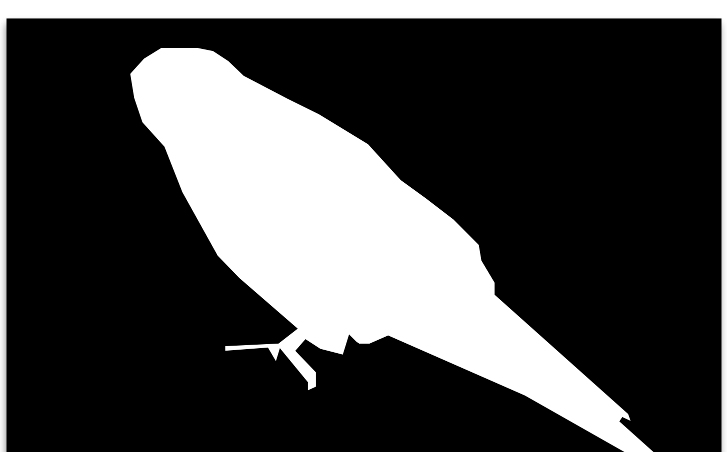
$$\theta(y_1, ..., y_n)$$

Inference in machine learning

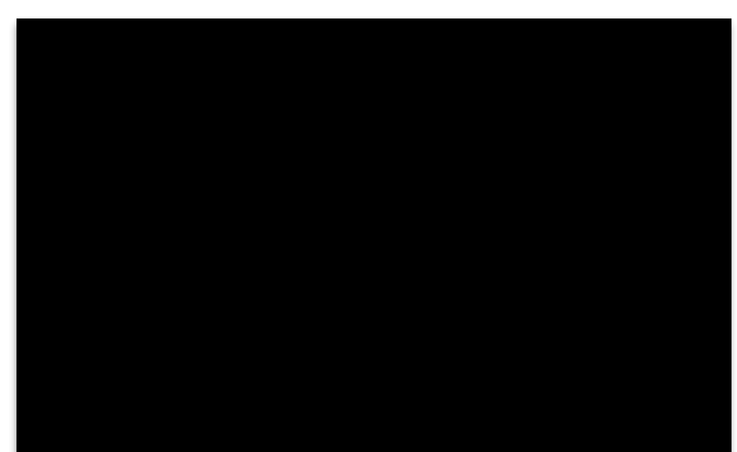
- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$
  - potential function that scores these structures

$$\theta(y_1, ..., y_n)$$

high score



low score

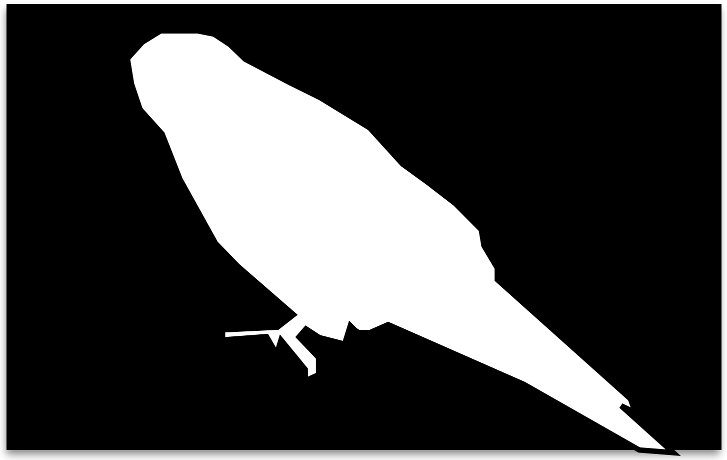


Inference in machine learning

- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$
  - potential function that scores these structures

$$\theta(y_1, ..., y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$

high score



low score

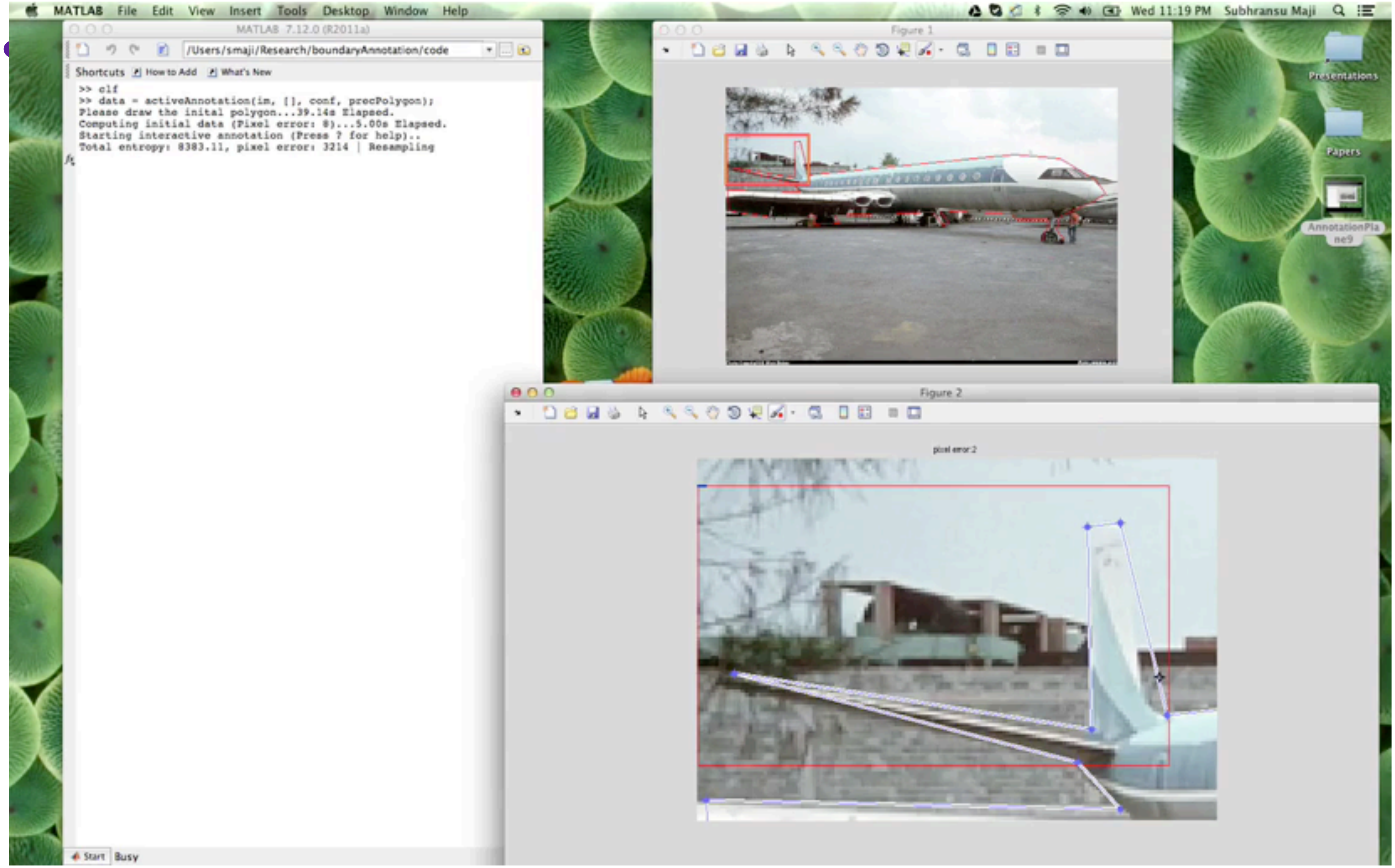
Inference in machine learning

- machine learning applications are characterized by:
  - complex structures $y = (y_1, ..., y_n)$
  - potential function that scores these structures

$$\theta(y_1, ..., y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$

- For machine learning we need to efficiently infer from distributions over complex structures.

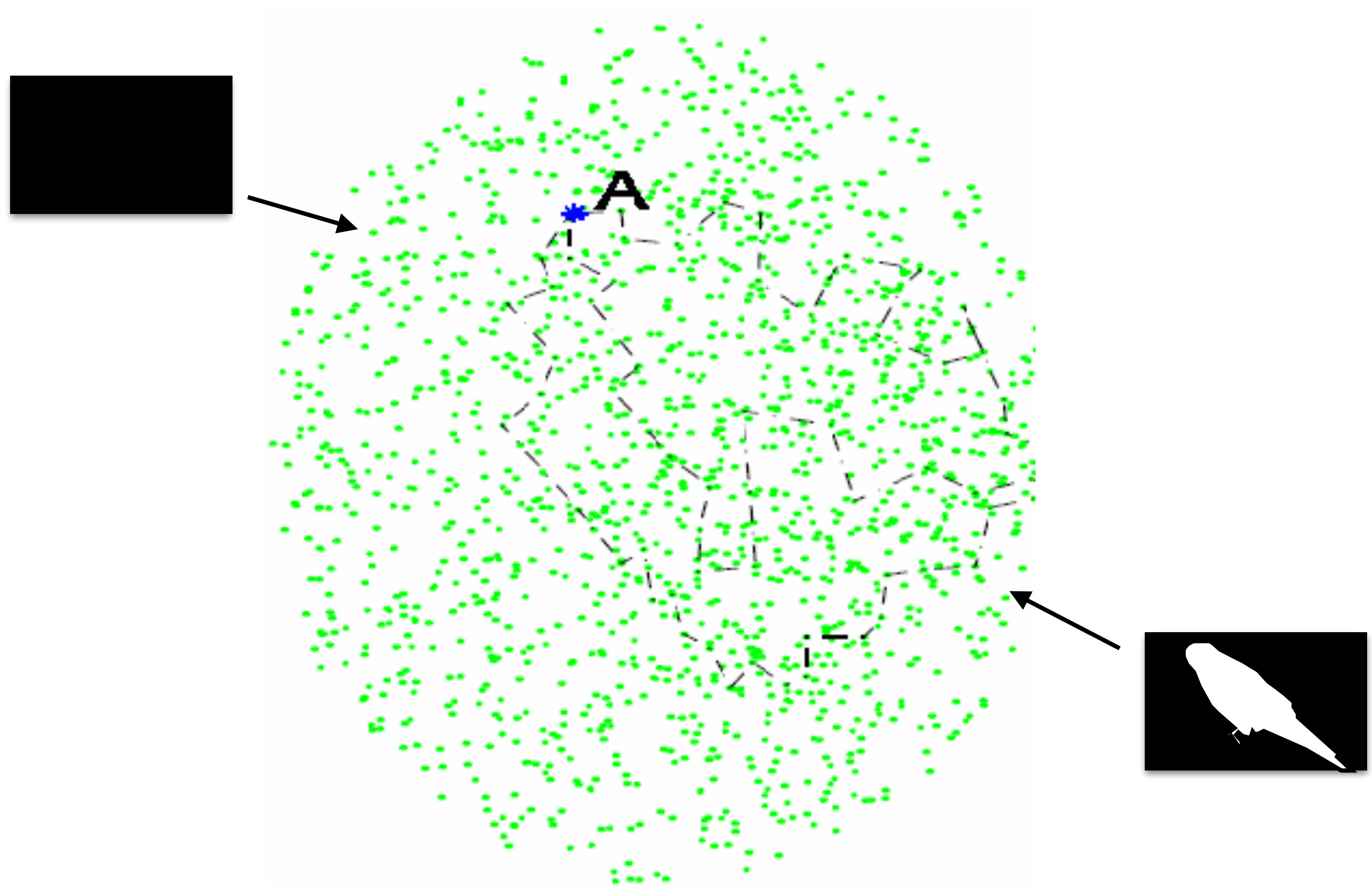
Inference in machine learning



Gibbs distribution

$$p(y_1, ..., y_n) = \frac{1}{Z} \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) \Big)$$

- MCMC samplers:



Gibbs distribution

$$p(y_1, ..., y_n) = \frac{1}{Z} \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) \right)$$

- MCMC samplers:
  - Gibbs sampling, Metropolis-Hastings, Swendsen-Wang
- Many efficient sampling algorithms for special cases:
  - Counting bi-partite matchings in planar graphs (Kasteleyn 61)
  - Ising models (Jerrum 93)
  - Approximating the permanent (Jerrum 04)
  - Many others…

Gibbs distribution

- Gibbs distribution has a significant impact on statistics and computer science

Gibbs distribution

- Gibbs distribution has a significant impact on statistics and computer science
  - Efficient sampling in Ising models (Jerrum 93)

$$p(y) \propto \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\Big)$$

Gibbs distribution

- Gibbs distribution has a significant impact on statistics and computer science
  - Efficient sampling in Ising models (Jerrum 93)
  - Attractive pairwise potentials

$$\theta_{i,j}(y_i, y_j) = \begin{cases} w_{i,j} & \text{if } y_i = y_j \\ -w_{i,j} & \text{otherwise} \end{cases}$$

$$w_{i,j} \geq 0$$

  - No data terms

$$\theta_i(y_i) = 0$$

$$p(y) \propto \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\Big)$$

Gibbs distribution

- Gibbs distribution has a significant impact on statistics and computer science
  - Efficient sampling in Ising models (Jerrum 93)
  - Attractive pairwise potentials

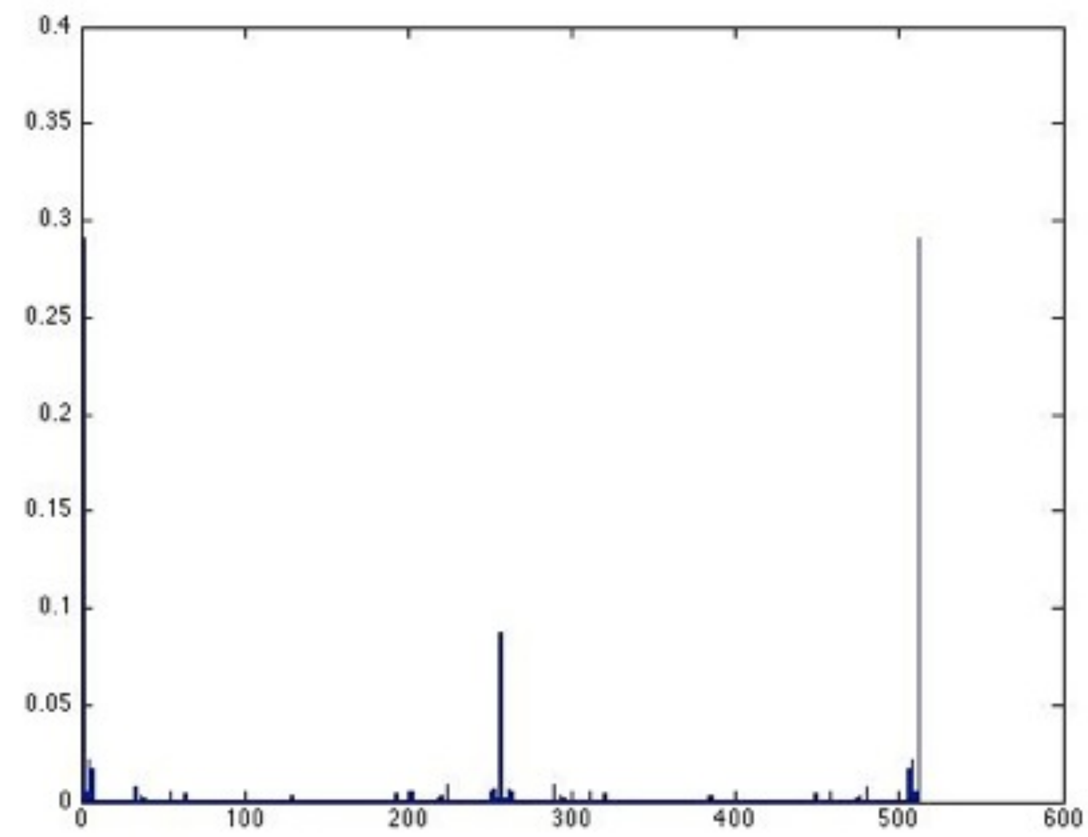
$$\theta_{i,j}(y_i, y_j) = \begin{cases} w_{i,j} & \text{if } y_i = y_j \\ -w_{i,j} & \text{otherwise} \end{cases}$$

$$w_{i,j} \geq 0$$

  - No data terms

$$\theta_i(y_i) = 0$$

$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



Gibbs distribution

- Gibbs distribution has a significant impact on statistics and computer science
  - Efficient sampling in Ising models (Jerrum 93)
  - Attractive pairwise potentials

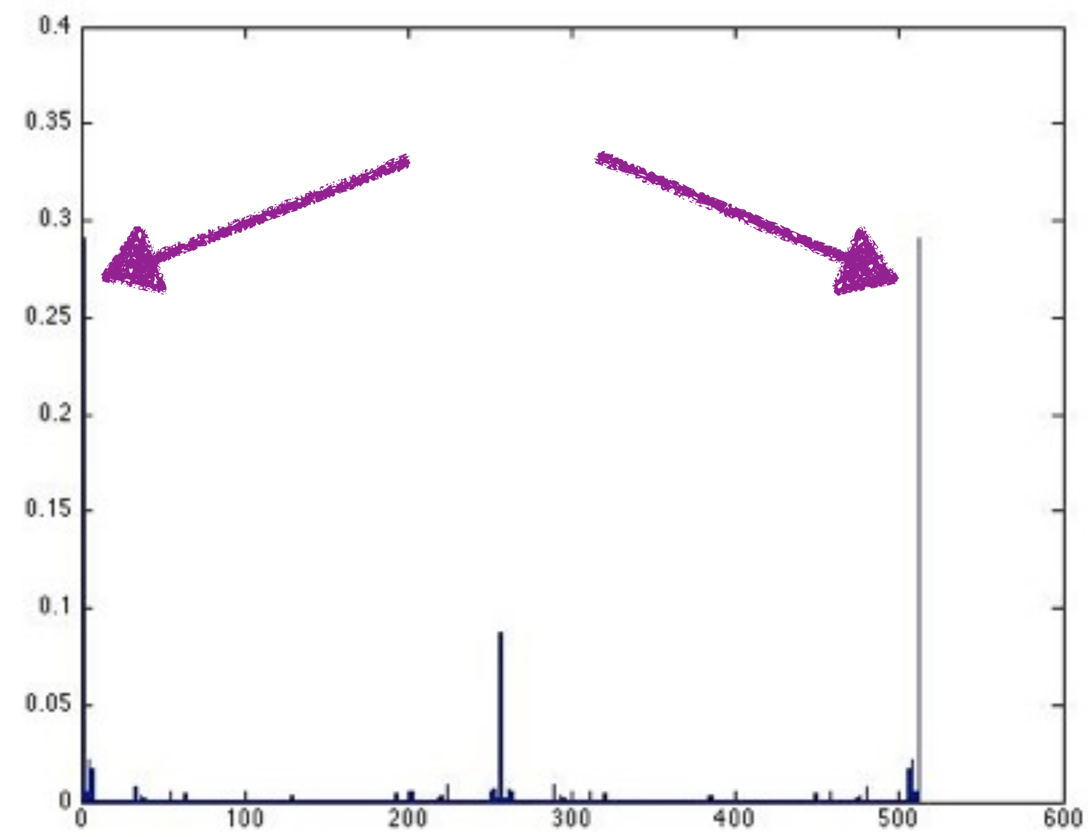
$$\theta_{i,j}(y_i, y_j) = \begin{cases} w_{i,j} & \text{if } y_i = y_j \\ -w_{i,j} & \text{otherwise} \end{cases}$$

$$w_{i,j} \geq 0$$

  - No data terms

$$\theta_i(y_i) = 0$$

$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



- Nicely behaved distribution that is centered around the (1,…,1) or (0,…,0)

Sampling likely structures

- Sampling from the Gibbs distribution is provably hard in AI applications (Goldberg 05, Jerrum 93)

- x_i RGB color of pixel i

$$\theta_i(y_i) = \log p(y_i | x_i)$$

$$p(y) \propto \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\Big)$$

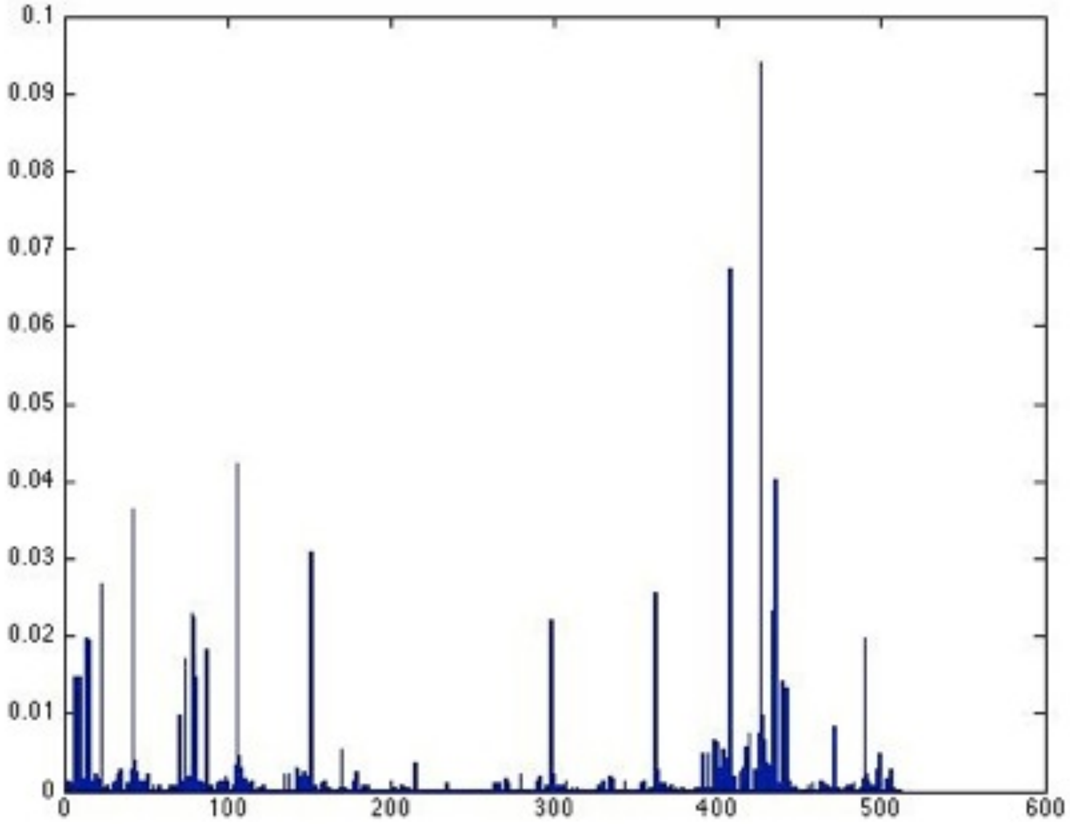
Sampling likely structures

- Sampling from the Gibbs distribution is provably hard in AI applications (Goldberg 05, Jerrum 93)

- x_i RGB color of pixel i

$$\theta_i(y_i) = \log p(y_i | x_i)$$

$$p(y) \propto \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\Big)$$

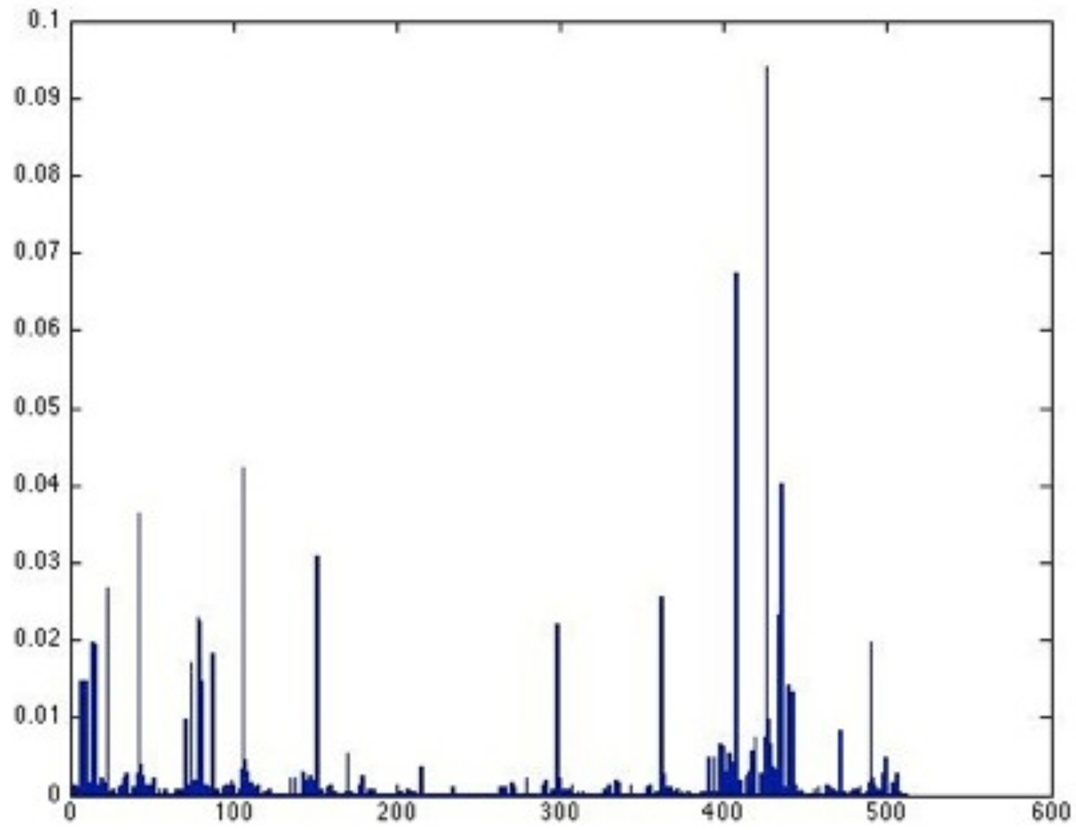


Sampling likely structures

- Sampling from the Gibbs distribution is provably hard in AI applications (Goldberg 05, Jerrum 93)

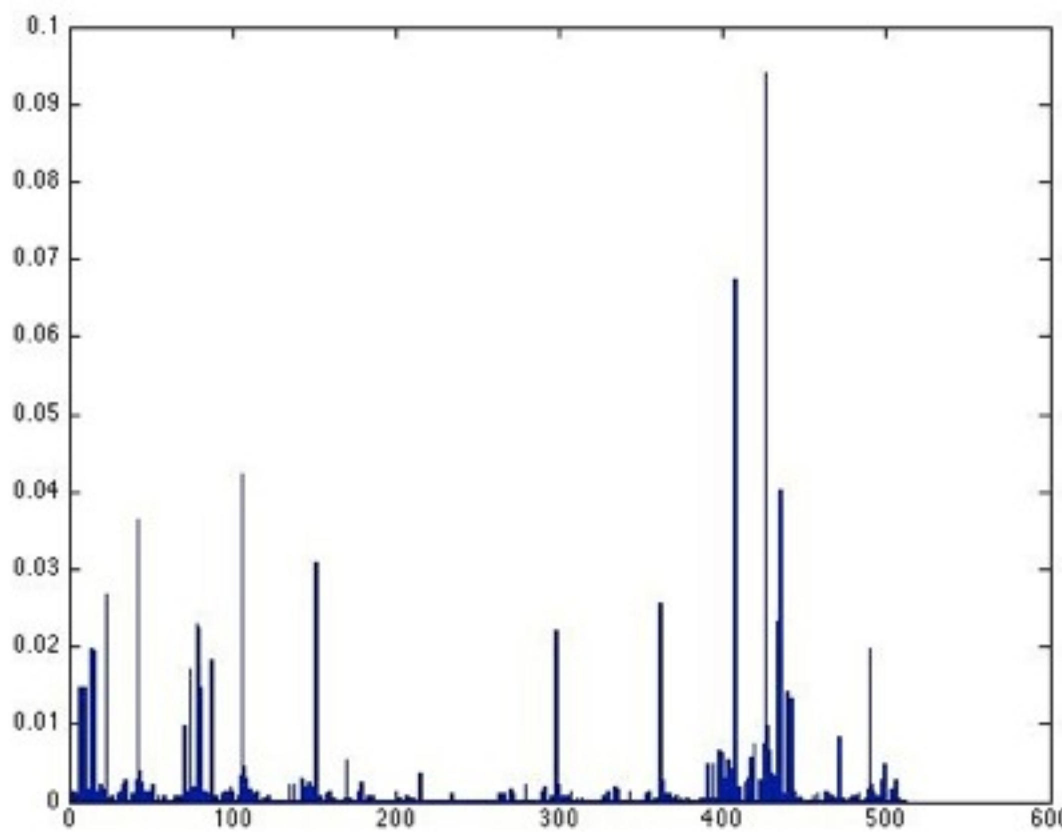
$$\theta_i(y_i) = \log p(y_i|x_i)$$

$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



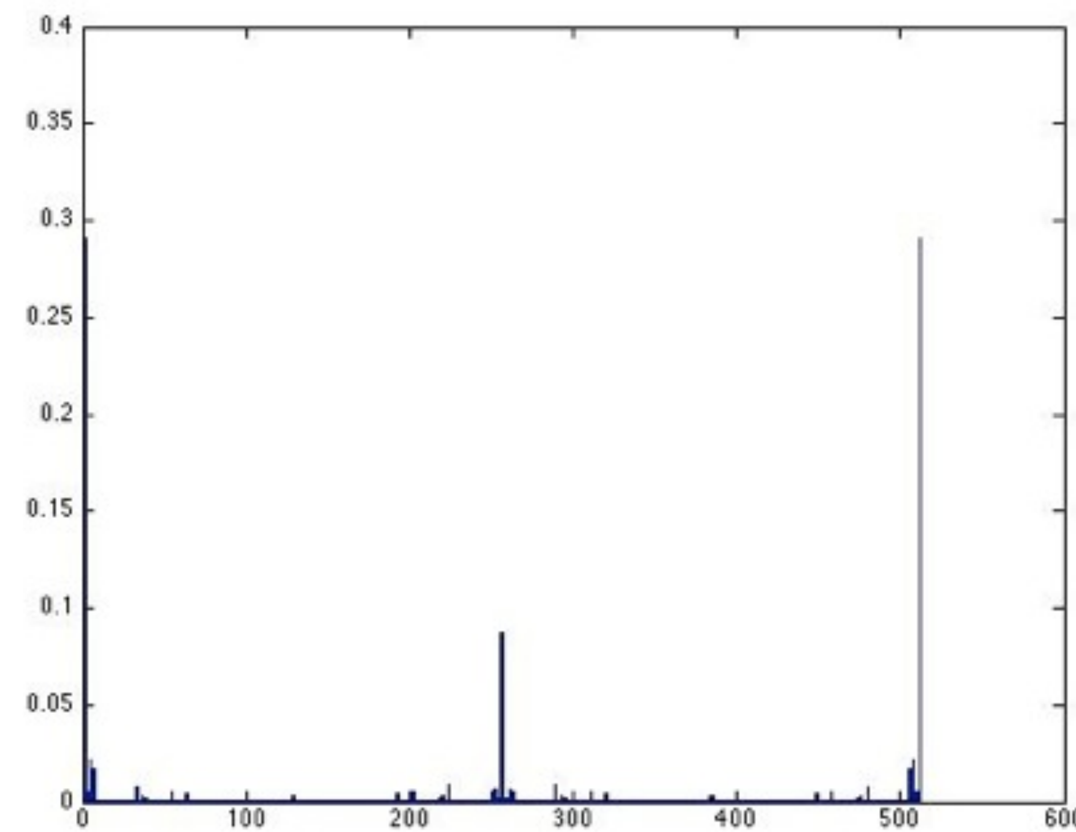
Sampling likely structures

- Sampling from the Gibbs distribution is provably hard in AI applications (Goldberg 05, Jerrum 93)



$$\theta_i(y_i) = \log p(y_i|x_i)$$

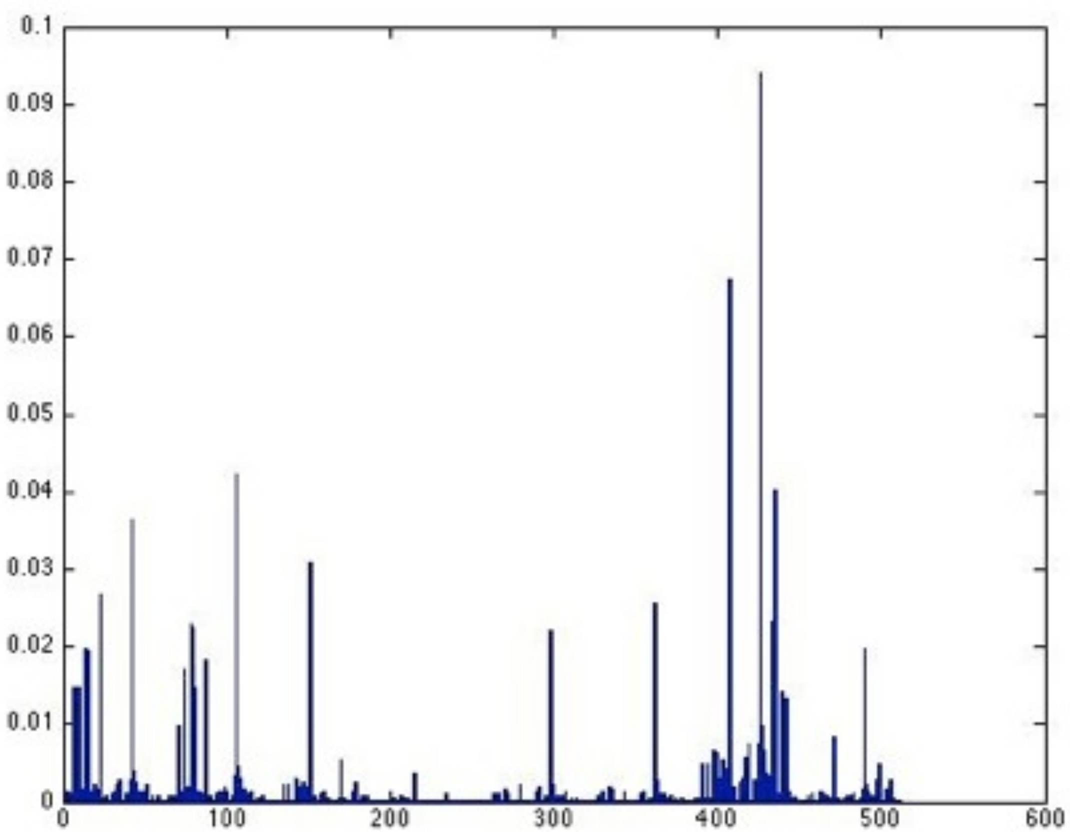
- Recall: sampling from the Gibbs distribution is easy in Ising models (Jerrum 93)



$$\theta_i(y_i) = 0$$

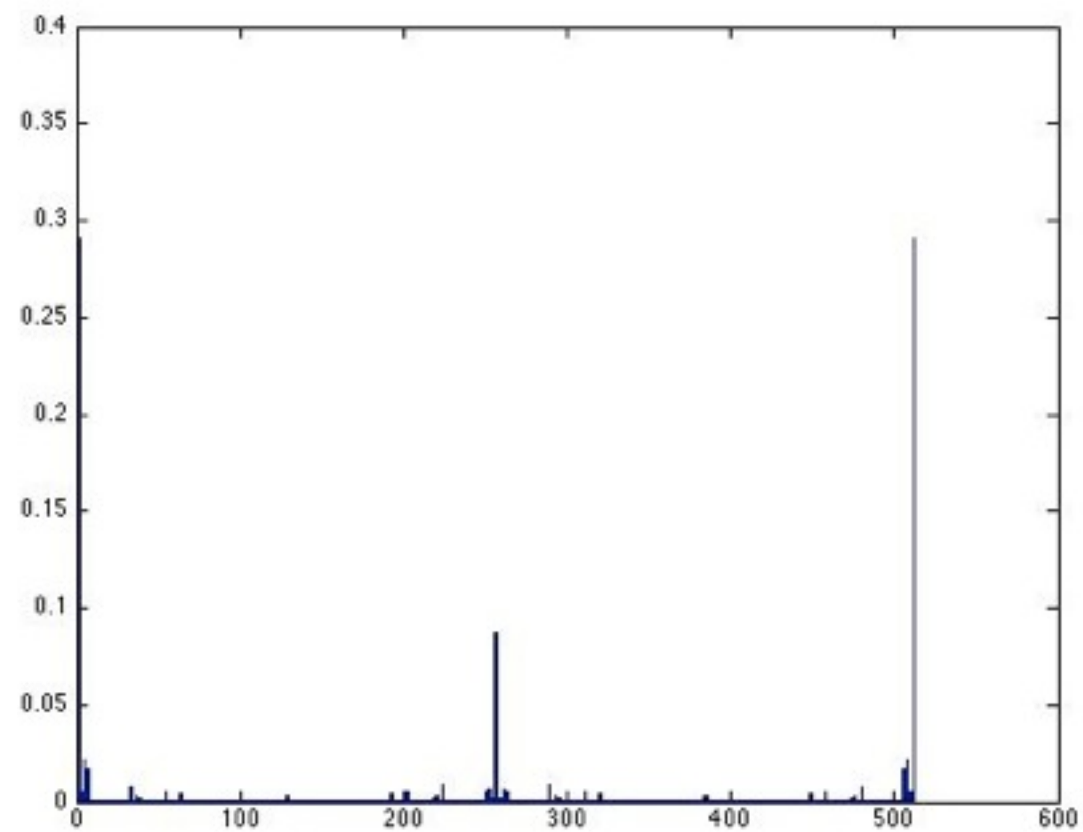
Sampling likely structures

- Sampling from the Gibbs distribution is provably hard in AI applications (Goldberg 05, Jerrum 93)



$$\theta_i(y_i) = \log p(y_i|x_i)$$

- Recall: sampling from the Gibbs distribution is easy in Ising models (Jerrum 93)



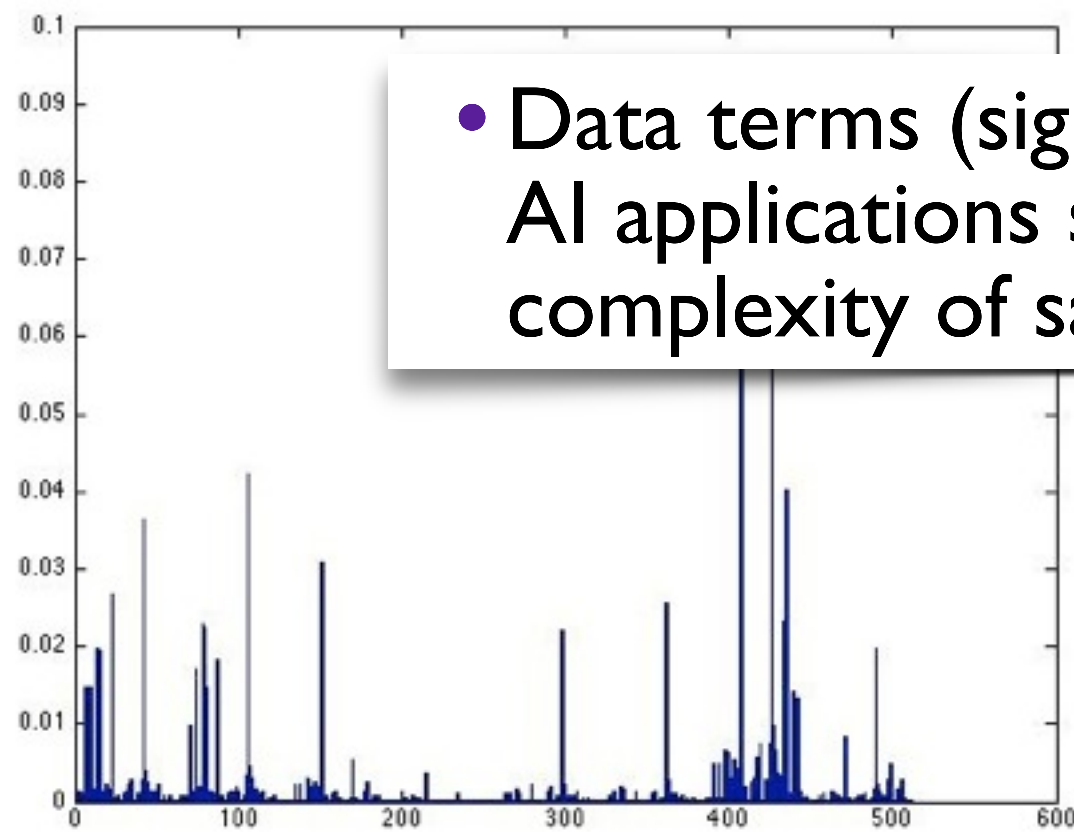
$$\theta_i(y_i) = 0$$

Sampling likely structures

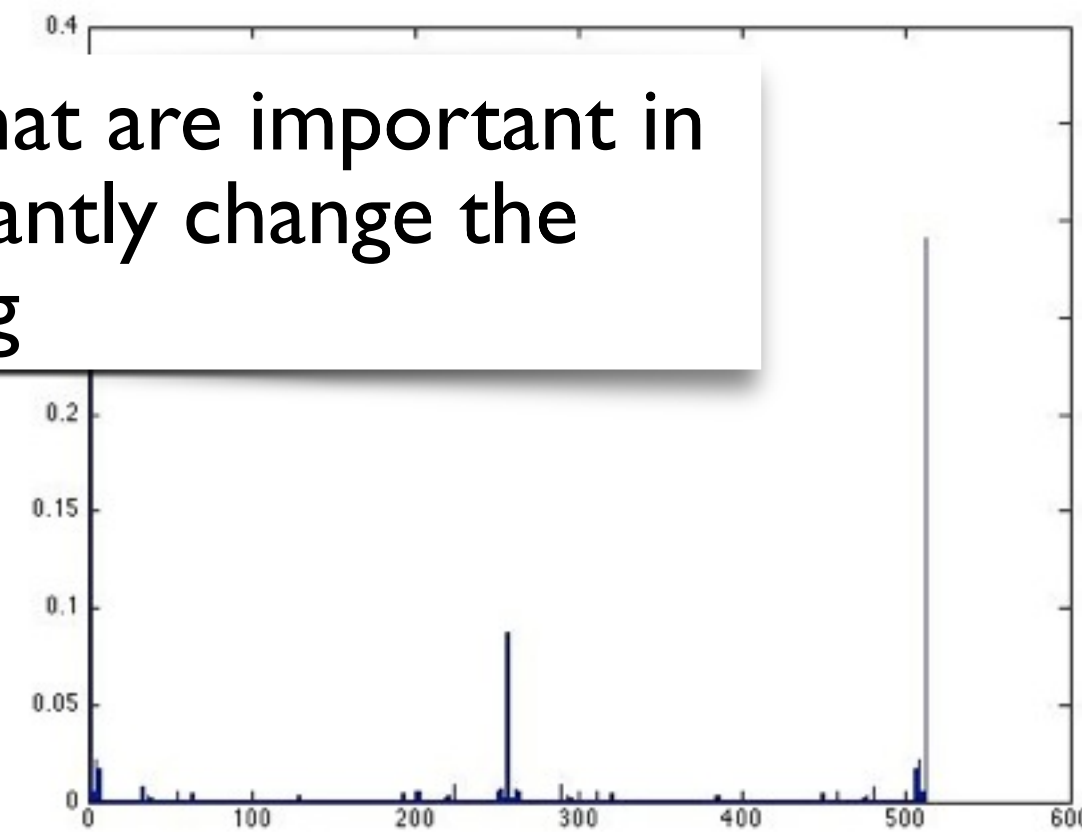
- Sampling from the Gibbs distribution is provably hard in AI applications (Goldberg 05, Jerrum 93)

- Recall: sampling from the Gibbs distribution is easy in Ising models (Jerrum 93)

- Data terms (signals) that are important in AI applications significantly change the complexity of sampling



$\theta_i(y_i) = \log p(y_i|x_i)$



$\theta_i(y_i) = 0$

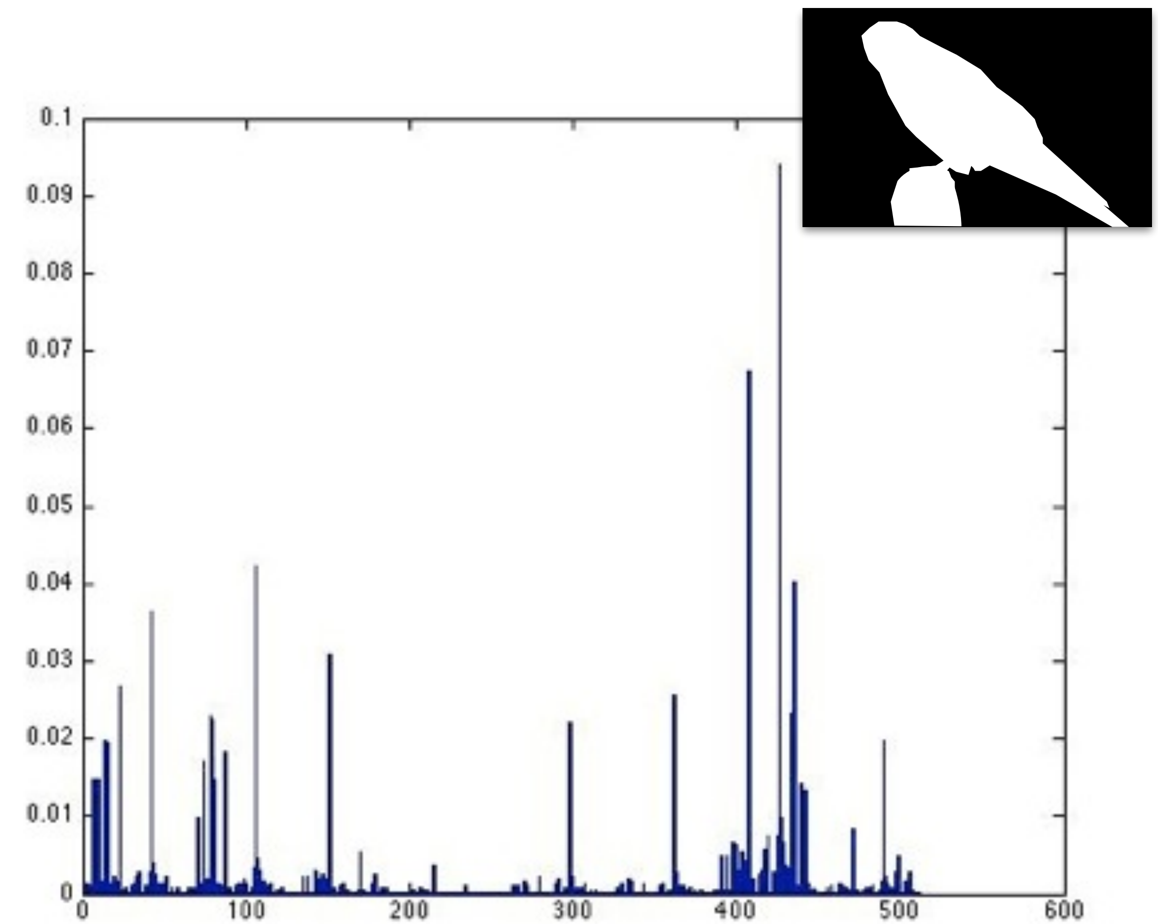
Most likely structure

- Instead of sampling, it may be significantly faster to find the most likely structure

Most likely structure

- Instead of sampling, it may be significantly faster to find the most likely structure

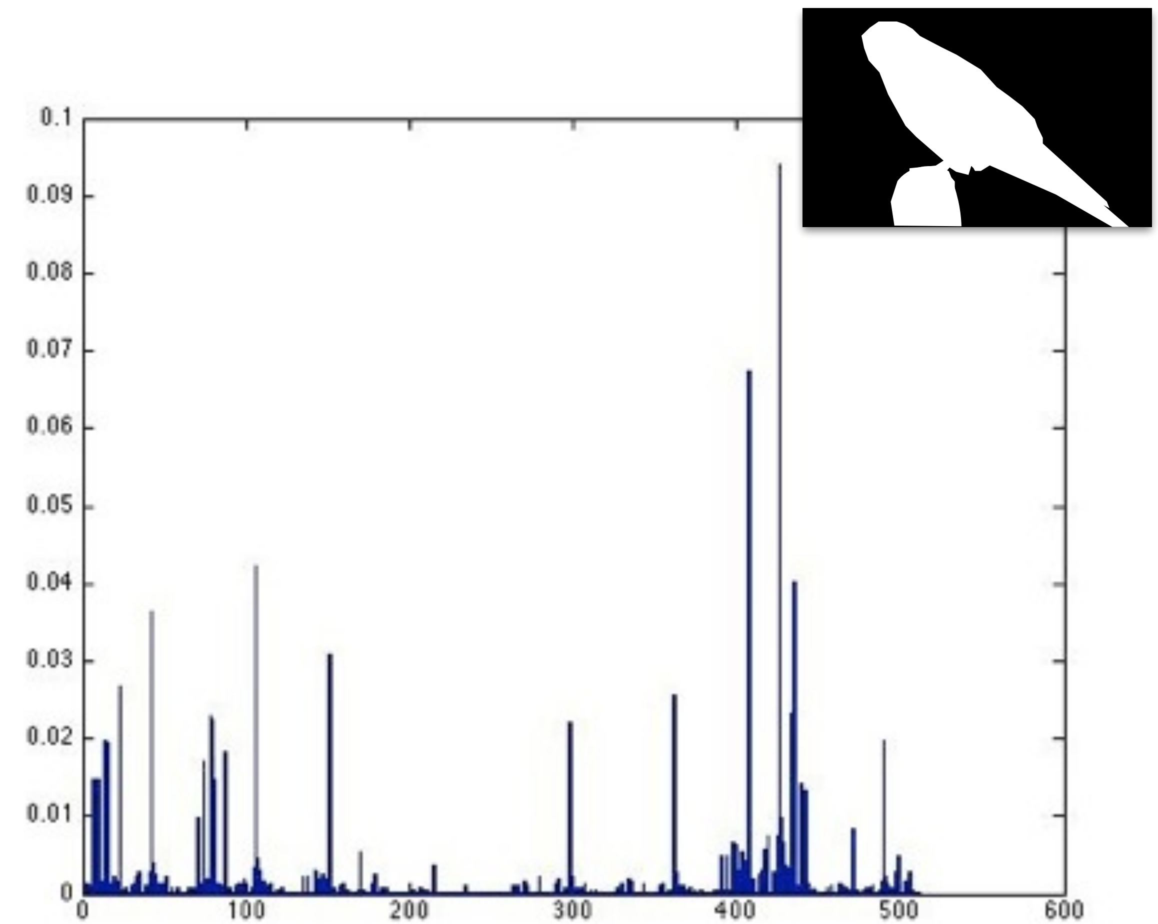
- The most likely structure



Most likely structure

- Instead of sampling, it may be significantly faster to find the most likely structure
  - Graph-cuts

- The most likely structure



Most likely structure

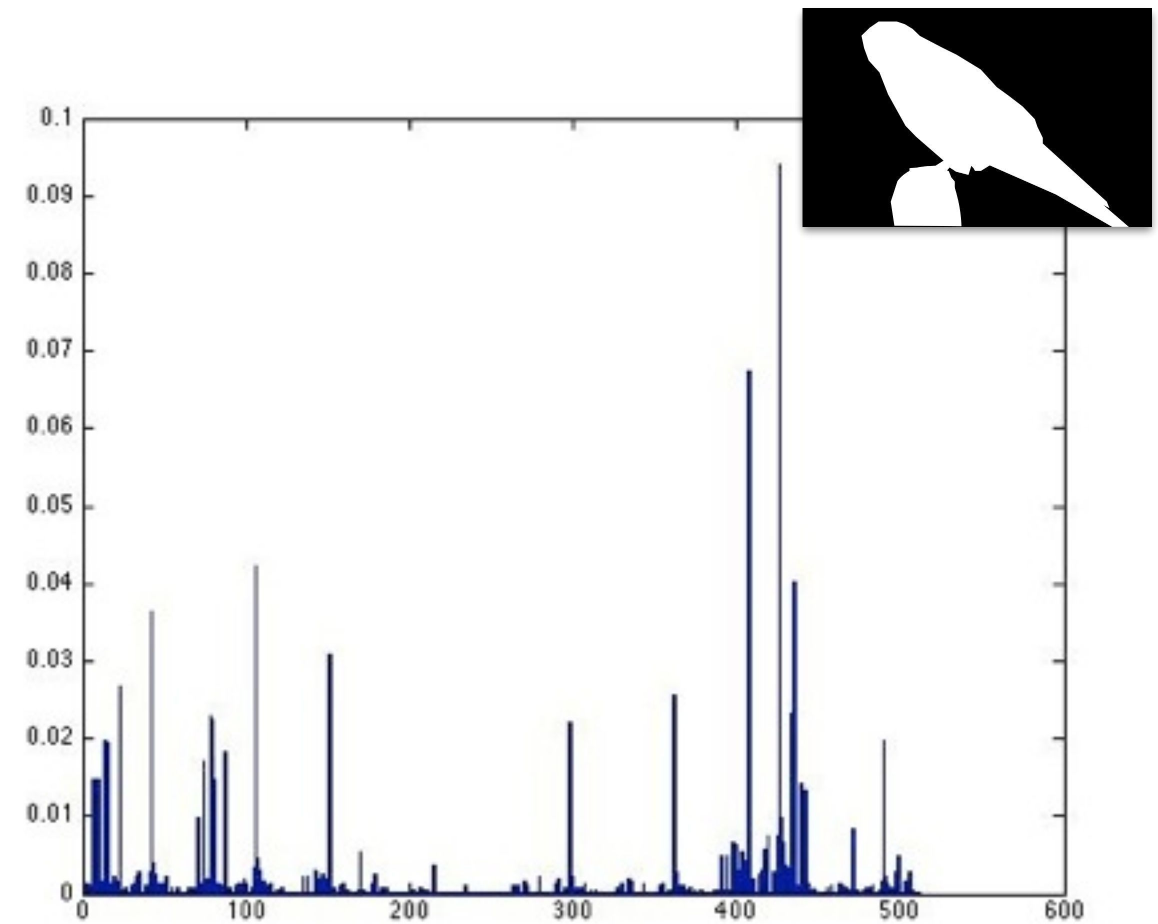
- Instead of sampling, it may be significantly faster to find the most likely structure
  - Graph-cuts

$$\theta_{i,j}(y_i, y_j) = \begin{cases} w_{i,j} & \text{if } y_i = y_j \\ -w_{i,j} & \text{otherwise} \end{cases}$$

$$w_{i,j} \geq 0$$

$$\theta_i(y_i) = \log p(y_i | x_i)$$

- The most likely structure



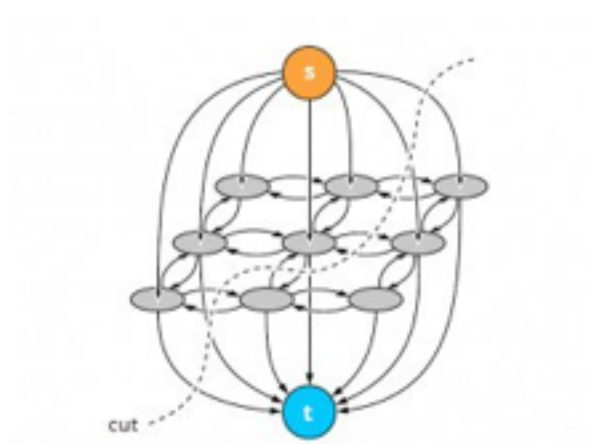
Most likely structure

- Instead of sampling, it may be significantly faster to find the most likely structure
  - Graph-cuts

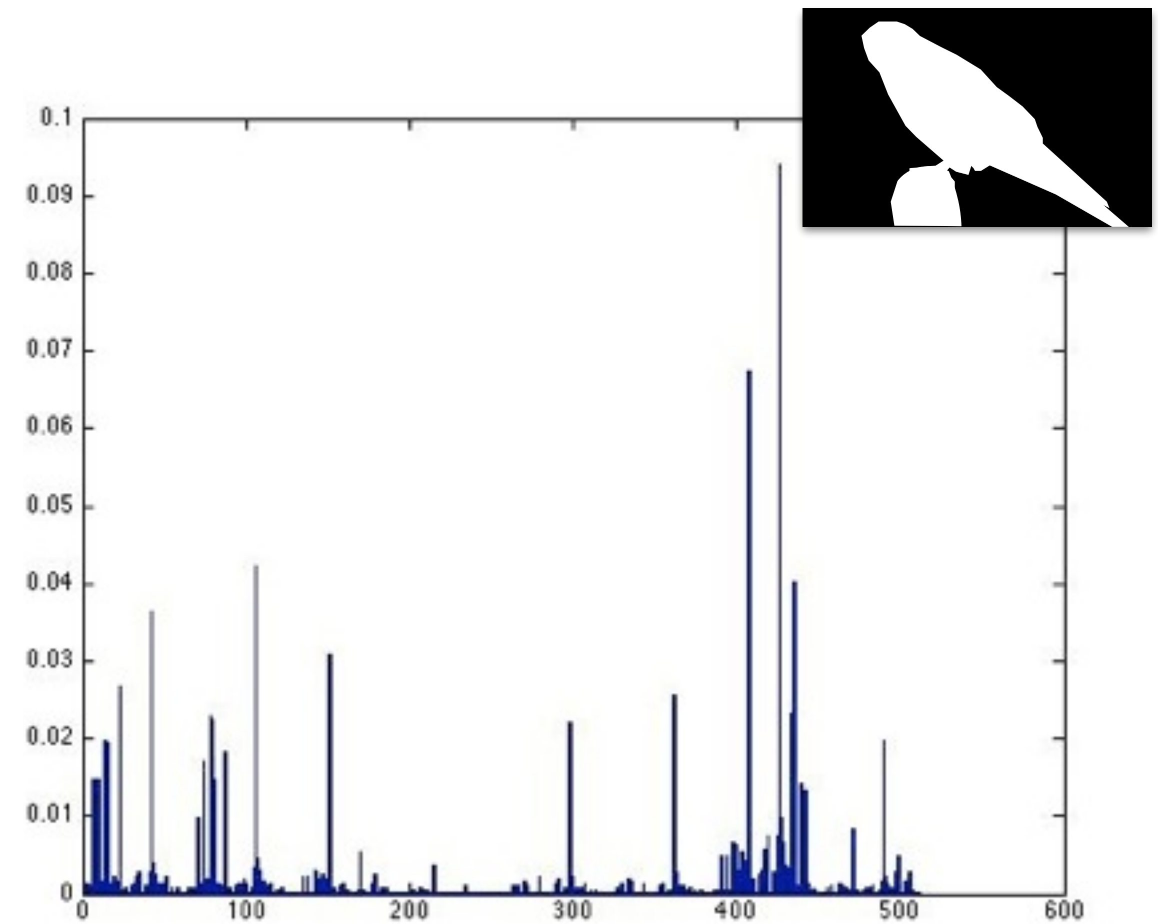
$$\theta_{i,j}(y_i, y_j) = \begin{cases} w_{i,j} & \text{if } y_i = y_j \\ -w_{i,j} & \text{otherwise} \end{cases}$$

$$w_{i,j} \geq 0$$

$$\theta_i(y_i) = \log p(y_i | x_i)$$



- The most likely structure



Most likely structure

$$y^* = \arg \max_{y_1,\dots,y_n} \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)$$

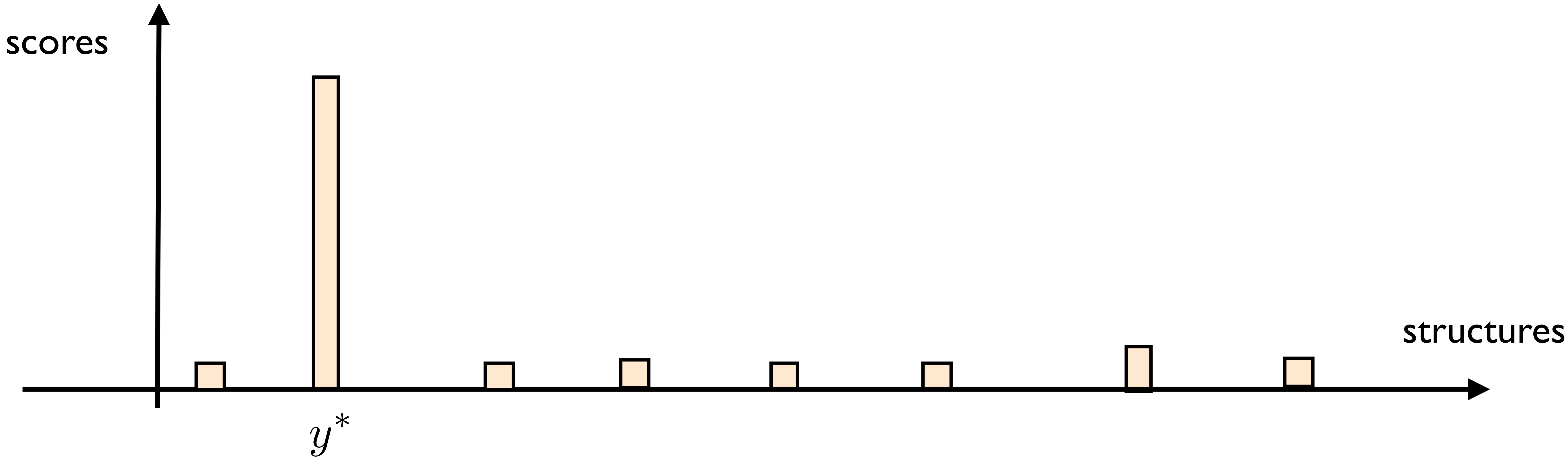
- Maximum a-posterior (MAP) inference.
- Many efficient optimization algorithms for special cases:
  - Beliefs propagation: trees (Pearl 88), perfect graphs (Jebara 10),
  - Graph-cuts for image segmentation
  - branch and bound (Rother 09), branch and cut (Gurobi)
  - Linear programming relaxations (Schlesinger 76, Wainwright 05, Kolmogorov 06, Werner 07, Sontag 08, Hazan 10, Batra 10, Nowozin 10, Pletscher 12, Kappes 13, Savchynskyy13, Tarlow 13, Kohli 13, Jancsary 13, Schwing 13)
  - CKY for parsing
  - Many others…

The challenge

Sampling from the likely high dimensional structures (with millions of variables, e.g., image segmentation with 12 million pixels) as efficient as optimizing

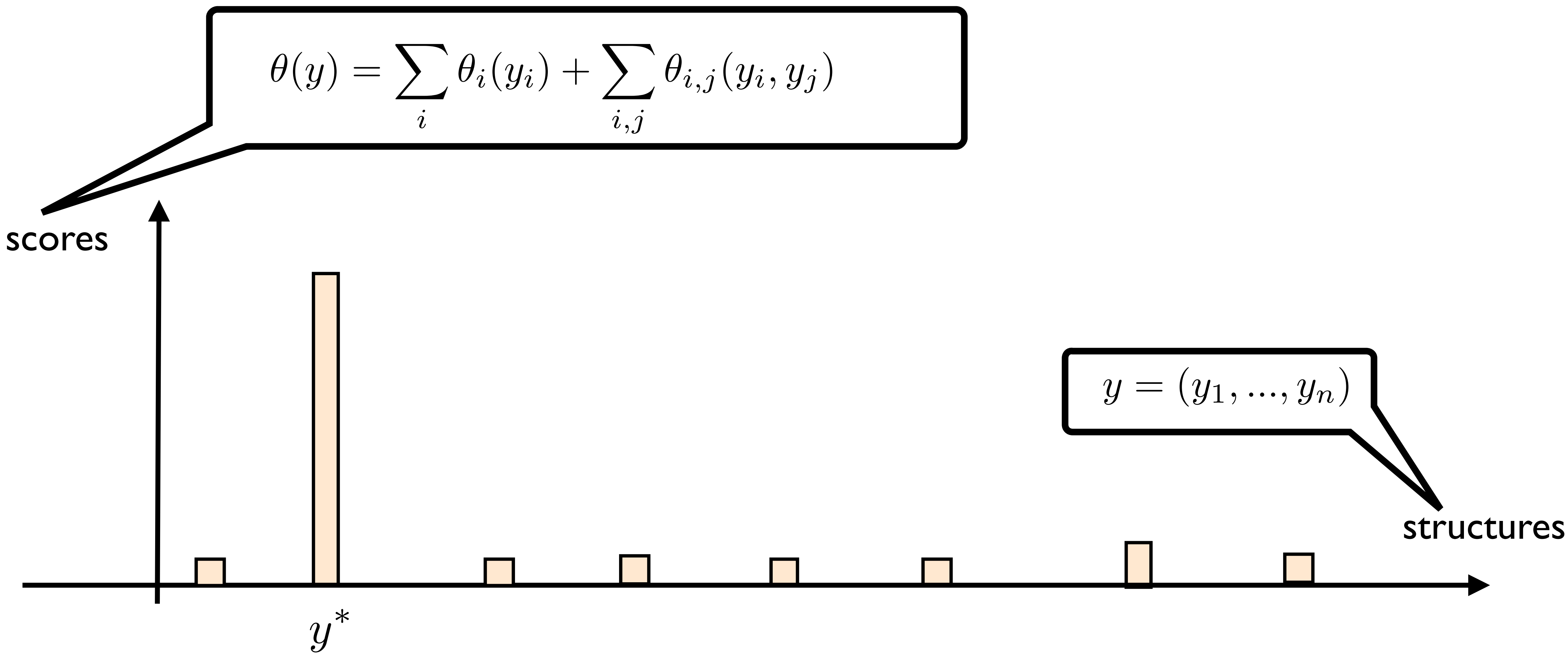
Most likely structure

- Selecting the maximizing structure is appropriate when one structure (e.g., segmentation / parse) dominates others



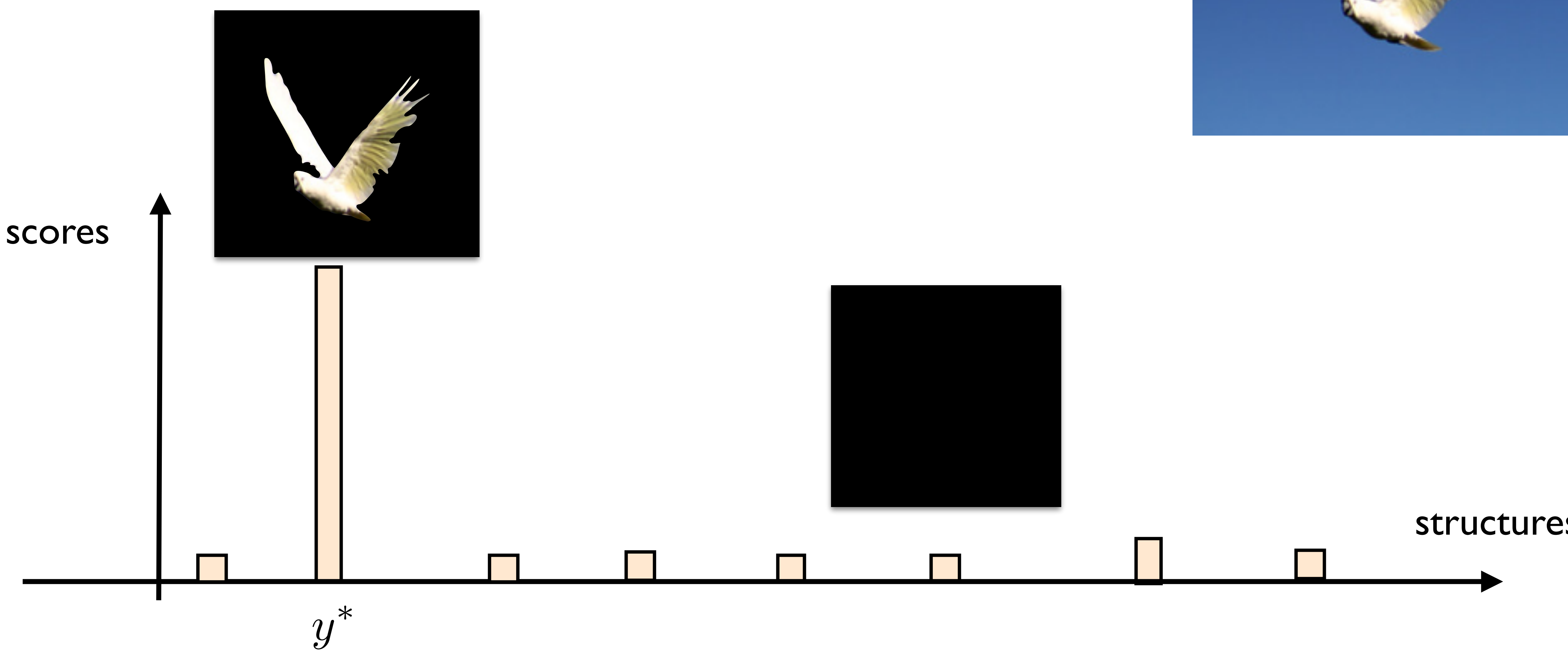
Most likely structure

- Selecting the maximizing structure is appropriate when one structure (e.g., segmentation / parse) dominates others



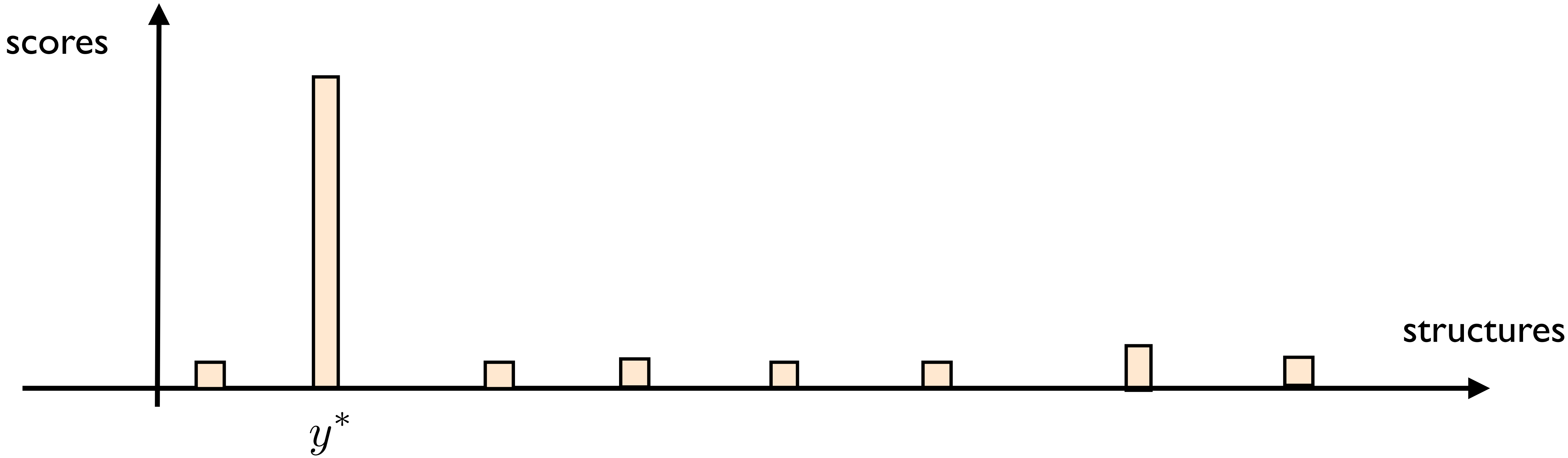
Most likely structure

- Selecting the maximizing structure is appropriate when one structure (e.g., segmentation / parse) dominates others



Most likely structure

- Selecting the maximizing structure is appropriate when one structure (e.g., segmentation / parse) dominates others

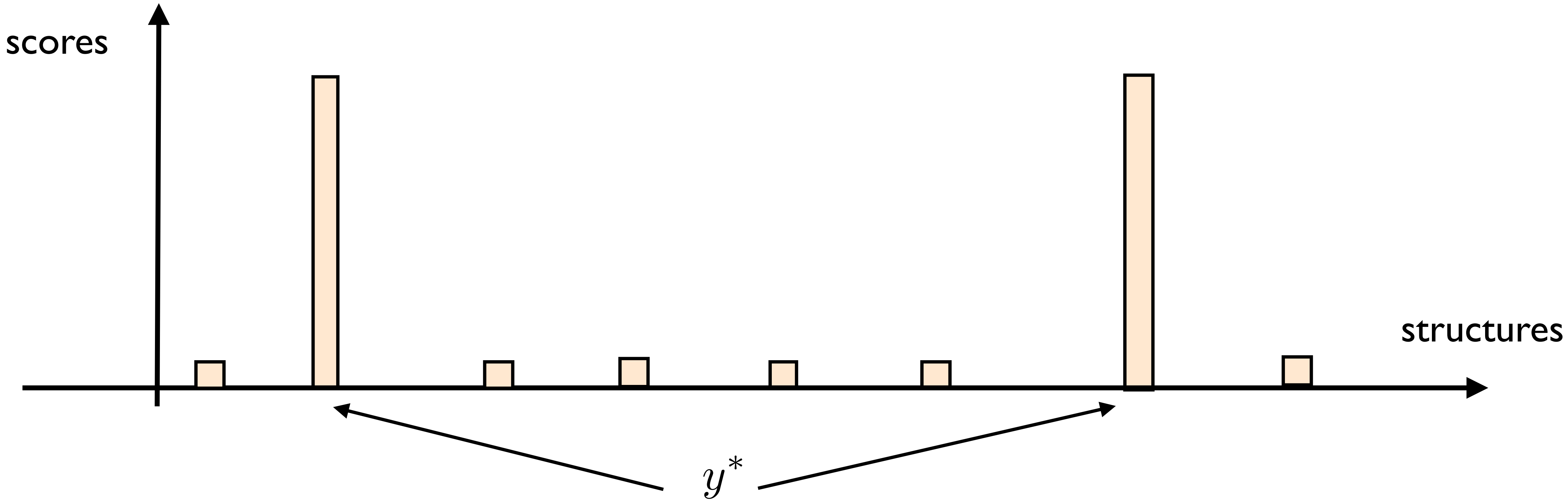


Most likely structure

- The maximizing structure is not robust in case of multiple high scoring alternatives

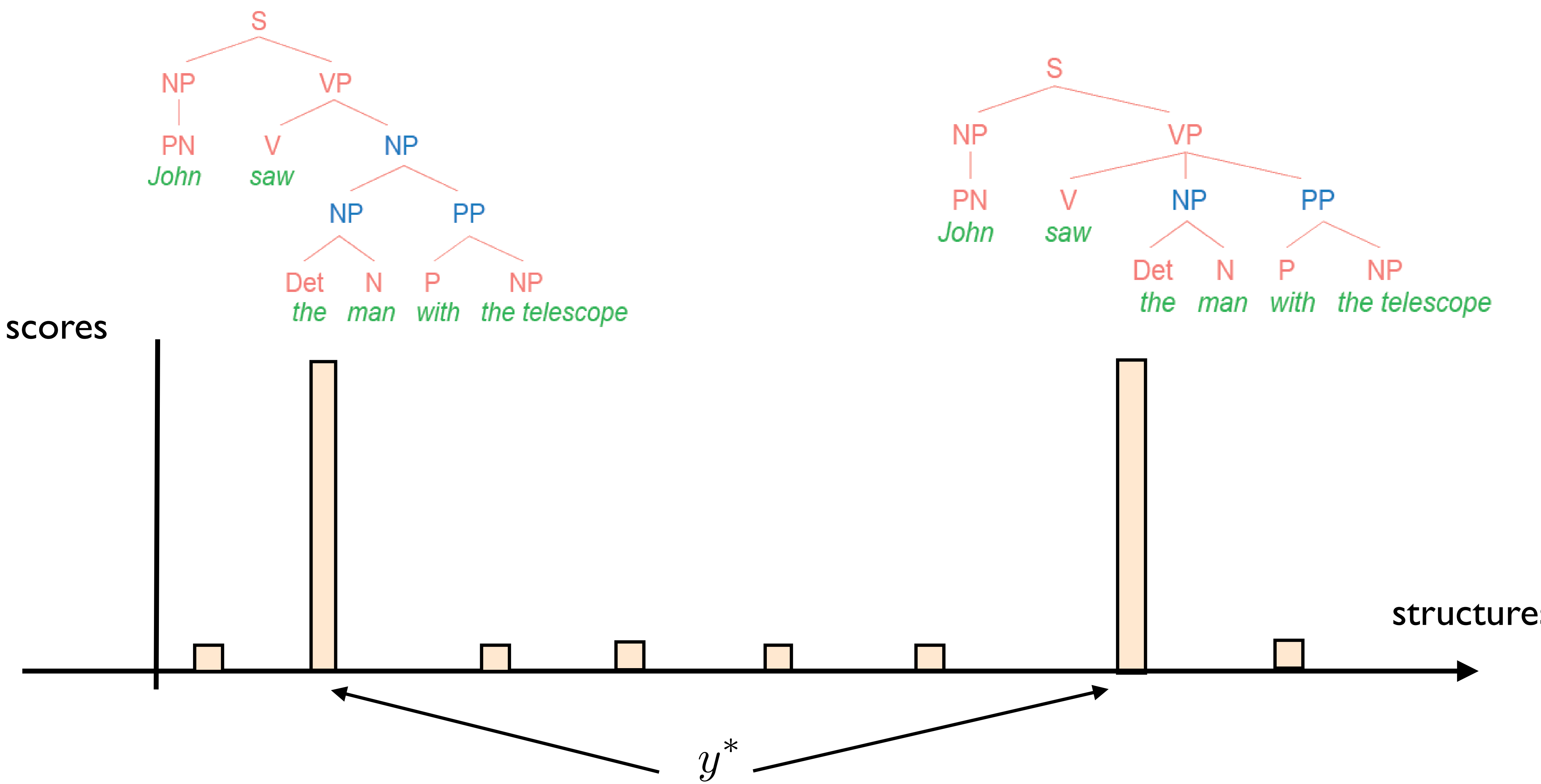
Most likely structure

- The maximizing structure is not robust in case of ambiguities



Most likely structure

- The maximizing structure is not robust in case of ambiguities

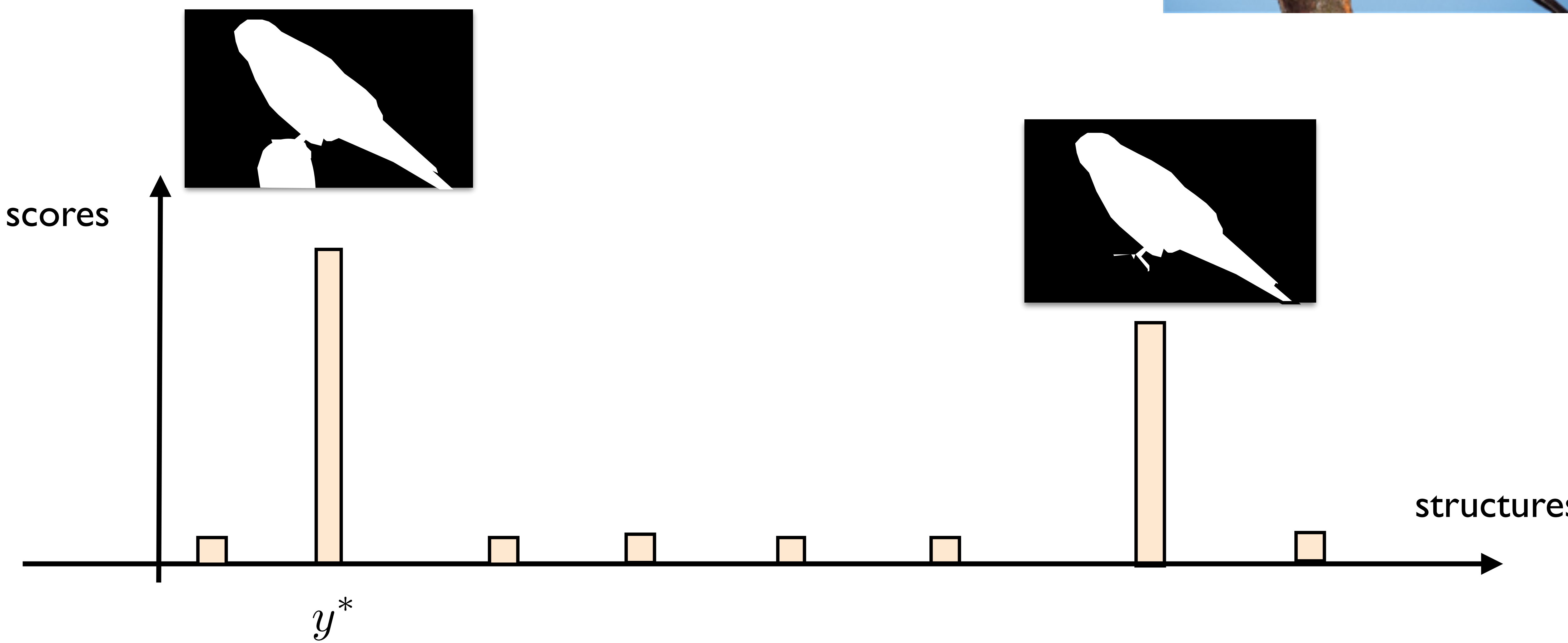


Most likely structure

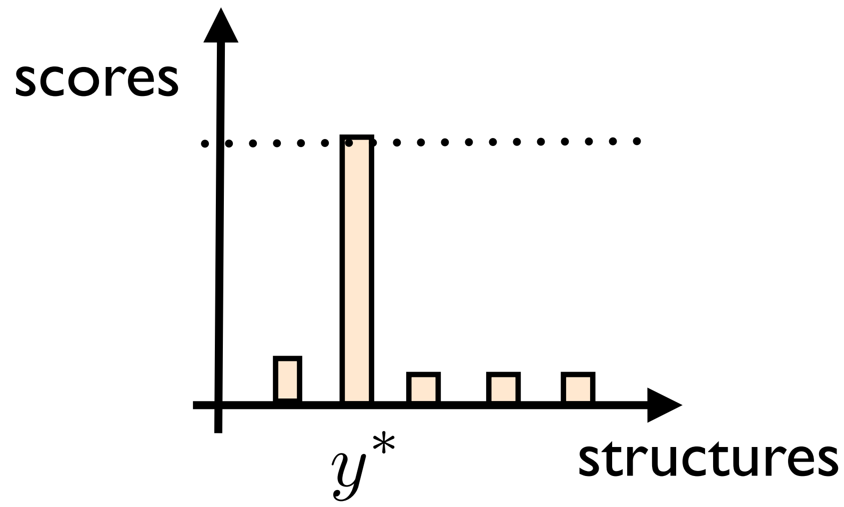
- The maximizing structure is not robust in case of computationally limited models

Most likely structure

- The maximizing structure is not robust in case of computationally limited models

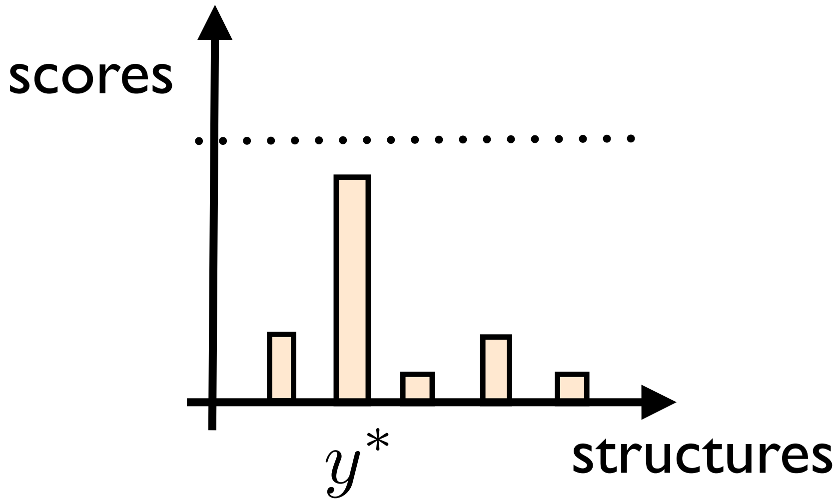


Random perturbations



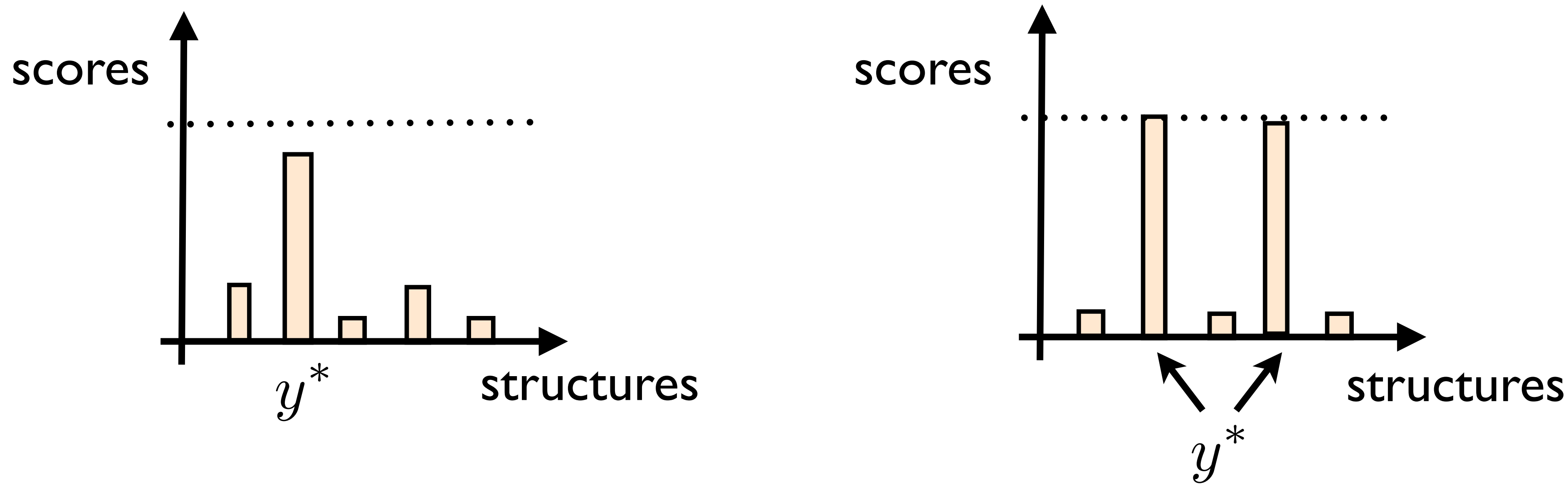
- Randomly perturbing the system reveals its complexity
  - little effect when the maximizing structure is "evident"

Random perturbations



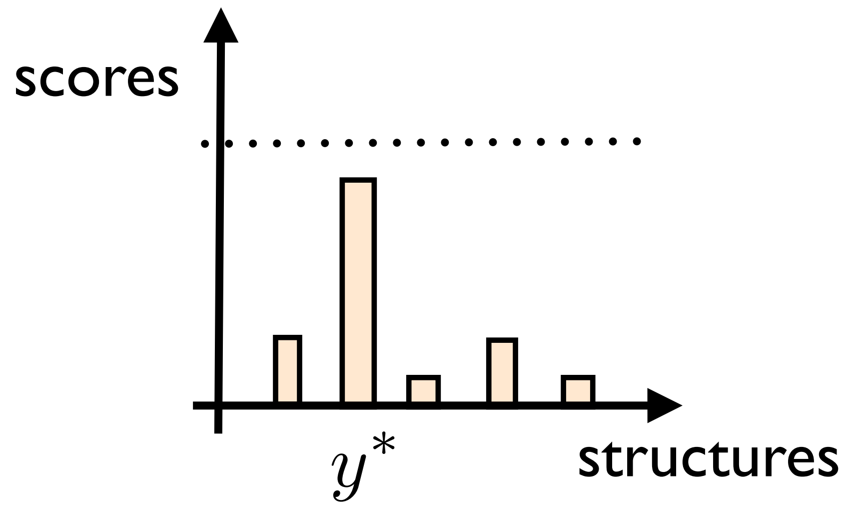
- Randomly perturbing the system reveals its complexity
  - little effect when the maximizing structure is "evident"

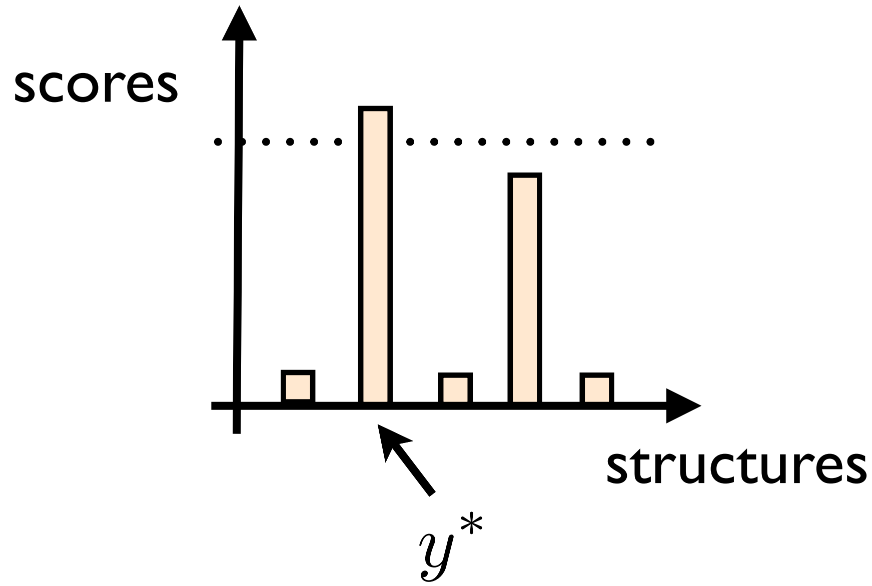
Random perturbations



- Randomly perturbing the system reveals its complexity
  - little effect when the maximizing structure is "evident"
  - substantial effect when there are alternative high scoring structures

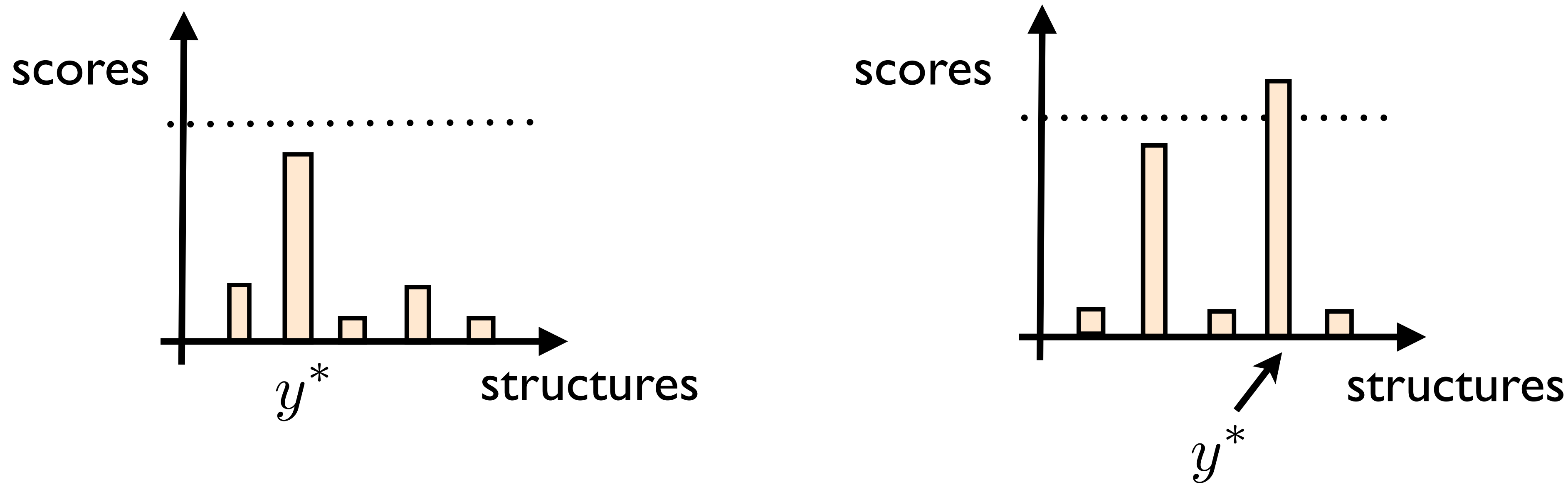
Random perturbations





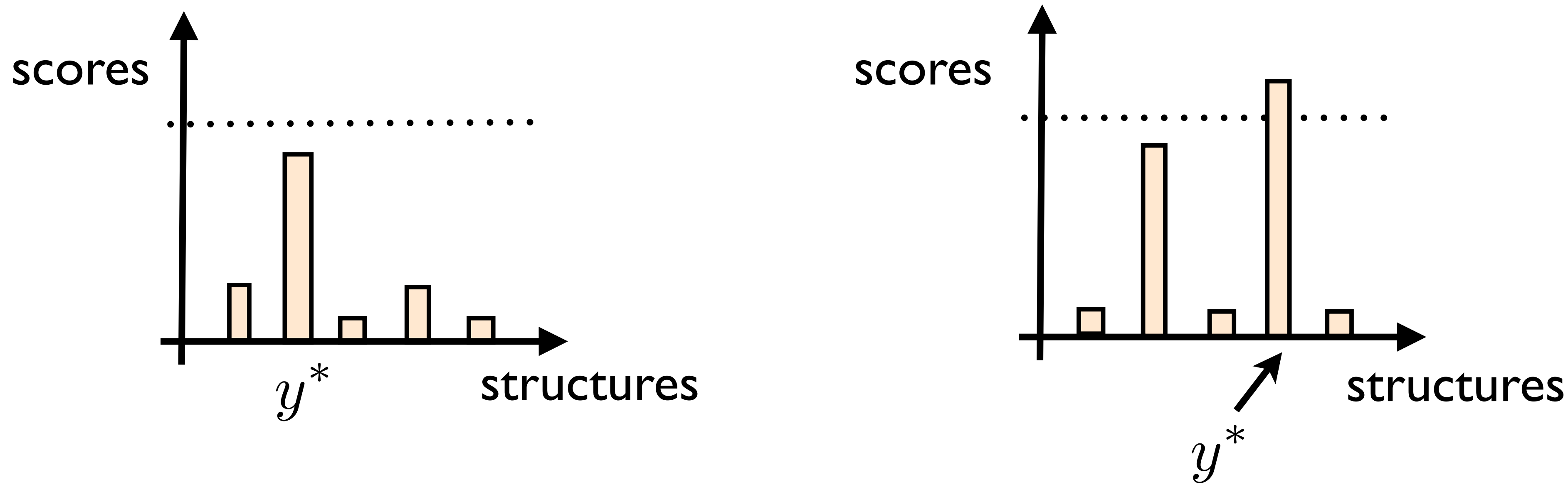
- Randomly perturbing the system reveals its complexity
  - little effect when the maximizing structure is "evident"
  - substantial effect when there are alternative high scoring structures

Random perturbations



- Randomly perturbing the system reveals its complexity
  - little effect when the maximizing structure is "evident"
  - substantial effect when there are alternative high scoring structures

Random perturbations

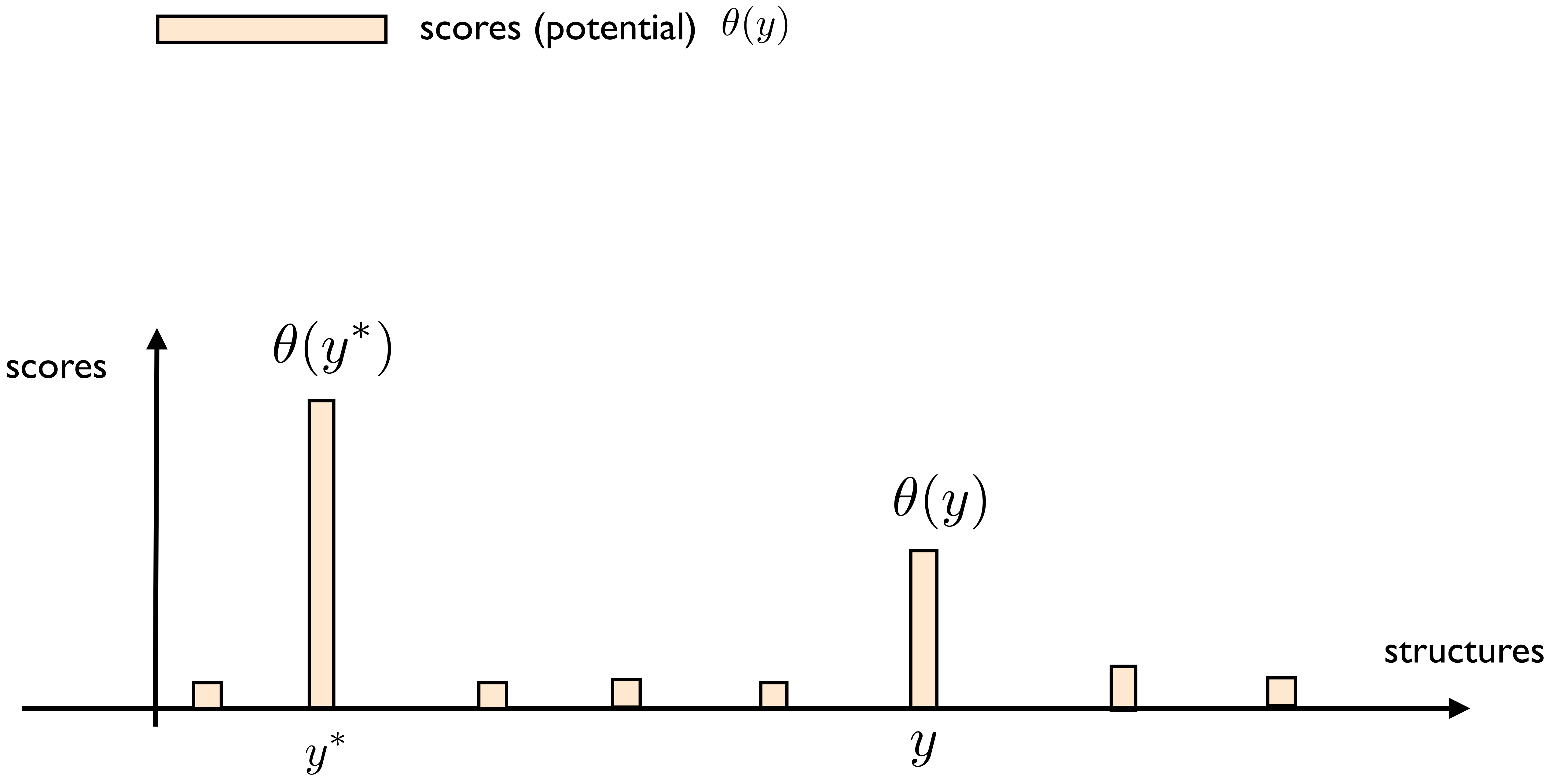


- Randomly perturbing the system reveals its complexity
  - little effect when the maximizing structure is "evident"
  - substantial effect when there are alternative high scoring structures

- Related work:
  - McFadden 74 (Discrete choice theory)
  - Talagrand 94 (Canonical processes)

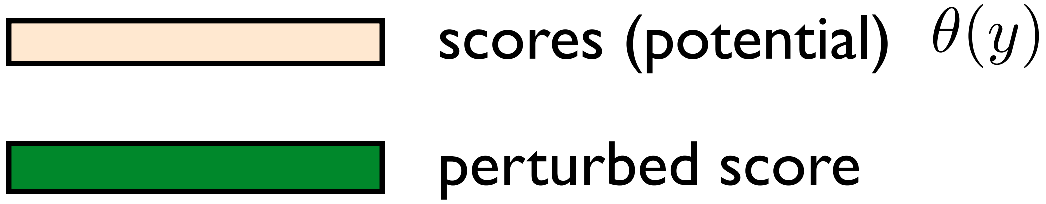
Random perturbations

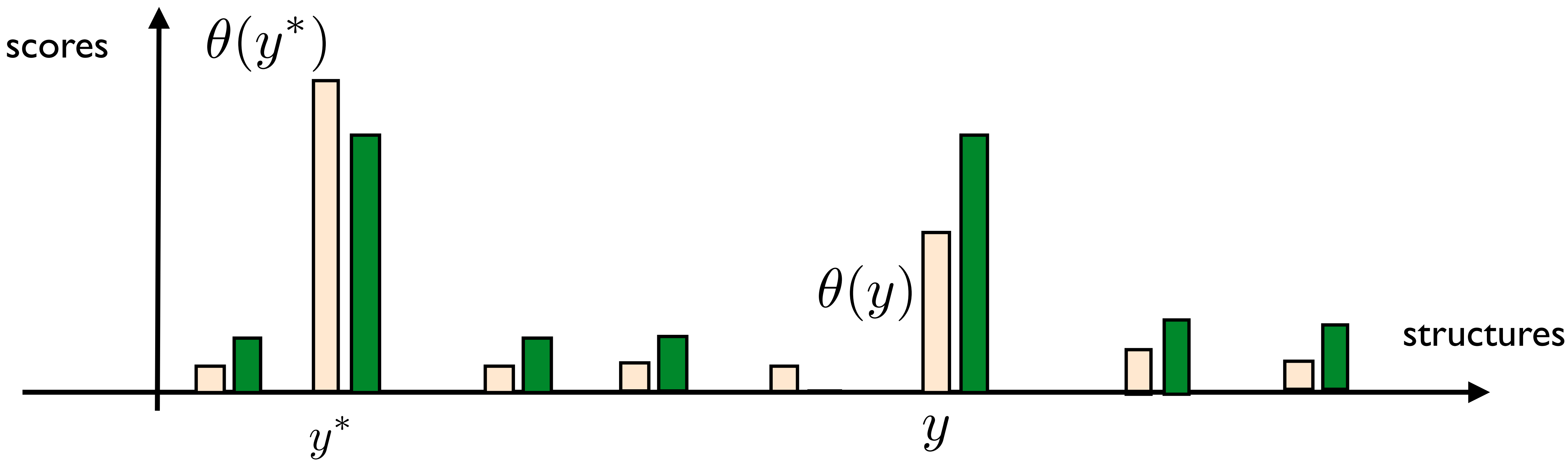
- Notation:



Random perturbations

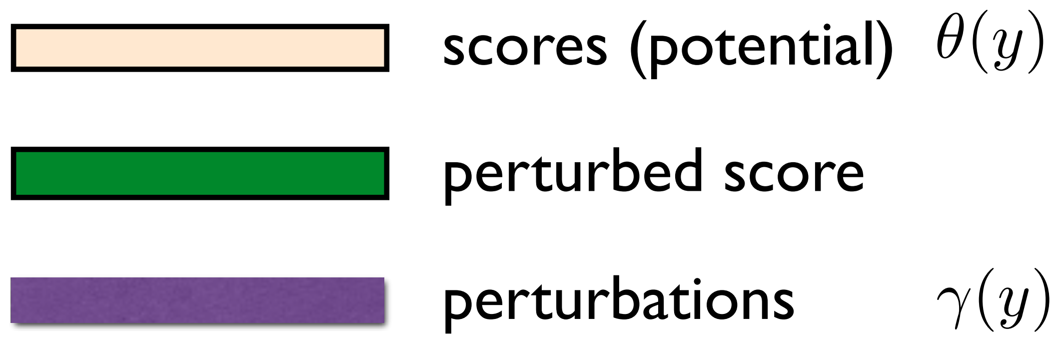
- Notation:

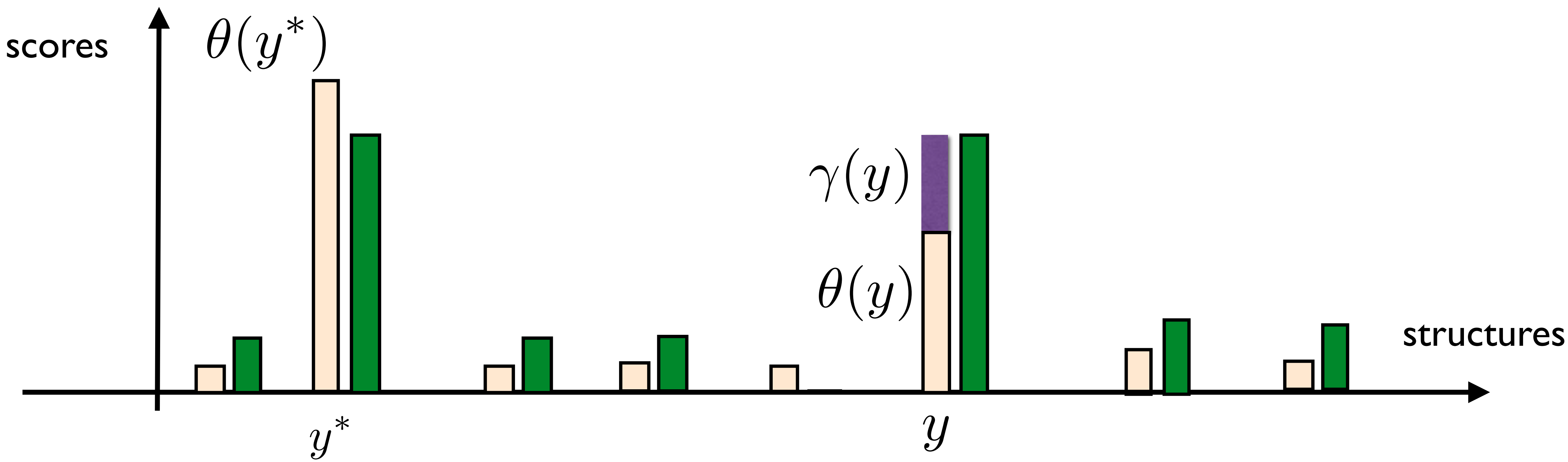




Random perturbations

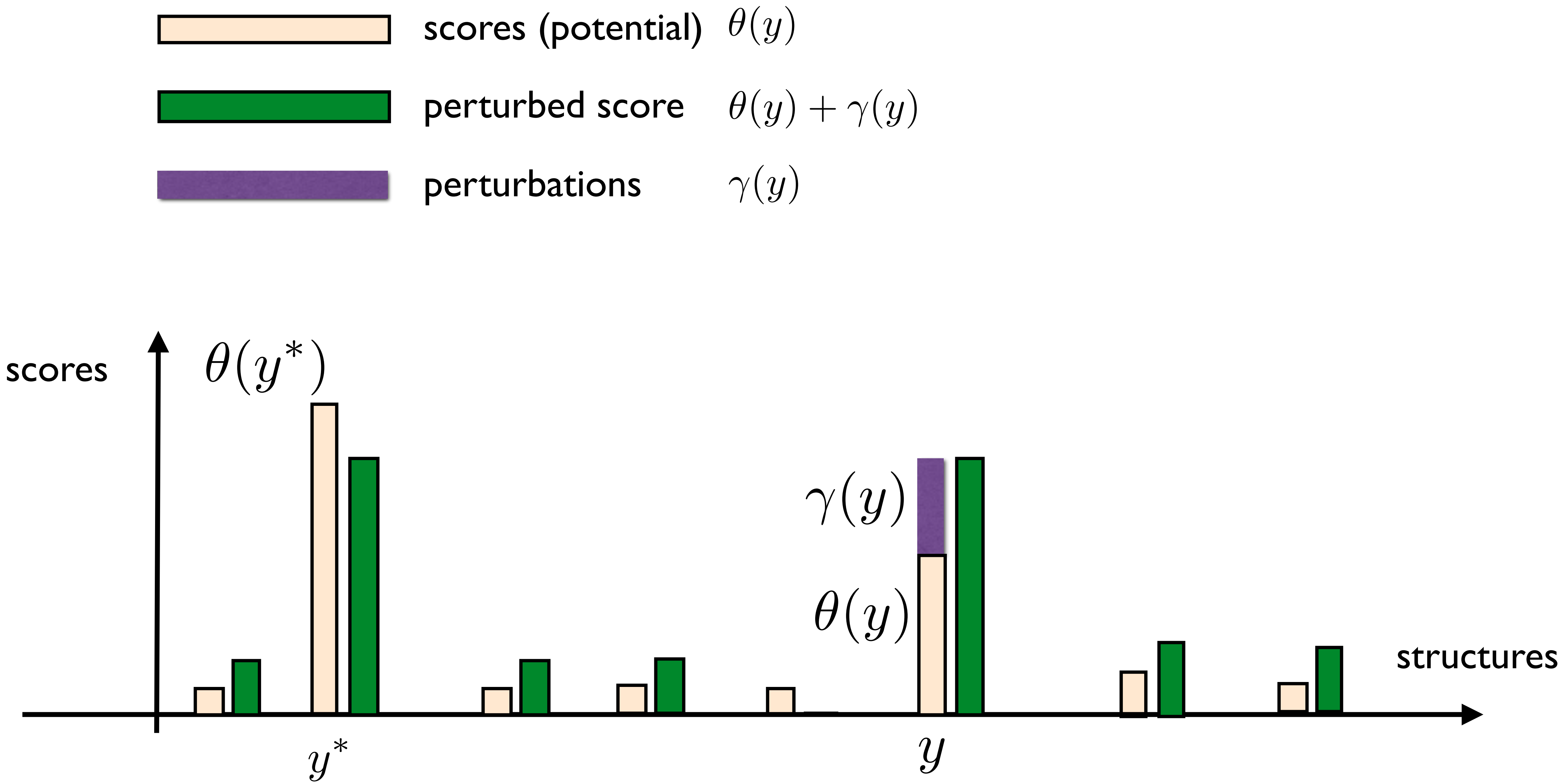
- Notation:





Random perturbations

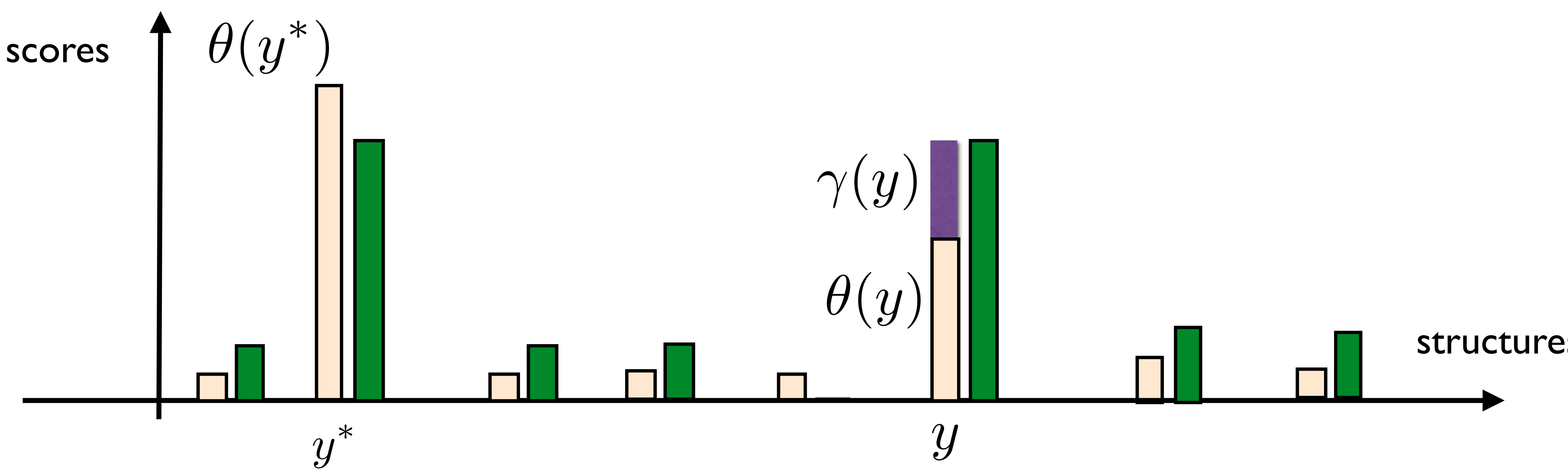
- Notation:



Random perturbations

- For every structure y, the perturbation value $\gamma(y)$ is a random variable (y is an index, traditional notation is γ_y).

- Perturb-max models: how stable is the maximal structure to random changes in the potential function.



Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs with perturb-max
- Application: interactive annotation.
  - New entropy bounds for perturb-max models.

Perturb-max models

- **Theorem**

Let $\gamma(y)$ be i.i.d. with Gumbel distribution with zero mean

$$F(t) \stackrel{def}{=} P[\gamma(y) \leq t] = \exp(-\exp(-t))$$

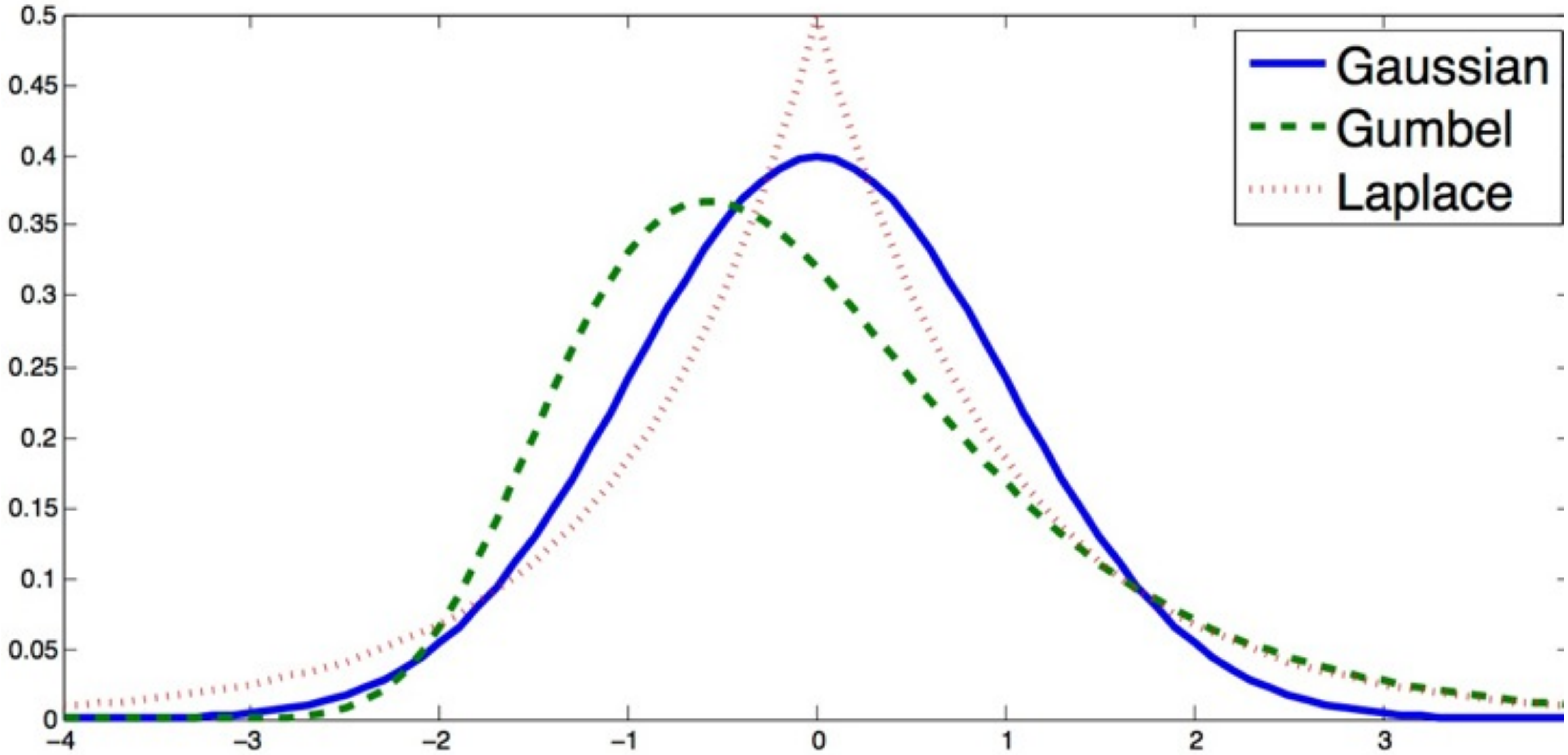
Perturb-max models

- **Theorem**

Let $\gamma(y)$ be i.i.d. with Gumbel distribution with zero mean

$$F(t) \stackrel{def}{=} P[\gamma(y) \leq t] = \exp(-\exp(-t))$$

$$f(t) = F'(t) = \exp(-t)F(t)$$



Perturb-max models

- **Theorem**

Let $\gamma(y)$ be i.i.d. with Gumbel distribution with zero mean

$$F(t) \stackrel{def}{=} P[\gamma(y) \leq t] = \exp(-\exp(-t))$$

then the perturb-max model is the Gibbs distribution

$$\frac{1}{Z}\exp(\theta(y)) = P_{\gamma \sim Gumbel}[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \gamma(\hat{y})\}]$$

Perturb-max models

- Why Gumbel distribution? $F(t) = \exp(-\exp(-t))$
- Since maximum of Gumbel variables is a Gumbel variable.

Let $\gamma(y)$ be i.i.d Gumbel ($P[\gamma(y) \leq t] = F(t)$). Then

$$\max_y \{\theta(y) + \gamma(y)\}$$

mean is $\log Z$

$$Z = \sum_y \exp(\theta(y))$$

Perturb-max models

- Why Gumbel distribution? $F(t) = \exp(-\exp(-t))$
- Since maximum of Gumbel variables is a Gumbel variable.

Let $\gamma(y)$ be i.i.d Gumbel ($P[\gamma(y) \leq t] = F(t)$). Then

$$\max_y \{\theta(y) + \gamma(y)\}$$

has Gumbel distribution whose mean is $\log Z$

Perturb-max models

- Why Gumbel distribution? $F(t) = \exp(-\exp(-t))$
- Since maximum of Gumbel variables is a Gumbel variable.

Let $\gamma(y)$ be i.i.d Gumbel ($P[\gamma(y) \leq t] = F(t)$). Then

$$\max_y \{\theta(y) + \gamma(y)\}$$

has Gumbel distribution whose mean is $\log Z$

- **Proof:** $P_\gamma[\max_y \{\theta(y) + \gamma(y)\} \leq t] = \prod_y F(t - \theta(y))$

$$= \exp(-\sum_y \exp(-(t - \theta(y)))) = F(t - \log Z)$$

Perturb-max models

- Max stability:

$$\log\Big(\sum_y \exp(\theta(y))\Big) = E_{\gamma\sim Gumbel}\Big[\max_y\{\theta(y)+\gamma(y)\}\Big]$$

- Implications (taking gradients):

$$\frac{1}{Z}\exp(\theta(y)) = P_{\gamma\sim Gumbel}[y = \arg\max_{\hat{y}}\{\theta(\hat{y})+\gamma(\hat{y})\}]$$

Perturb-max models

- Representing the Gibbs distribution using perturb-max models may require exponential number of perturbations

Perturb-max models

- Representing the Gibbs distribution using perturb-max models may require exponential number of perturbations

$$P_\gamma[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \gamma(\hat{y})\}]$$

Perturb-max models

- Representing the Gibbs distribution using perturb-max models may require exponential number of perturbations

$$P_\gamma[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \gamma(\hat{y})\}]$$

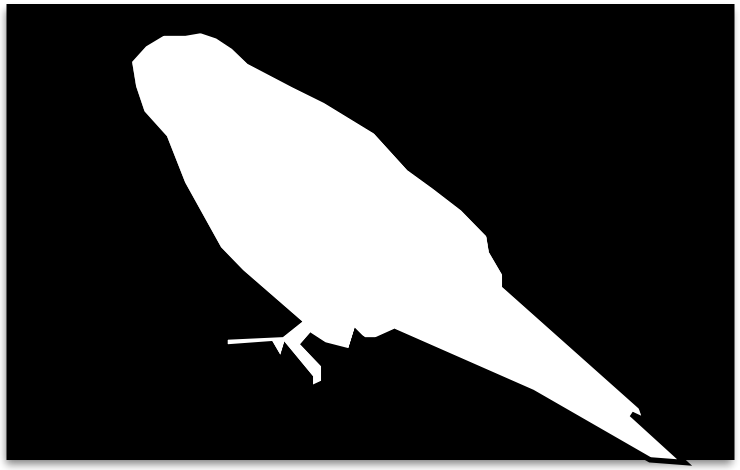
$$y = (y_1, ..., y_n)$$

Perturb-max models

- Representing the Gibbs distribution using perturb-max models may require exponential number of perturbations

$$P_\gamma[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \gamma(\hat{y})\}]$$

$$y = (y_1, ..., y_n)$$



Perturb-max models

- Representing the Gibbs distribution using perturb-max models may require exponential number of perturbations

$$P_\gamma[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \gamma(\hat{y})\}]$$

- Use low dimension perturbations [Papandreou & Yuille11, Tarlow et. al12]

$$P_\gamma[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}]$$

Outline

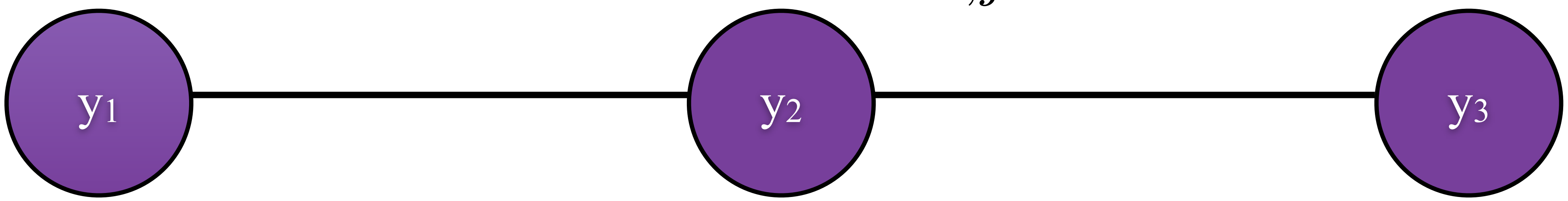
- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs with perturb-max
- Application: interactive annotation.
  - New entropy bounds for perturb-max models.

The marginal polytope

$$\theta(y_1, ..., y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$

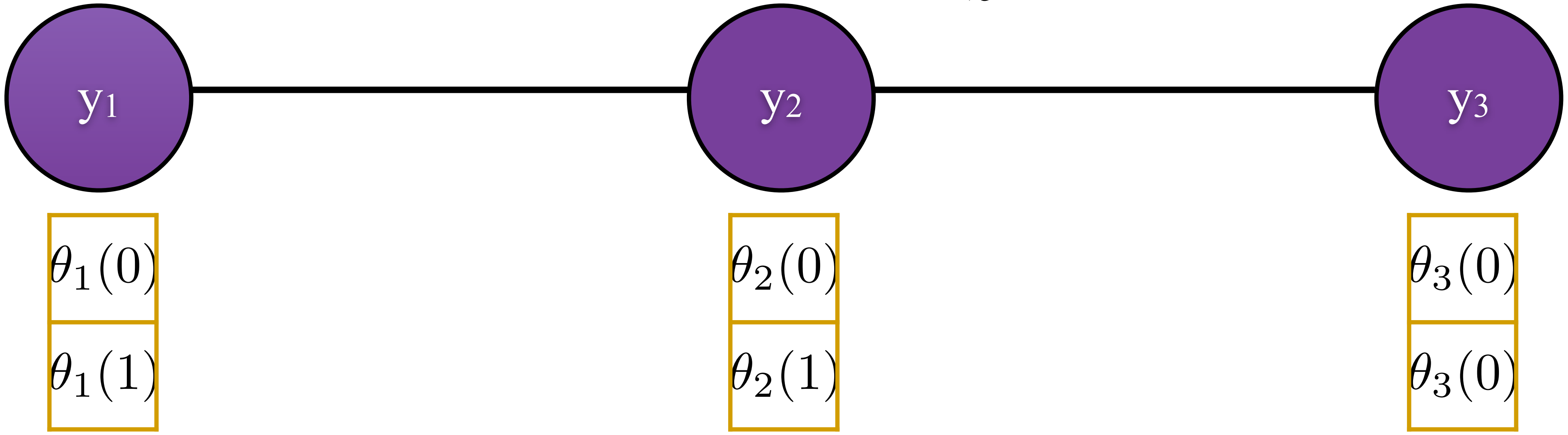
The marginal polytope

$$\theta(y_1, ..., y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$



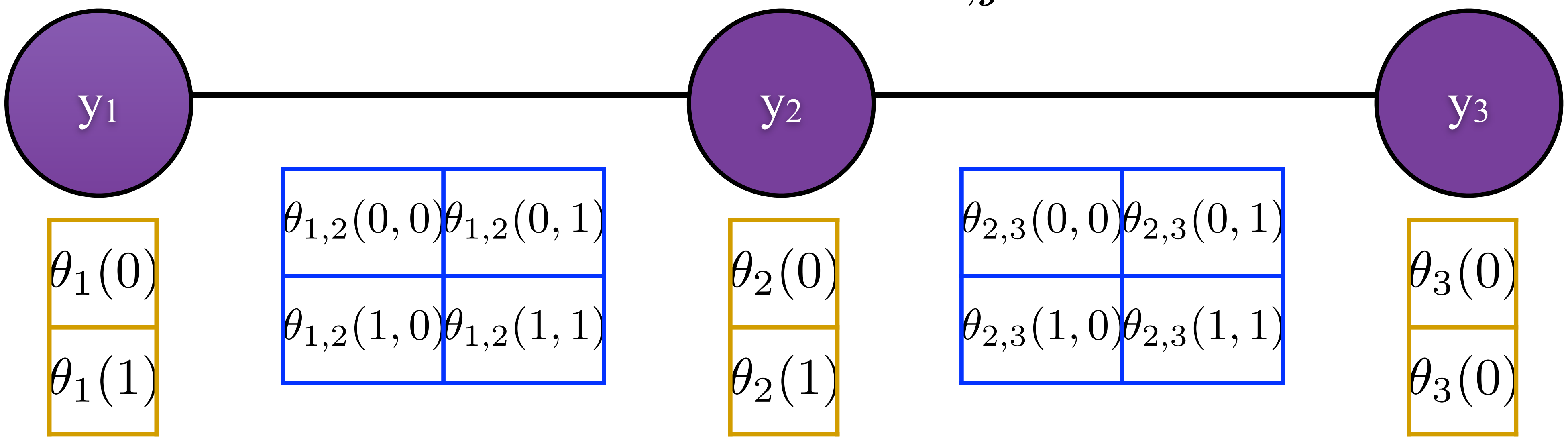
The marginal polytope

$$\theta(y_1, ..., y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$



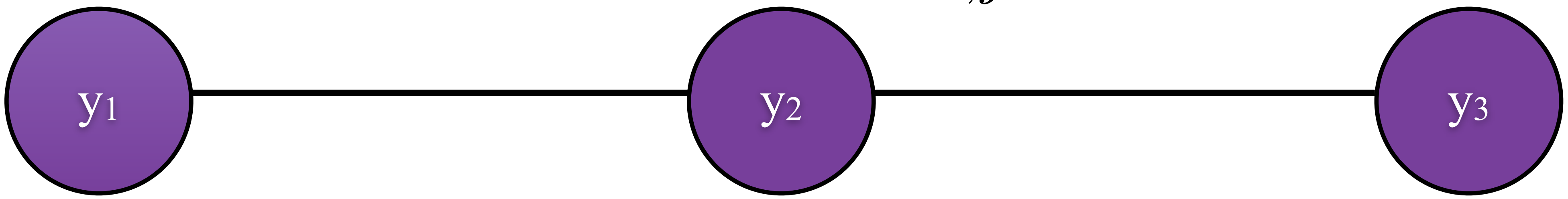
The marginal polytope

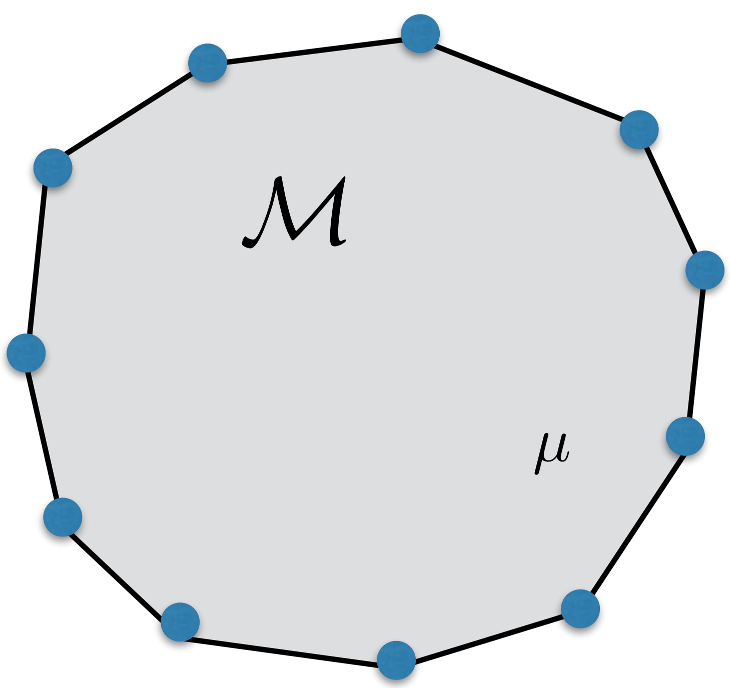
$$\theta(y_1, \ldots, y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$



The marginal polytope

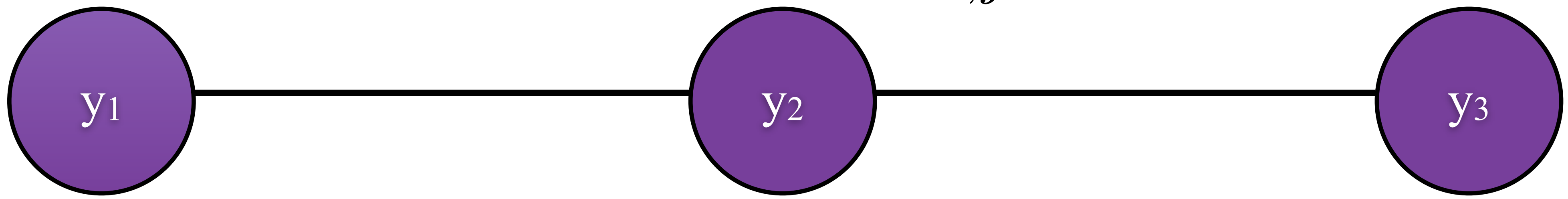
$$\theta(y_1, \ldots, y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$

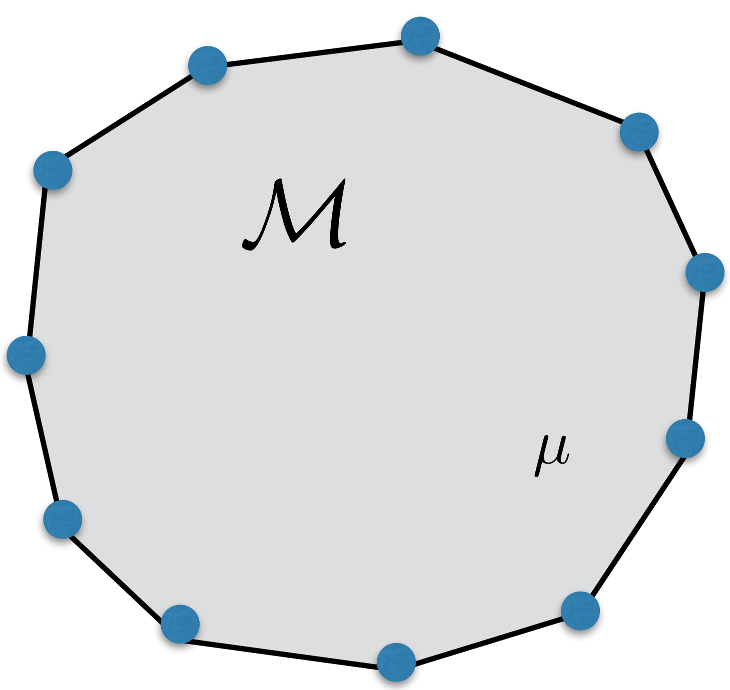




The marginal polytope

$$\theta(y_1, ..., y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$

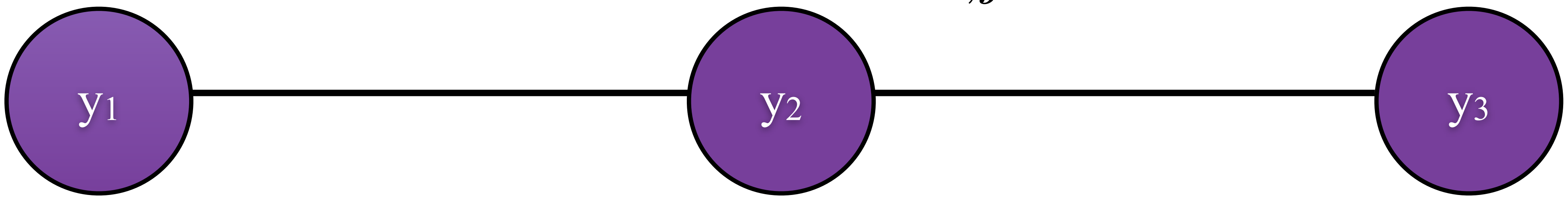


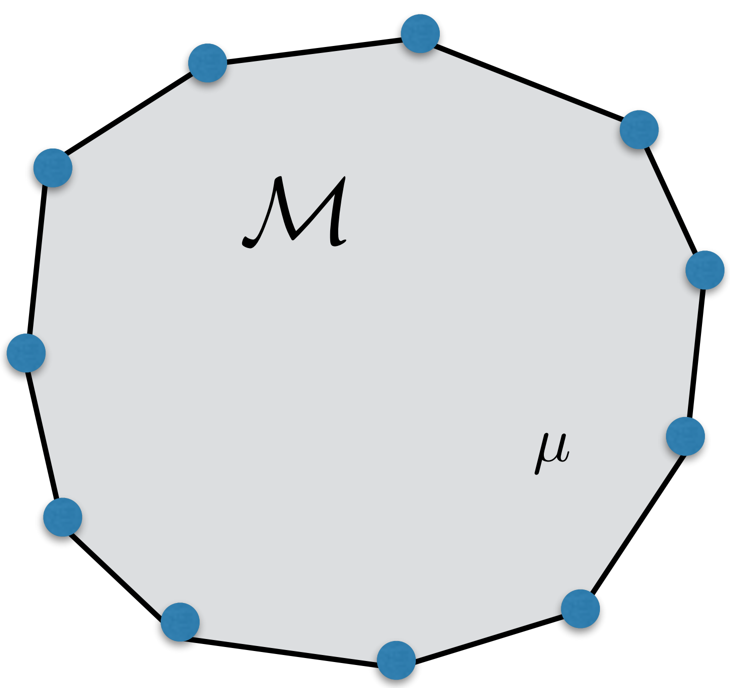


$$\mu = \begin{pmatrix} \mu_1(0), \mu_1(1), \mu_2(0), \mu_2(1), \mu_3(0), \mu_3(1), \\ \mu_{1,2}(0,0), \mu_{1,2}(0,1), \mu_{1,2}(1,0), \mu_{1,2}(1,1), \\ \mu_{2,3}(0,0), \mu_{2,3}(0,1), \mu_{2,3}(1,0), \mu_{2,3}(1,1)) \end{pmatrix}$$

The marginal polytope

$$\theta(y_1, \dots, y_n) = \sum_{i \in V} \theta_i(y_i) + \sum_{i,j \in E} \theta_{i,j}(y_i, y_j)$$



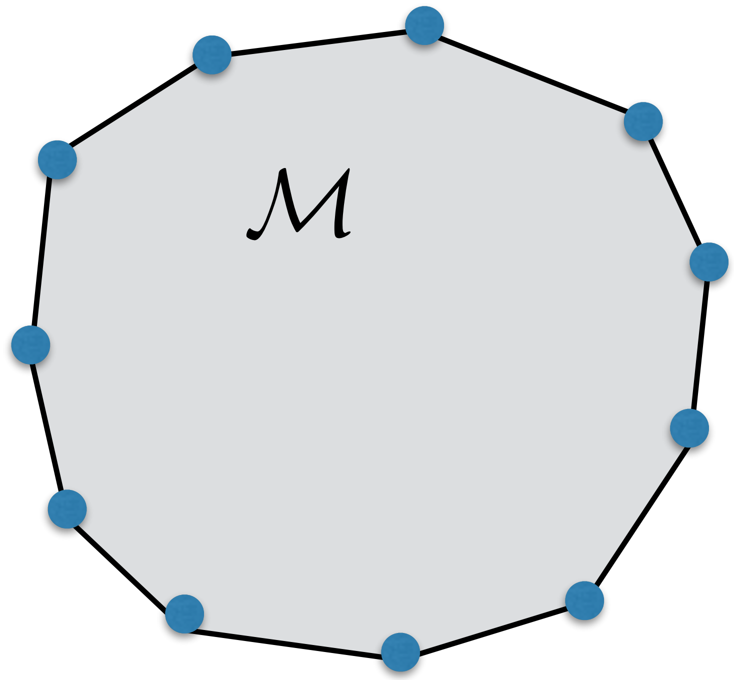


$$\mu = \begin{pmatrix} \mu_1(0), \mu_1(1), \mu_2(0), \mu_2(1), \mu_3(0), \mu_3(1), \\ \mu_{1,2}(0,0), \mu_{1,2}(0,1), \mu_{1,2}(1,0), \mu_{1,2}(1,1), \\ \mu_{2,3}(0,0), \mu_{2,3}(0,1), \mu_{2,3}(1,0), \mu_{2,3}(1,1)) \end{pmatrix}$$

$$\exists p(y_1, y_2, y_3) \text{ s.t. } \mu_1(y_1) = \sum_{y_2, y_3} p(y_1, y_2, y_3), \dots$$

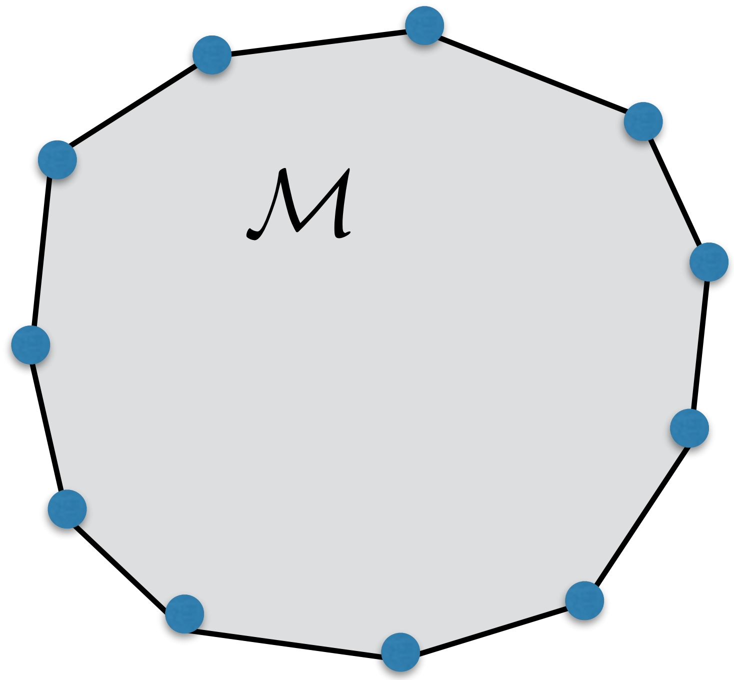
$$\mu_{1,2}(y_1, y_2) = \sum_{y_3} p(y_1, y_2, y_3), \dots$$

The marginal polytope



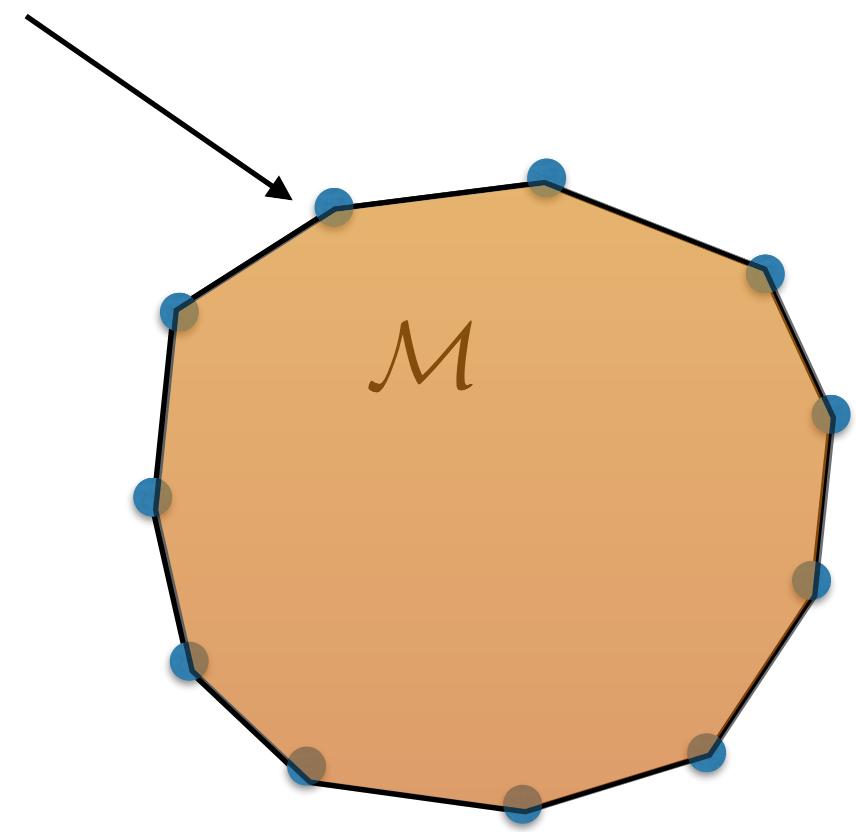
The marginal polytope

$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) \right)$$



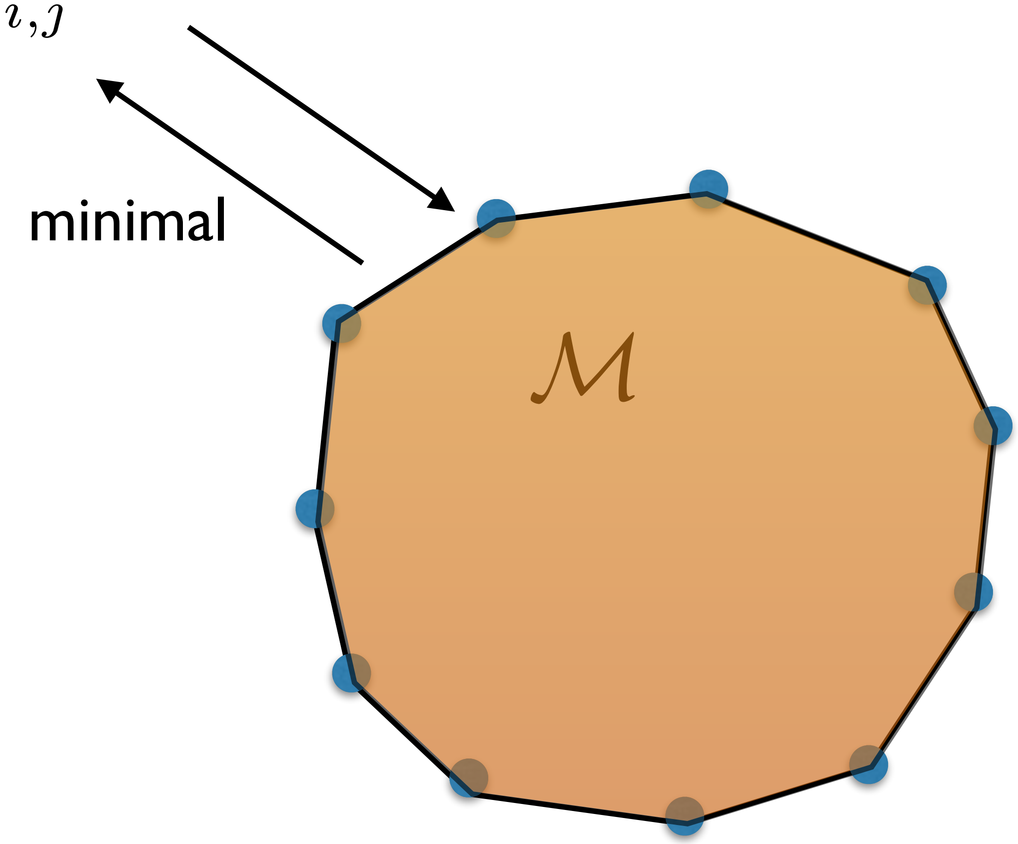
The marginal polytope

$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



The marginal polytope

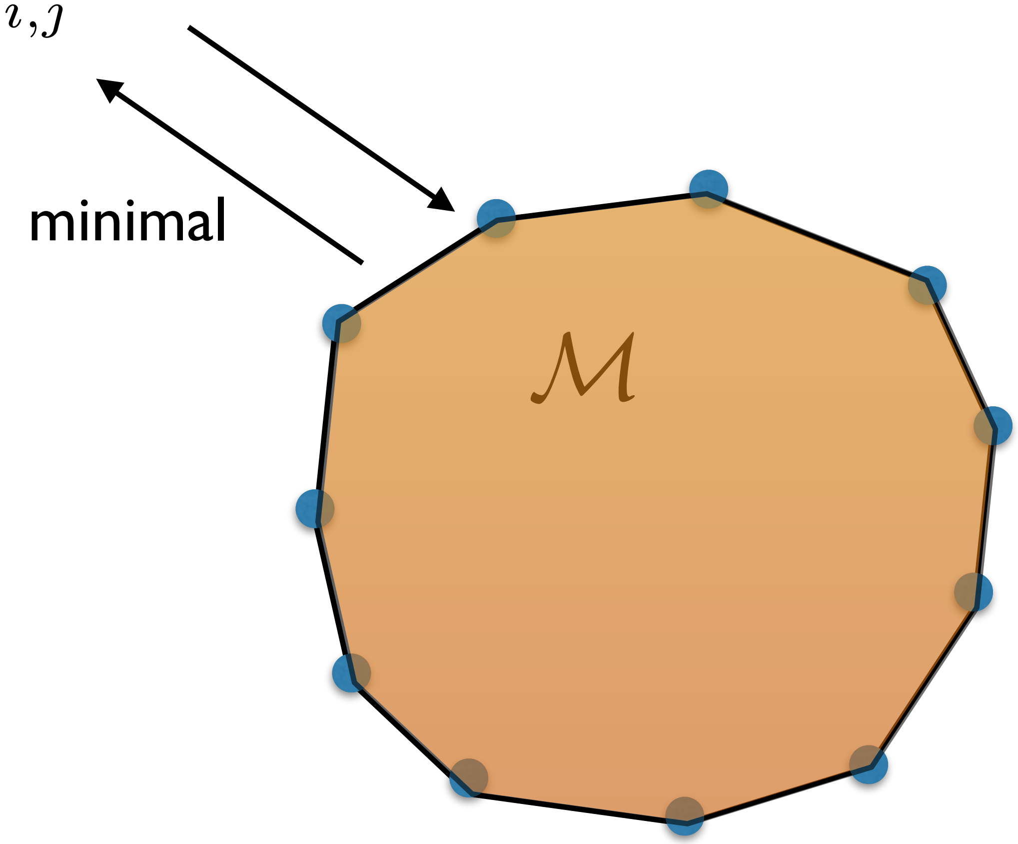
$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



The marginal polytope

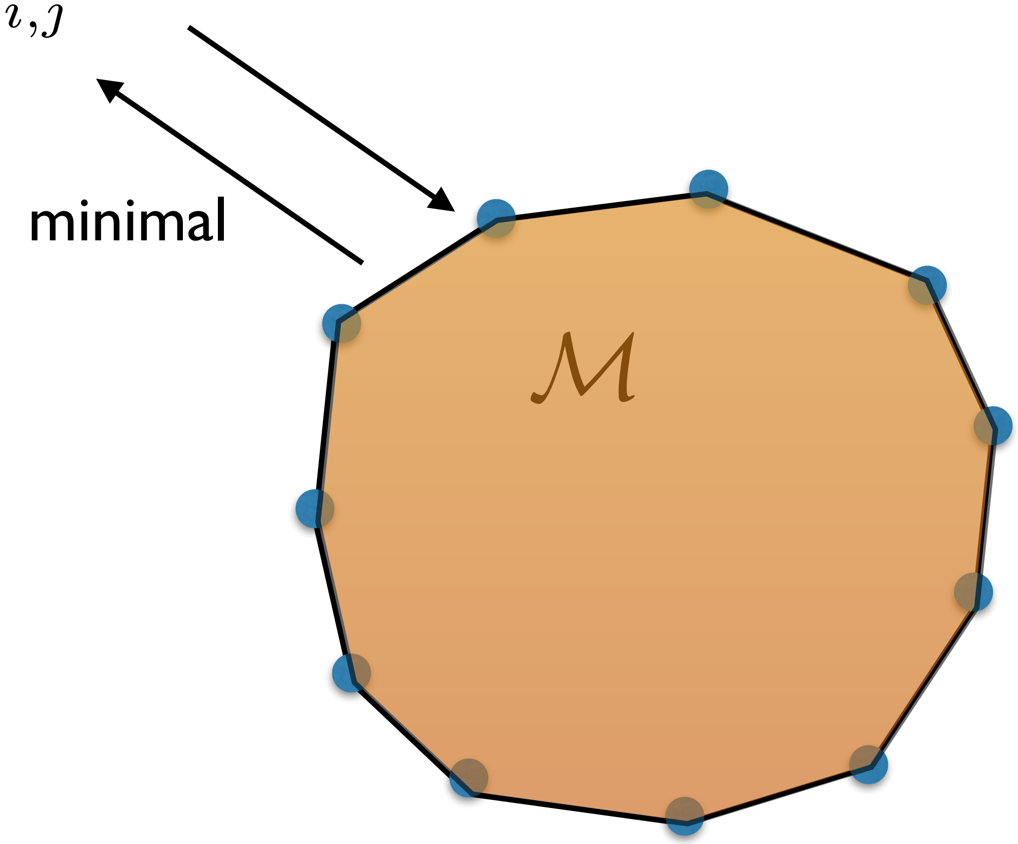
$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$

[Wainwright & Jordan 08]



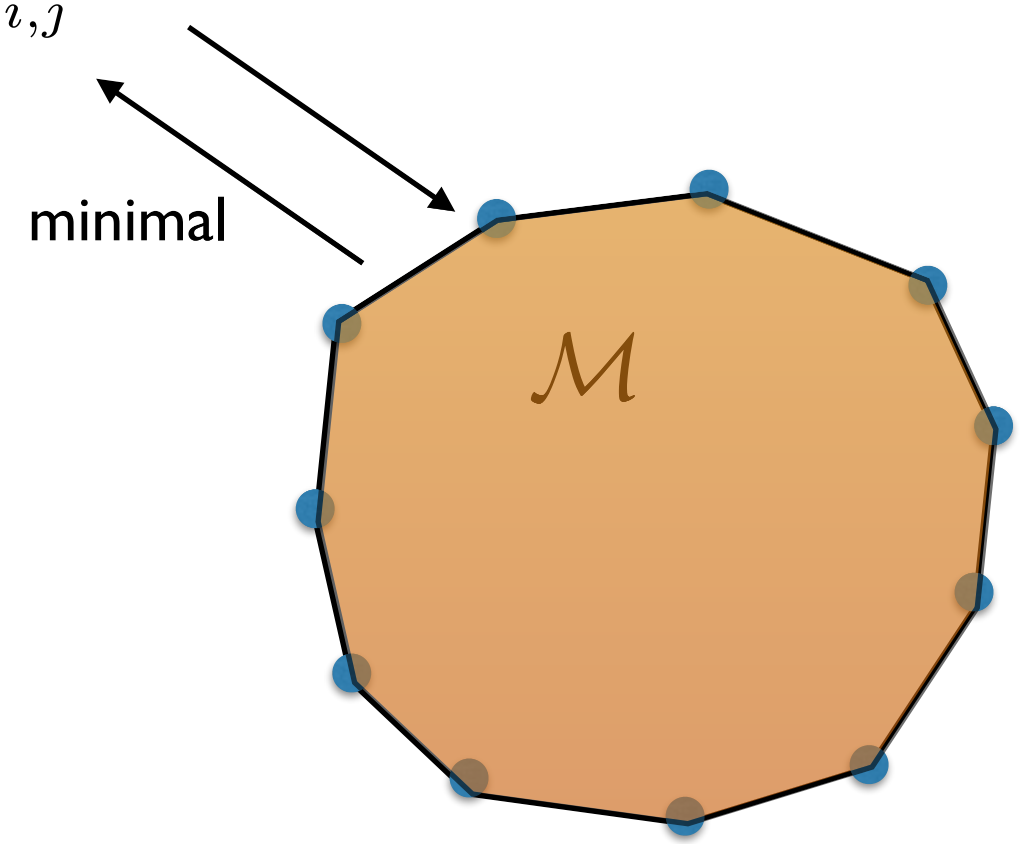
The marginal polytope

$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



The marginal polytope

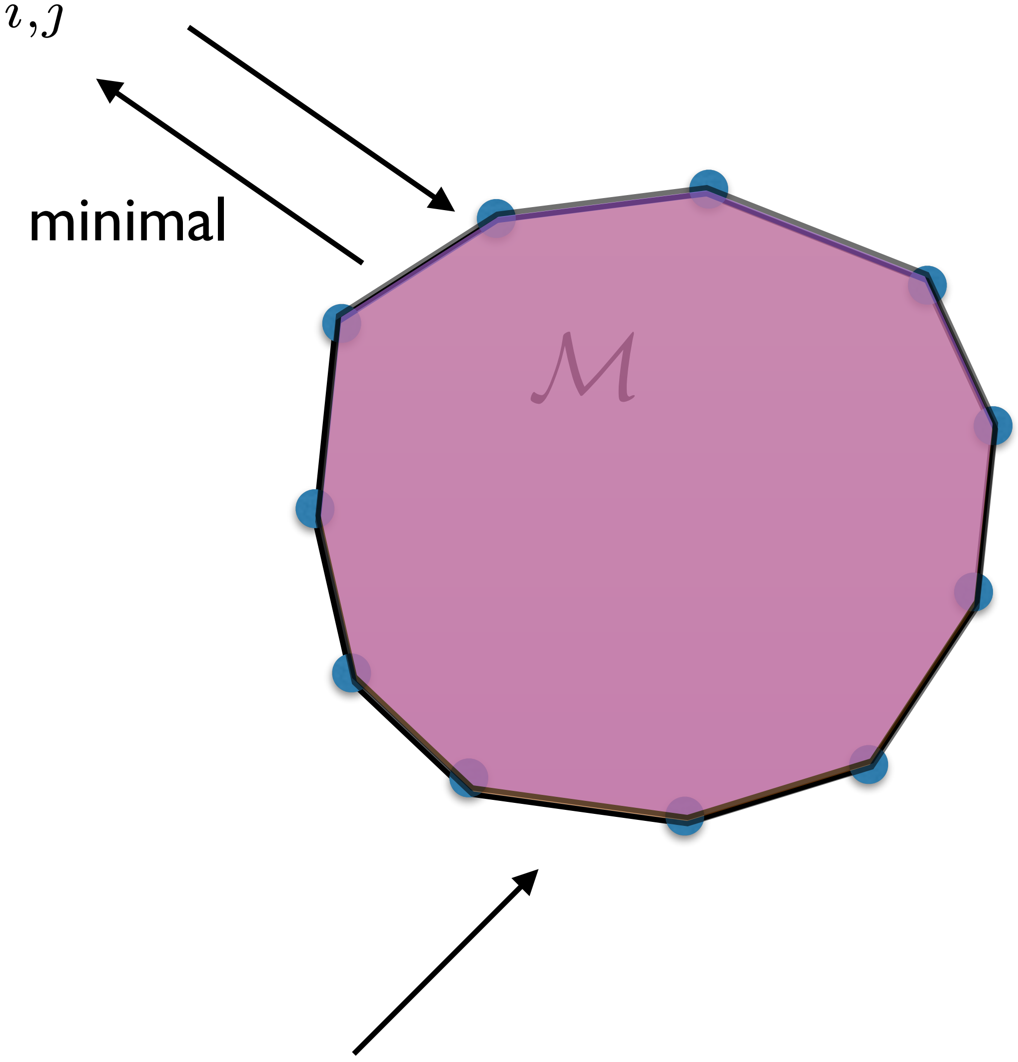
$$p(y) \propto \exp\left(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\right)$$



$$p(y) = P_\gamma\left[y = \arg\max_y \left\{\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i)\right\}\right]$$

The marginal polytope

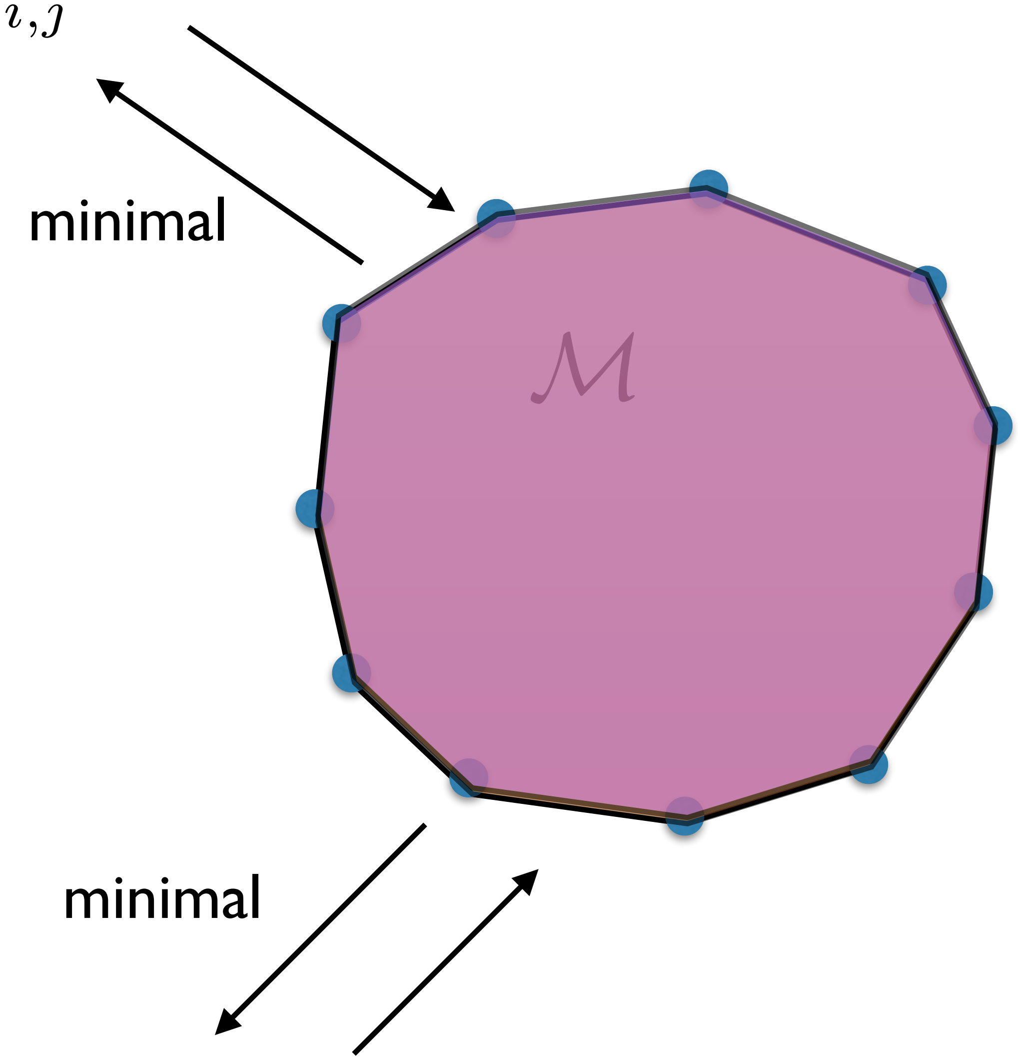
$$p(y) \propto \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\Big)$$



$$p(y) = P_\gamma\Big[y = \arg\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \Big\}\Big]$$

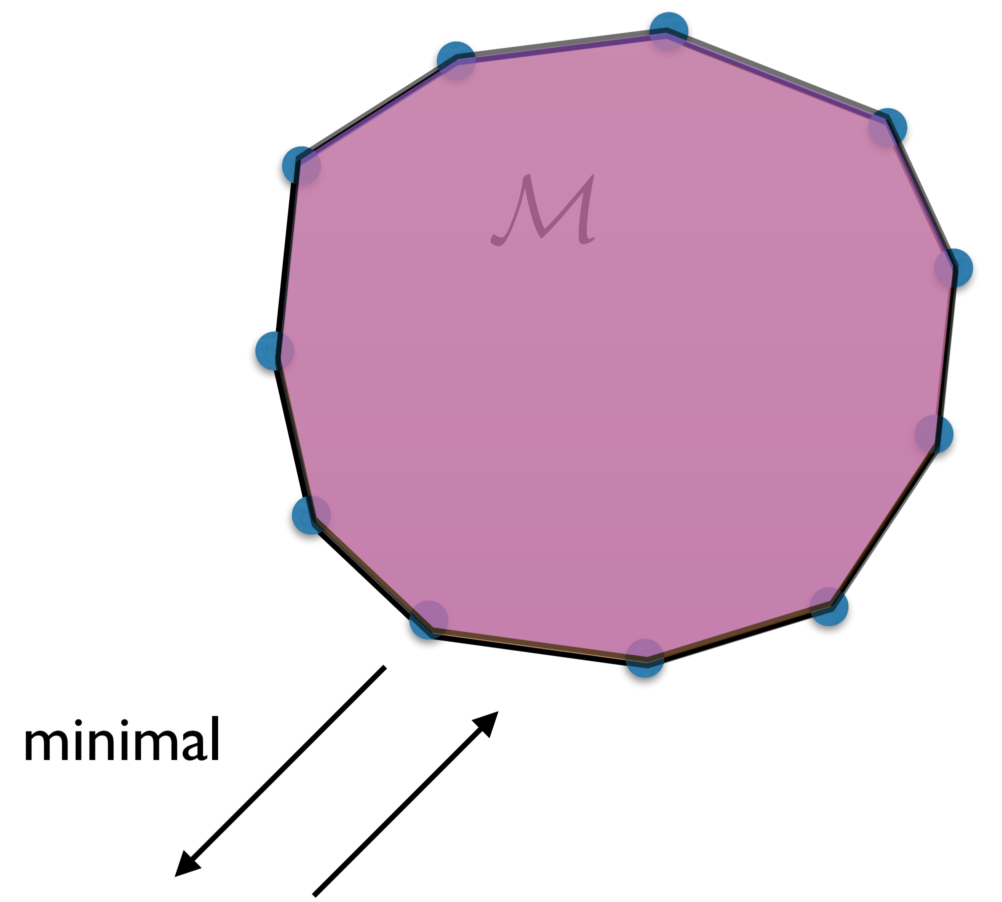
The marginal polytope

$$p(y) \propto \exp\Big(\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)\Big)$$



$$p(y) = P_\gamma\Big[y = \arg\max_y \Big\{\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i)\Big\}\Big]$$

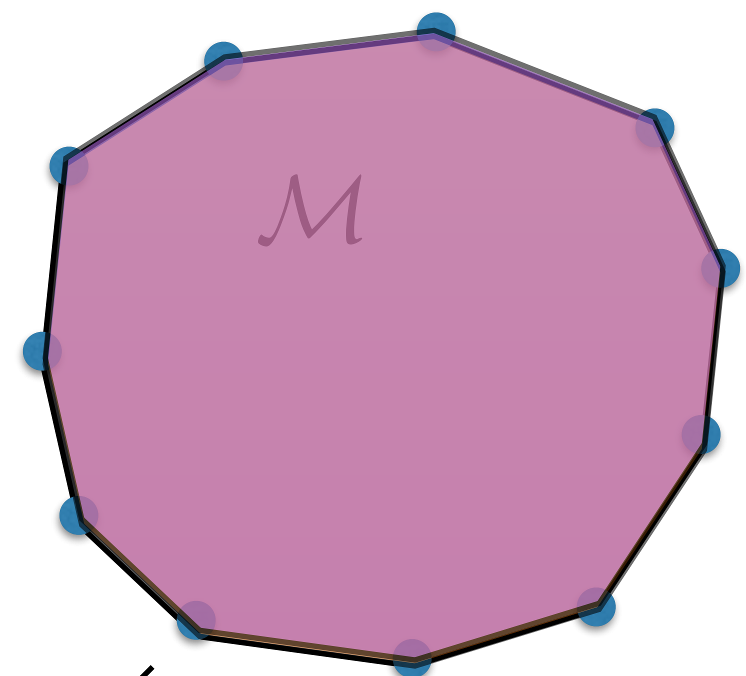
The marginal polytope



$$p(y) = P_\gamma\Big[y = \arg\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \Big\}\Big]$$

The marginal polytope

$$\mu = \begin{pmatrix} \mu_1(0), \mu_1(1), \mu_2(0), \mu_2(1), \mu_3(0), \mu_3(1), \\ \mu_{1,2}(0,0), \mu_{1,2}(0,1), \mu_{1,2}(1,0), \mu_{1,2}(1,1), \\ \mu_{2,3}(0,0), \mu_{2,3}(0,1), \mu_{2,3}(1,0), \mu_{2,3}(1,1)) \end{pmatrix}$$

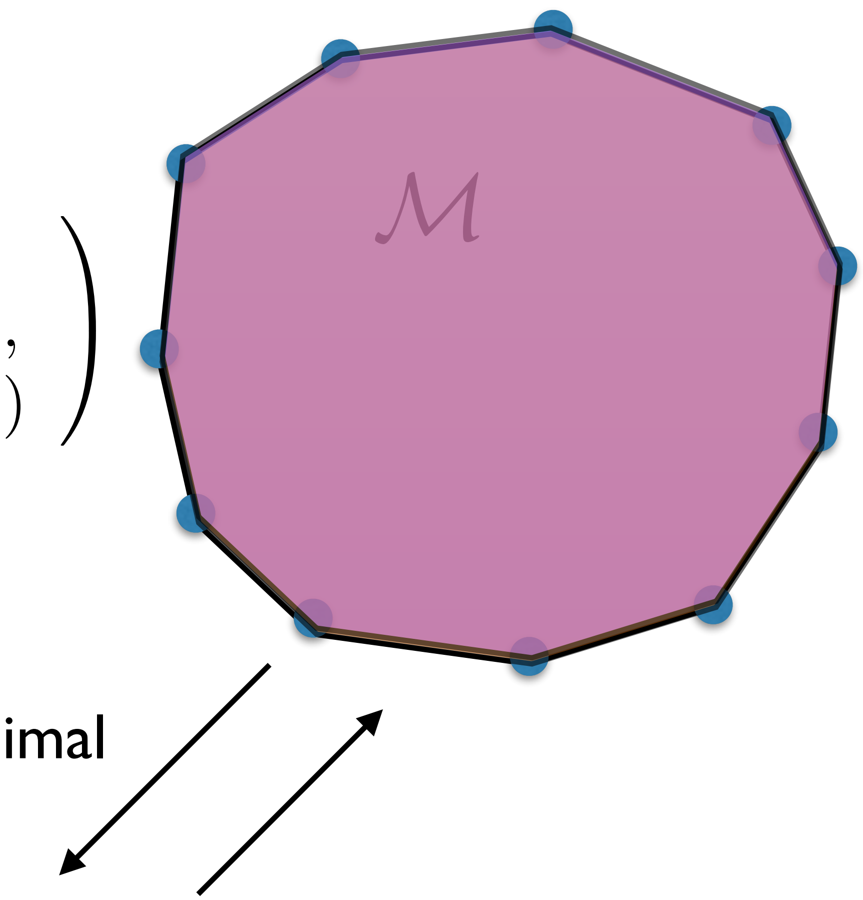


minimal

$$p(y) = P_\gamma\Big[y = \arg\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \Big\}\Big]$$

The marginal polytope

$$\mu = \begin{pmatrix} \mu_1(0), \mu_1(1), \mu_2(0), \mu_2(1), \mu_3(0), \mu_3(1), \\ \mu_{1,2}(0,0), \mu_{1,2}(0,1), \mu_{1,2}(1,0), \mu_{1,2}(1,1), \\ \mu_{2,3}(0,0), \mu_{2,3}(0,1), \mu_{2,3}(1,0), \mu_{2,3}(1,1)) \end{pmatrix}$$



minimal

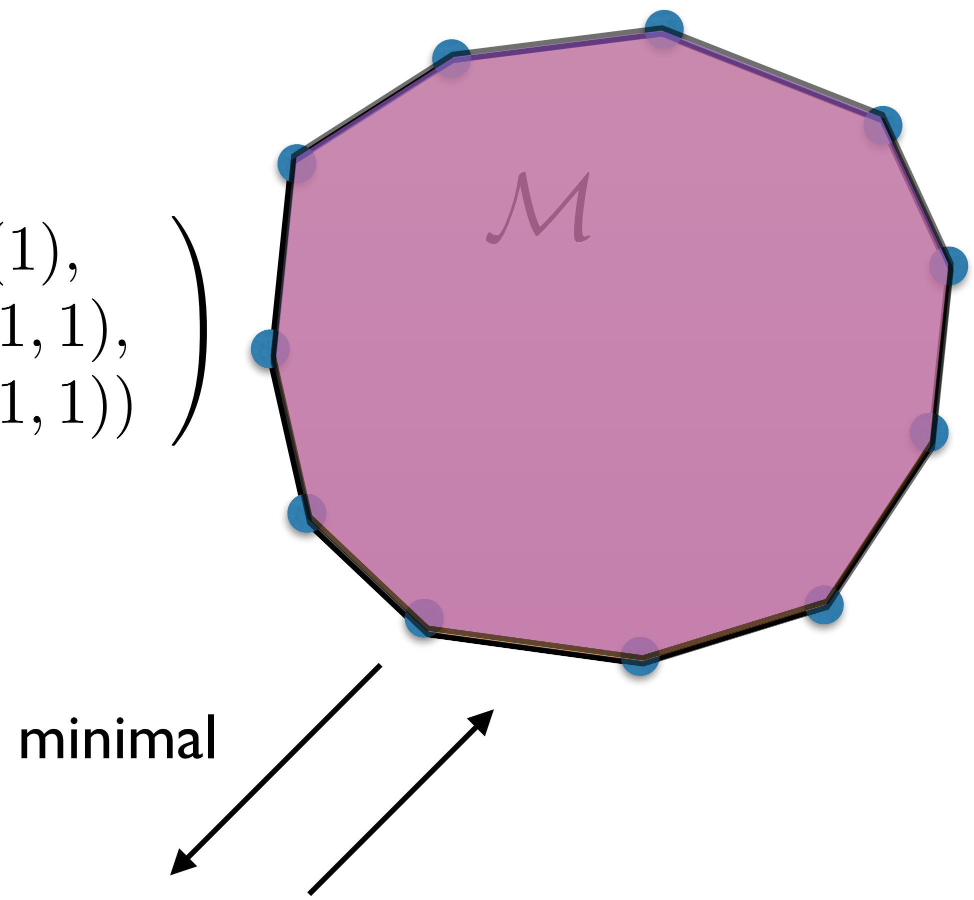
$$p(y) = P_\gamma \Big[y = \arg\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \Big\} \Big]$$

- Proof idea:

$$\mu_i(y_i) = \frac{\partial E_\gamma \Big[\max_y \big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \big\} \Big]}{\partial \theta_i(y_i)}$$

The marginal polytope

$$\mu = \begin{pmatrix} \mu_1(0), \mu_1(1), \mu_2(0), \mu_2(1), \mu_3(0), \mu_3(1), \\ \mu_{1,2}(0,0), \mu_{1,2}(0,1), \mu_{1,2}(1,0), \mu_{1,2}(1,1), \\ \mu_{2,3}(0,0), \mu_{2,3}(0,1), \mu_{2,3}(1,0), \mu_{2,3}(1,1)) \end{pmatrix}$$



$$p(y) = P_\gamma \Big[y = \arg\max_y \Big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \Big\} \Big]$$

- Proof idea:

$$\mu_i(y_i) = \frac{\partial E_\gamma \Big[\max_y \big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \big\} \Big]}{\partial \theta_i(y_i)}$$

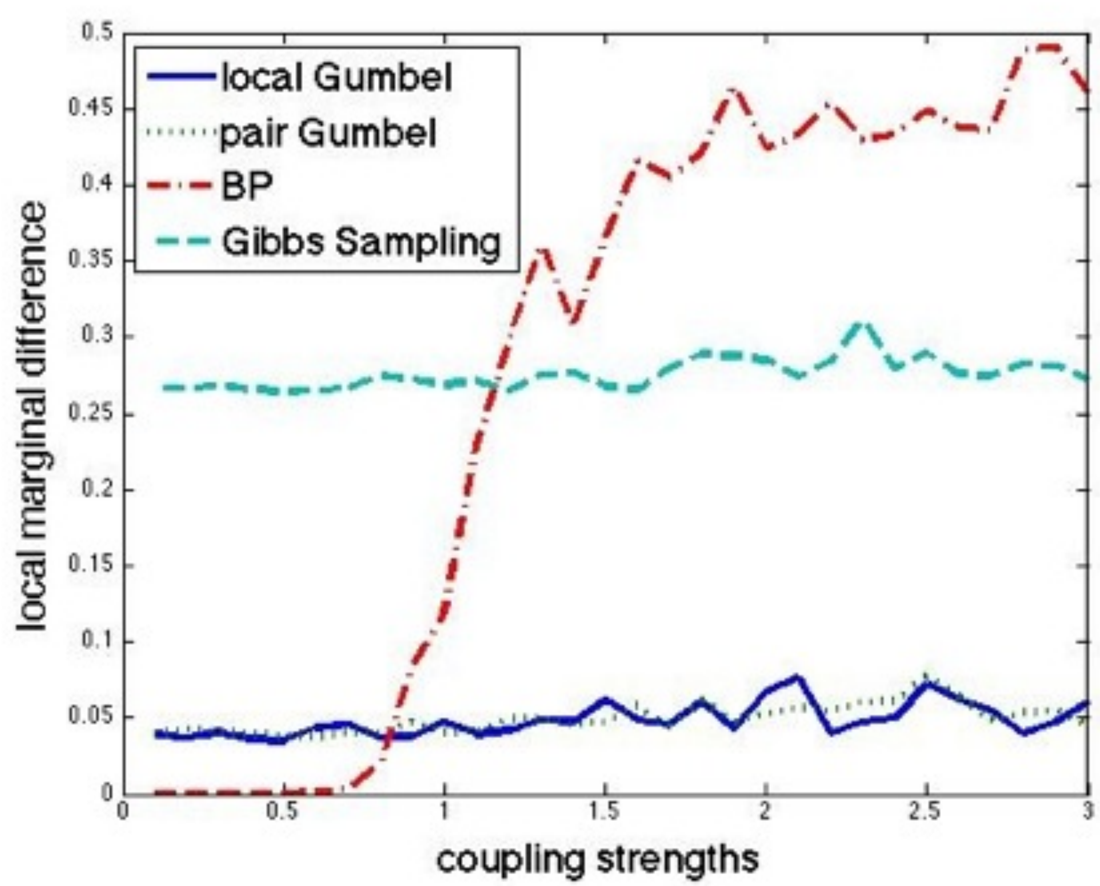
$$\mu_{i,j}(y_i, y_j) = \frac{\partial E_\gamma \Big[\max_y \big\{ \sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j) + \sum_i \gamma_i(y_i) \big\} \Big]}{\partial \theta_{i,j}(y_i, y_j)}$$

Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs with perturb-max
- Application: interactive annotation.
  - New entropy bounds for perturb-max models.

Non-MCMC sampling

- Perturb-max sample from tree-shaped Gibbs distribution [Gane, H, Jaakkola 14].
- Perturb-max + rejections sample from the Gibbs distribution on general graphs [H, Maji, Jaakkola 13].
- In practice, perturb-max marginals approximate the Gibbs marginals for general graphs [Papandreou & Yuille 11].



Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs distributions with perturb-max
- Application: interactive annotation.
  - New entropy bounds for perturb-max models.

Image annotation

- Image annotation is a time consuming (and tedious) task. Can computers do it for us?

Image annotation

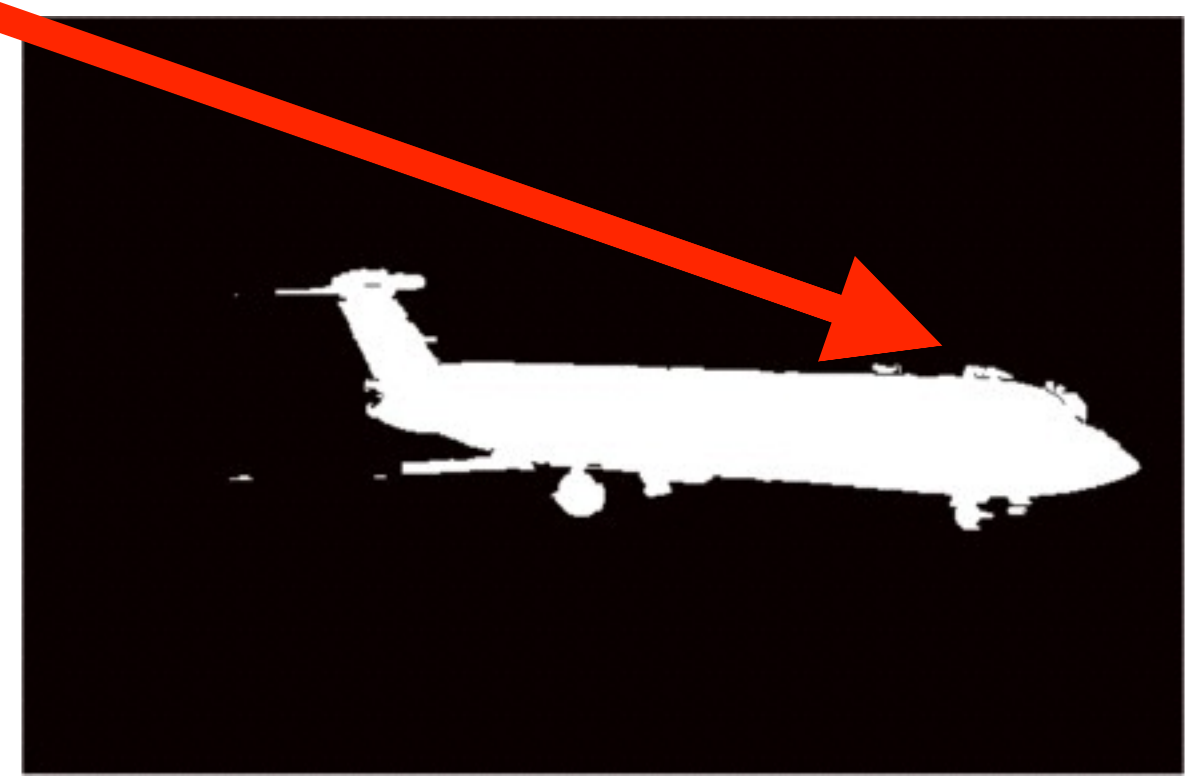
- Why not to use the most likely annotation instead?

Image annotation

- Why not to use the most likely annotation instead?
- Most likely annotation is inaccurate around
  - “thin” areas

Image annotation

- Why not to use the most likely annotation instead?
- Most likely annotation is inaccurate around
  - "thin" areas
  - clutter



Interactive image annotation

- Perturb-max models show the boundary of decision.

Interactive image annotation

- Perturb-max models show the boundary of decision.

Interactive image annotation

- Perturb-max models show the boundary of decision.

- Interactive annotation directs the human annotator to areas of uncertainty - significantly reduces annotation time [Maji, H., Jaakkola 14].

Uncertainty

- Entropy $$H(p_\theta) = -\sum_y p_\theta(y) \log p_\theta(y)$$

- Entropy = uncertainty
  - It is a nonnegative function over probability distributions.
  - It attains its maximal value for the uniform distribution.
  - It attains its minimal value for the zero-one distribution.

- Computing the entropy requires summing over exponential many configurations $y = (y_1, ..., y_n)$

- Can we bound it with perturb-max approach?

Uncertainty

- Perturb-max models

$$p_\theta(y) \stackrel{def}{=} P_\gamma[y = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}]$$

- Entropy

$$H(p_\theta) = -\sum_{y} p_\theta(y) \log p_\theta(y)$$

- Entropy bound

$$H(p_\theta) \leq E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

Uncertainty

$$U(p_\theta) = E_\gamma \Big[\sum_{i=1}^{n} \gamma_i(y_i^*) \Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

Uncertainty

$$U(p_\theta) = E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big] \qquad y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- $U(p_\theta)$ is an uncertainty measure

Uncertainty

$$U(p_\theta) = E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big] \qquad y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- $U(p_\theta)$ is an uncertainty measure
  - $U(p_\theta)$ is nonnegative since $0 \leq H(p_\theta) \leq U(p_\theta)$

Uncertainty

$$U(p_\theta) = E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big] \qquad y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- $U(p_\theta)$ is an uncertainty measure
  - $U(p_\theta)$ is nonnegative since $0 \leq H(p_\theta) \leq U(p_\theta)$
  - $U(\text{zero-one distribution}) = 0$

Uncertainty

$$U(p_\theta) = E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big] \qquad y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- $U(p_\theta)$ is an uncertainty measure
  - $U(p_\theta)$ is nonnegative since $0 \leq H(p_\theta) \leq U(p_\theta)$
  - $U(\text{zero-one distribution}) = 0$

$$\theta(\hat{y}) = 0, \;\; \forall y \neq \hat{y} \;\; \theta(y) = -\infty$$

$$E[\gamma_i(\hat{y}_i)] = 0$$

Uncertainty

$$U(p_\theta) = E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big] \qquad\qquad y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n}\gamma_i(\hat{y}_i)\}$$

- $U(p_\theta)$ is an uncertainty measure
  - $U(p_\theta)$ is nonnegative since $0 \leq H(p_\theta) \leq U(p_\theta)$
  - $U(\text{zero-one distribution}) = 0$
  - $U(\text{uniform distribution}) = \text{maximal}$

Uncertainty

$$U(p_\theta) = E_\gamma\Big[\sum_{i=1}^{n} \gamma_i(y_i^*)\Big] \qquad y^* = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- $U(p_\theta)$ is an uncertainty measure
  - $U(p_\theta)$ is nonnegative since $0 \leq H(p_\theta) \leq U(p_\theta)$
  - $U(\text{zero-one distribution}) = 0$
  - $U(\text{uniform distribution}) = \text{maximal}$

$$\theta(y) \equiv 0$$

higher $\theta(y)$ favor lower $\gamma(y)$ at the expanse of higher $\gamma(\hat{y})$

Uncertainty

- How does it compare to standard entropy bounds?
- Perturb-max entropy bound:

$$H(p_\theta) \leq E\Big[\sum_i \gamma_i(y_i^*)\Big] = \sum_i E\Big[\gamma_i(y_i^*)\Big]$$

- Standard entropy independence bound:

$$H(p_\theta) \leq \sum_i H(p_\theta(y_i))$$

$$p_\theta(y_i) = P_\gamma[y_i = \arg\max_{\hat{y}}\{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}]$$

- Perturb-max entropy bound requires less samples since sampled average tail decreases exponentially.

Perturb-max entropy bounds

- Spin glass, 5x5 grid

$$\sum_i \theta_i(y_i) + \sum_{i,j} \theta_{i,j}(y_i, y_j)$$

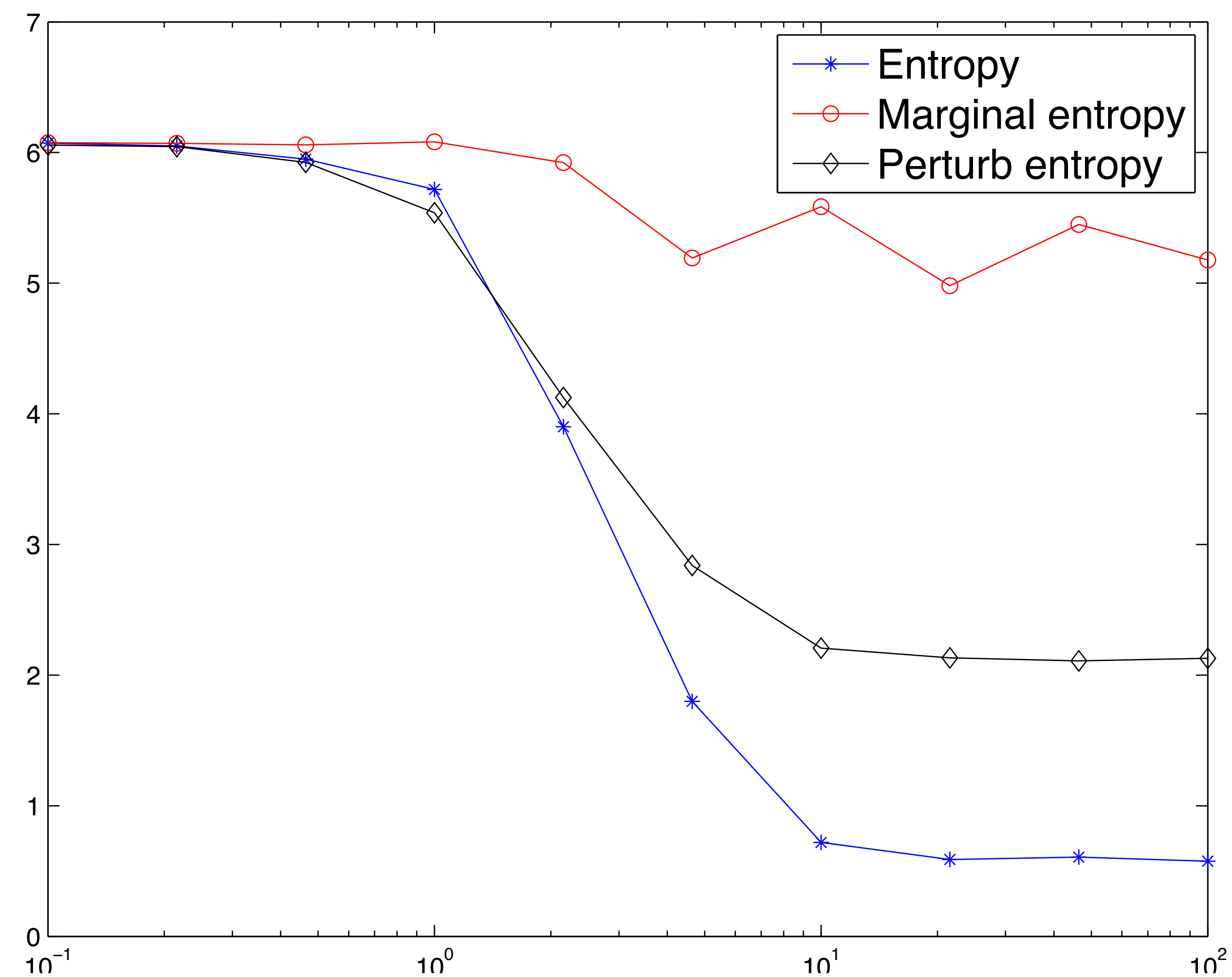
$$y_i \in \{-1, 1\}$$

$$\theta_i(y_i) = w_i y_i$$

$$w_i \sim N(0,1)$$

$$\theta_{i,j}(y_i, y_j) = w_{i,j} y_i y_j$$

- attractive $w_{i,j} \geq 0$. Graph-cuts.



Uncertainty*

- Theorem: $$H(p_\theta) \leq E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

Uncertainty*

- Theorem: $H(p_\theta) \le E\Big[\sum_i \gamma_i(y_i^*)\Big]$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y) p(y) \Big\}$$

Uncertainty*

- Theorem: $$H(p_\theta) \leq E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$\log Z(\hat{\theta}) \leq E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big]$$

The flashback slide

- Max stability:

$$\log\Big(\sum_y \exp(\theta(y))\Big) = E_{\gamma\sim Gumbel}\Big[\max_y\{\theta(y)+\gamma(y)\}\Big]$$

Uncertainty*

- Theorem: $$H(p_\theta) \le E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y)p(y) \Big\}$$

$$\log Z(\hat{\theta}) \le E_\gamma\Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\}\Big]$$

$$H(p) \le \min_{\hat{\theta}} \Big\{ E_\gamma\Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\}\Big] - \sum_y \hat{\theta}(y)p(y) \Big\}$$

Uncertainty*

- Theorem: $$H(p_\theta) \leq E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$\log Z(\hat{\theta}) \leq E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big]$$

$$H(\overset{p_\theta}{p}) \leq \min_{\hat{\theta}} \Big\{ E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big] - \sum_y \hat{\theta}(y) \overset{p_\theta}{p}(y) \Big\}$$

Uncertainty*

- Theorem: $$H(p_\theta) \le E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$\log Z(\hat{\theta}) \le E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big]$$

$$H(\overset{p_\theta}{p}) \le \min_{\hat{\theta}} \Big\{ E_\gamma \Big[\max_y \{\overset{\hat{\theta}^* = \theta}{\hat{\theta}(y)} + \sum_i \gamma_i(y_i)\} \Big] - \sum_y \overset{\hat{\theta}^* = \theta}{\hat{\theta}(y)} \overset{p_\theta}{p(y)} \Big\}$$

Uncertainty*

- Theorem: $$H(p_\theta) \le E\Big[\sum_i \gamma_i(y_i^*)\Big]$$
$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$\log Z(\hat{\theta}) \le E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big]$$

p_θ $\quad \hat{\theta}^* = \theta \quad \hat{\theta}^* = \theta \quad p_\theta$

$$H(p) \le \min_{\hat{\theta}} \Big\{ E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big] - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$H(p_\theta) \le E_\gamma \Big[\max_y \{\theta(y) + \sum_i \gamma_i(y_i)\} \Big] - \sum_y \theta(y) p_\theta(y)$$

Uncertainty*

- Theorem: $$H(p_\theta) \le E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- Proof idea: conjugate duality

$$H(p) = \min_{\hat{\theta}} \Big\{ \log Z(\hat{\theta}) - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$\log Z(\hat{\theta}) \le E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big]$$

$$\color{blue}{p_\theta} \qquad \color{red}{\hat{\theta}^* = \theta} \qquad \color{red}{\hat{\theta}^* = \theta} \qquad \color{blue}{p_\theta}$$

$$H(p) \le \min_{\hat{\theta}} \Big\{ E_\gamma \Big[\max_y \{\hat{\theta}(y) + \sum_i \gamma_i(y_i)\} \Big] - \sum_y \hat{\theta}(y) p(y) \Big\}$$

$$H(p_\theta) \le E_\gamma \Big[\cancel{\max_y \{\theta(y}) + \sum_i \gamma_i(y_i)\} \Big] - \sum_y \theta\cancel{(y) p_\theta(y)}$$

Sample complexity*

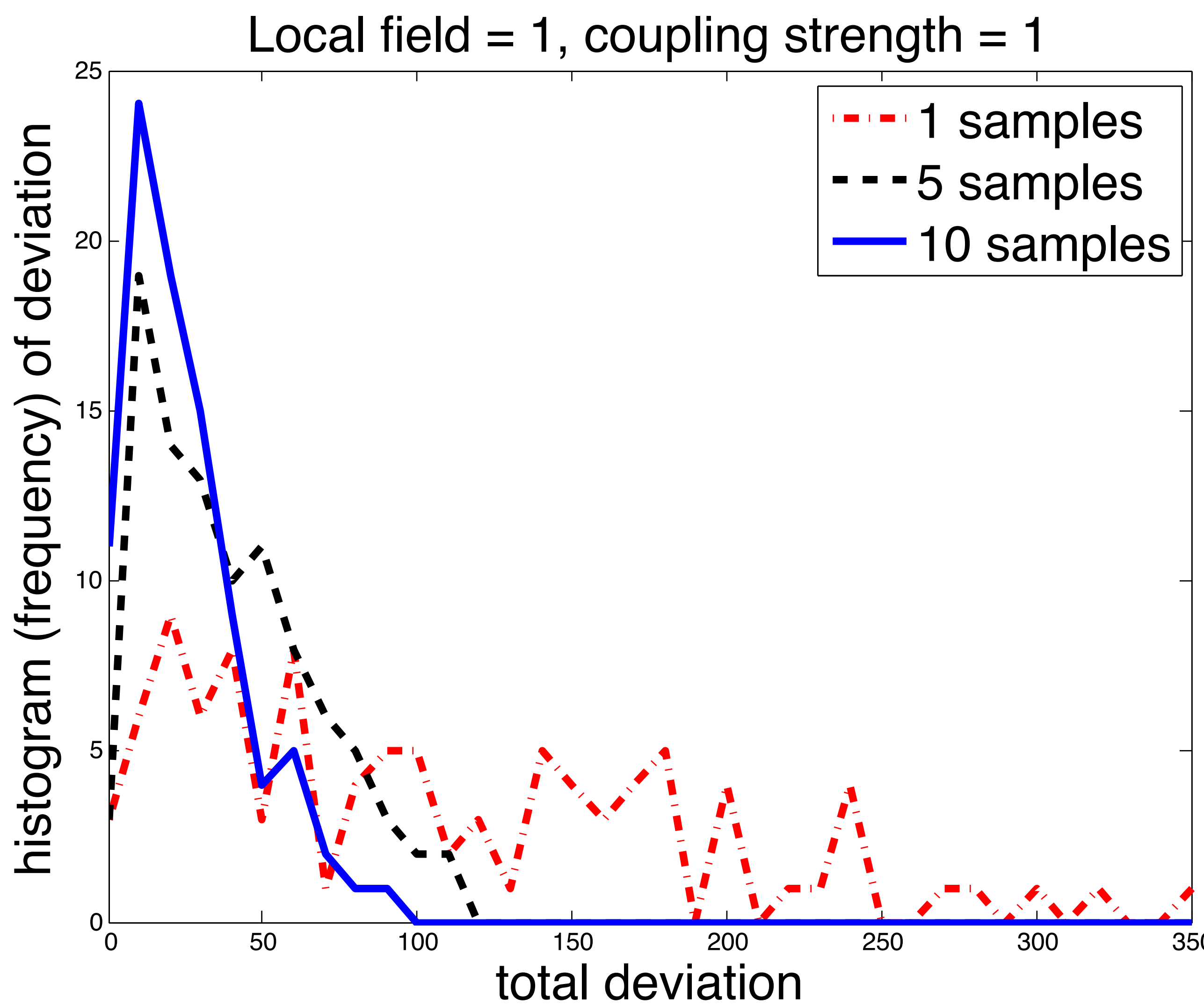
- The upper bounds hold in expectation.

$$H(p_\theta) \le E\Big[\sum_i \gamma_i(y_i^*)\Big]$$

$$y^* = \arg\max_{\hat{y}} \{\theta(\hat{y}) + \sum_{i=1}^{n} \gamma_i(\hat{y}_i)\}$$

- The distance between the sampled average and the true expectation decays exponentially

Sample complexity*



$$P[\text{avg of M samples} \le \text{expectation} + r] \le \exp\Big(-\frac{M}{20}\min(r, \frac{r^2}{n})\Big)$$

[Orabona, H., Sarwate, Jaakkola 14], [Nguyen 14]

Sample complexity*

- Why is it hard to get exponential decay?

Sample complexity*

- Why is it hard to get exponential decay?

$$P\left[\sum_i \gamma_i(y_i^*) > r\right] \leq \frac{E\left[\exp(\sum_i \gamma_i(y_i^*))\right]}{\exp(r)}$$

Sample complexity*

- Why it is hard to get exponential decay?

moment generating function

$$P\Big[\sum_i \gamma_i(y_i^*) > r\Big] \leq \frac{E\big[\exp(\sum_i \gamma_i(y_i^*))\big]}{\exp(r)}$$

- Measure concentration requires to bound the moment generating function
  - Hoeffding concentration requires bounded perturbations.
  - McDiarmid concentration requires bounded differences.
  - Our perturbations are unbounded with exponential tail.

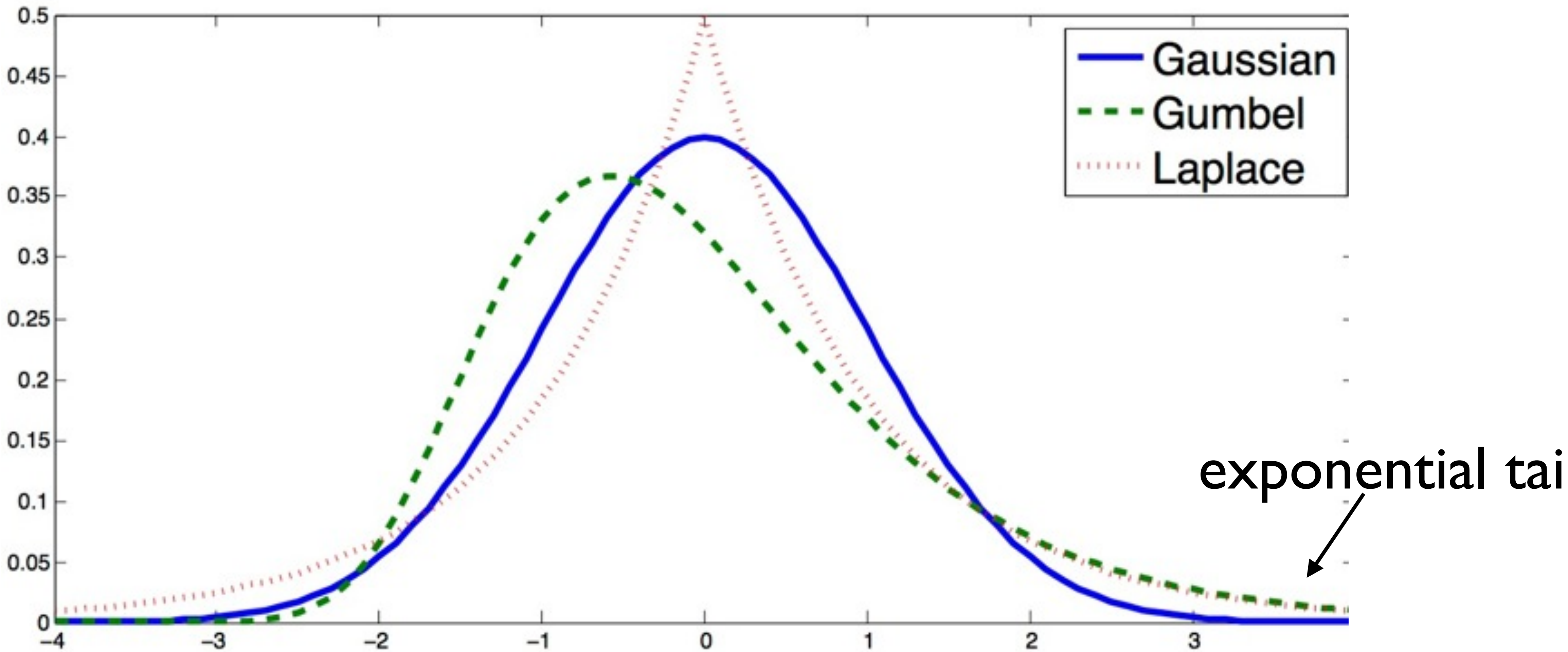
Sample complexity*

- The exponential tail of Gumbel distribution

$$E\Big[\exp\Big(\sum_i \gamma_i(y_i^*)\Big)\Big] = \int q(\gamma)\exp\Big(\sum_i \gamma_i(y_i^*)\Big)$$

$q(\gamma_i(y_i)) \to \exp(-\gamma_i(y_i))$

$\sim \exp(\gamma_i(y_i))$



Sample complexity*

- A function concentrates around its expectation if it does not change too much.

Sample complexity*

- A function concentrates around its expectation if it does not change too much.
  - Use tensorization to deal with one dimension at a time

Sample complexity*

- A function concentrates around its expectation if it does not change too much.
  - Use tensorization to deal with one dimension at a time

$$Var[\sum_i \gamma_i(y_i^*)] = \sum_{j,y_j} Var_{\gamma_{j,y_j}}[\sum_i \gamma_i(y_i^*)]$$

Sample complexity*

- A function concentrates around its expectation if it does not change too much.
  - Use tensorization to deal with one dimension at a time

$$Var[\sum_i \gamma_i(y_i^*)] = \sum_{j,y_j} Var_{\gamma_{j,y_j}}[\sum_i \gamma_i(y_i^*)]$$

  - Bound any dimension's variance with its perturb-max probability (a Poincare inequality)

$$Var_{j,y_j}[\cdot] \leq P_{\gamma_j(y_j)}[y_j = \arg\max_y\{\theta(y) + \sum_i \gamma_i(y_i)\}]$$

Outline

- Random perturbation - why and how?
  - Sampling likely structures as fast as finding the most likely one.
- Connections and Alternatives to Gibbs distribution:
  - the marginal polytope
  - non-MCMC sampling for Gibbs distributions with perturb-max
- Application: interactive annotation.
  - New entropy bounds for perturb-max models.

Open problems

- Perturb-max models:
  - How do perturb-max models generalize - Follow the Perturbed Leader [Manfred Warmuth, Jacob Abernethy]
  - Adversarial learning objective [Ian Goodfellow]
  - Perturb-max models stabilize the prediction. Do they connect computational and statistical stability [Yury Makarychev]?
  - Perturb-max models in continuous space [Maddison et. al 14]
  - When does fixing variables in the max-function amount to statistical conditioning?
  - When do perturb-max models preserve the most likely assignment?
  - How do the perturbations dimension affect the model properties?
  - How to encourage diverse sampling?

Thank you