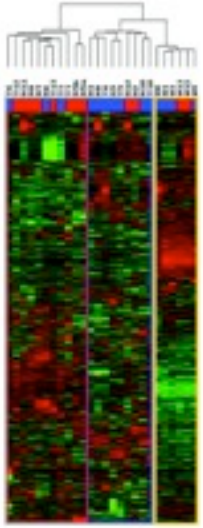
What cannot be learned with belief propagation

Amir Globerson and Uri Heinemann (Hebrew Univ.)

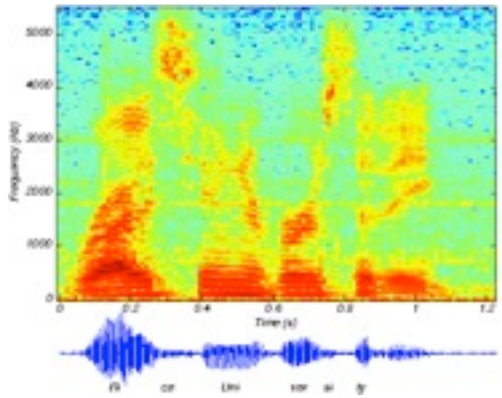
Multivariate Signals

- High dimensional signals are everywhere!
- Need a principled way for modeling distributions over those.



The New York Times.

ARMISTICE SIGNED, END OF THE WAR!
BERLIN SEIZED BY REVOLUTIONISTS;
NEW CHANCELLOR BEGS FOR ORDER;
OUSTED KAISER FLEES TO HOLLAND



Modeling Multivariate Distributions

- Goal: Model distributions over $x_1, \ldots, x_n$
- Problem: For large n, this requires an exponential number of parameters
- Approach: Model distribution as a product of "local" factors

$$p(x_1, \ldots, x_n) \propto \prod_c \psi_c(x_c) \qquad c \subset \{1, \ldots, n\}$$

Example: $p(x_1, \ldots, x_n) \propto \psi(x_1, x_2, x_3)\psi(x_2, x_4)\psi(x_2, x_6, x_8) \ldots$

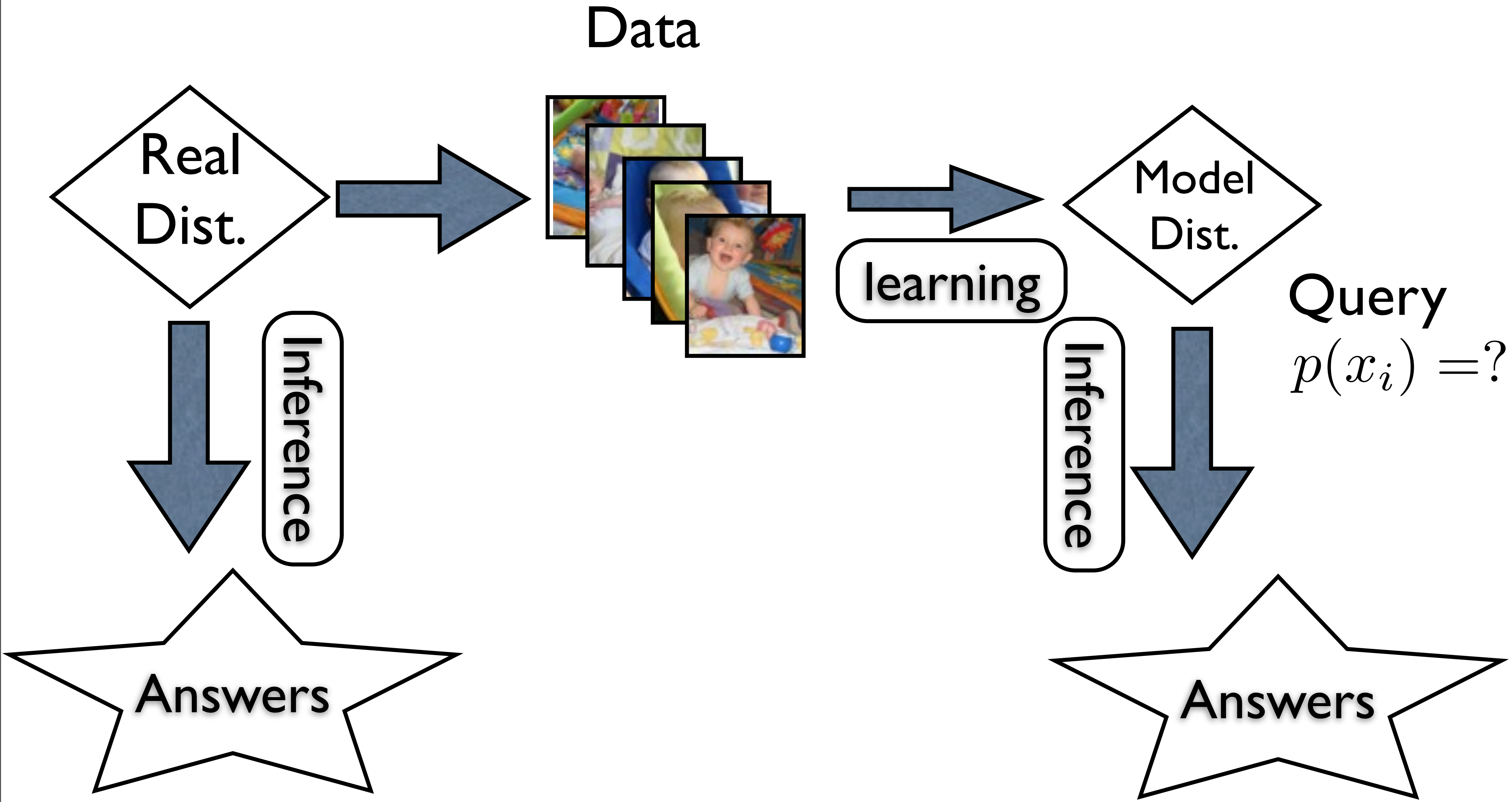
- Focus on pairwise factors

Pairwise Graphical Models

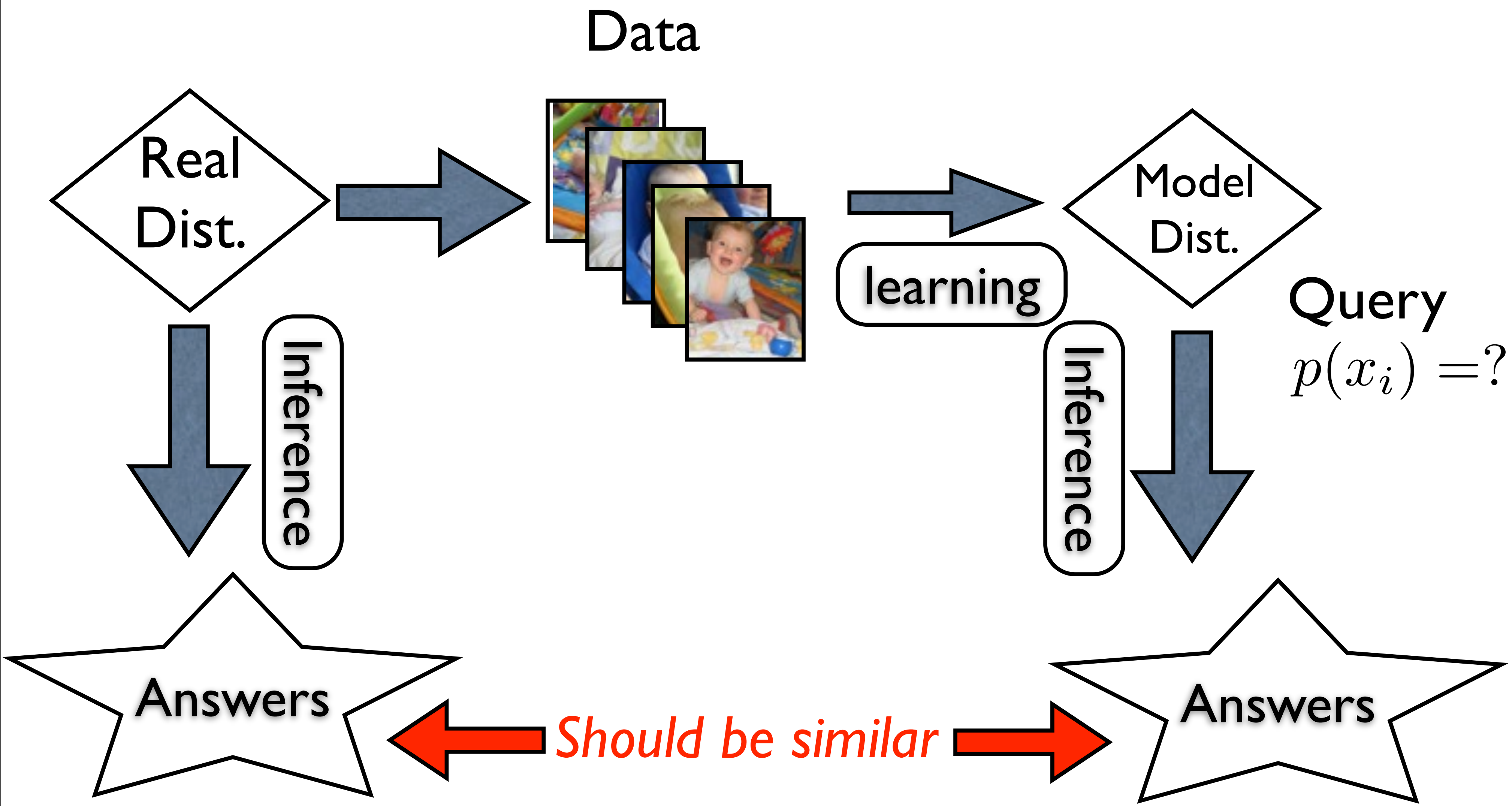
- Consider graph G=(V,E) with n nodes
- Functions on E,V: $\theta_{ij}(x_i, x_j), \theta_i(x_i)$
- Defines a distribution over n variables

$$p(x_1, \ldots, x_n; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{\sum_{ij \in E} \theta_{ij}(x_i, x_j) + \sum_{i \in V} \theta_i(x_i)}$$

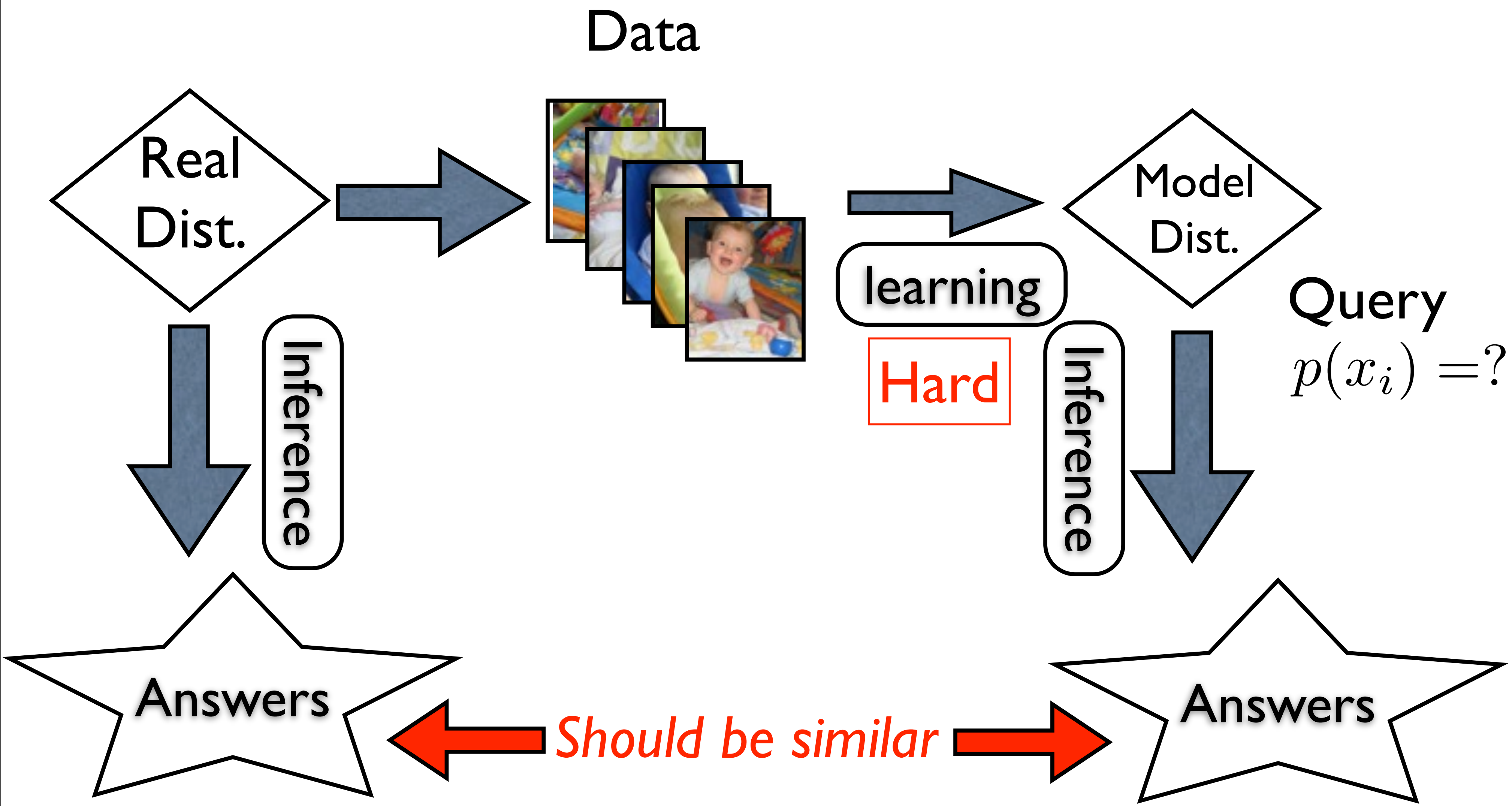
The Learning Problem



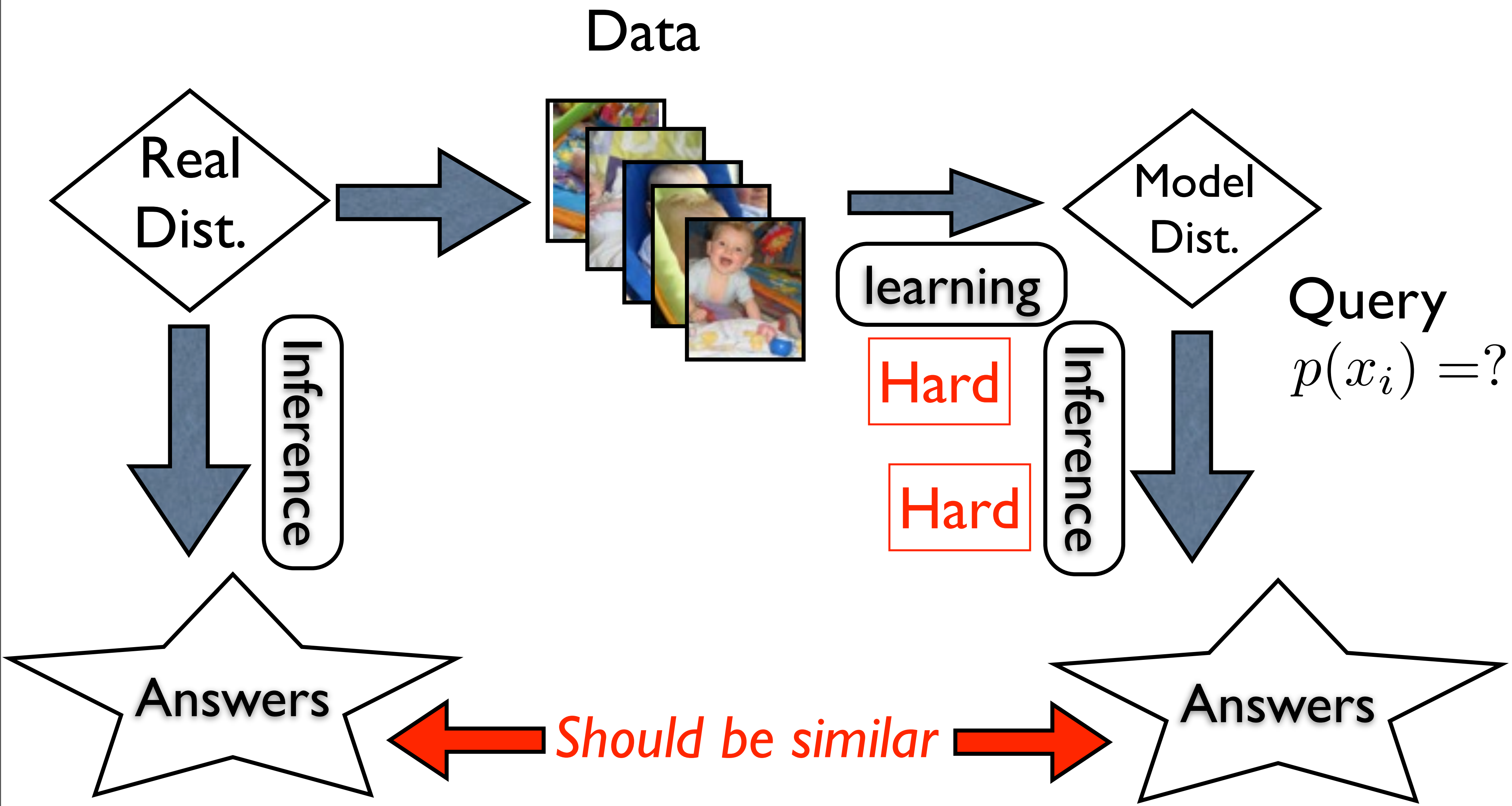
The Learning Problem



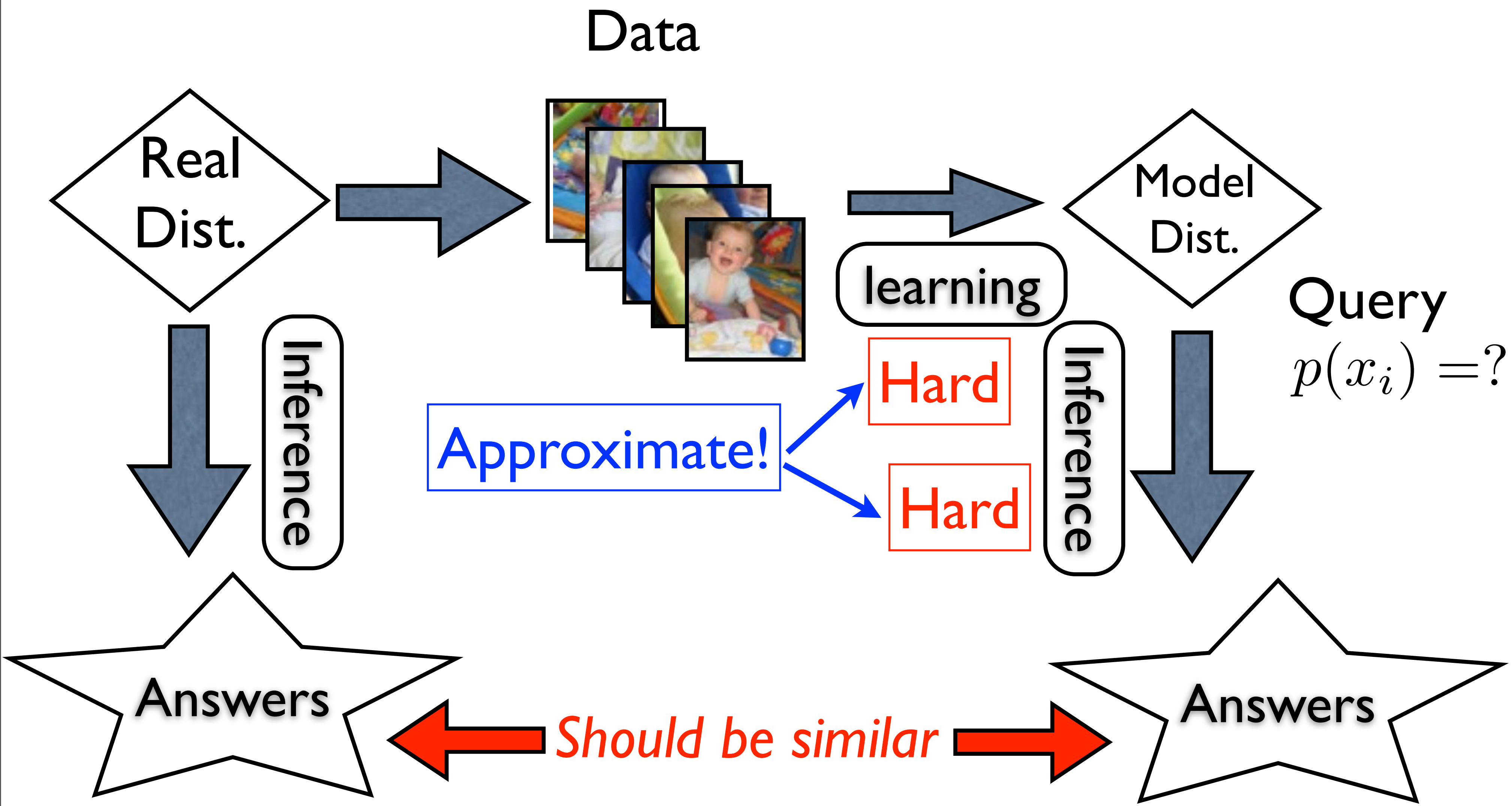
The Learning Problem



The Learning Problem



The Learning Problem



Approximate Learning

- Goal: Understand how well we can learn with approximate learning and inference?.
- Focus on approximation using loopy belief propagation
  - Good approximation for marginals.
  - Learning with it is poorly understood.

Results

- BP has “spectacular failure modes” for learning.
- Characterize those.
- Well correlated with empirical behavior.
- Suggests which models to use when learning with BP.
- New insights on BP fixed points.

Maximum Likelihood

- Given M training instances: $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}$
- Each instance is an assignment to n variables:

$$\boldsymbol{x}^{(i)} = \left[x_1^{(i)}, \ldots, x_n^{(i)}\right]$$

- Find $\boldsymbol{\theta}$ that maximizes the likelihood:

$$\ell(\boldsymbol{\theta}) = \frac{1}{M} \sum_m \log p(x^{(m)}; \boldsymbol{\theta})$$

Maximum Likelihood

- Rewrite the likelihood in a simpler form.
- Define empirical marginals:

$$\bar{\mu}_i(x_i) = \frac{1}{M}\sum_m \delta_{x_i^{(m)},x_i}$$

$$\bar{\mu}_{ij}(x_i, x_j) = \frac{1}{M}\sum_m \delta_{x_i^{(m)},x_i}\delta_{x_j^{(m)},x_j}$$

- Then:

$$\ell(\boldsymbol{\theta}) = \sum_{ij} \bar{\mu}_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) + \sum_i \bar{\mu}_i(x_i)\theta_i(x_i) - \log Z(\boldsymbol{\theta})$$

- Or:

Maximum Likelihood

- Rewrite the likelihood in a simpler form.
- Define empirical marginals:

$$\bar{\mu}_i(x_i) = \frac{1}{M}\sum_m \delta_{x_i^{(m)}, x_i}$$

$$\bar{\mu}_{ij}(x_i, x_j) = \frac{1}{M}\sum_m \delta_{x_i^{(m)}, x_i}\delta_{x_j^{(m)}, x_j}$$

- Then:

$$\ell(\boldsymbol{\theta}) = \sum_{ij} \bar{\mu}_{ij}(x_i, x_j)\theta_{ij}(x_i, x_j) + \sum_i \bar{\mu}_i(x_i)\theta_i(x_i) - \log Z(\boldsymbol{\theta})$$

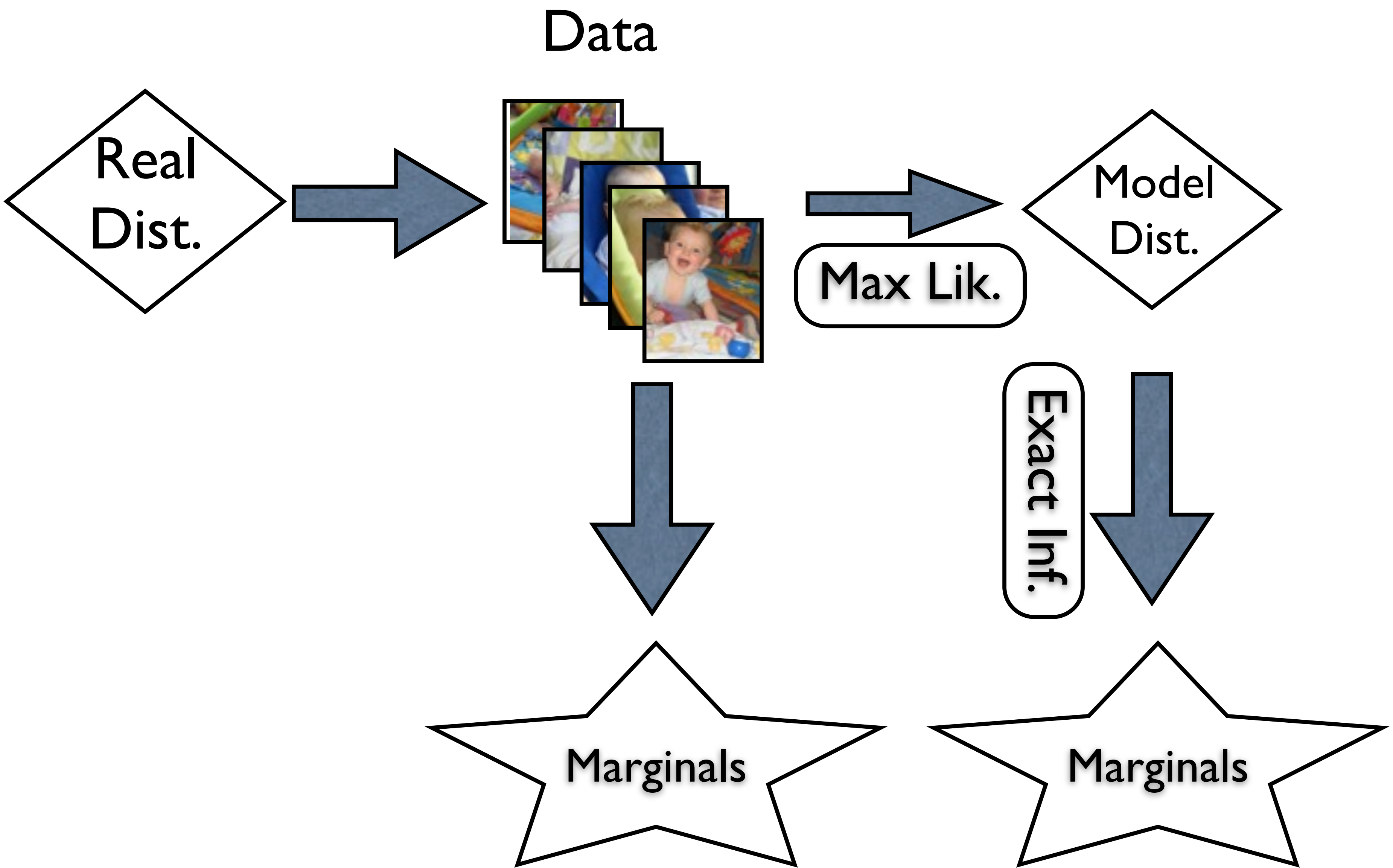
- Or:

$$\boxed{\ell(\boldsymbol{\theta}) = \bar{\mu}\cdot\boldsymbol{\theta} - \log Z(\boldsymbol{\theta})}$$

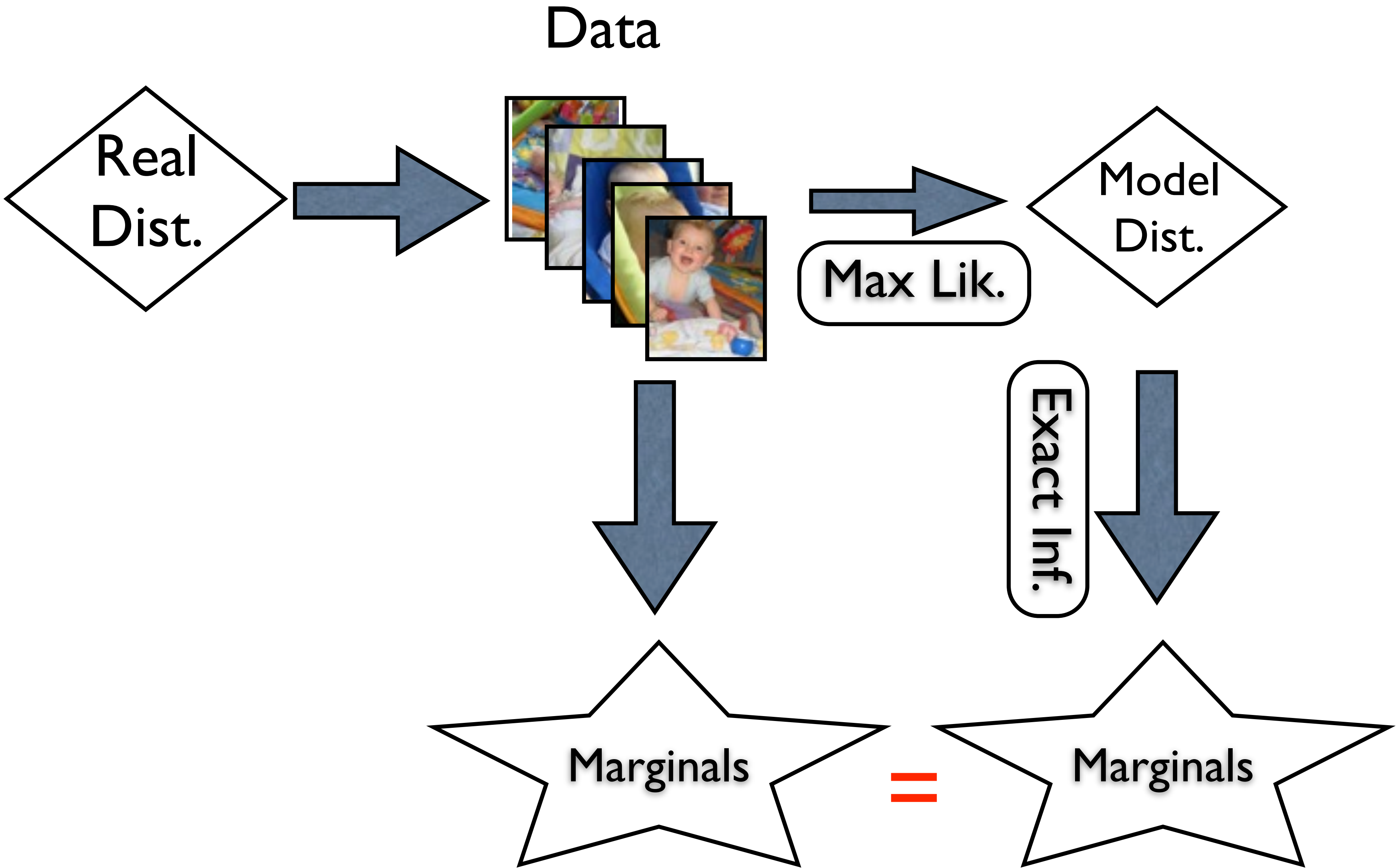
Maximum Likelihood

- Goal is to maximize: $\boxed{\ell(\boldsymbol{\theta}) = \bar{\mu} \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta})}$
- Difficulty is to calculate the partition function and gradient (marginals).
- Say we can maximize it efficiently...
- The optimum parameter has a simple characterization: moment matching.

Moment Matching



Moment Matching



Moment Matching

- Define the marginals for parameter $\boldsymbol{\theta}$ as:

$$\boldsymbol{\mu}_i^{\boldsymbol{\theta}}(x_i) = p(x_i; \boldsymbol{\theta})$$

$$\boldsymbol{\mu}_{ij}^{\boldsymbol{\theta}}(x_i, x_j) = p(x_i, x_j; \boldsymbol{\theta})$$

- The maximum likelihood parameters satisfy:

$$\boldsymbol{\mu}^{\boldsymbol{\theta}_{ML}} = \bar{\boldsymbol{\mu}}$$

Moment Matching

- Define the marginals for parameter $\boldsymbol{\theta}$ as:

$$\boldsymbol{\mu}_i^{\boldsymbol{\theta}}(x_i) = p(x_i; \boldsymbol{\theta})$$

$$\boldsymbol{\mu}_{ij}^{\boldsymbol{\theta}}(x_i, x_j) = p(x_i, x_j; \boldsymbol{\theta})$$

- The maximum likelihood parameters satisfy:

$$\boldsymbol{\mu}^{\boldsymbol{\theta}_{ML}} = \bar{\boldsymbol{\mu}}$$

Moment Matching

Moment Matching

- Define the marginals for parameter $\boldsymbol{\theta}$ as:

$$\boldsymbol{\mu}_i^{\boldsymbol{\theta}}(x_i) = p(x_i; \boldsymbol{\theta})$$

$$\boldsymbol{\mu}_{ij}^{\boldsymbol{\theta}}(x_i, x_j) = p(x_i, x_j; \boldsymbol{\theta})$$

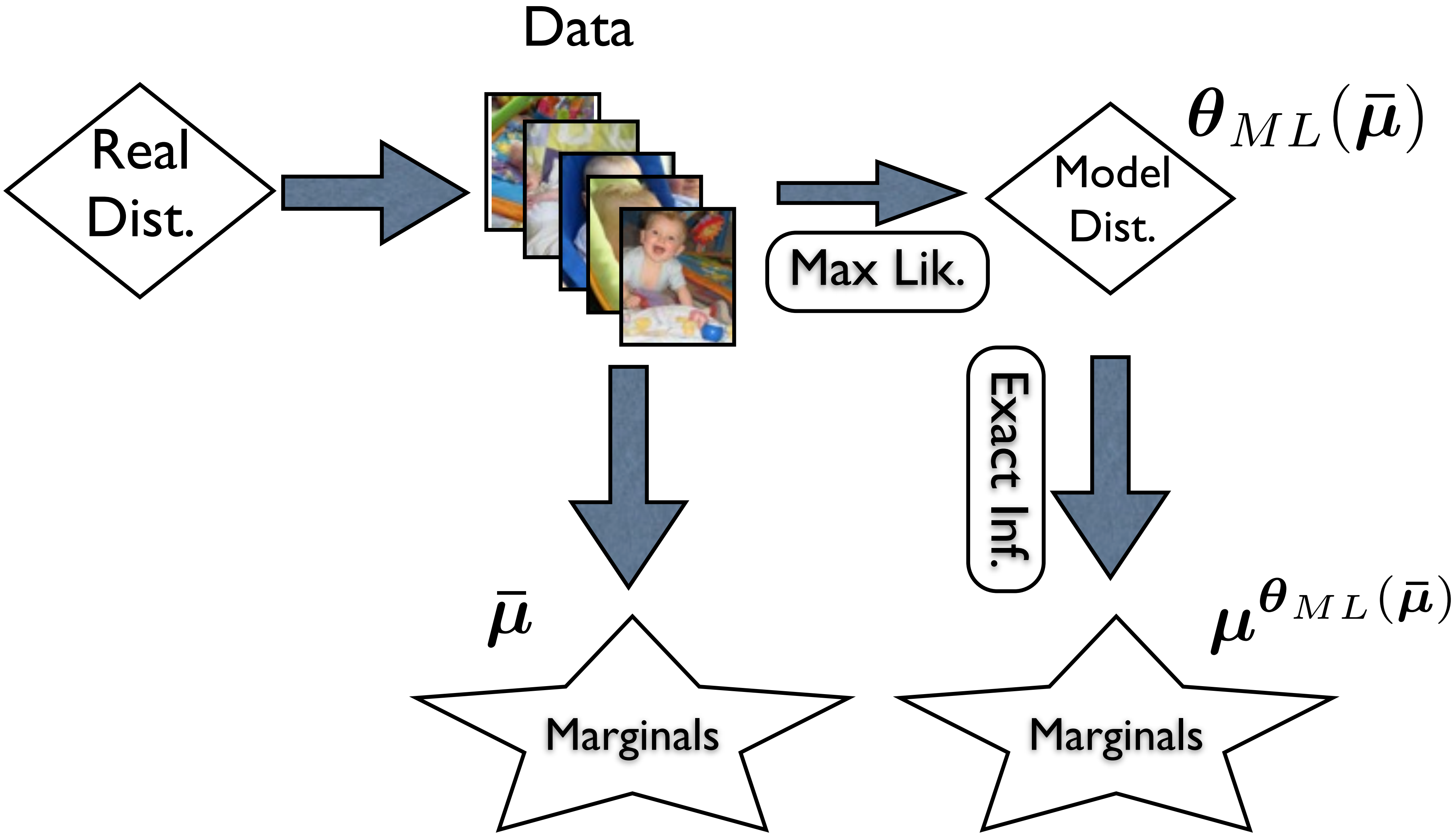
- The maximum likelihood parameters satisfy:

$$\boldsymbol{\mu}^{\boldsymbol{\theta}_{ML}} = \bar{\boldsymbol{\mu}}$$

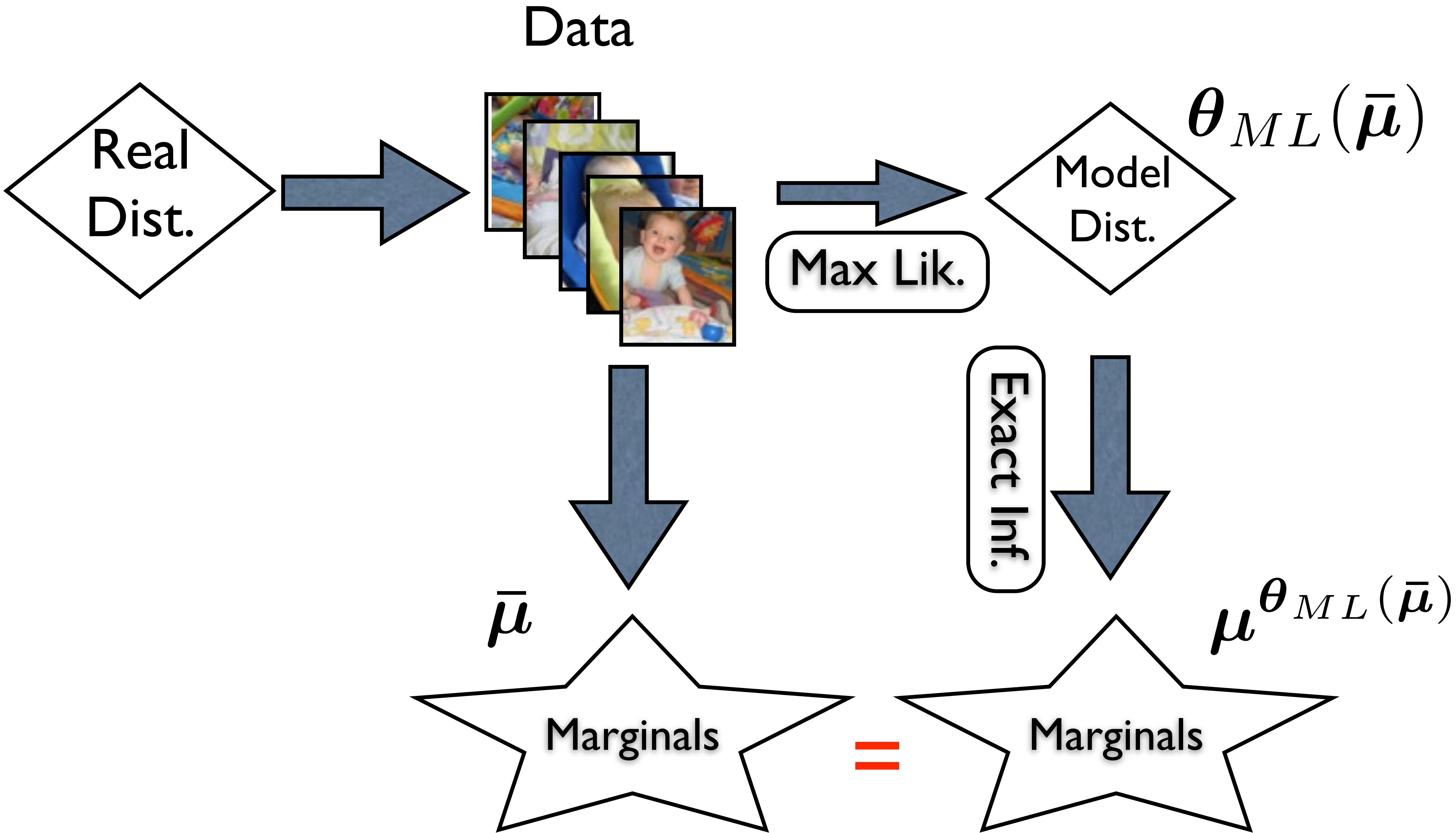
Moment Matching

- The marginals of the optimal model agree with the empirical ones!

Moment Matching



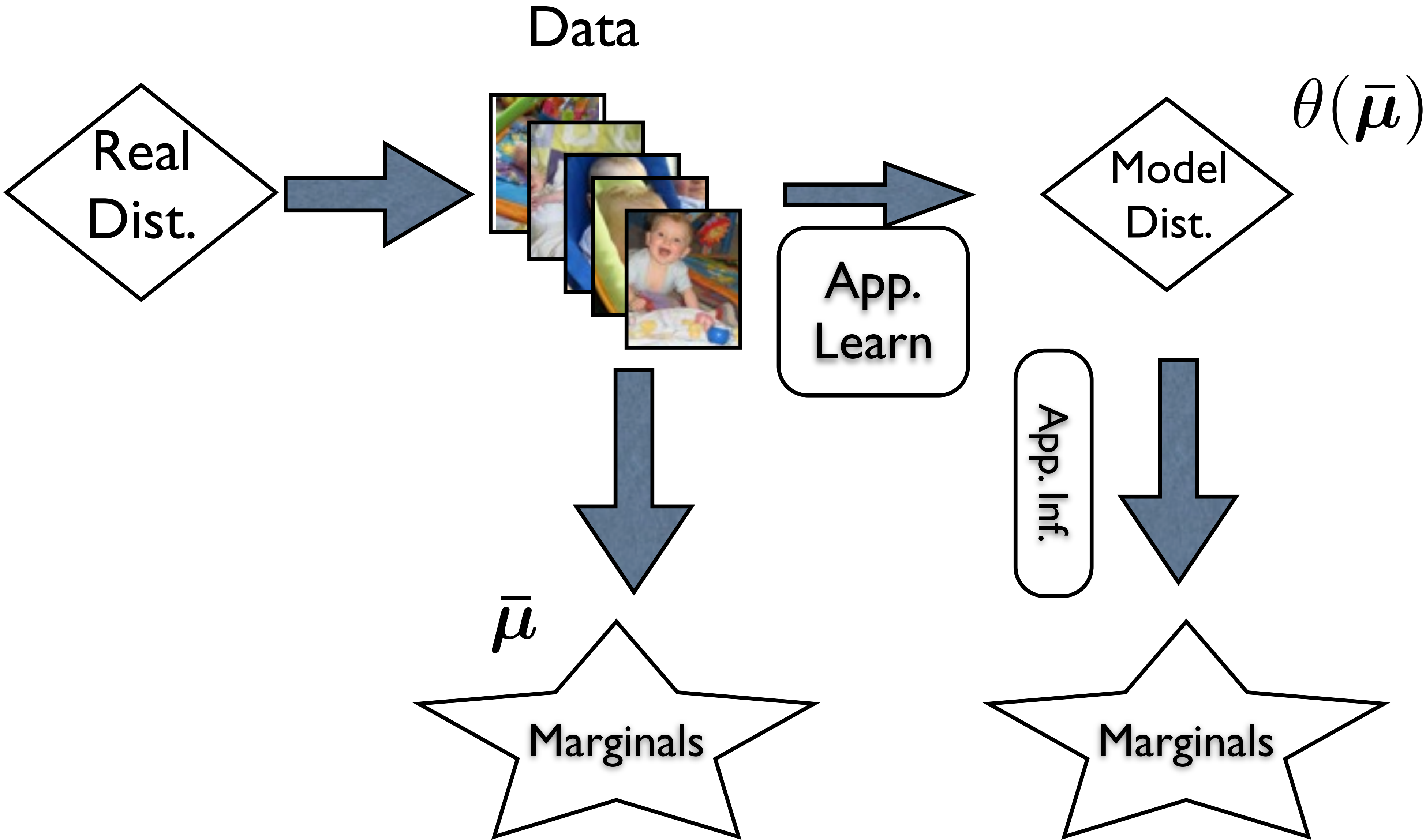
Moment Matching



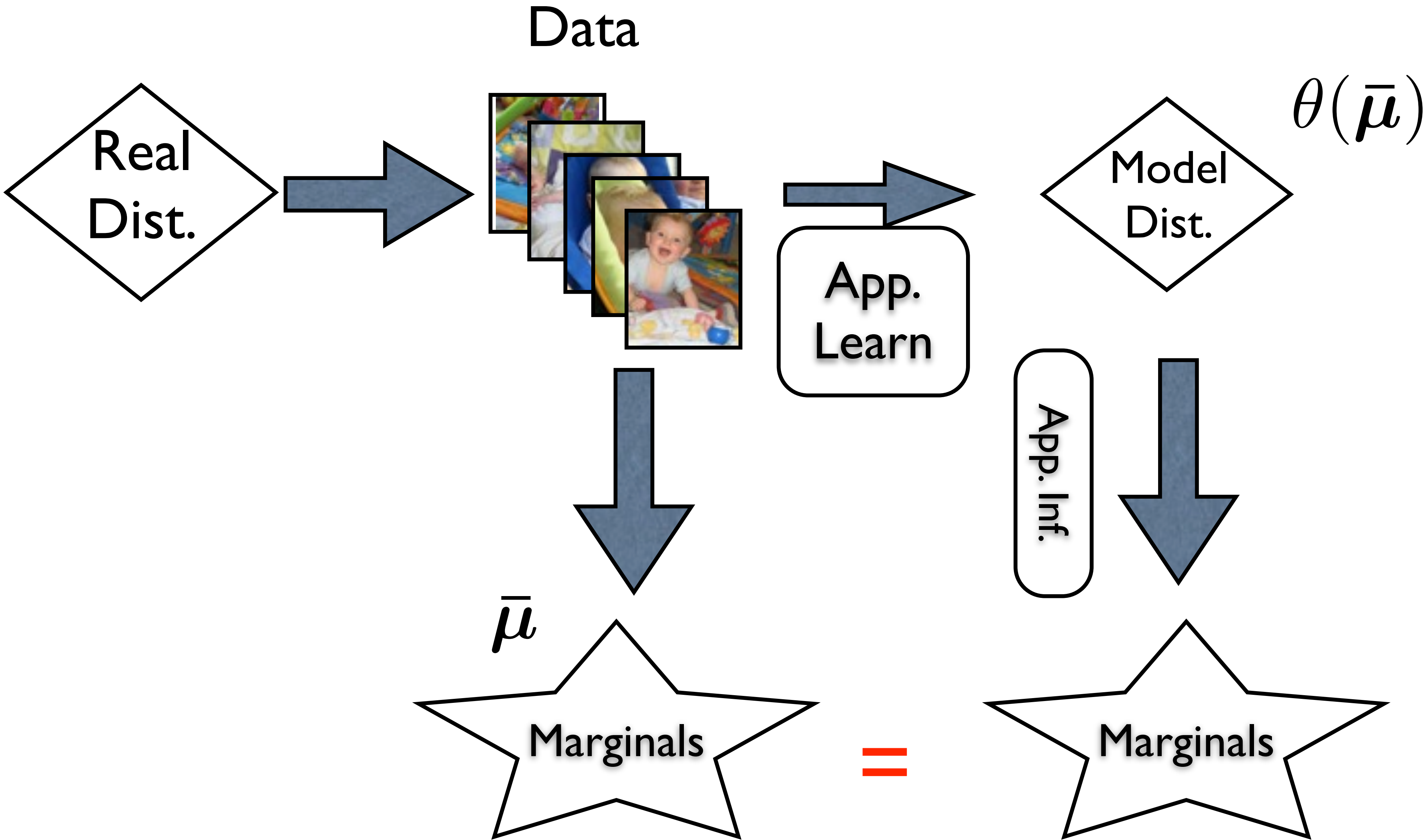
Moment Matching

- Makes sense. Means that the sufficient statistics of the model fit the empirical ones.
- If all we care about are these statistics, we don't really need to learn (e.g., *Wainwright 06*).
- Holds for exact learning.
- What happens if we approximate?
- For certain approximations (e.g., convex free energies) we get moment matching.
- What about Bethe/BP approaches?

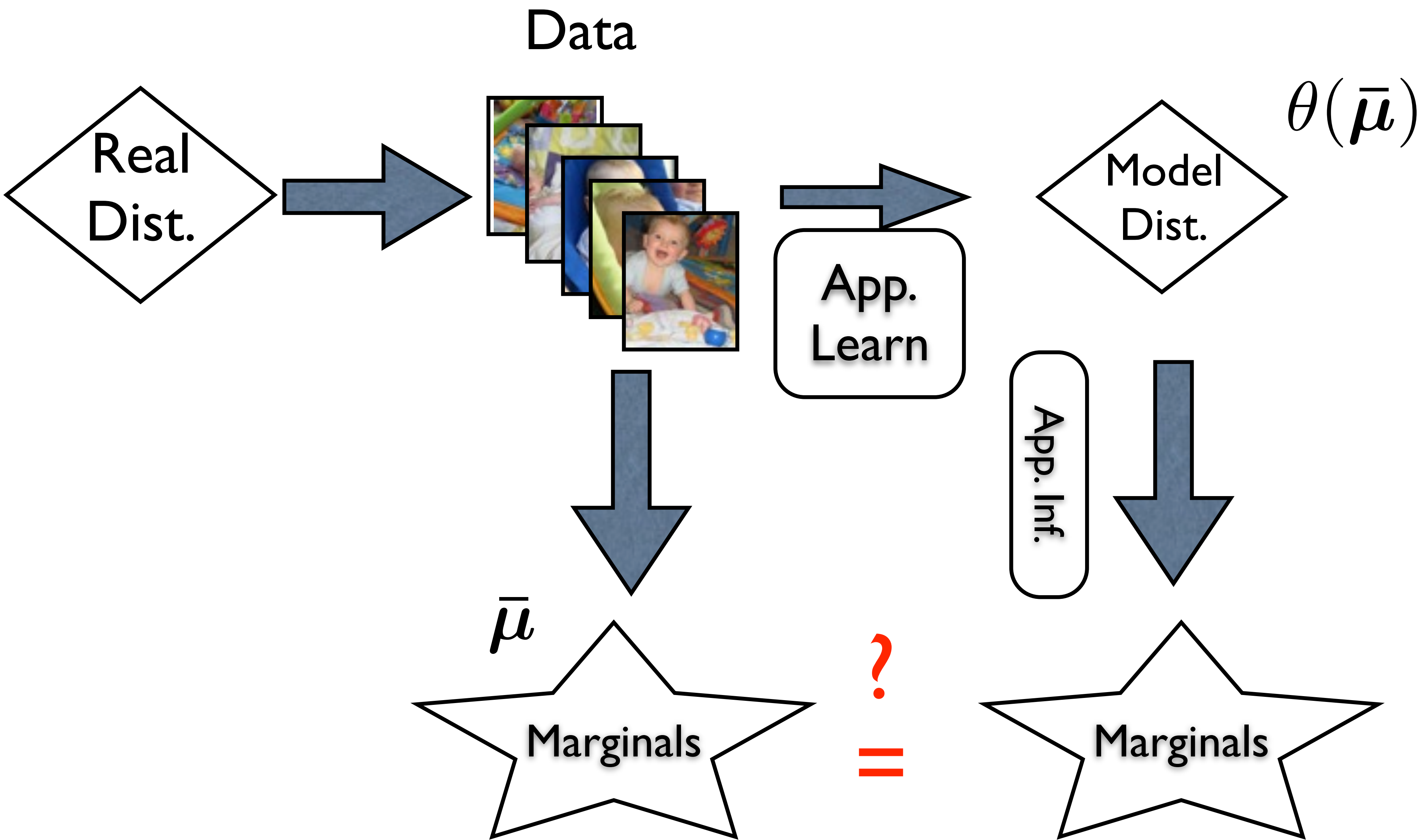
Approximate Learning



Approximate Learning



Approximate Learning



Approximate ML

- Recall the likelihood: $\ell(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}} \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta})$
- To maximize it we need to calculate:
  - Objective. Requires: $\log Z(\boldsymbol{\theta})$
  - Gradient. Requires: $\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\mu}^{\boldsymbol{\theta}}$

Approximate ML

- Recall the likelihood: $\ell(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}} \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta})$
- To maximize it we need to calculate:
  - Objective. Requires: $\log Z(\boldsymbol{\theta})$ *Hard!*
  - Gradient. Requires: $\dfrac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\mu}^{\boldsymbol{\theta}}$

Approximate ML

- Recall the likelihood: $\ell(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}} \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta})$
- To maximize it we need to calculate:
  - Objective. Requires: $\log Z(\boldsymbol{\theta})$ *Hard!*
  - Gradient. Requires: $\dfrac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\mu}^{\boldsymbol{\theta}}$ *Hard!*

Approximate ML

- Recall the likelihood: $\ell(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}} \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta})$
- To maximize it we need to calculate:
  - Objective. Requires: $\log Z(\boldsymbol{\theta})$ *Hard!*
  - Gradient. Requires: $\dfrac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{\mu}^{\boldsymbol{\theta}}$ *Hard!*
- Approximate both using a variational approach.

Variational view of Z

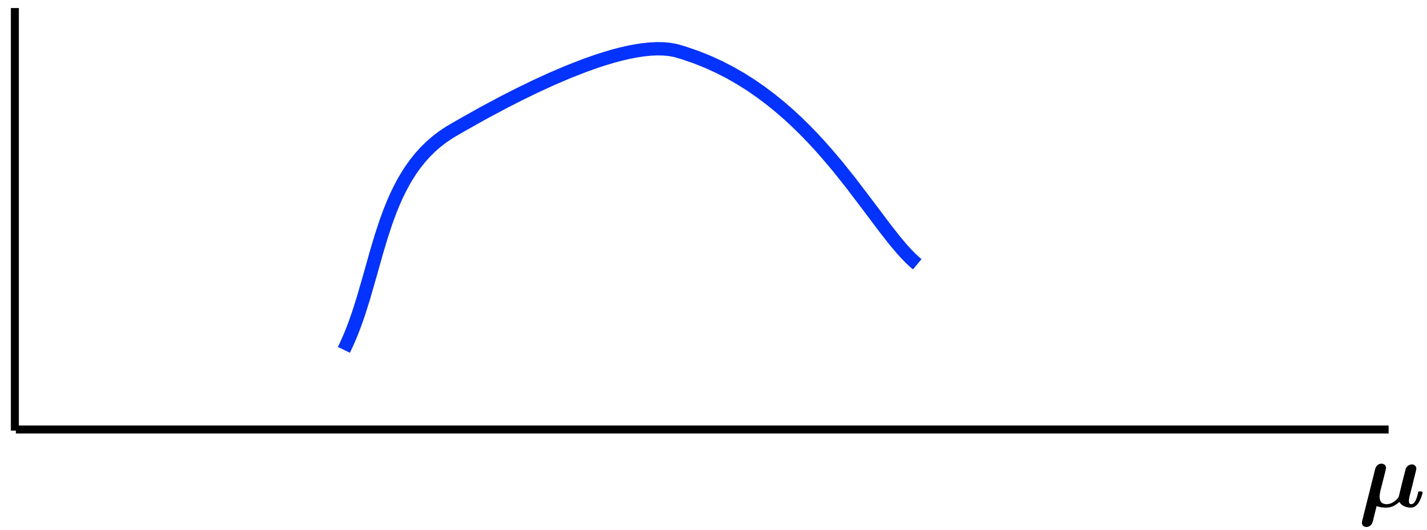
- Both partition function and marginals can be cast as solutions to optimization problem.

Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.

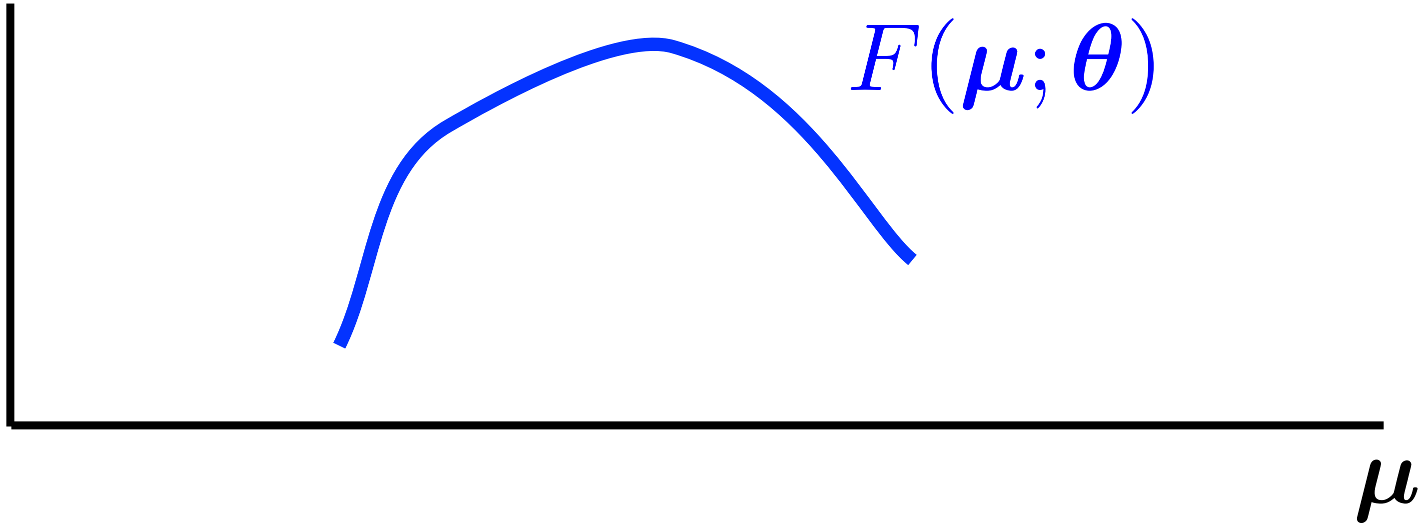
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



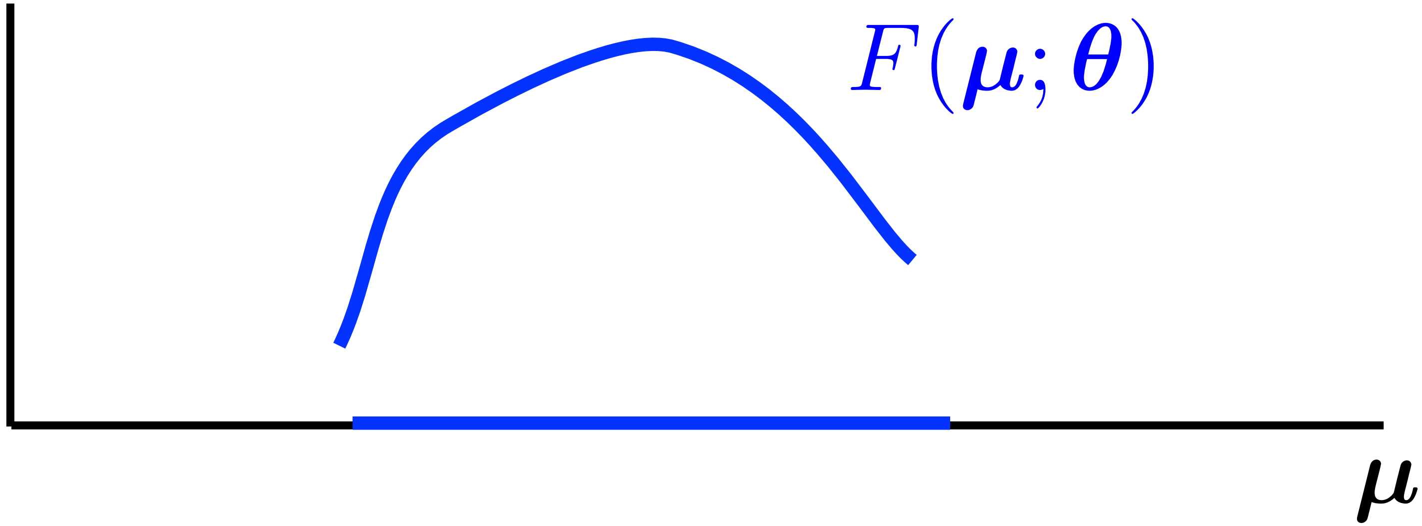
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



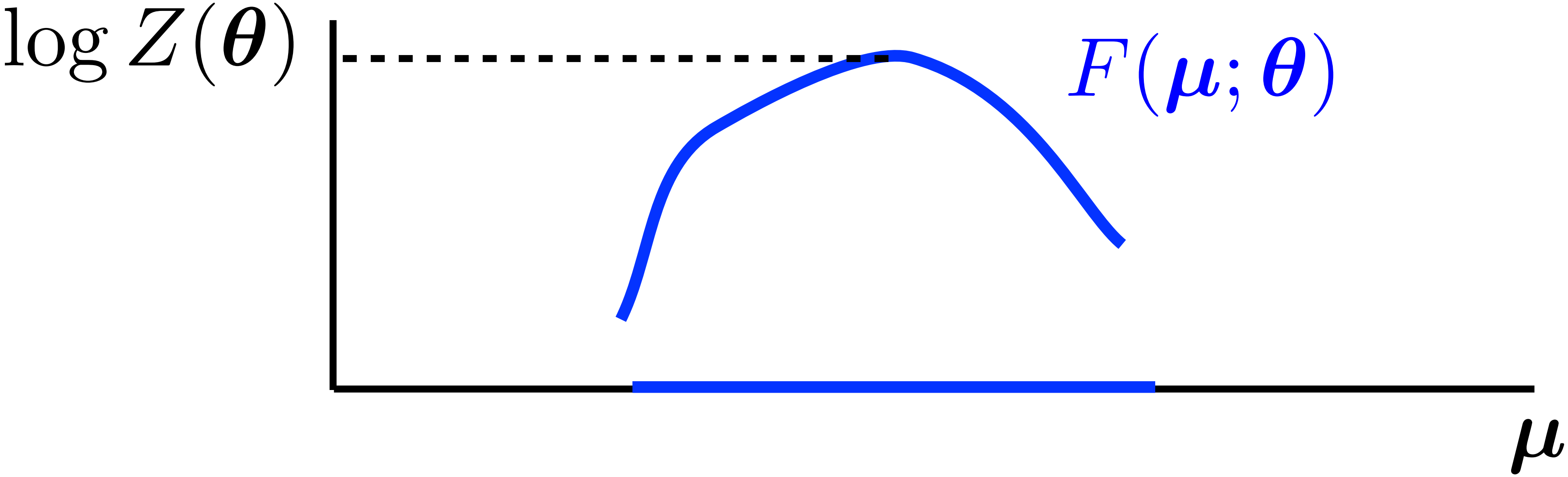
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



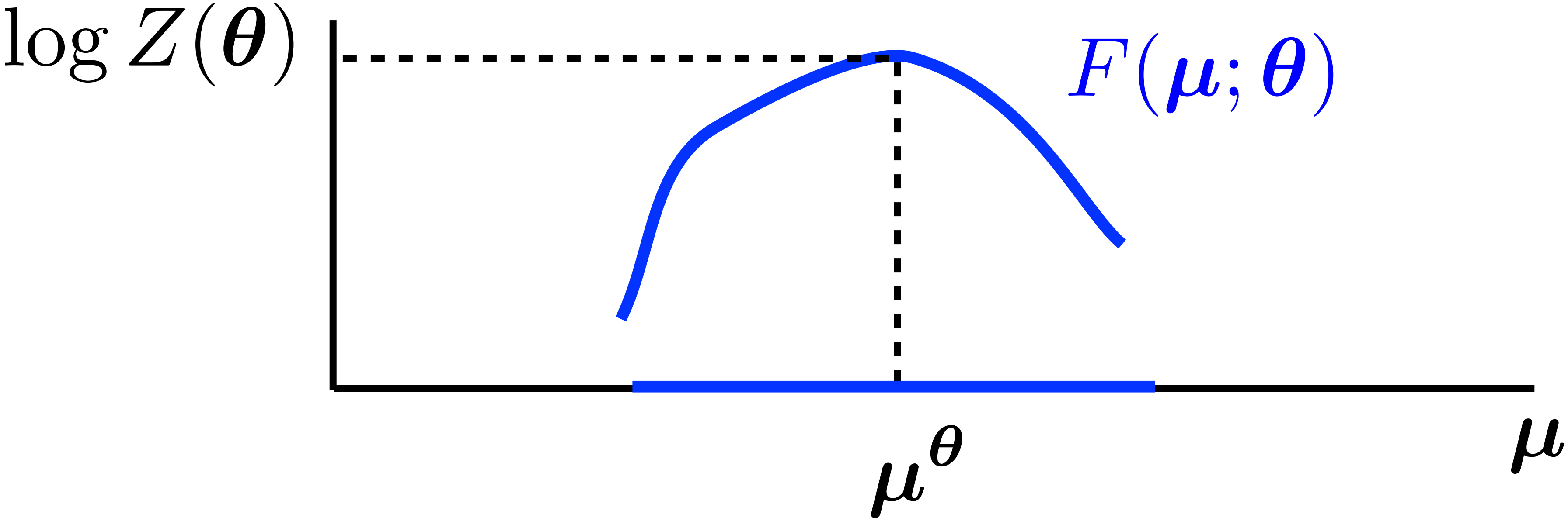
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



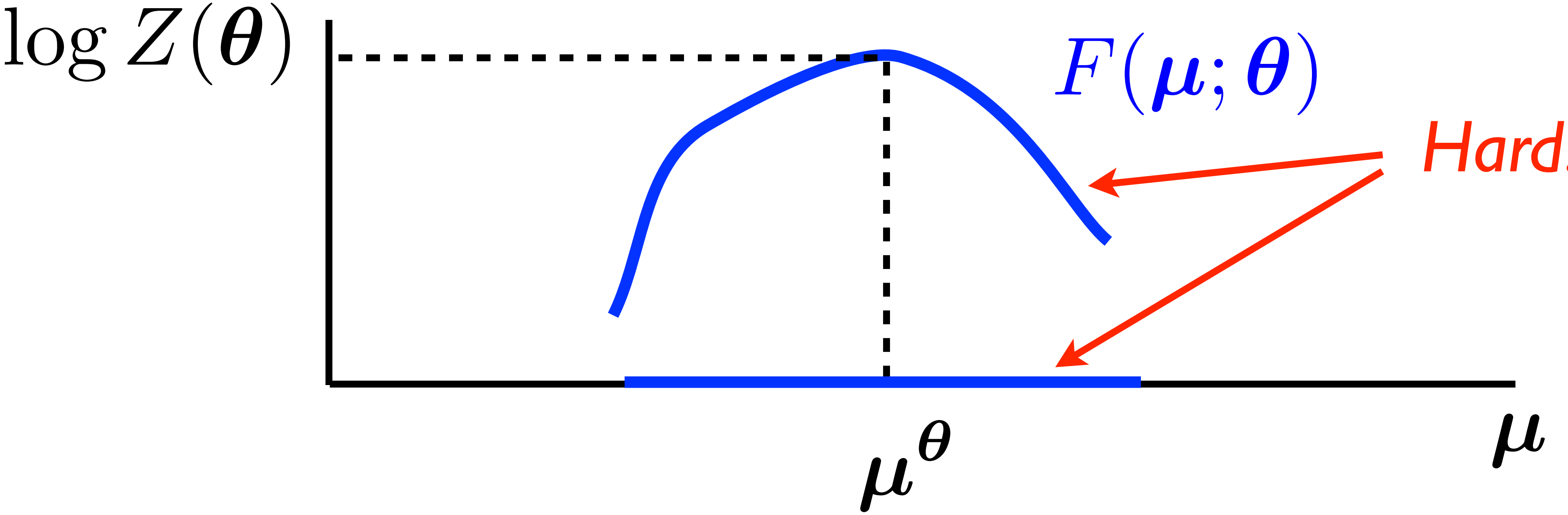
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



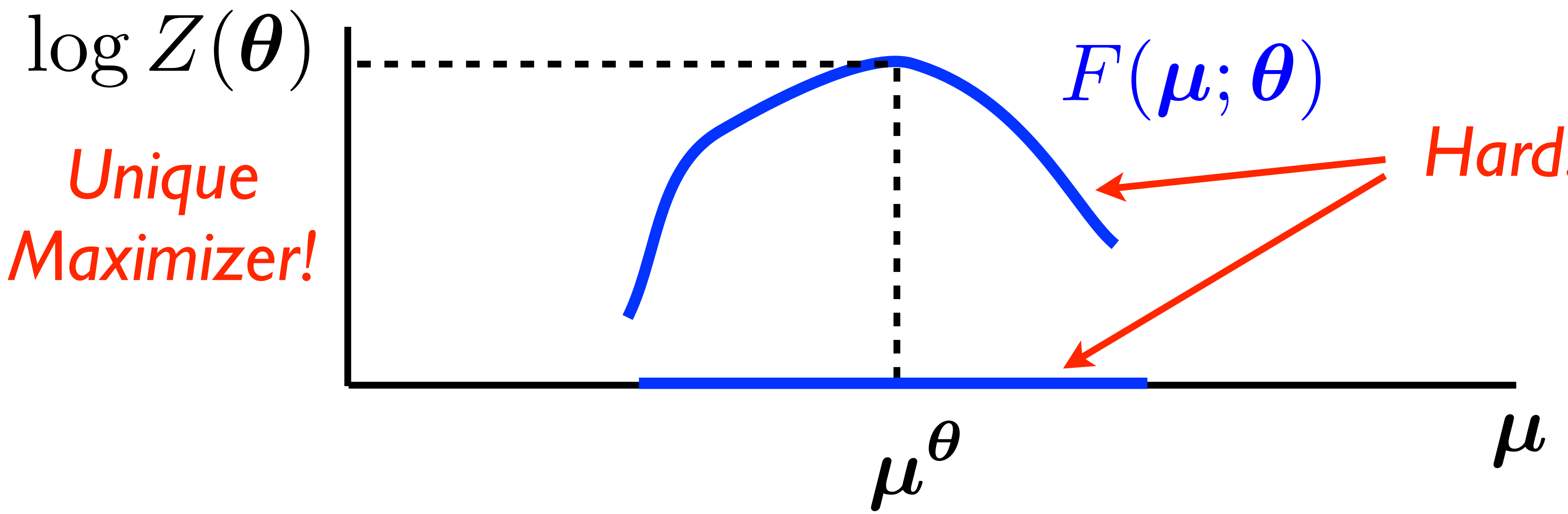
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



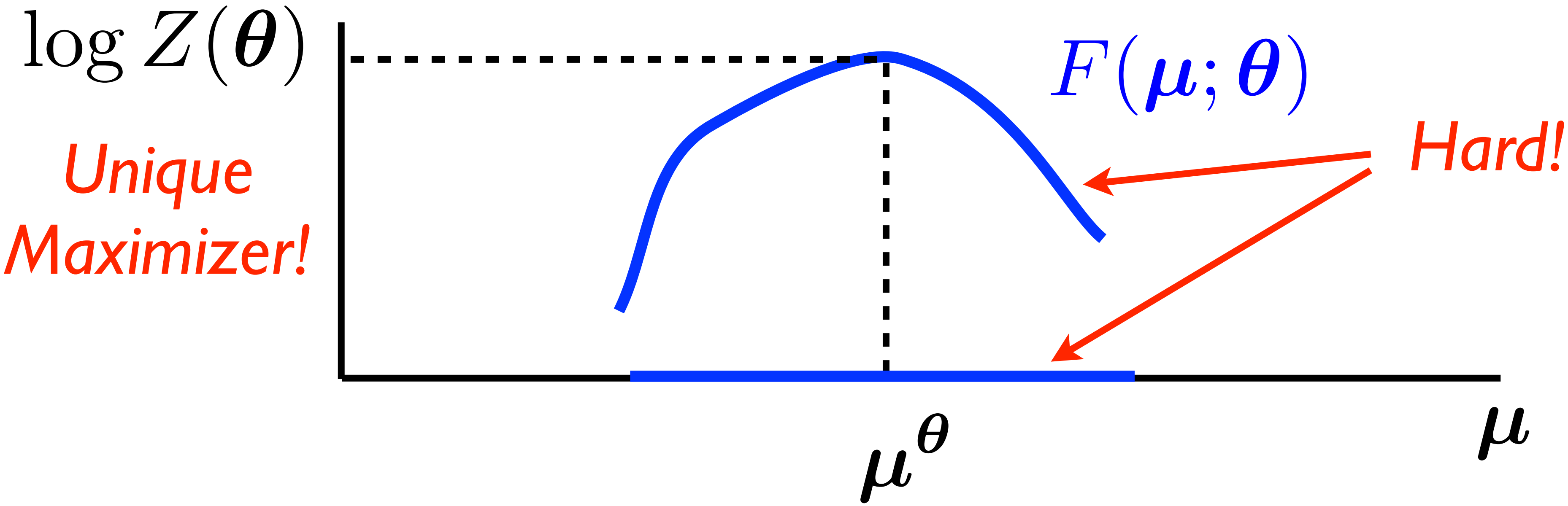
Variational view of Z

- Both partition function and marginals can be cast as solutions to optimization problem.



Variational view of Z

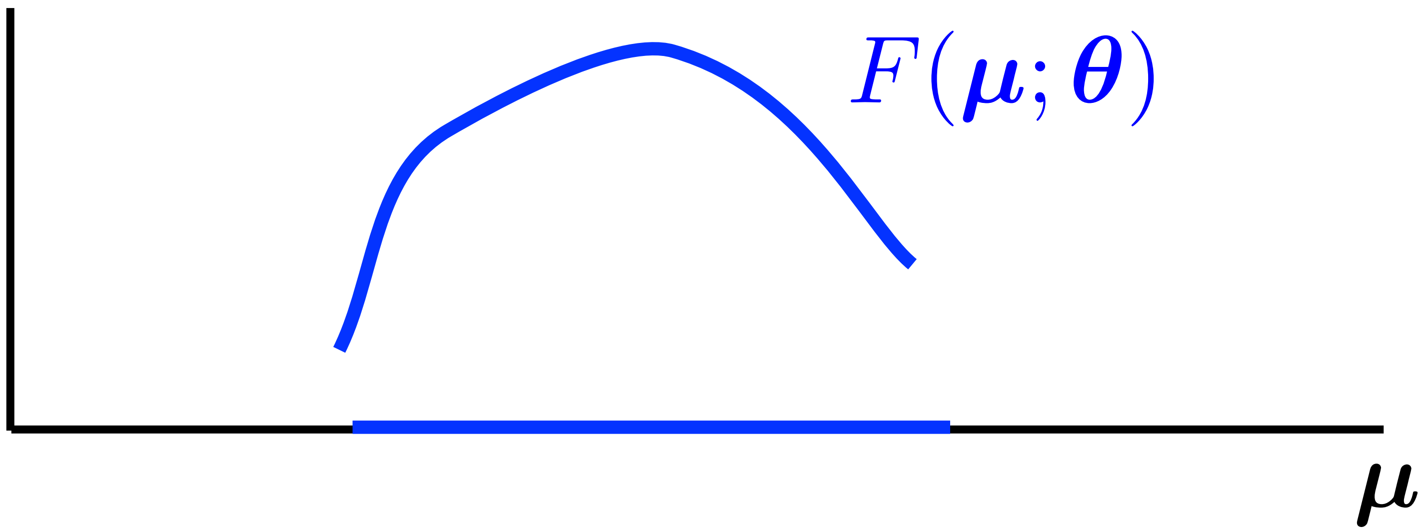
- Both partition function and marginals can be cast as solutions to optimization problem.



$$F(\boldsymbol{\mu};\boldsymbol{\theta}) = \boldsymbol{\mu}\cdot\boldsymbol{\theta} + H(\boldsymbol{\mu})$$

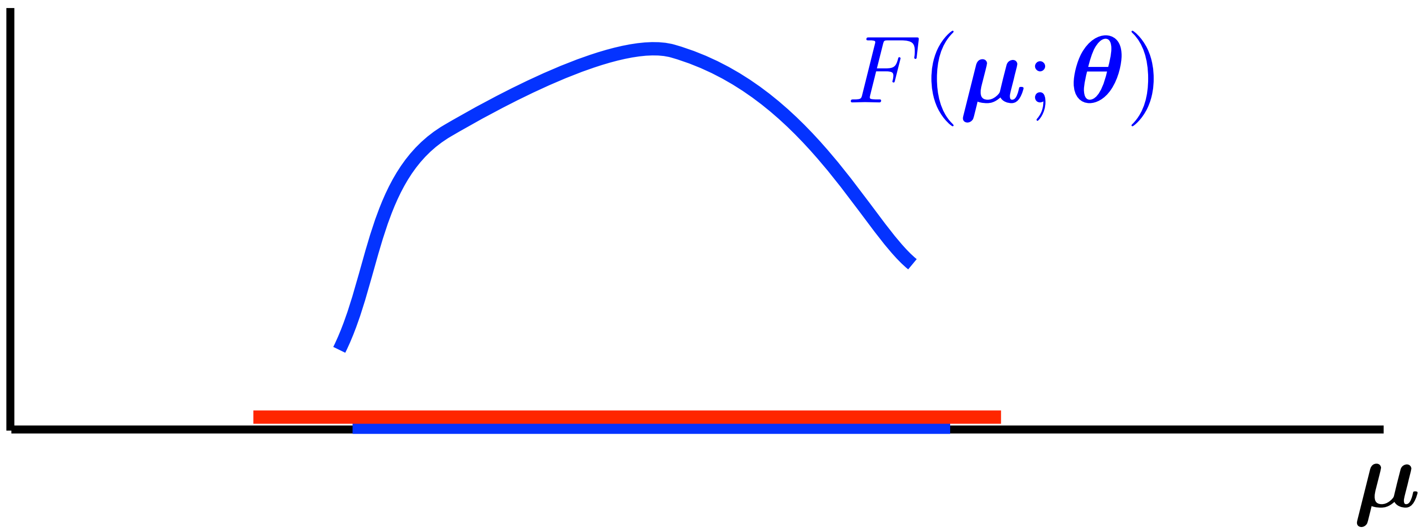
Bethe approximations

- Replace both constraints and objective with approximations.



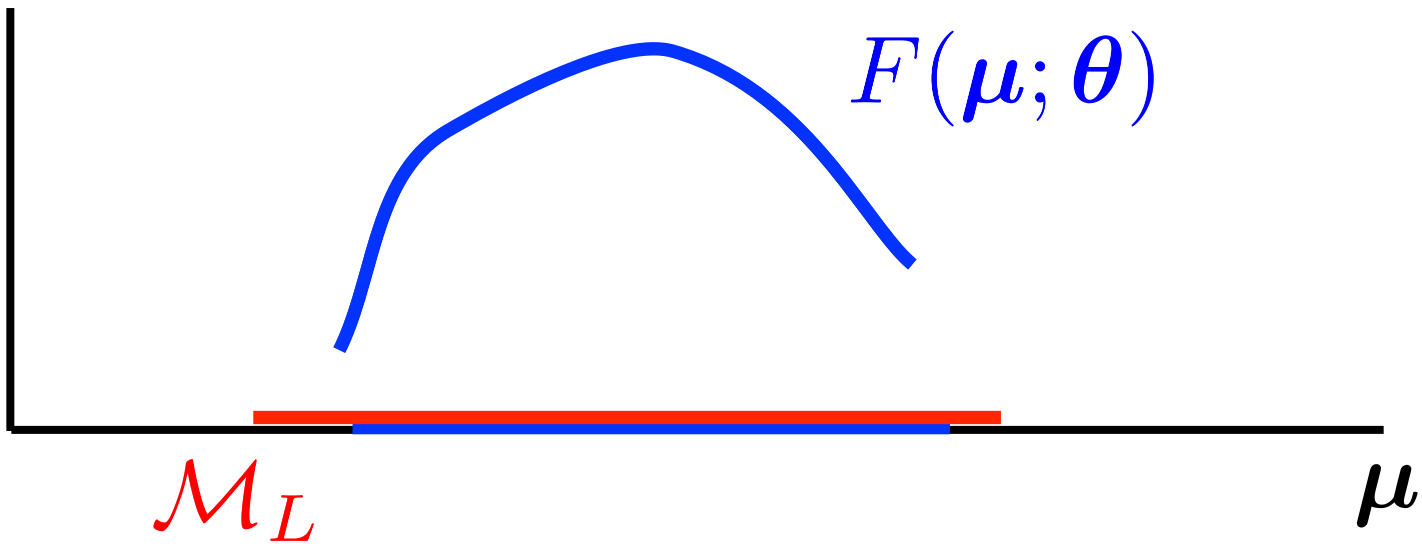
Bethe approximations

- Replace both constraints and objective with approximations.



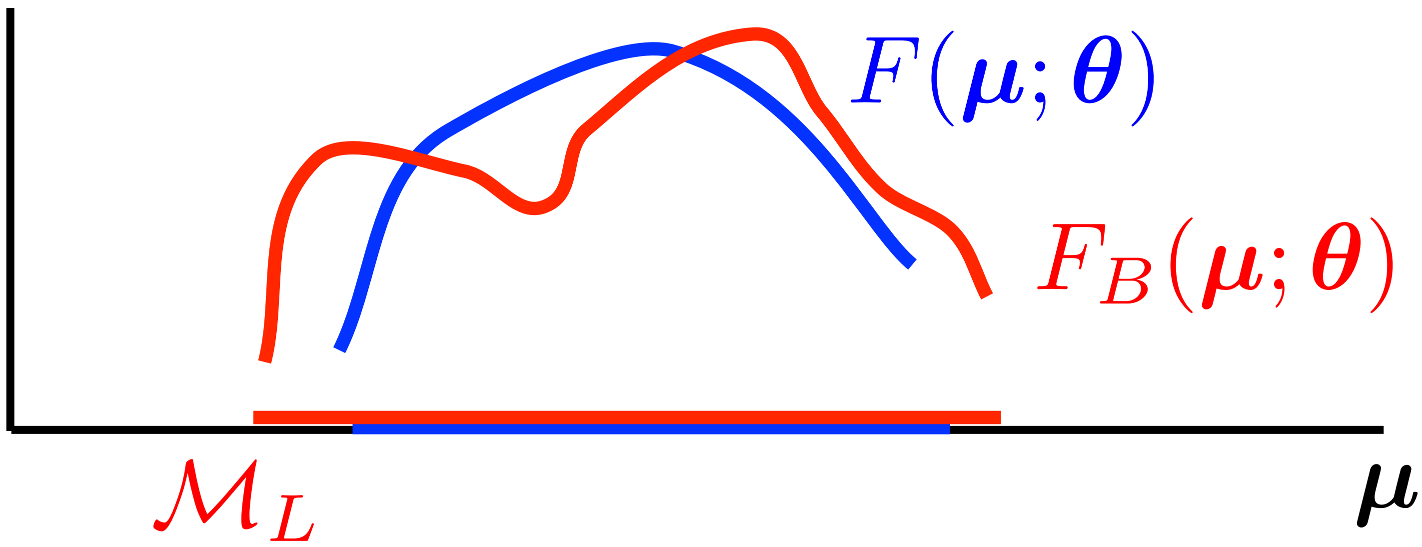
Bethe approximations

- Replace both constraints and objective with approximations.



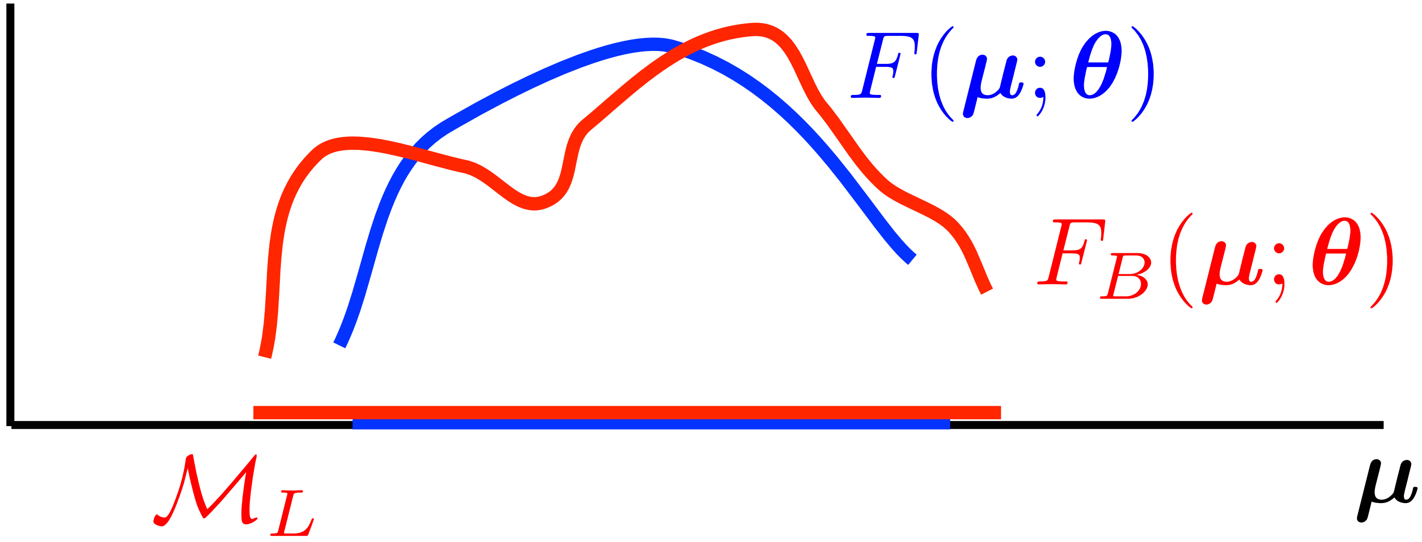
Bethe approximations

- Replace both constraints and objective with approximations.



Bethe approximations

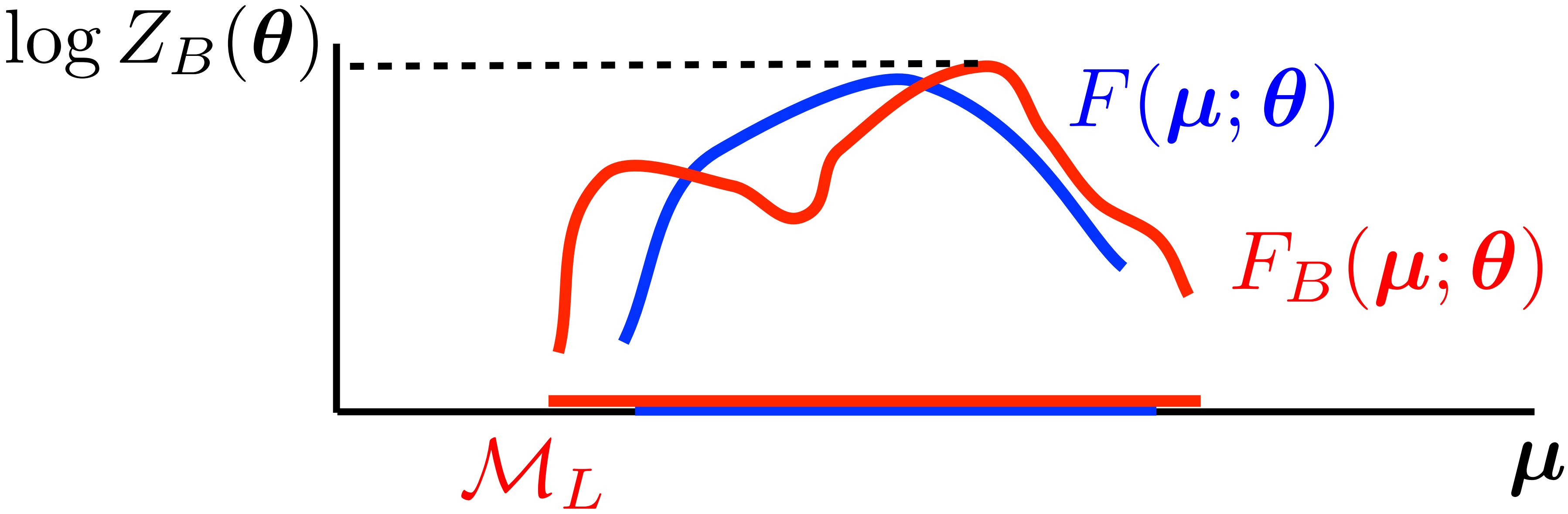
- Replace both constraints and objective with approximations.



$$F_B(\boldsymbol{\mu};\boldsymbol{\theta}) = \boldsymbol{\mu}\cdot\boldsymbol{\theta} + H_B(\boldsymbol{\mu})$$

Bethe approximations

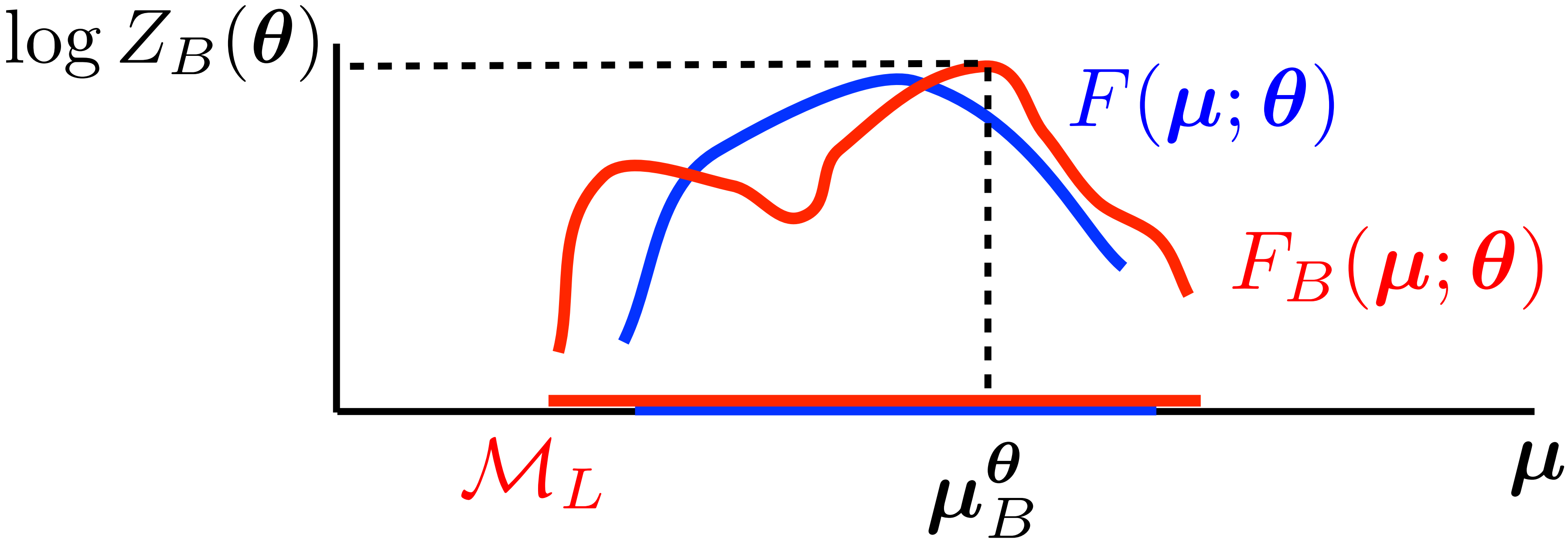
- Replace both constraints and objective with approximations.



$$F_B(\boldsymbol{\mu};\boldsymbol{\theta}) = \boldsymbol{\mu}\cdot\boldsymbol{\theta} + H_B(\boldsymbol{\mu})$$

Bethe approximations

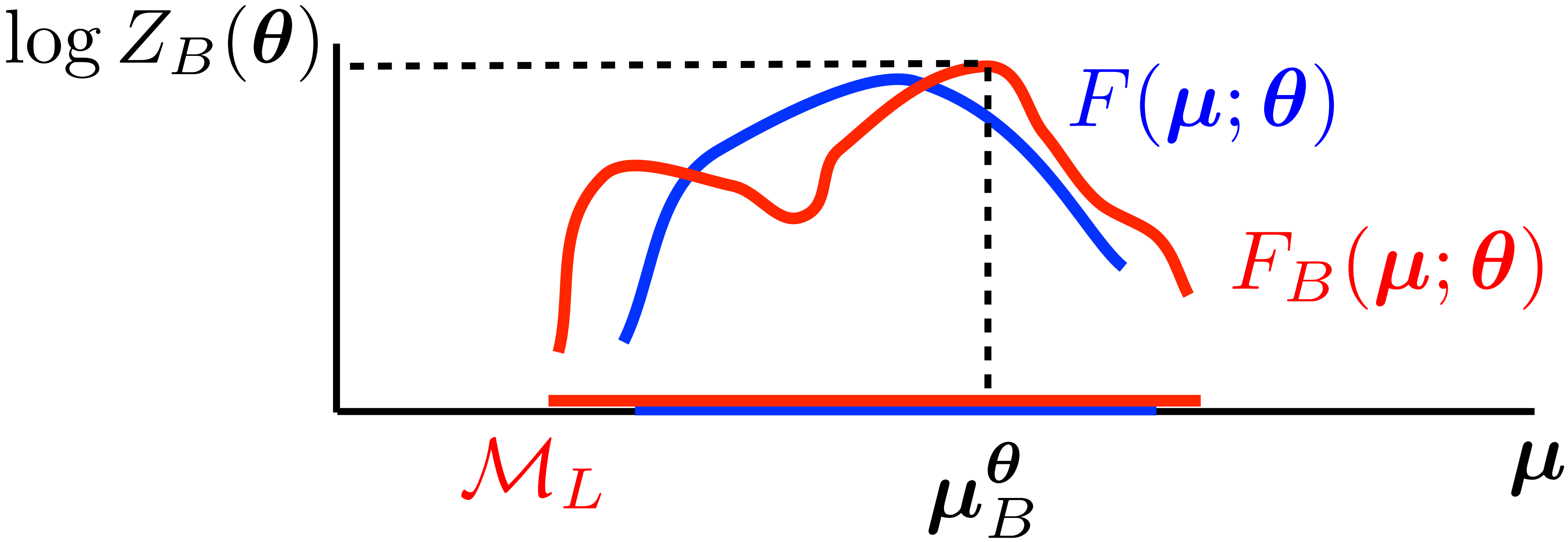
- Replace both constraints and objective with approximations.



$$F_B(\boldsymbol{\mu};\boldsymbol{\theta}) = \boldsymbol{\mu}\cdot\boldsymbol{\theta} + H_B(\boldsymbol{\mu})$$

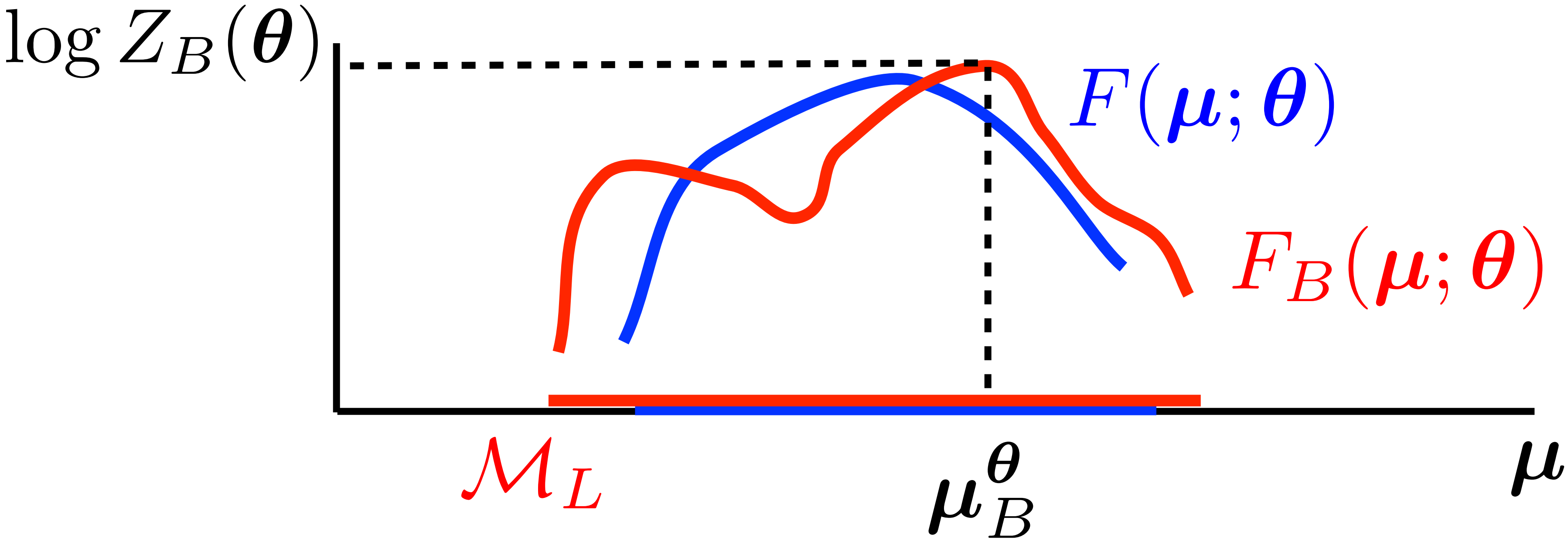
Bethe approximations

- Replace both constraints and objective with approximations.

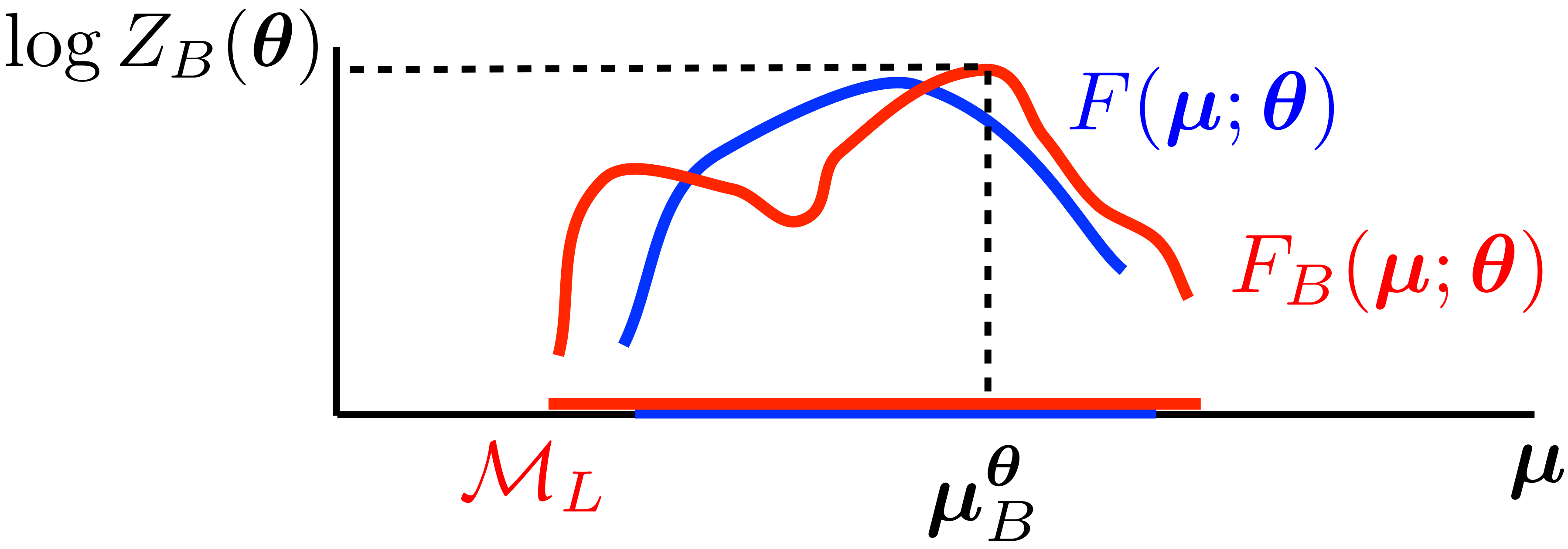


$$F_B(\boldsymbol{\mu};\boldsymbol{\theta}) = \boldsymbol{\mu}\cdot\boldsymbol{\theta} + H_B(\boldsymbol{\mu})$$

Bethe approximations



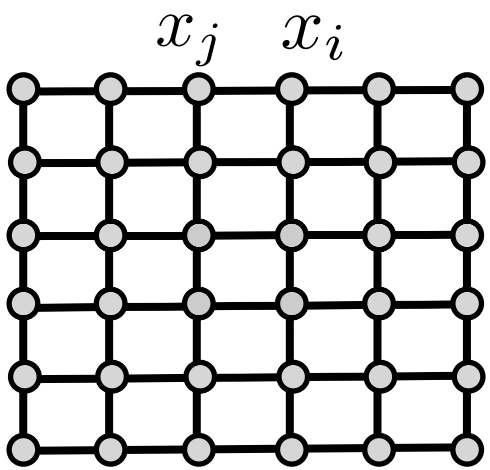
Bethe approximations



- Local maxima of the approximation correspond to stable fixed points of loopy belief propagation (Yedidia Freeman and Weiss, Heskes).

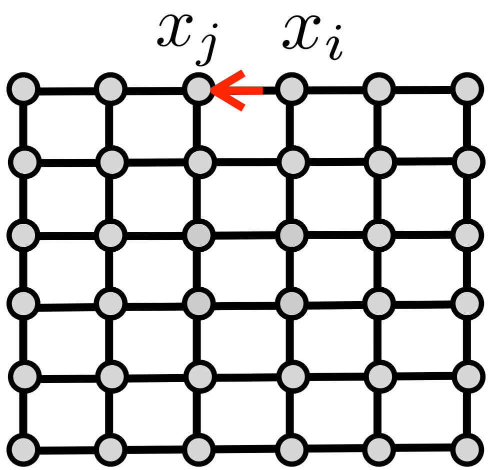
Loopy BP

- Protocol for passing messages along edges of the graph.



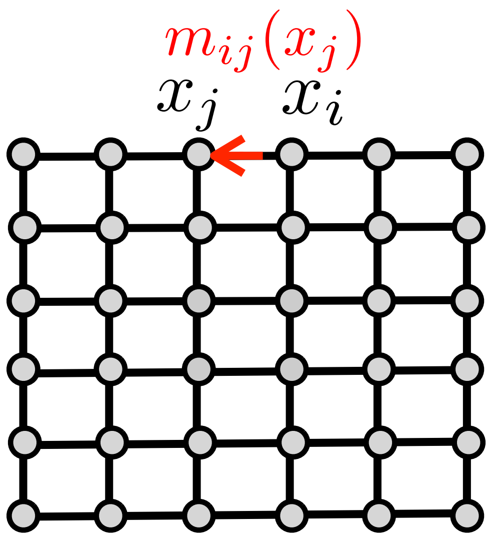
Loopy BP

- Protocol for passing messages along edges of the graph.



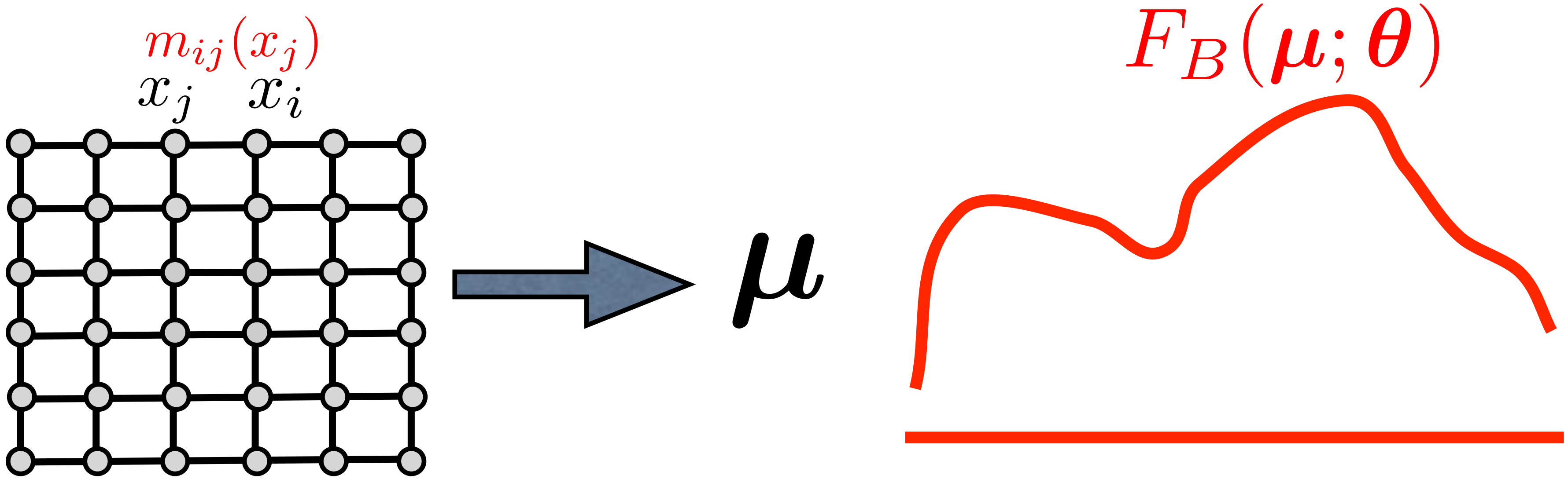
Loopy BP

- Protocol for passing messages along edges of the graph.



Loopy BP

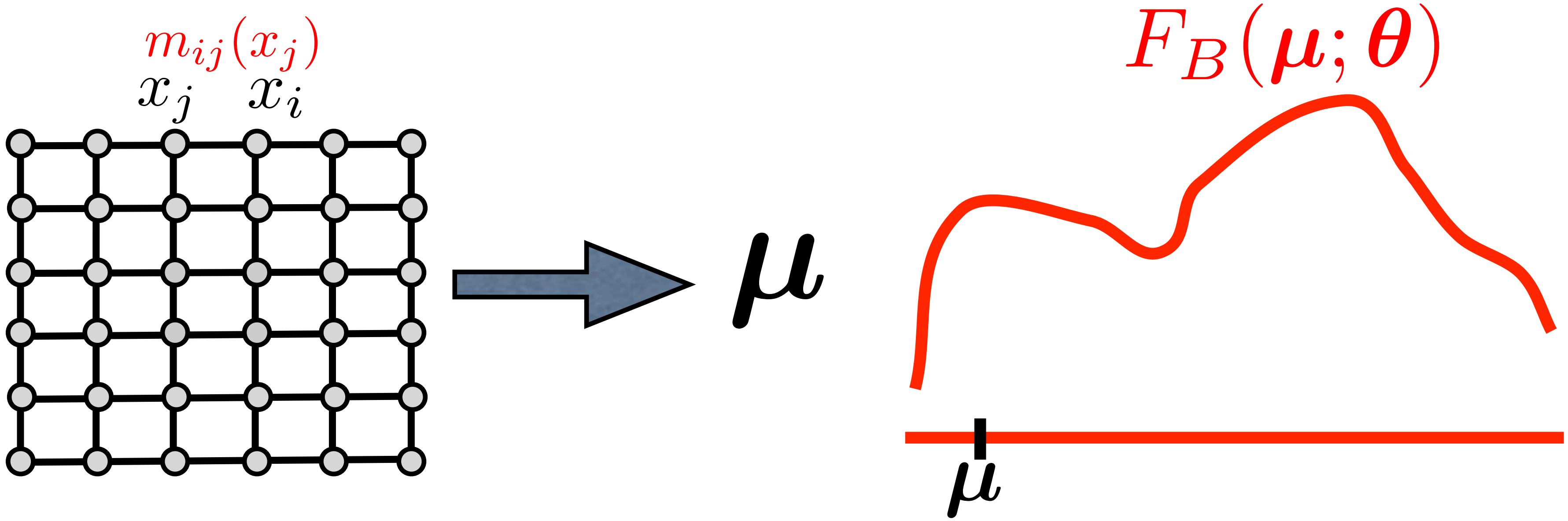
- Protocol for passing messages along edges of the graph.



- Returns marginals that are stationary points of the function $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$ (typically maxima).

Loopy BP

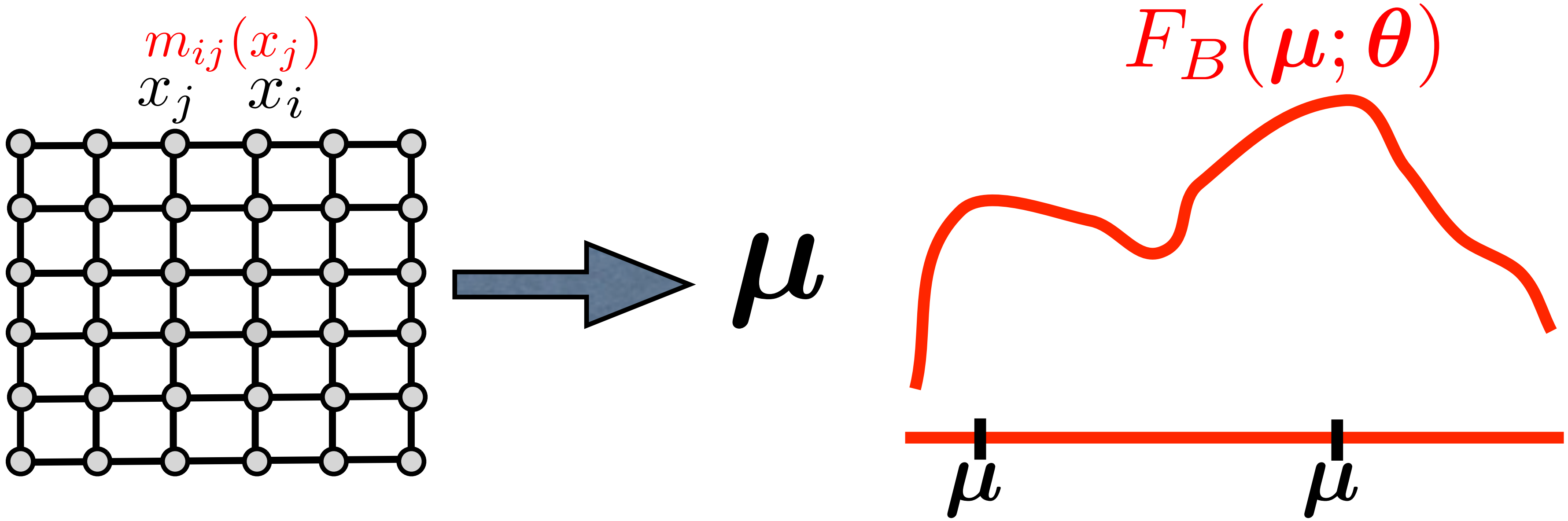
- Protocol for passing messages along edges of the graph.



- Returns marginals that are stationary points of the function $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$ (typically maxima).

Loopy BP

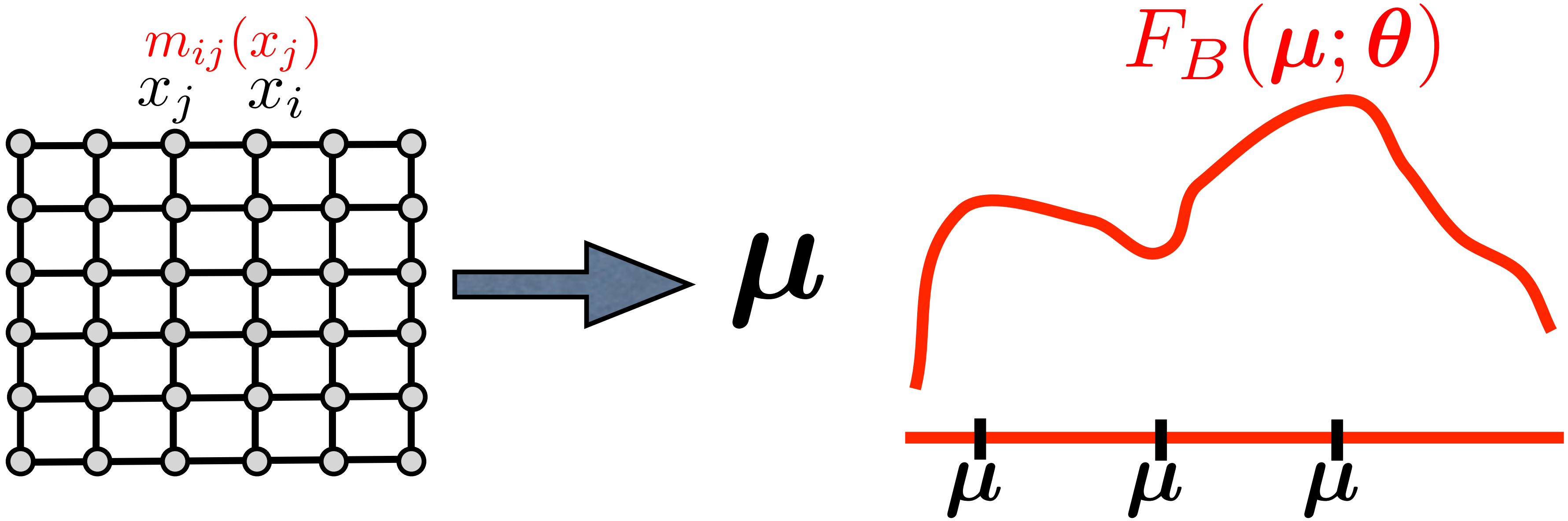
- Protocol for passing messages along edges of the graph.



- Returns marginals that are stationary points of the function $F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$ (typically maxima).

Loopy BP

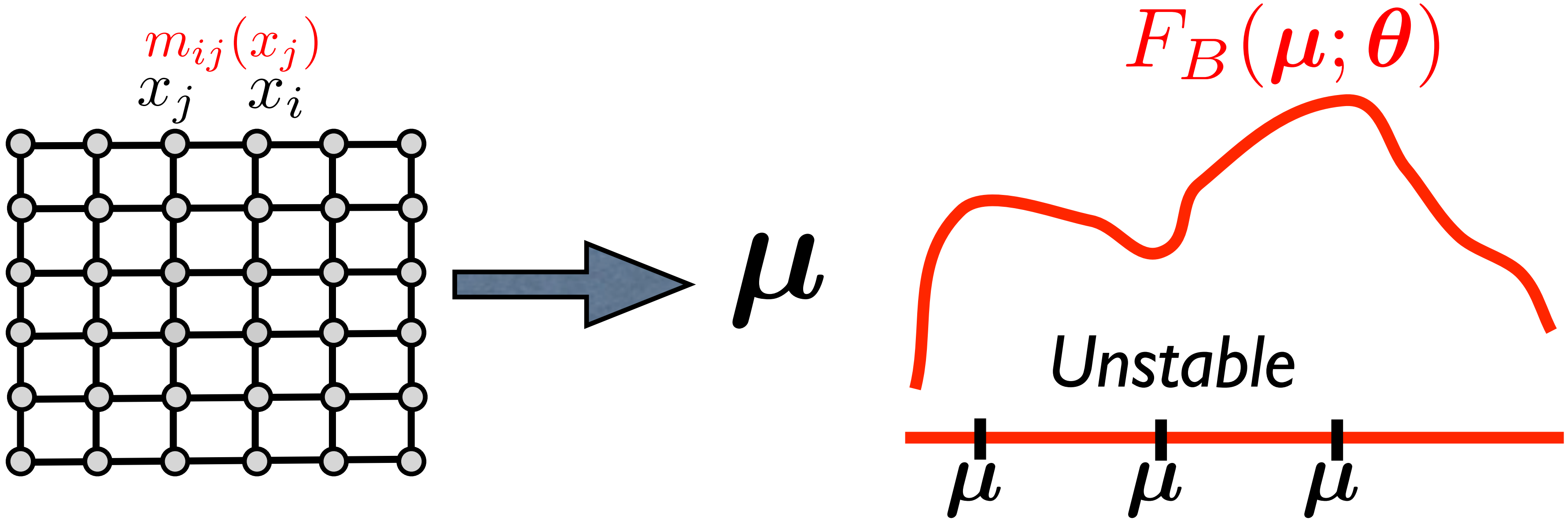
- Protocol for passing messages along edges of the graph.



- Returns marginals that are stationary points of the function $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$ (typically maxima).

Loopy BP

- Protocol for passing messages along edges of the graph.



- Returns marginals that are stationary points of the function $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$ (typically maxima).

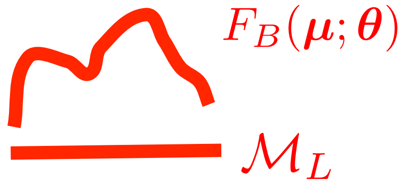
Loopy BP

- Typically an effective approximation of the partition function and marginals.
- Exact for tree graphs.
- Works well in many cases.
- Caveat: can return local optima so hard to analyze. Assume for now we can find the global maximum.
- Lets use it in learning...

Bethe ML

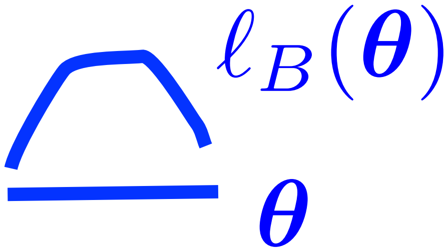
- Recall the likelihood: $\ell(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}} \cdot \boldsymbol{\theta} - \log Z(\boldsymbol{\theta})$
- Approximate: $Z(\boldsymbol{\theta}) \approx Z_B(\boldsymbol{\theta})$

$$\log Z_B(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu} \in \mathcal{M}_L} F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$$



- Maximize the Bethe likelihood:

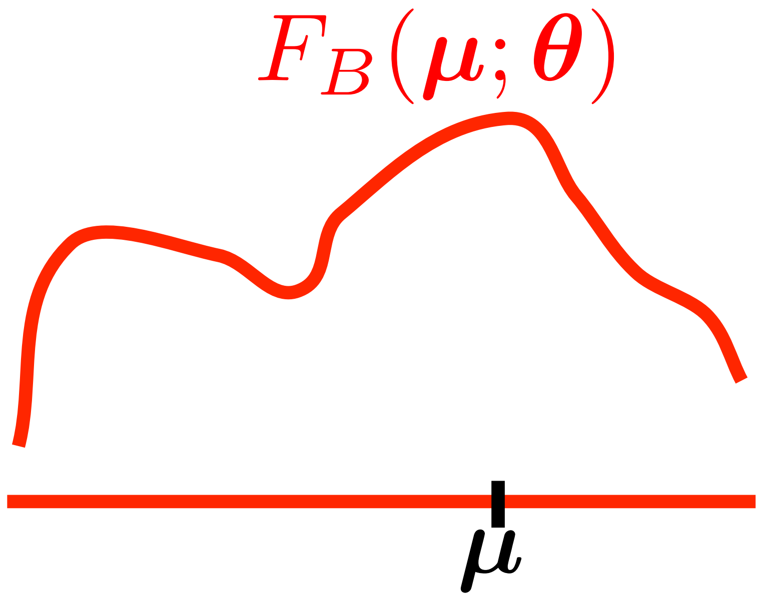
$$\ell_B(\boldsymbol{\theta}) = \bar{\boldsymbol{\mu}} \cdot \boldsymbol{\theta} - \max_{\boldsymbol{\mu} \in \mathcal{M}_L} \left[\boldsymbol{\mu} \cdot \boldsymbol{\theta} + H_B(\boldsymbol{\mu})\right]$$



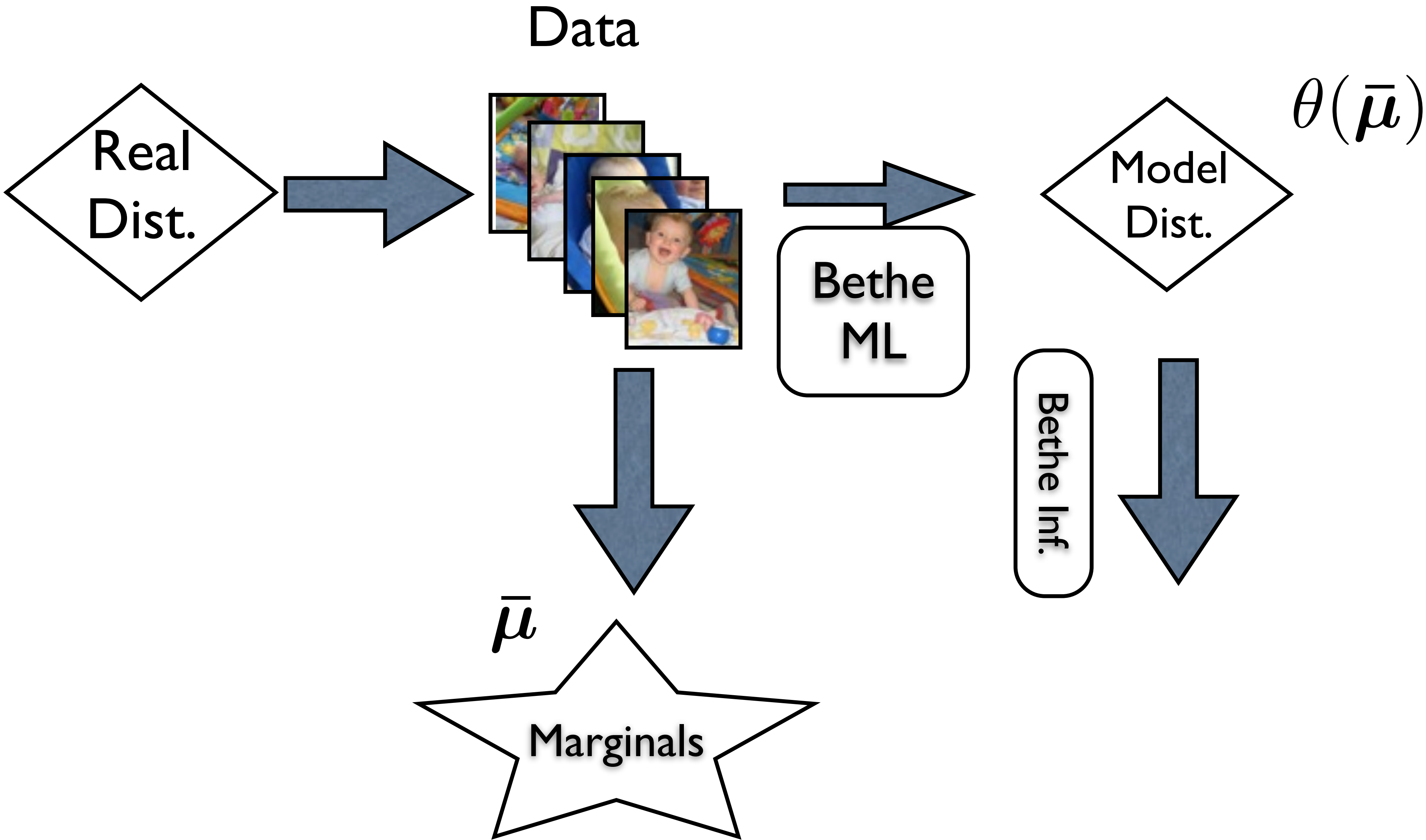
- A concave function of $\boldsymbol{\theta}$!

Bethe Inference

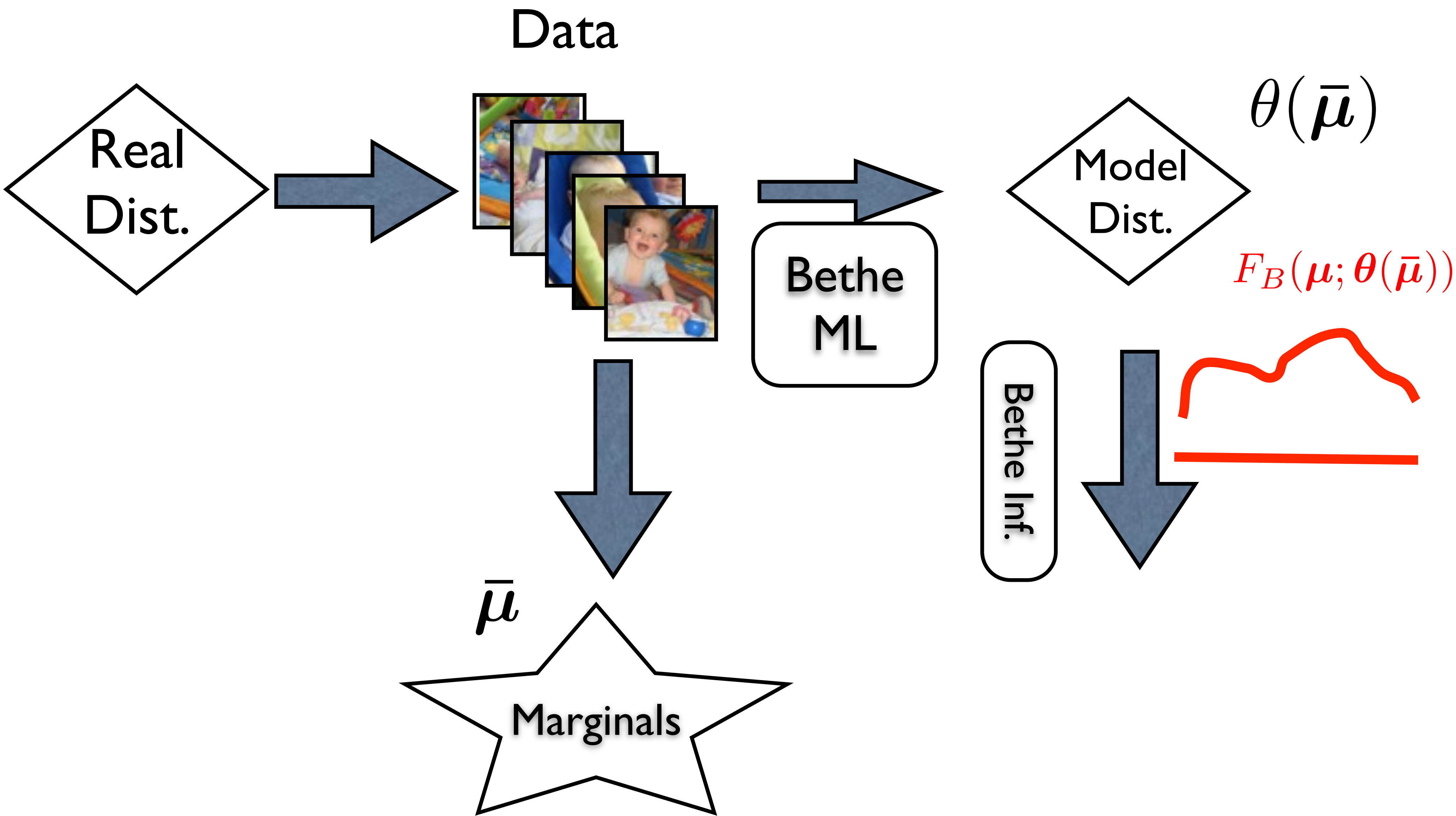
- Given a parameter vector $\boldsymbol{\theta}$, take its marginal to be the maximum of $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$.
- Assume there are no issues with local optima.
- We will see that the serious problem is of non-unique maximizers.



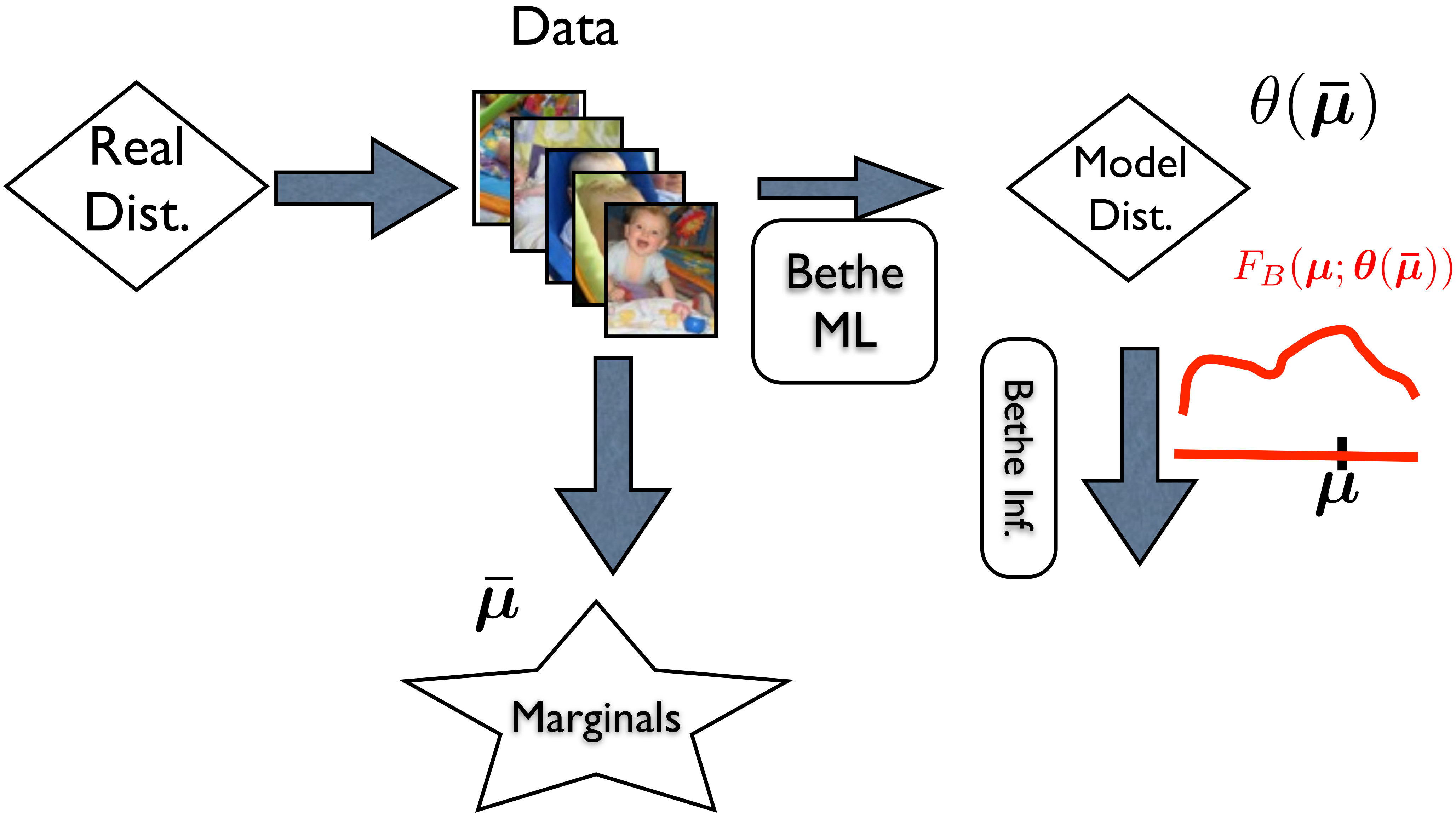
Approximate Learning



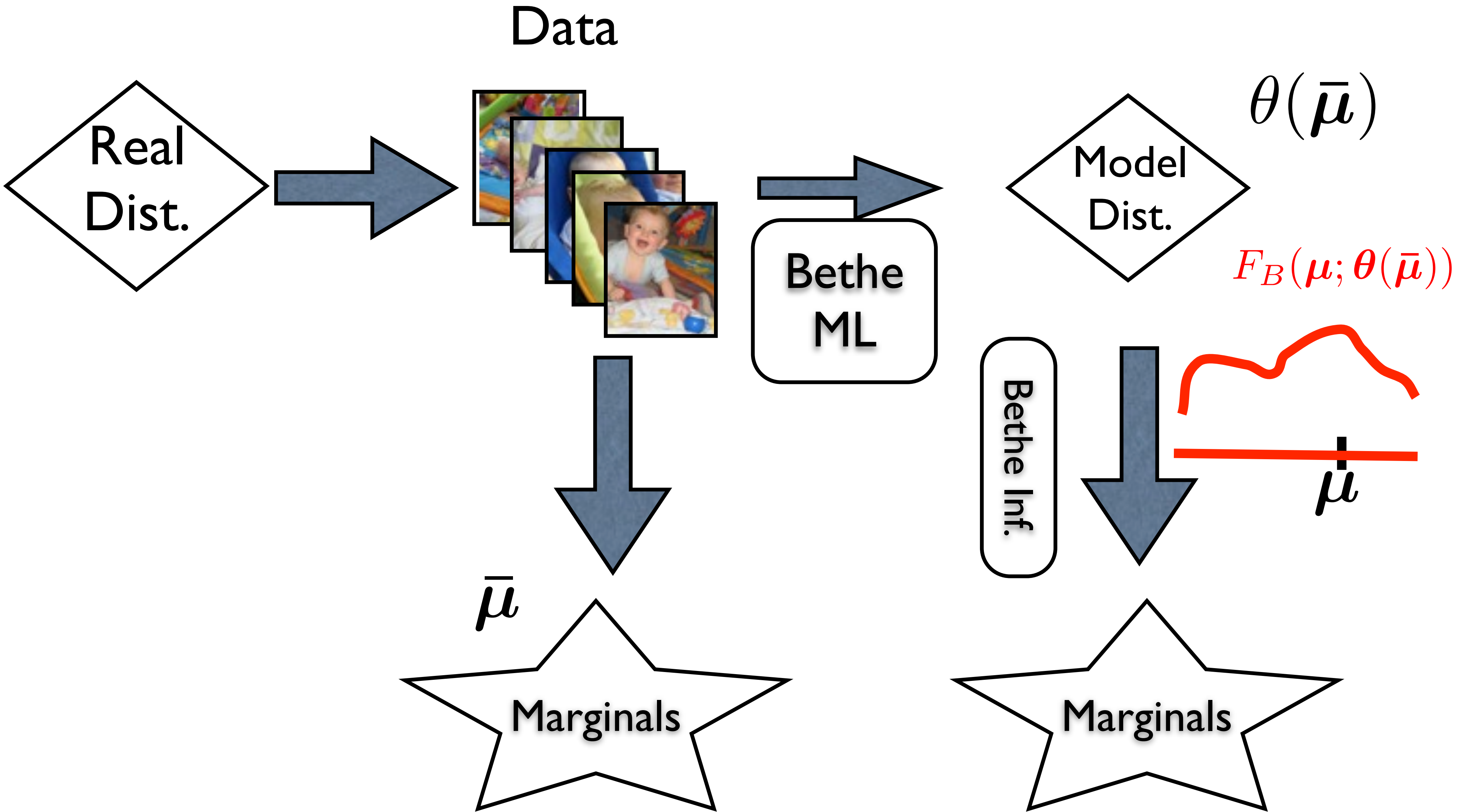
Approximate Learning



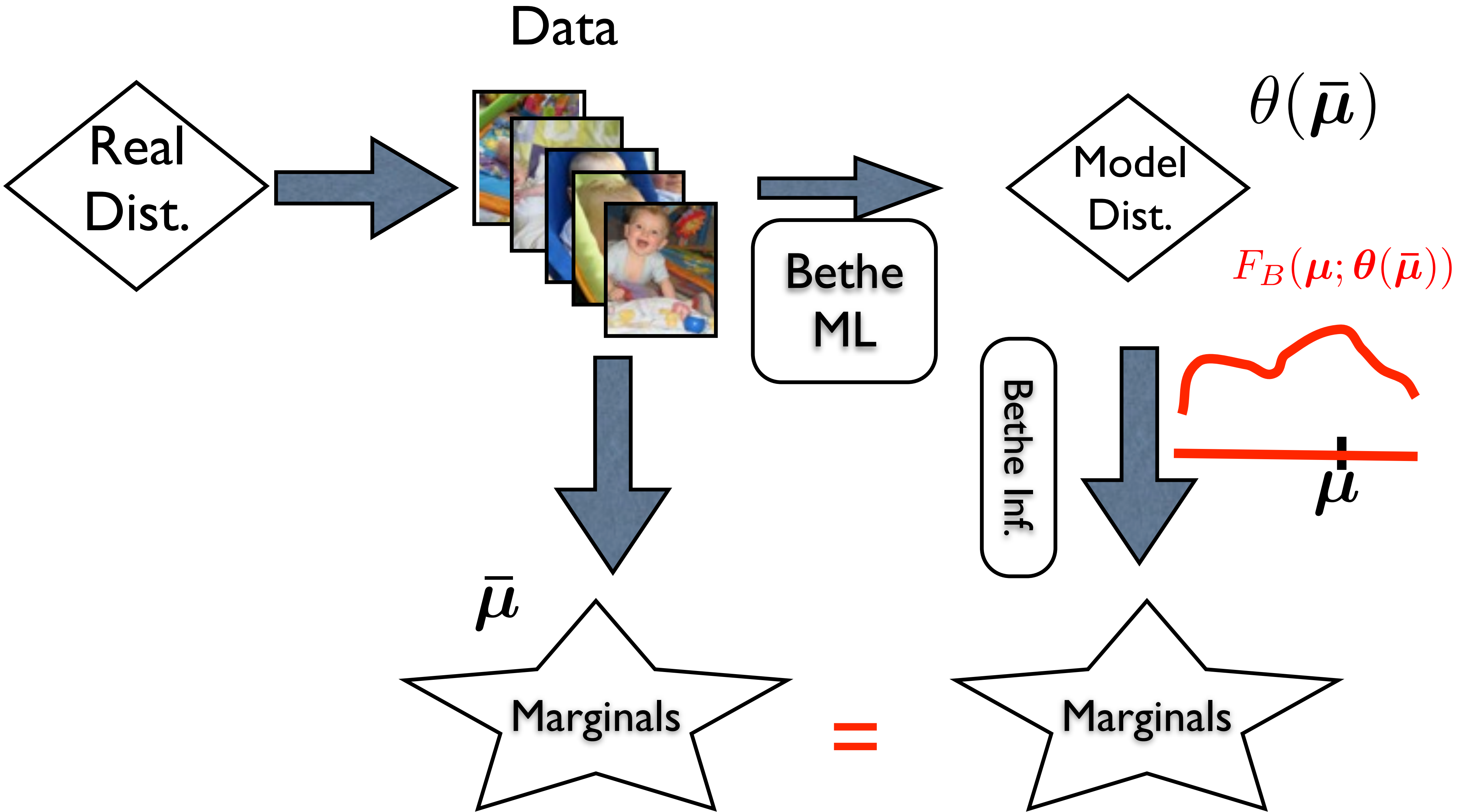
Approximate Learning



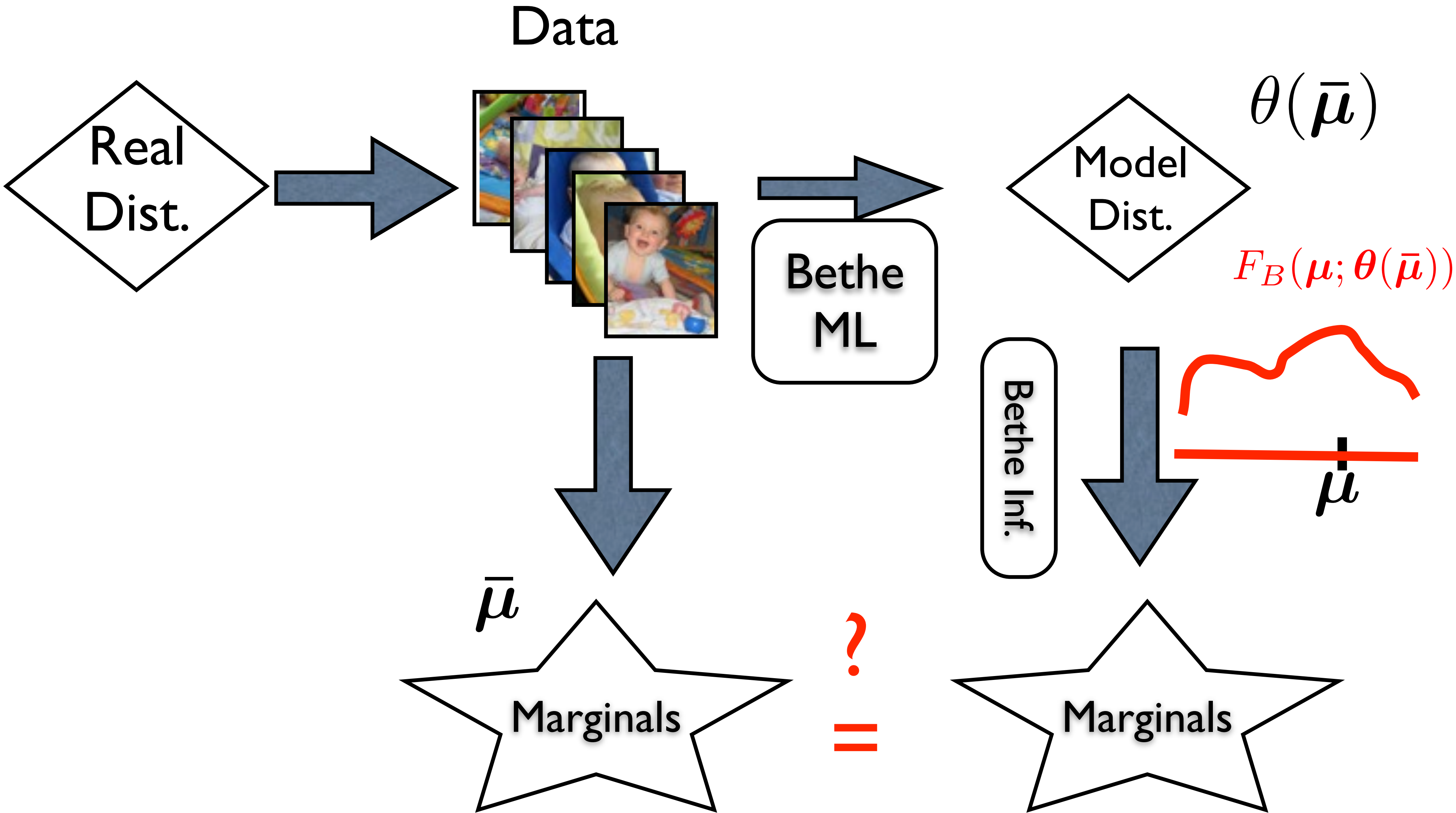
Approximate Learning



Approximate Learning



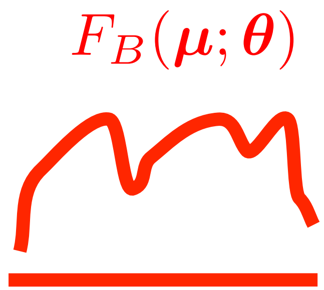
Approximate Learning



Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

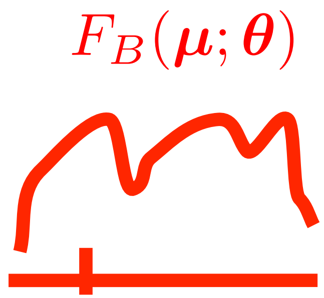
$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

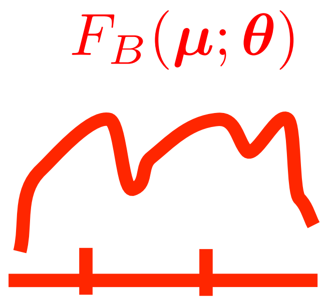
$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$

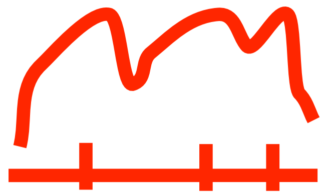


Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$

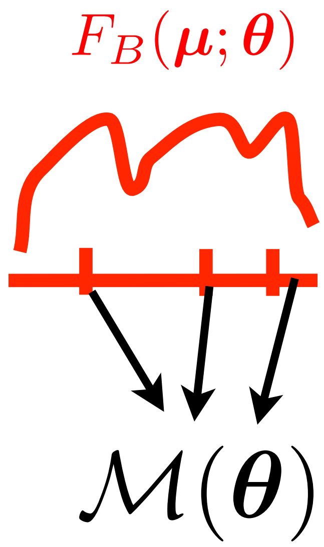
$F_B(\boldsymbol{\mu};\boldsymbol{\theta})$



Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

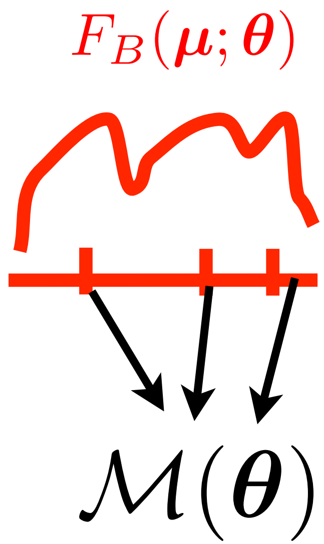
$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$

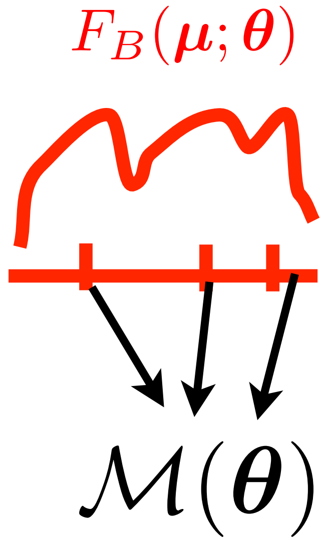


- θ maximizes Bethe likelihood if:

Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



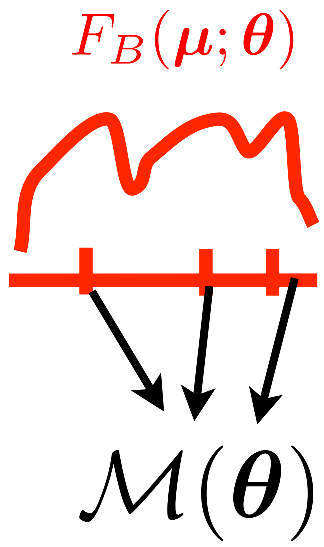
- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$

Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



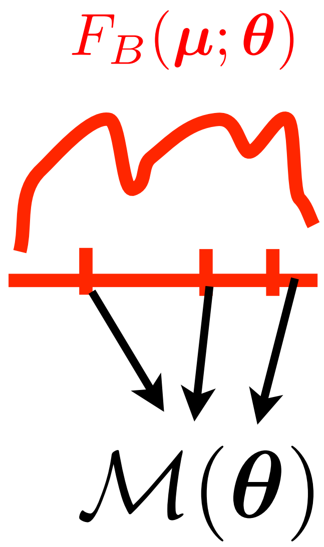
- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$

Optimality in Bethe ML

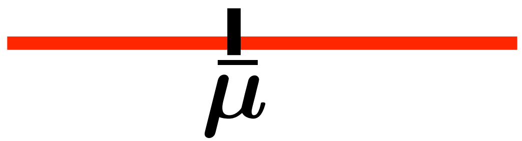
- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



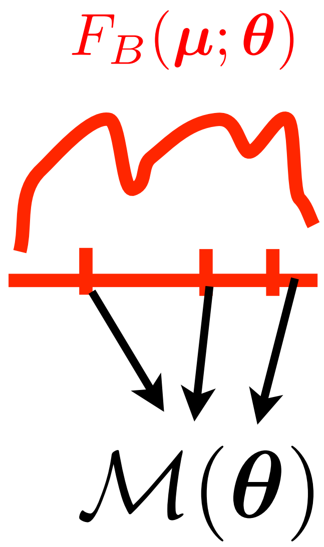
- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$



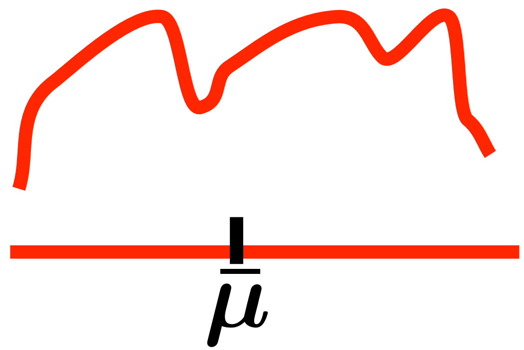
Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:



$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$

- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

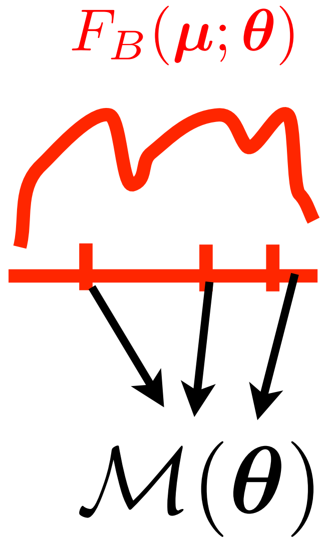


$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$

Optimality in Bethe ML

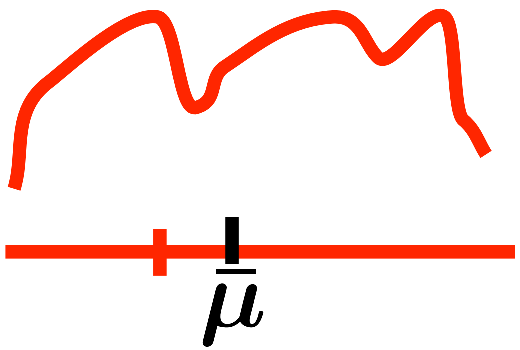
- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$$



- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

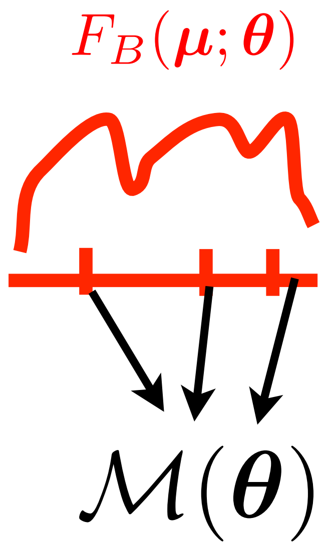
$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$



Optimality in Bethe ML

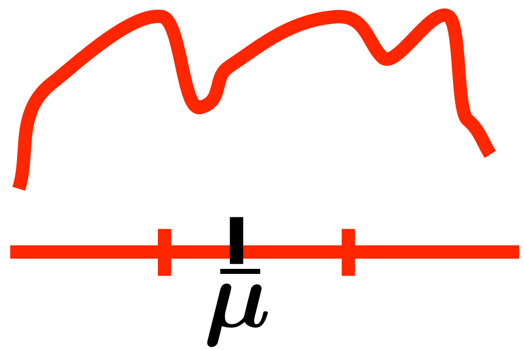
- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$$



- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

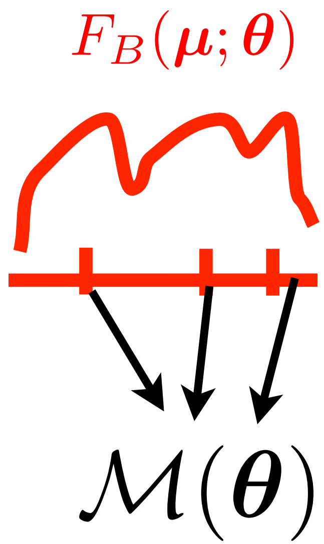
$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$



Optimality in Bethe ML

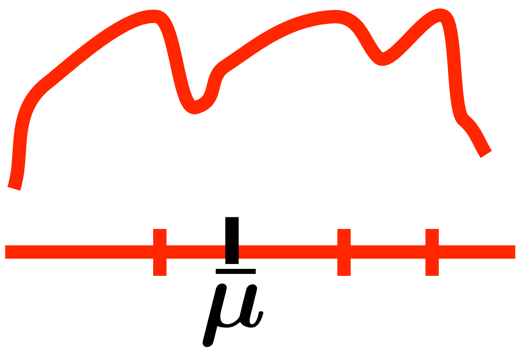
- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

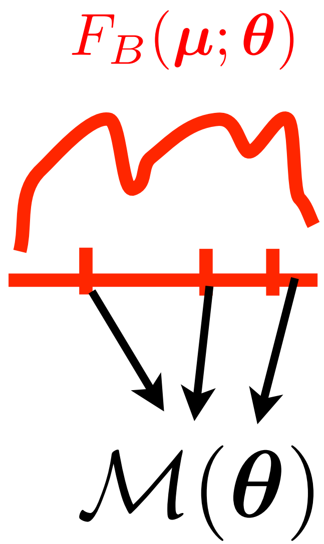
$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$



Optimality in Bethe ML

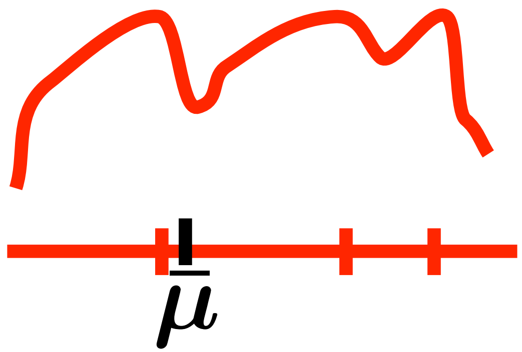
- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$

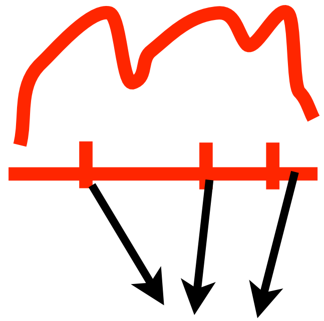


Optimality in Bethe ML

- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$

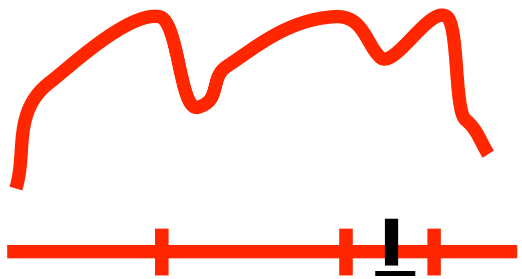
$F_B(\boldsymbol{\mu};\boldsymbol{\theta})$



$\mathcal{M}(\boldsymbol{\theta})$

- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$

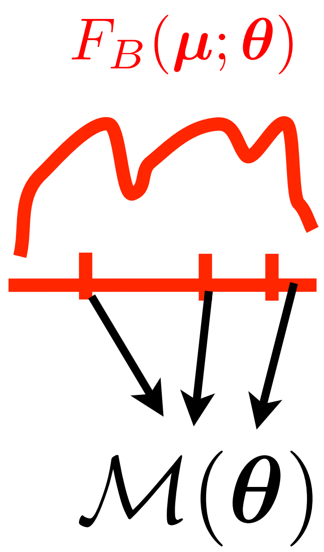


$\bar{\boldsymbol{\mu}}$

Optimality in Bethe ML

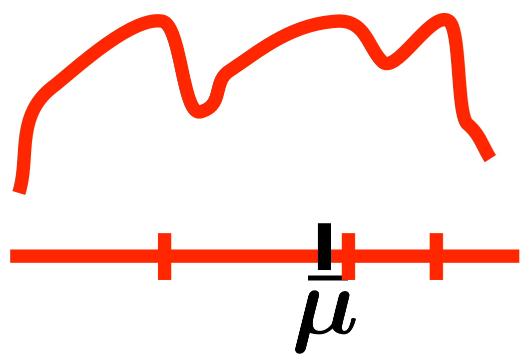
- Given parameter $\boldsymbol{\theta}$ define:

$$\mathcal{M}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$



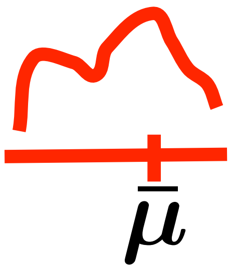
- $\boldsymbol{\theta}$ maximizes Bethe likelihood if:

$$\bar{\boldsymbol{\mu}} \in \mathrm{Conv}\left\{\mathcal{M}(\boldsymbol{\theta})\right\}$$



Optimality in Bethe ML

- If there is a $\boldsymbol{\theta}$ with *single maximizer* such that:

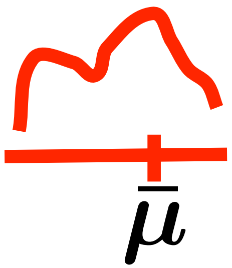


$$\bar{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$$

- This will be a maximum Bethe likelihood optimum.
- The marginals are recoverable from the parameter.

Optimality in Bethe ML

- If there is a $\boldsymbol{\theta}$ with *single maximizer* such that:



$$\bar{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu};\boldsymbol{\theta})$$

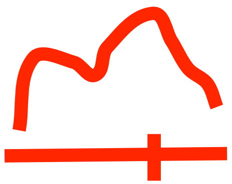
- This will be a maximum Bethe likelihood optimum.

- The marginals are recoverable from the parameter.

Moment Matching!

Optimality in Bethe ML

- If there is a $\boldsymbol{\theta}$ with *single maximizer* such that:



$\bar{\boldsymbol{\mu}}$

$$\bar{\boldsymbol{\mu}} = \arg\max_{\boldsymbol{\mu}} F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$$

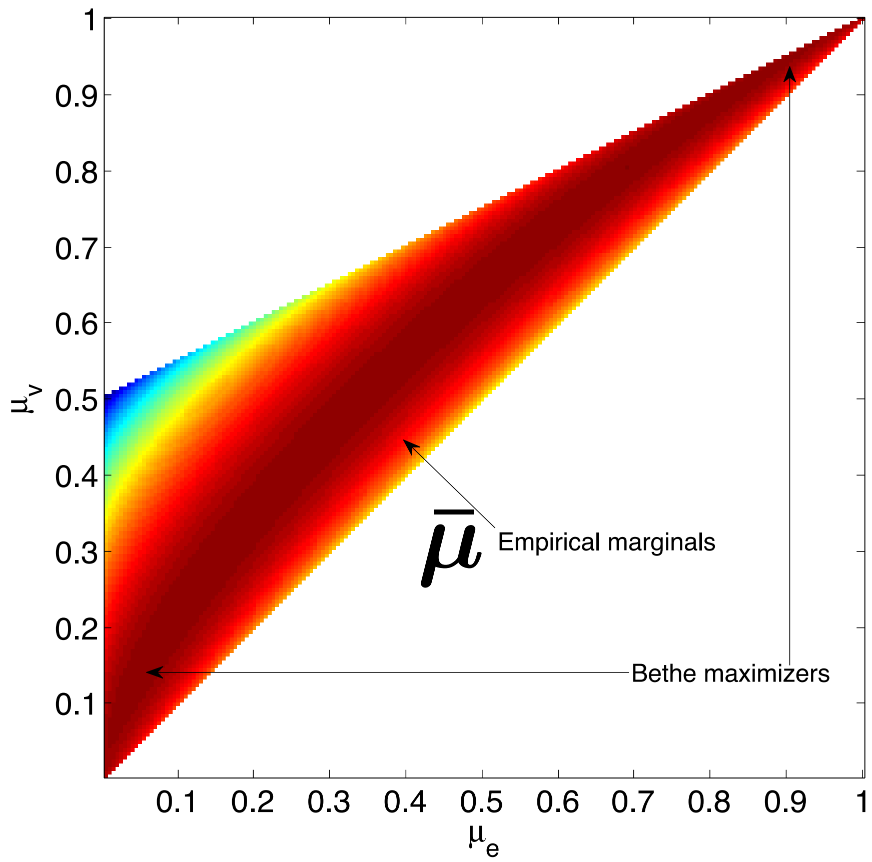
- This will be a maximum Bethe likelihood optimum.
- The marginals are recoverable from the parameter.

Moment Matching!

- What if there is no such parameter?

A two maxima case

- Here is $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$ for a 2D case



- $\bar{\boldsymbol{\mu}}$ is not a maximizer, but at a convex hull of maximizers.
- It cannot be recovered from $\boldsymbol{\theta}(\bar{\boldsymbol{\mu}})$
- Non moment matching...

Bethe Learnable Marginals

- Definition: A marginal $\bar{\mu}$ is Bethe learnable if learning with Bethe achieves moment matching.

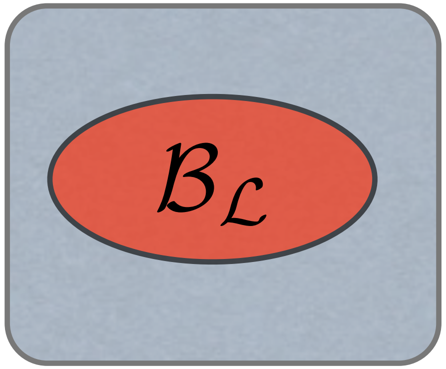
Bethe Learnable Marginals

- Definition: A marginal $\bar{\mu}$ is Bethe learnable if learning with Bethe achieves moment matching.

$\bar{\mu}$ space

Bethe Learnable Marginals

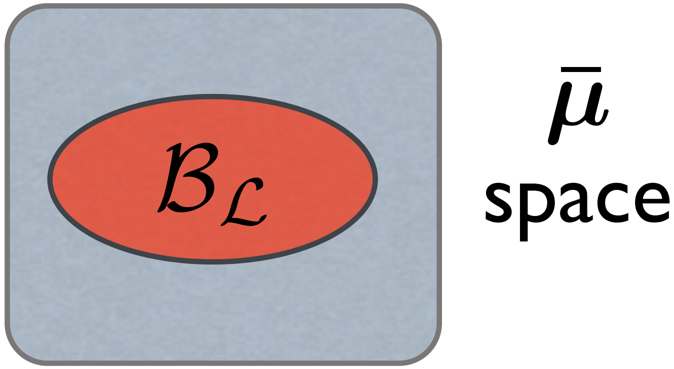
- Definition: A marginal $\bar{\mu}$ is Bethe learnable if learning with Bethe achieves moment matching.



$\bar{\mu}$
space

Bethe Learnable Marginals

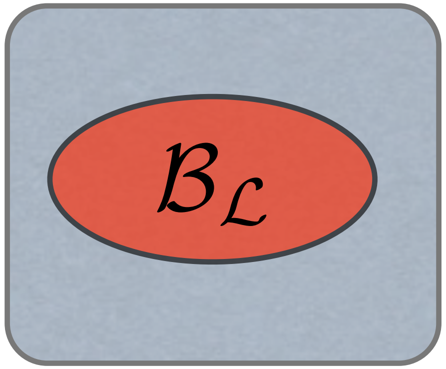
- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.



$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$

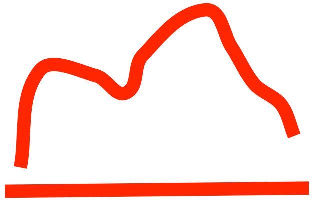
Bethe Learnable Marginals

- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.



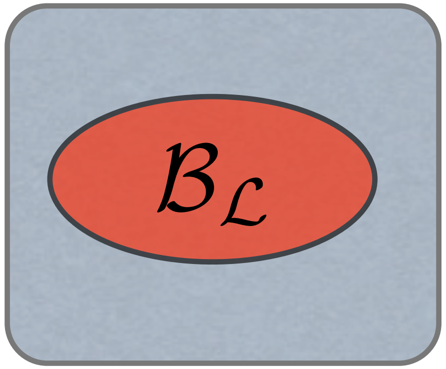
$\bar{\boldsymbol{\mu}}$ space

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$



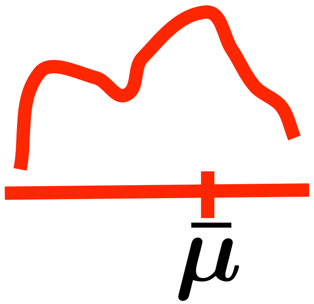
Bethe Learnable Marginals

- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.



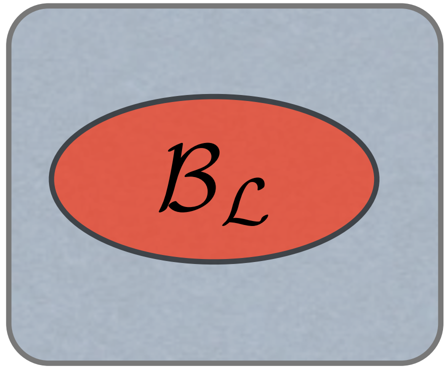
$\bar{\boldsymbol{\mu}}$ space

$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$



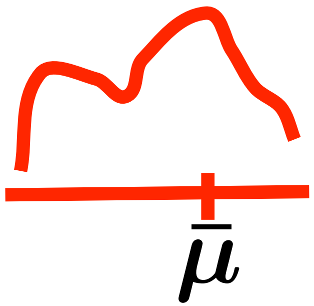
Bethe Learnable Marginals

- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.



$\bar{\boldsymbol{\mu}}$ space

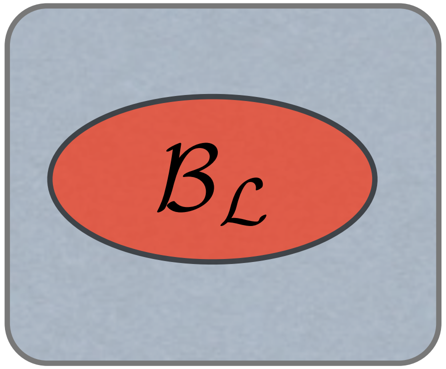
$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$



Learnable!

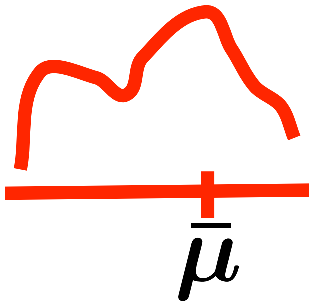
Bethe Learnable Marginals

- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.

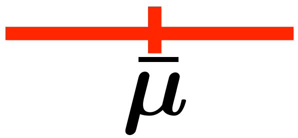


$\bar{\boldsymbol{\mu}}$ space

$$F_B(\boldsymbol{\mu};\boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$

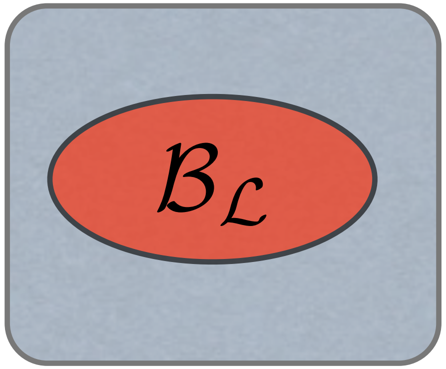


Learnable!



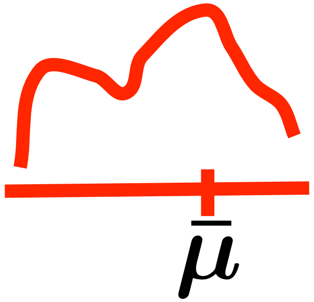
Bethe Learnable Marginals

- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.

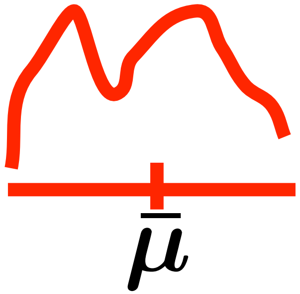


$\bar{\boldsymbol{\mu}}$
space

$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$

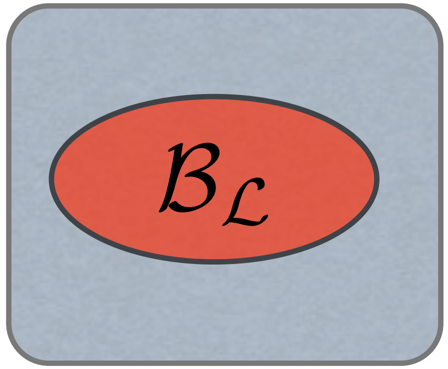


Learnable!



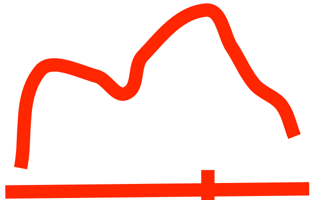
Bethe Learnable Marginals

- Definition: A marginal $\bar{\boldsymbol{\mu}}$ is Bethe learnable if learning with Bethe achieves moment matching.



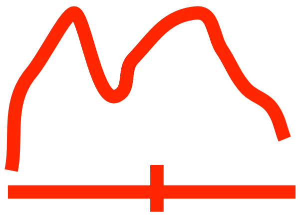
$\bar{\boldsymbol{\mu}}$ space

$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$



$\bar{\boldsymbol{\mu}}$

Learnable!



$\bar{\boldsymbol{\mu}}$

Unlearnable!

Bethe Learnable Marginals

- How do we characterize those?
- To check if $\bar{\boldsymbol{\mu}}$ is learnable:
  - Do Bethe ML. i.e., find $\theta(\bar{\boldsymbol{\mu}})$
  - Check if $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$ has a single maximum.
- We want something simpler.

Canonical Parameters

- When the graph is a tree, Bethe is exact, and the following are the Bethe ML parameters:

$$\theta_i^c(x_i; \bar{\mu}) = \log \bar{\mu}_i(x_i)$$

$$\theta_{ij}^c(x_i, x_j; \bar{\mu}) = \log \frac{\bar{\mu}_{ij}(x_i, x_j)}{\bar{\mu}_i(x_i)\bar{\mu}_j(x_j)}$$

- Generally $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

Canonical Parameters

- When the graph is a tree, Bethe is exact, and the following are the Bethe ML parameters:

$$\theta_i^c(x_i; \bar{\mu}) = \log \bar{\mu}_i(x_i)$$

$$\theta_{ij}^c(x_i, x_j; \bar{\mu}) = \log \frac{\bar{\mu}_{ij}(x_i, x_j)}{\bar{\mu}_i(x_i)\bar{\mu}_j(x_j)}$$

- Generally $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

Canonical Parameters

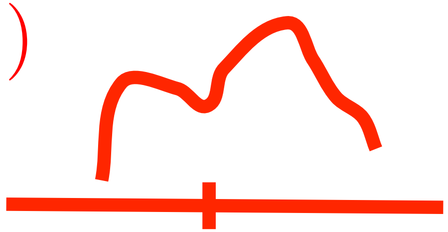
- When the graph is a tree, Bethe is exact, and the following are the Bethe ML parameters:

$$\theta_i^c(x_i; \bar{\mu}) = \log \bar{\mu}_i(x_i)$$

$$\theta_{ij}^c(x_i, x_j; \bar{\mu}) = \log \frac{\bar{\mu}_{ij}(x_i, x_j)}{\bar{\mu}_i(x_i)\bar{\mu}_j(x_j)}$$

- Generally $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



$\bar{\boldsymbol{\mu}}$

Canonical Parameters

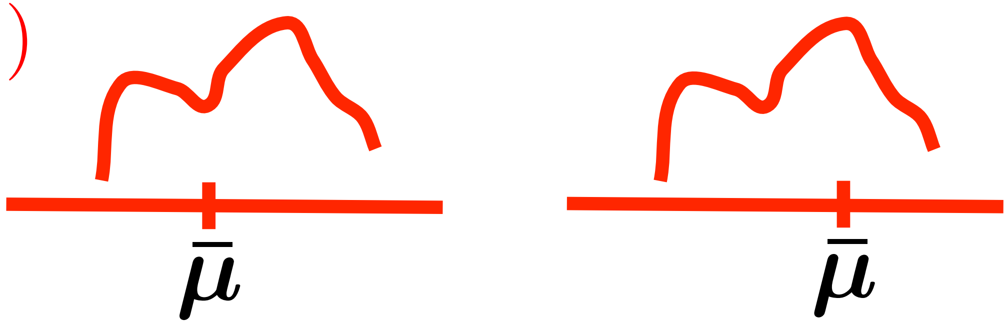
- When the graph is a tree, Bethe is exact, and the following are the Bethe ML parameters:

$$\theta_i^c(x_i;\bar{\mu}) = \log \bar{\mu}_i(x_i)$$

$$\theta_{ij}^c(x_i, x_j;\bar{\mu}) = \log \frac{\bar{\mu}_{ij}(x_i, x_j)}{\bar{\mu}_i(x_i)\bar{\mu}_j(x_j)}$$

- Generally $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

$F_B(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



Canonical Parameters

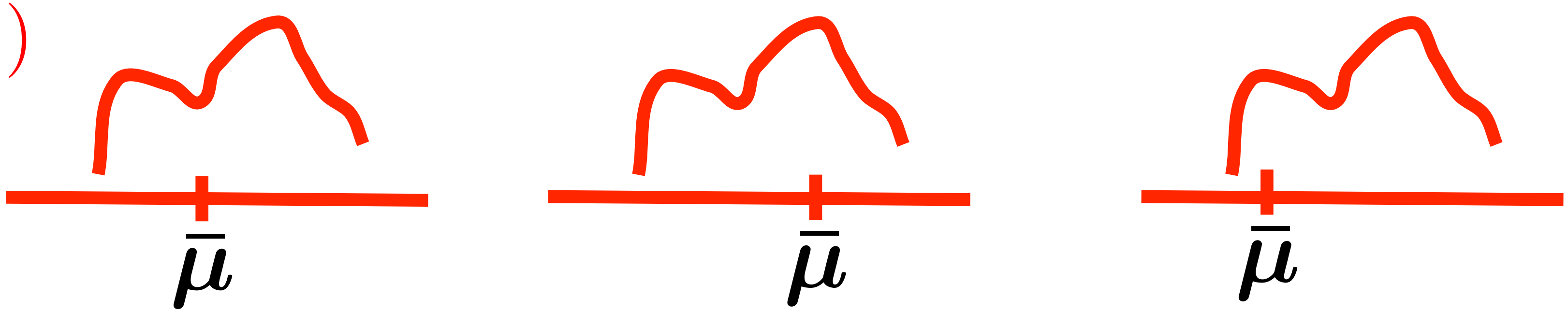
- When the graph is a tree, Bethe is exact, and the following are the Bethe ML parameters:

$$\theta_i^c(x_i;\bar{\mu}) = \log \bar{\mu}_i(x_i)$$

$$\theta_{ij}^c(x_i,x_j;\bar{\mu}) = \log \frac{\bar{\mu}_{ij}(x_i,x_j)}{\bar{\mu}_i(x_i)\bar{\mu}_j(x_j)}$$

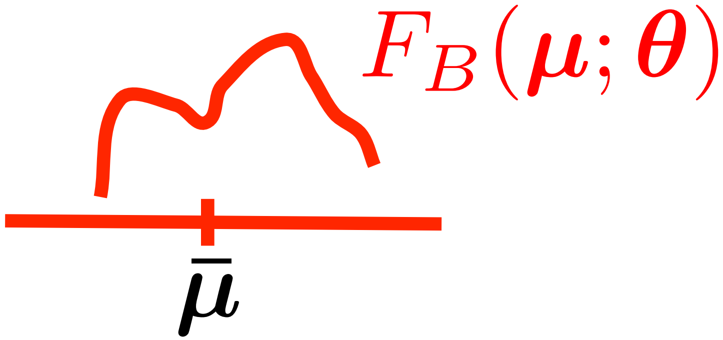
- Generally $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

$F_B(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



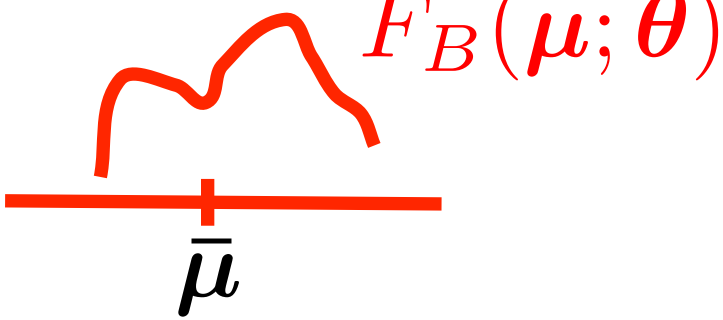
Stationary point invariance

- Say we have a non-canonical $\boldsymbol{\theta}$ s.t. $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$



Stationary point invariance

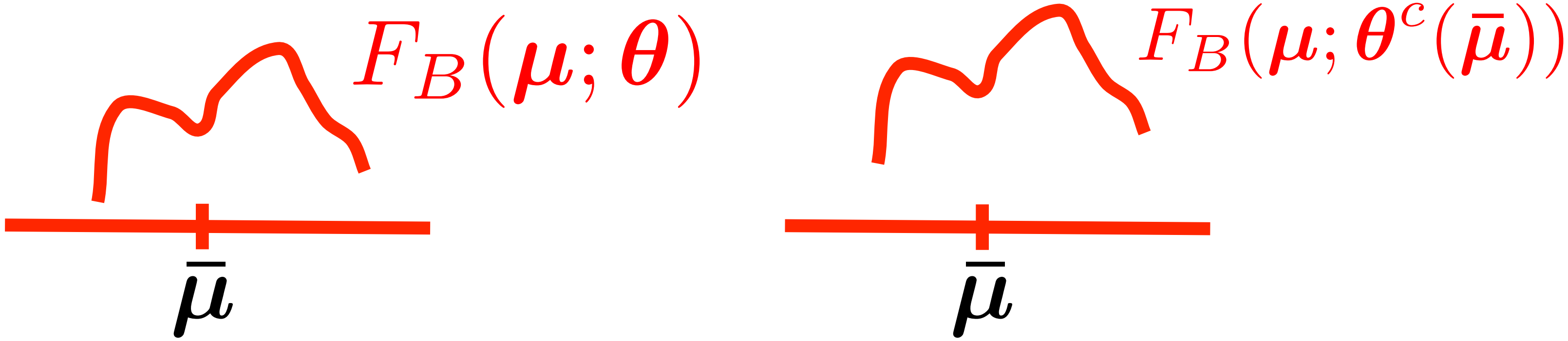
- Say we have a non-canonical $\boldsymbol{\theta}$ s.t. $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$



- The function for the canonical parameter will be the same up to a constant.

Stationary point invariance

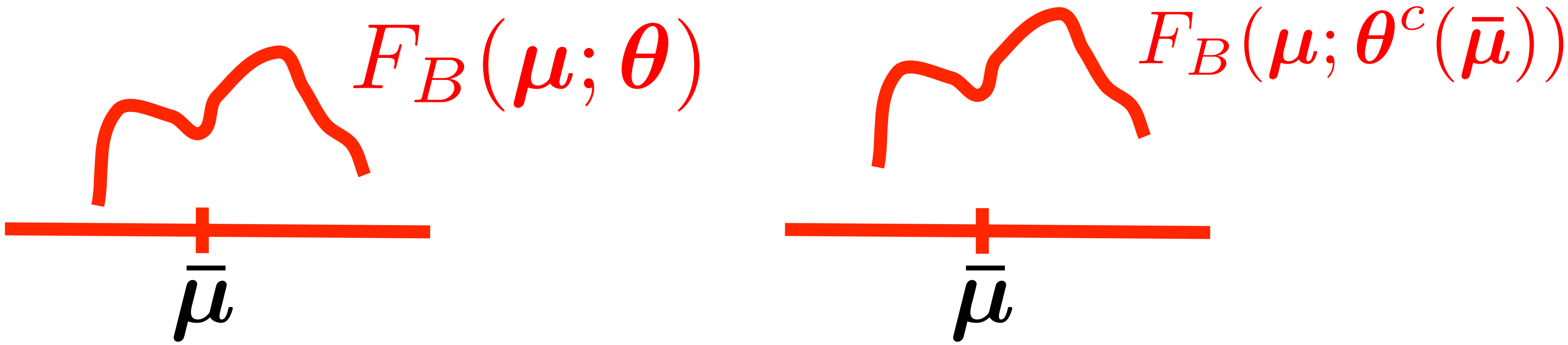
- Say we have a non-canonical $\boldsymbol{\theta}$ s.t. $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$



- The function for the canonical parameter will be the same up to a constant.

Stationary point invariance

- Say we have a non-canonical $\boldsymbol{\theta}$ s.t. $\bar{\boldsymbol{\mu}}$ is a stationary point of $F_B(\boldsymbol{\mu};\boldsymbol{\theta})$



- The function for the canonical parameter will be the same up to a constant.

- So, when looking for $\boldsymbol{\theta}$ s.t. $\bar{\boldsymbol{\mu}}$ is a single maximizer (learnable) it's enough to focus on canonical.

Message I

Use Canonical or don't use Anything!

Outer Bound I

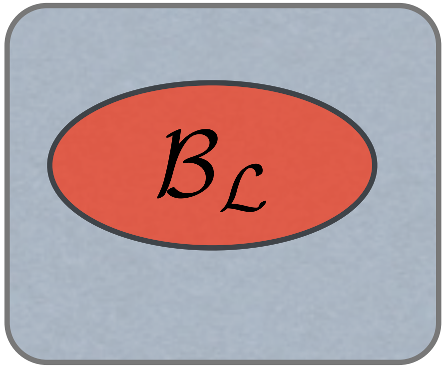
- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

Outer Bound I

- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

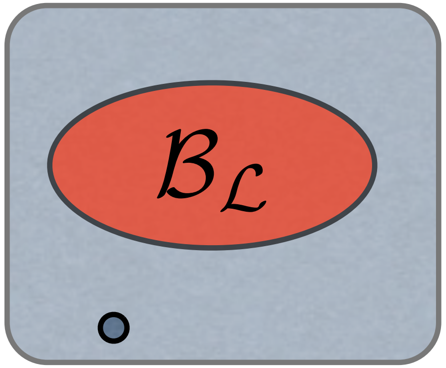
Outer Bound I

- Identifies cases where $\bar{\mu} \notin \mathcal{B}_{\mathcal{L}}$



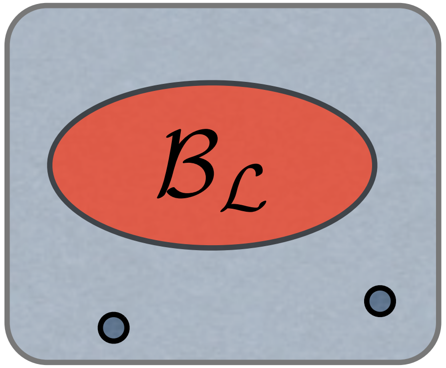
Outer Bound I

- Identifies cases where $\bar{\mu} \notin \mathcal{B}_{\mathcal{L}}$



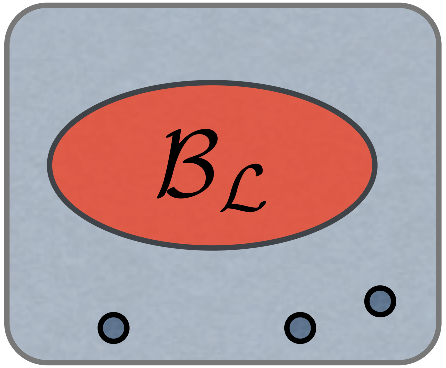
Outer Bound I

- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$



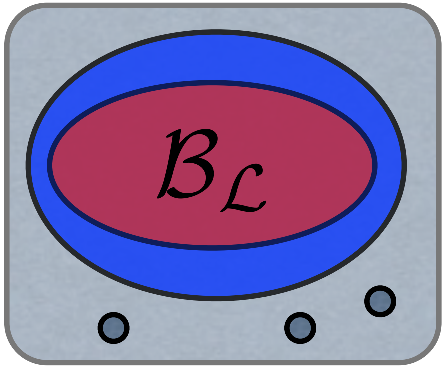
Outer Bound I

- Identifies cases where $\bar{\mu} \notin \mathcal{B}_{\mathcal{L}}$



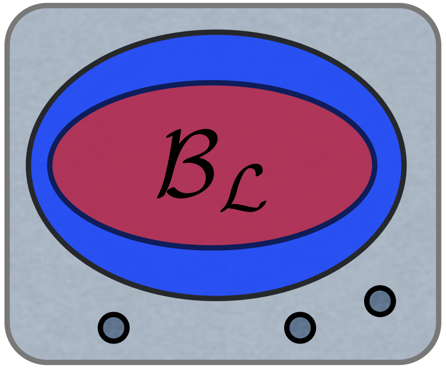
Outer Bound I

- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$



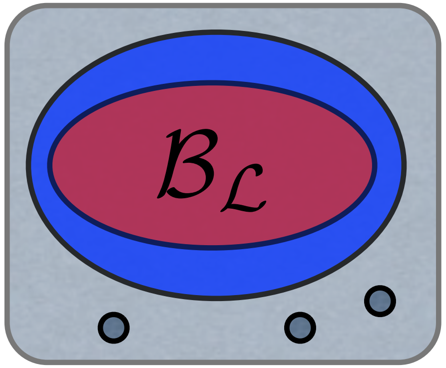
Outer Bound I

- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$
- Look at $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

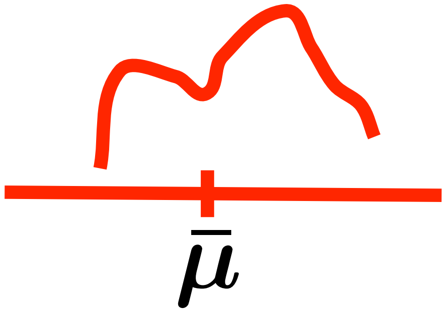


Outer Bound I

- Identifies cases where $\bar{\mu} \notin \mathcal{B}_{\mathcal{L}}$
- Look at $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

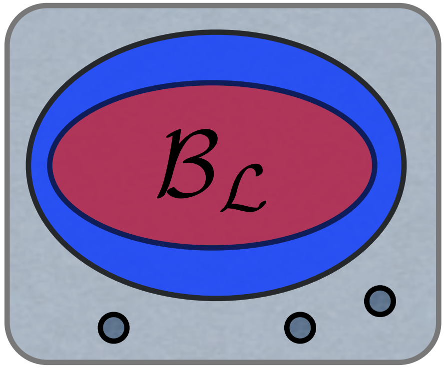


$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



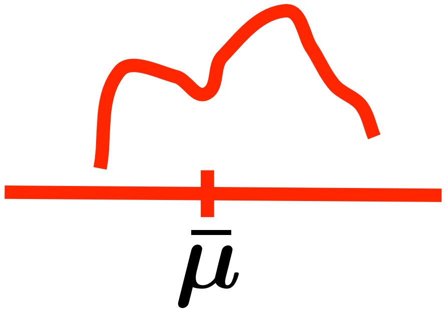
Outer Bound I

- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$
- Look at $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



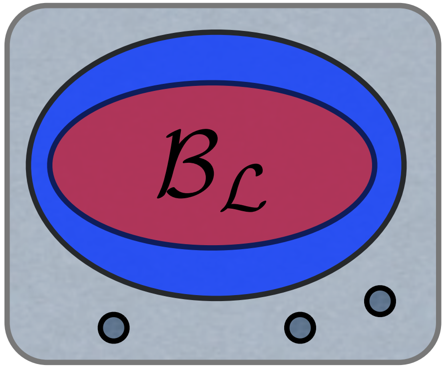
- If $\bar{\boldsymbol{\mu}}$ is not its global maximum, then $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



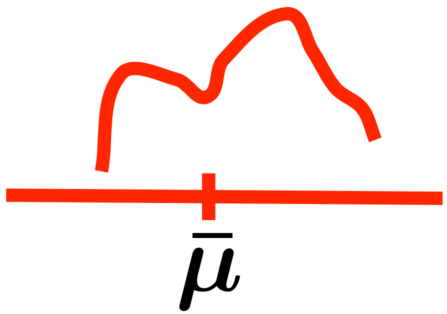
Outer Bound I

- Identifies cases where $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$
- Look at $F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



- If $\bar{\boldsymbol{\mu}}$ is not its global maximum, then $\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

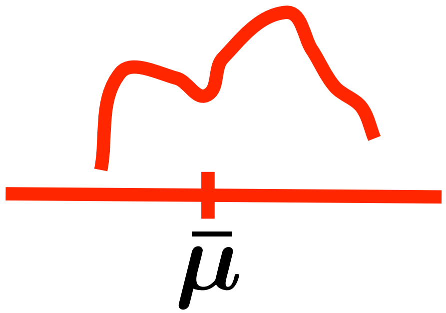
$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

Outer Bound I

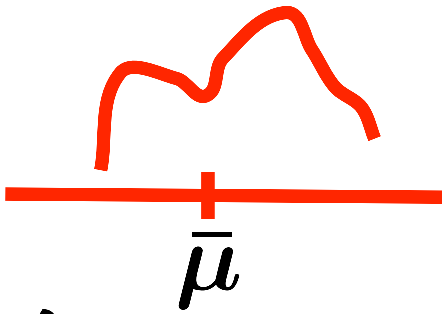
$F_B(\mu; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

Outer Bound I

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

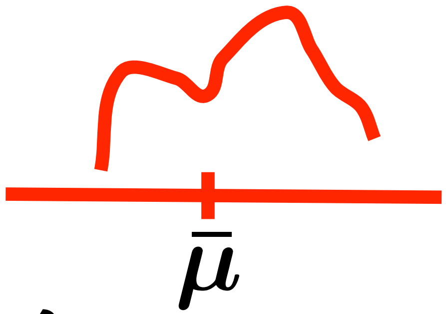


$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

- How do you check it?

Outer Bound I

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

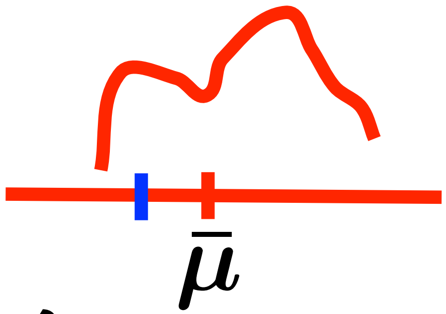


$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

- How do you check it?
- Run BP several time to find other optima and compare their values.

Outer Bound I

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

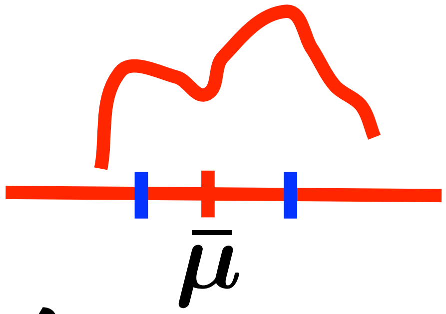


$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

- How do you check it?
- Run BP several time to find other optima and compare their values.

Outer Bound I

$F_B(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$

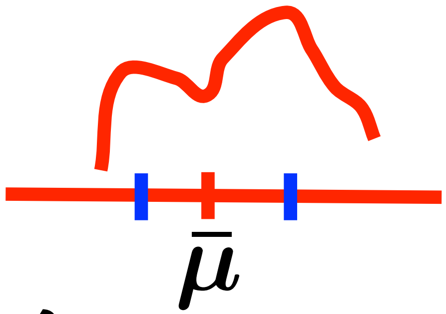


$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

- How do you check it?
- Run BP several time to find other optima and compare their values.

Outer Bound I

$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$



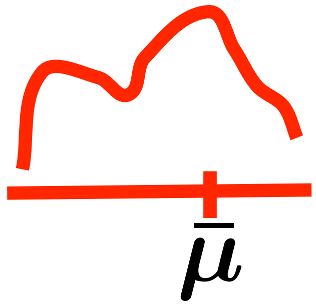
$\bar{\boldsymbol{\mu}} \notin \mathcal{B}_{\mathcal{L}}$

- How do you check it?
- Run BP several time to find other optima and compare their values.
- If we've discovered better maxima, then there is no chance that $\bar{\boldsymbol{\mu}}$ is learnable...

Outer Bound II

- Learnable marginals look like this:

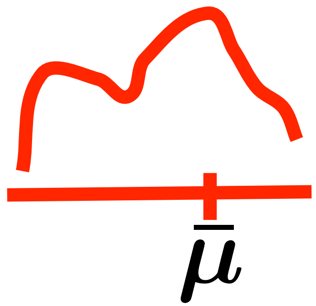
$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$



Outer Bound II

- Learnable marginals look like this:

$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$

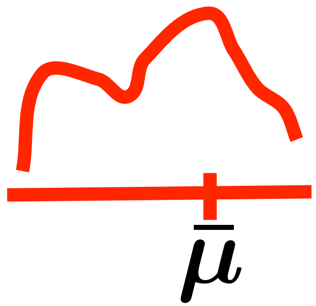


- If $\bar{\boldsymbol{\mu}}$ is not a maximum (even local) of $F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ then $\bar{\boldsymbol{\mu}}$ is not learnable.

Outer Bound II

- Learnable marginals look like this:

$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$

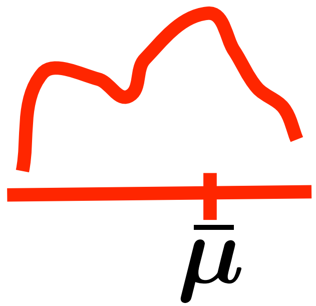


- If $\bar{\boldsymbol{\mu}}$ is not a maximum (even local) of $F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ then $\bar{\boldsymbol{\mu}}$ is not learnable.
- Do such marginals ever exist?!

Outer Bound II

- Learnable marginals look like this:

$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$



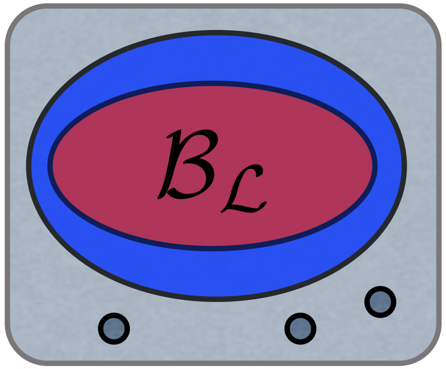
- If $\bar{\boldsymbol{\mu}}$ is not a maximum (even local) of $F_B(\boldsymbol{\mu}; \boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ then $\bar{\boldsymbol{\mu}}$ is not learnable.
- Do such marginals ever exist?!
- Yes! Many

Outer Bound II

- Consider marginals that are never local maxima of *any* Bethe free energy.
- They will also never be stable fixed points of BP (Heskes).
- Called unbelievable marginals in (Pitkow & Miller, 12)

Outer Bound II

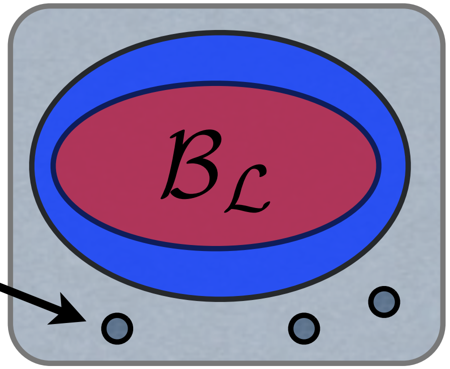
- Consider marginals that are never local maxima of *any* Bethe free energy.
- They will also never be stable fixed points of BP (Heskes).
- Called unbelievable marginals in (Pitkow & Miller, 12)



Outer Bound II

- Consider marginals that are never local maxima of *any* Bethe free energy.
- They will also never be stable fixed points of BP (Heskes).
- Called unbelievable marginals in (Pitkow & Miller, 12)

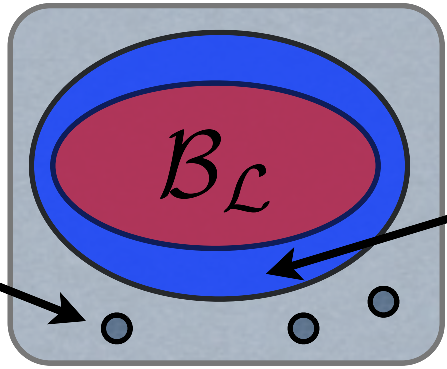
$\bar{\mu}$ that are not maxima of anything.



Outer Bound II

- Consider marginals that are never local maxima of *any* Bethe free energy.
- They will also never be stable fixed points of BP (Heskes).
- Called unbelievable marginals in (Pitkow & Miller, 12)

$\bar{\mu}$ *that are not maxima of anything.*



$\bar{\mu}$ *that are maxima but never global.*

Message II

Some marginals cannot be BP stable fixed points!

Outer Bound II

- How do you find marginals which can't maximize?
- Recall: $F(\boldsymbol{\mu};\boldsymbol{\theta}) = \boldsymbol{\mu}\cdot\boldsymbol{\theta} + H_B(\boldsymbol{\mu})$
- Hessian does not depend on $\boldsymbol{\theta}$ (roughly...)
- We only need to consider Hessian of $H_B(\bar{\boldsymbol{\mu}})$
- If it has non-negative eigenvalues, $\bar{\boldsymbol{\mu}}$ cannot be a local maximizer.
- For binary variables this is easy to test.

Homogenous Binary Case

- To get some intuition consider binary variables, and homogenous marginals:

$$\mu_i(x_i = 1) = \mu_v \qquad \forall i$$

$$\mu_{ij}(x_i = 1, x_j = 1) = \mu_e \qquad \forall ij$$

- Find a lower bound on the maximum eigenvalue of the Hessian, and check when it is non-negative.
- Closely related to the spectrum of the graph.

Homogenous Binary Case

- Following marginals are un-learnable:

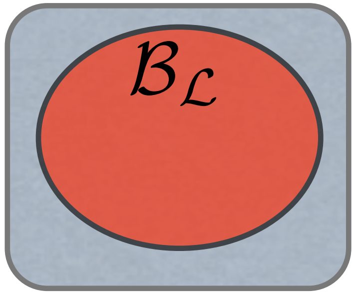
$$\bar{\mu}_e > \frac{(1 - \frac{V}{E})\bar{\mu}_v^2 + \frac{V}{2E}\bar{\mu}_v}{1 - \frac{V}{2E}}$$

- For complete graphs with infinite V this is:

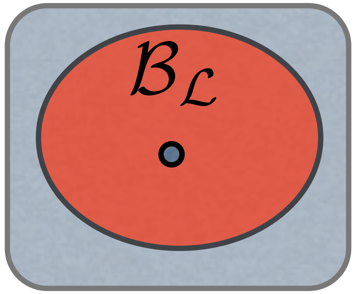
$$\bar{\mu}_e > \bar{\mu}_v^2$$

- All attractive Ising models are in this set!

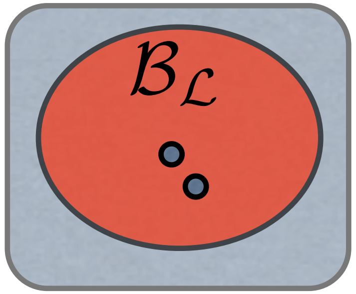
Inner Bounds



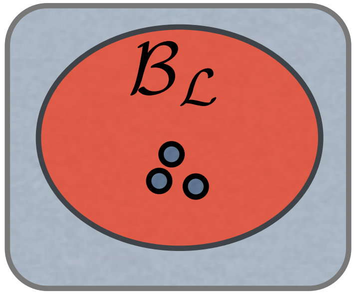
Inner Bounds



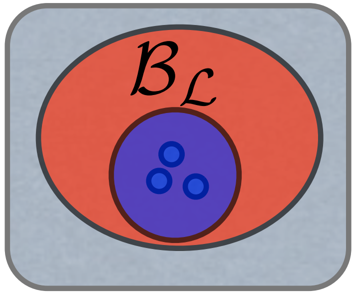
Inner Bounds



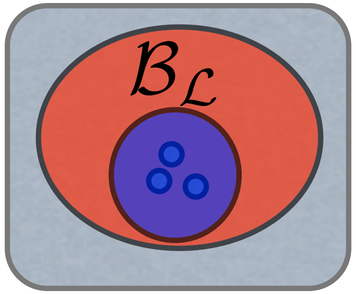
Inner Bounds



Inner Bounds

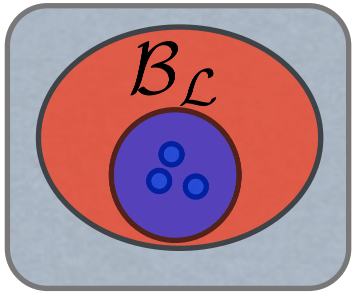


Inner Bounds



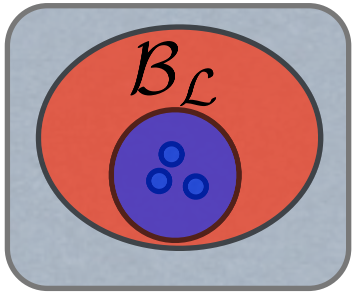
- How can we guarantee that $\bar{\mu}$ is learnable?

Inner Bounds

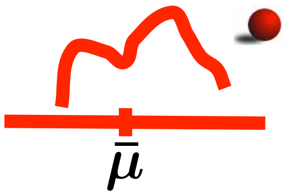


- How can we guarantee that $\bar{\boldsymbol{\mu}}$ is learnable?
- We know that it is a local optimum of the function $F(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$. When is it global?

Inner Bounds

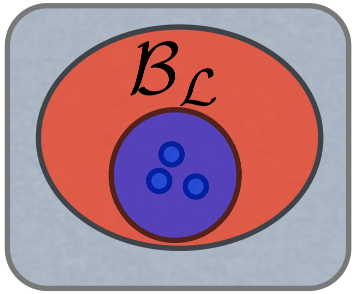


- How can we guarantee that $\bar{\boldsymbol{\mu}}$ is learnable?

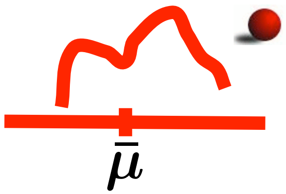


- We know that it is a local optimum of the function $F(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$. When is it global?

Inner Bounds

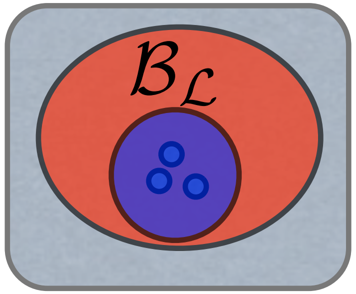


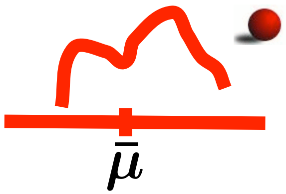
- How can we guarantee that $\bar{\boldsymbol{\mu}}$ is learnable?

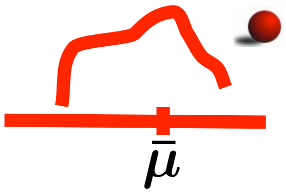


- We know that it is a local optimum of the function $F(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$. When is it global?
- If this function has a unique maximum point, then we have that $\bar{\boldsymbol{\mu}}$ is the global optimum!

Inner Bounds

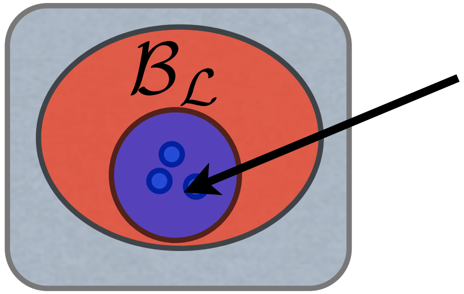






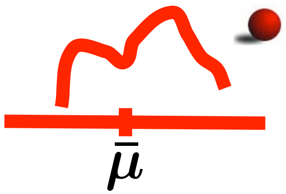
- How can we guarantee that $\bar{\boldsymbol{\mu}}$ is learnable?
- We know that it is a local optimum of the function $F(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$. When is it global?
- If this function has a unique maximum point, then we have that $\bar{\boldsymbol{\mu}}$ is the global optimum!

Inner Bounds

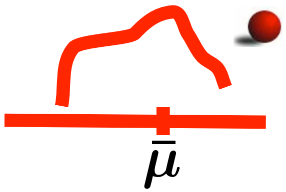


$\bar{\boldsymbol{\mu}}$ s.t. $F(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$ has single global maximum

- How can we guarantee that $\bar{\boldsymbol{\mu}}$ is learnable?

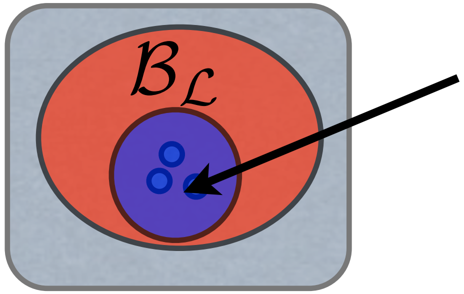


- We know that it is a local optimum of the function $F(\boldsymbol{\mu};\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$. When is it global?



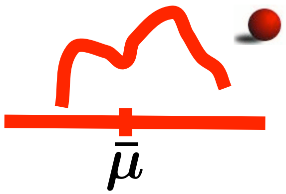
- If this function has a unique maximum point, then we have that $\bar{\boldsymbol{\mu}}$ is the global optimum!

Inner Bounds

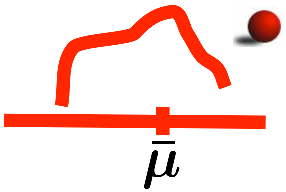


$\bar{\boldsymbol{\mu}}$ s.t. $F(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$ has single global maximum

- How can we guarantee that $\bar{\boldsymbol{\mu}}$ is learnable?



- We know that it is a local optimum of the function $F(\boldsymbol{\mu}; \boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}))$. When is it global?



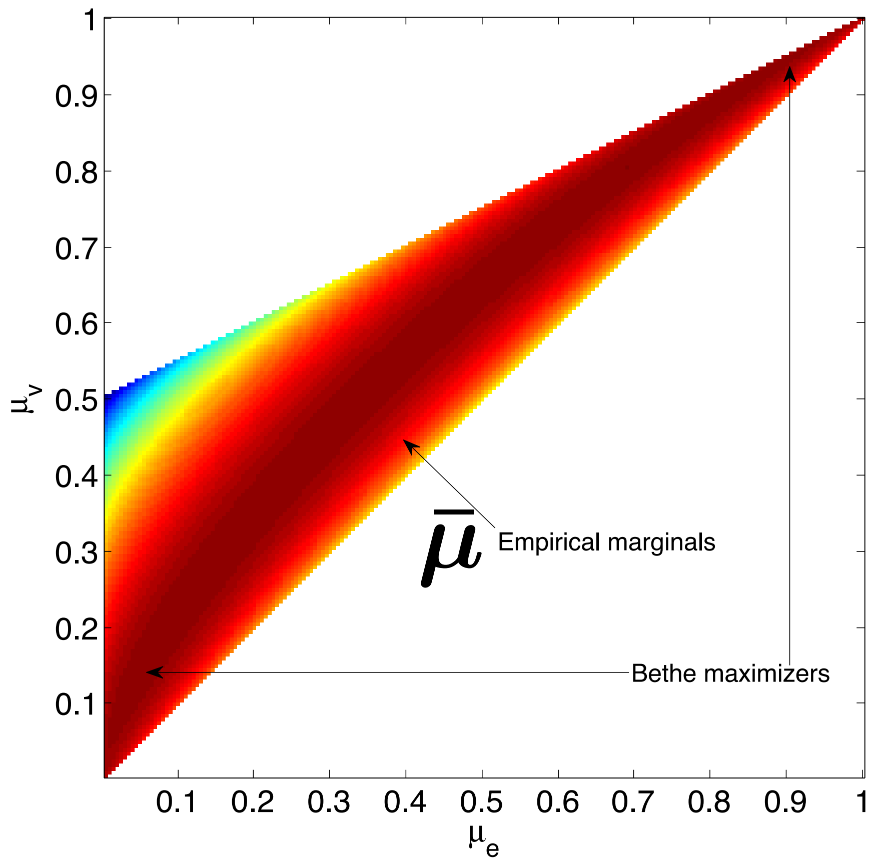
- If this function has a unique maximum point, then we have that $\bar{\boldsymbol{\mu}}$ is the global optimum!
- Multiple works on characterizing when BP has unique fixed points (*Mooij, Kappen 07; Roosta et al. 08*).

Experiments

- Focus on binary variables for ease of presentation.
- For homogenous case each marginal is characterized in 2D (depicting μ_v, μ_e).
- We also test empirically whether moment matching can be achieved (using gradient descent).

Experiments

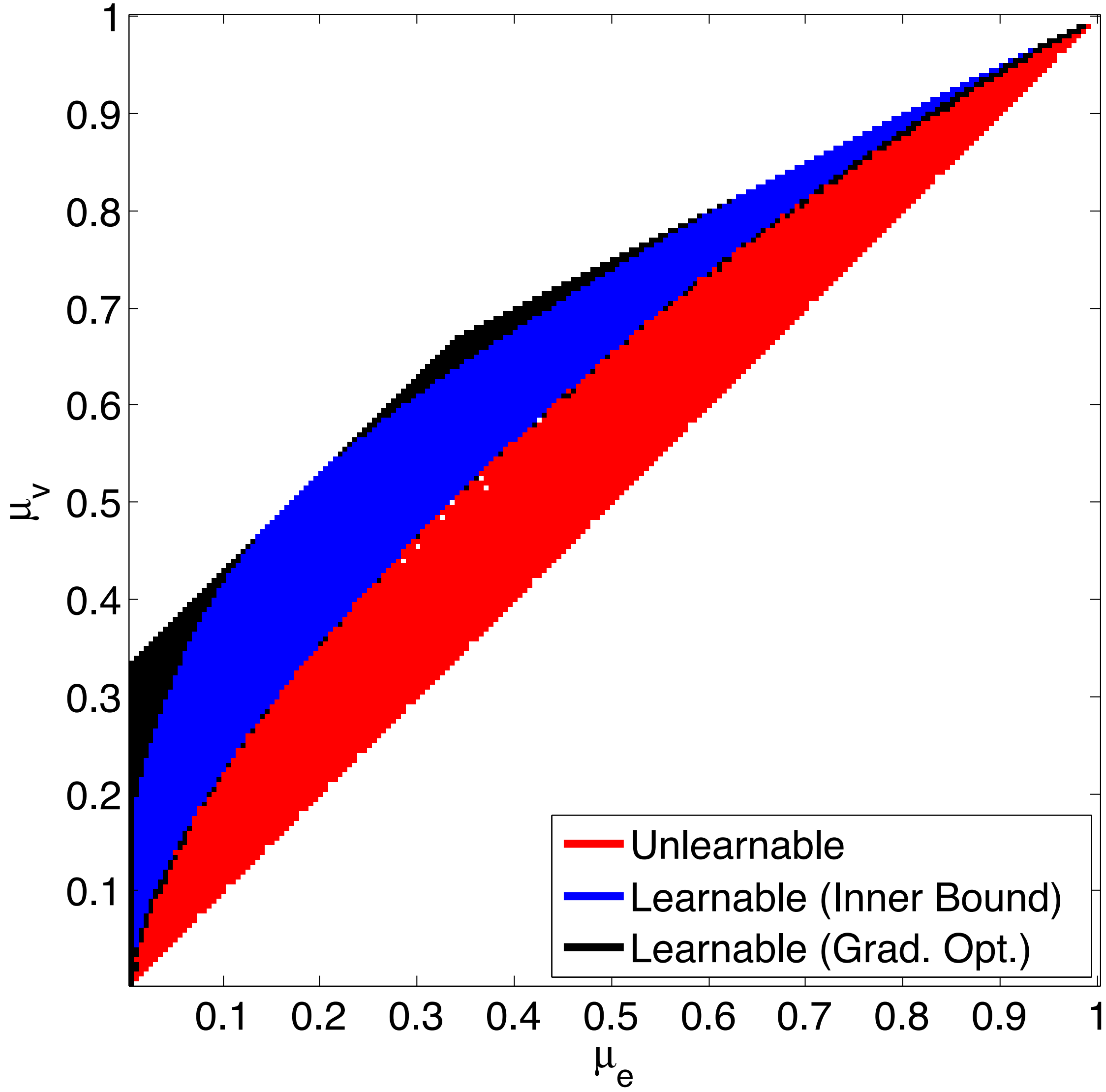
- What happens for unlearnable marginals?



$$F_B(\boldsymbol{\mu}; \boldsymbol{\theta}(\bar{\boldsymbol{\mu}}))$$

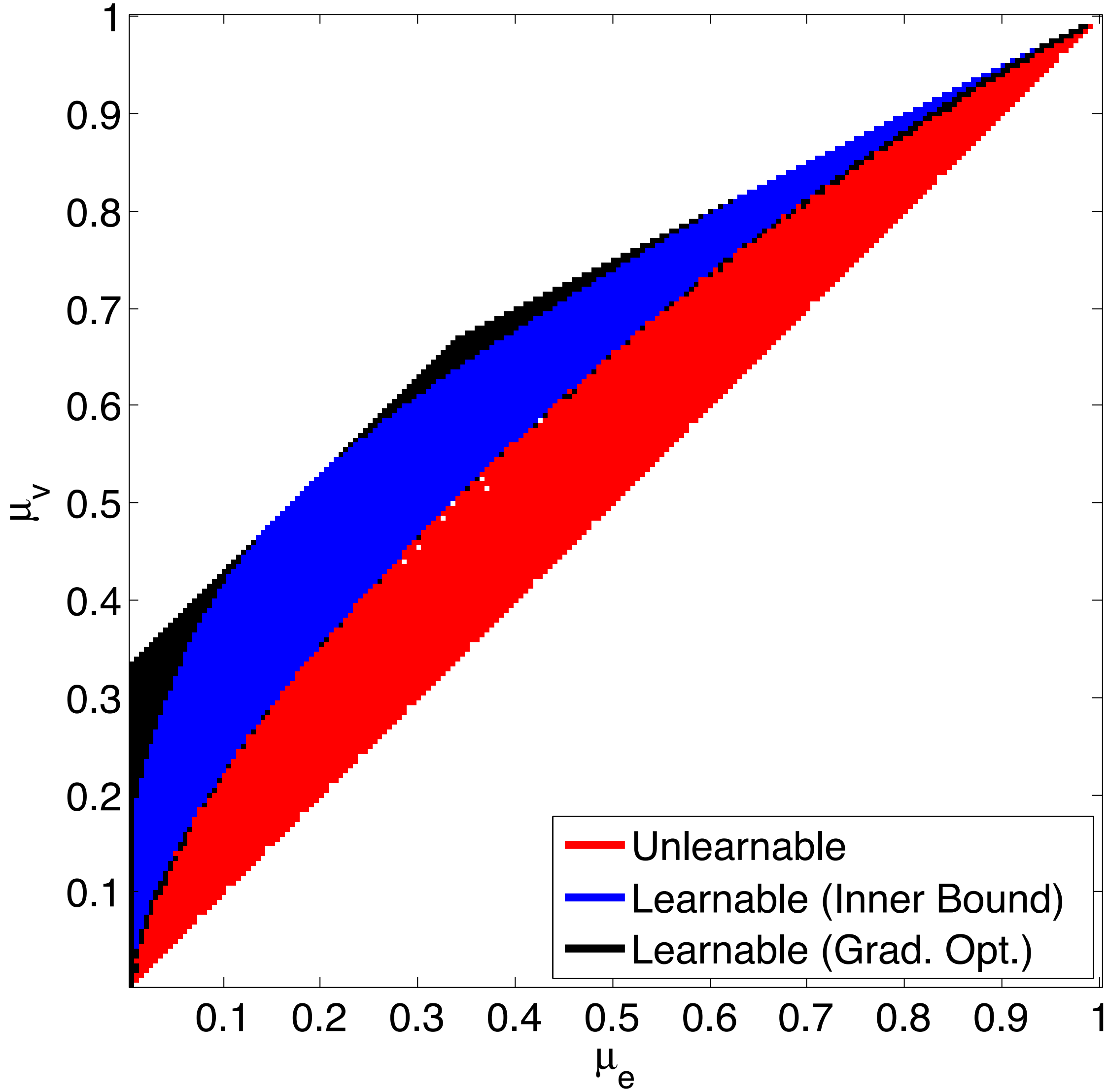
- $\bar{\boldsymbol{\mu}}$ is not a maximizer, but at a convex hull of maximizers.

3x3 Grid



1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
μv
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
μe
Unlearnable
Learnable (Inner Bound)
Learnable (Grad. Opt.)

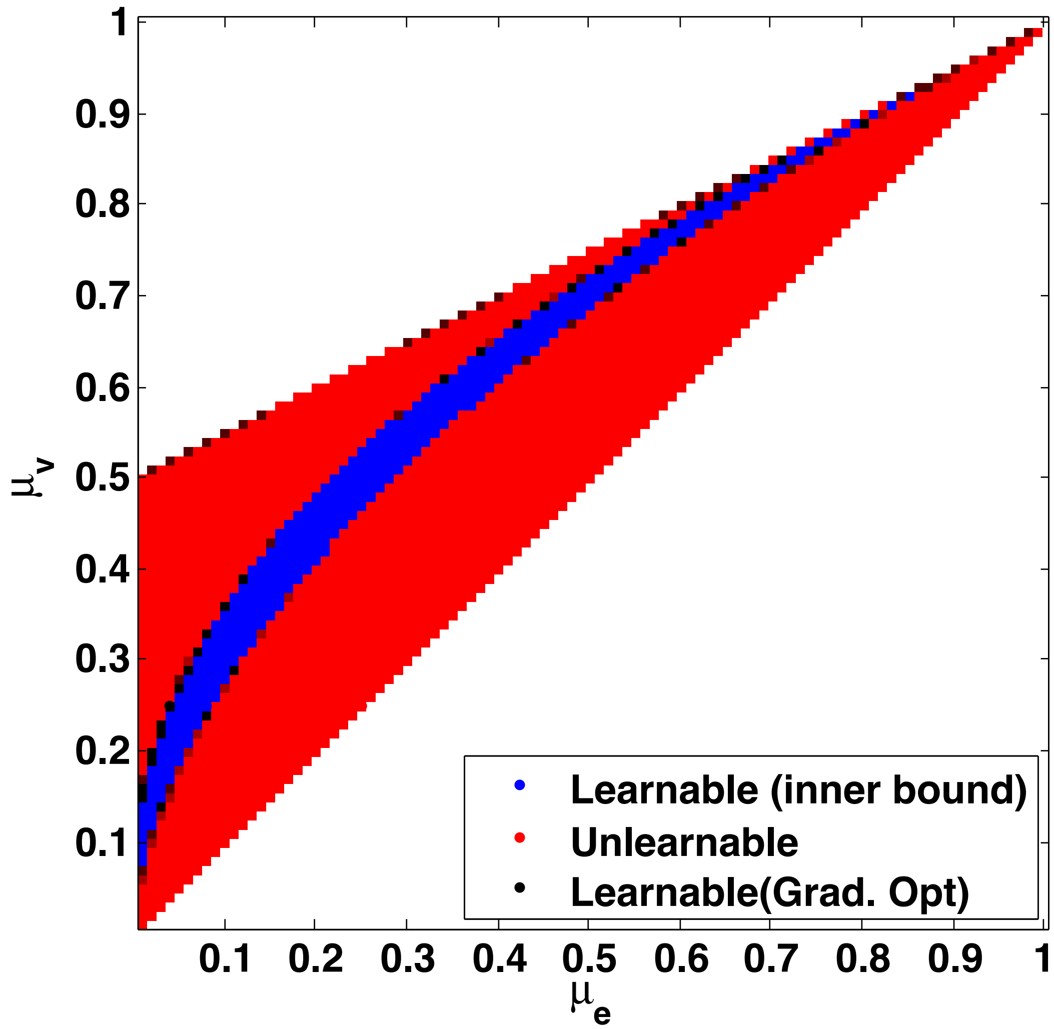
3x3 Grid



Outer bound is tight!

Bipartite 8x8

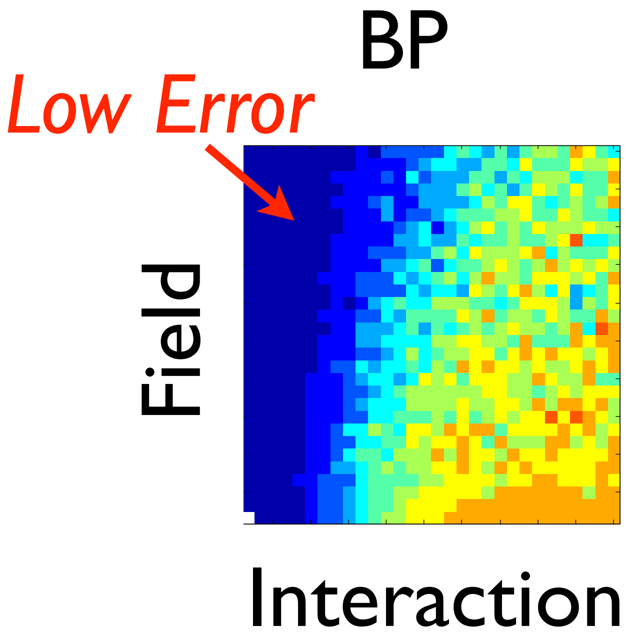
- Largely unlearnable
- Bad news for restricted Boltzmann Machines...



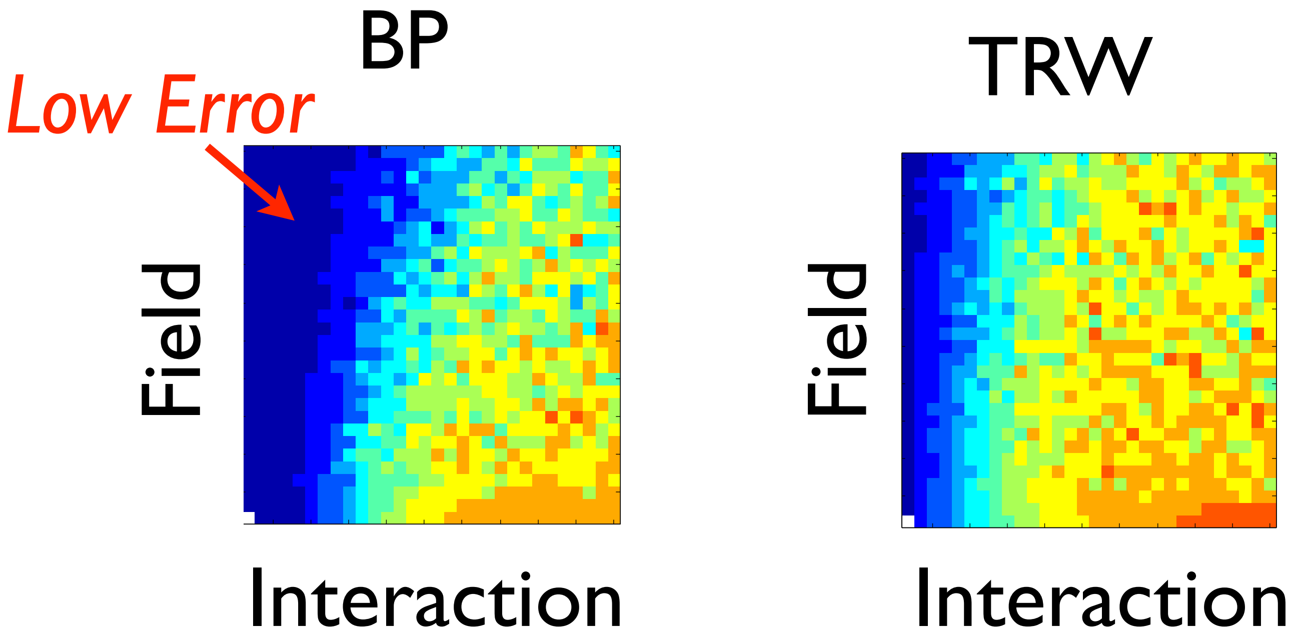
Learnability and Performance

- How well does BP perform in the learnable region?
- Test on new marginals (not those in $\bar{\mu}$).
- Use Ising grid graphs. Sample models with varying field and interaction strengths.
- Compare to TRW (Wainwright et al.)

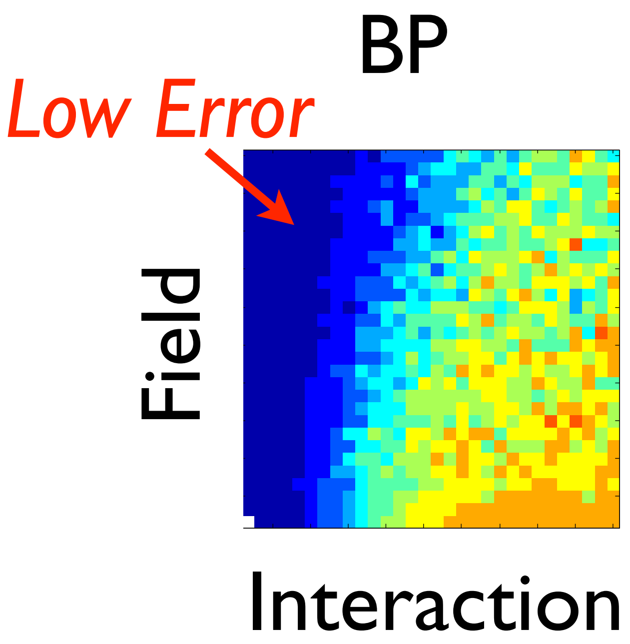
Learnability and Performance

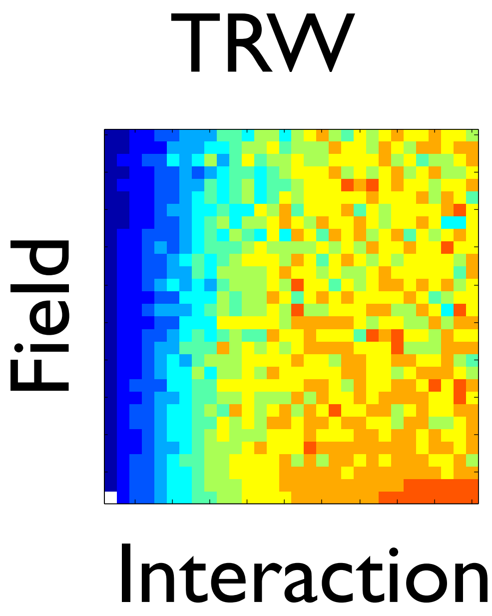


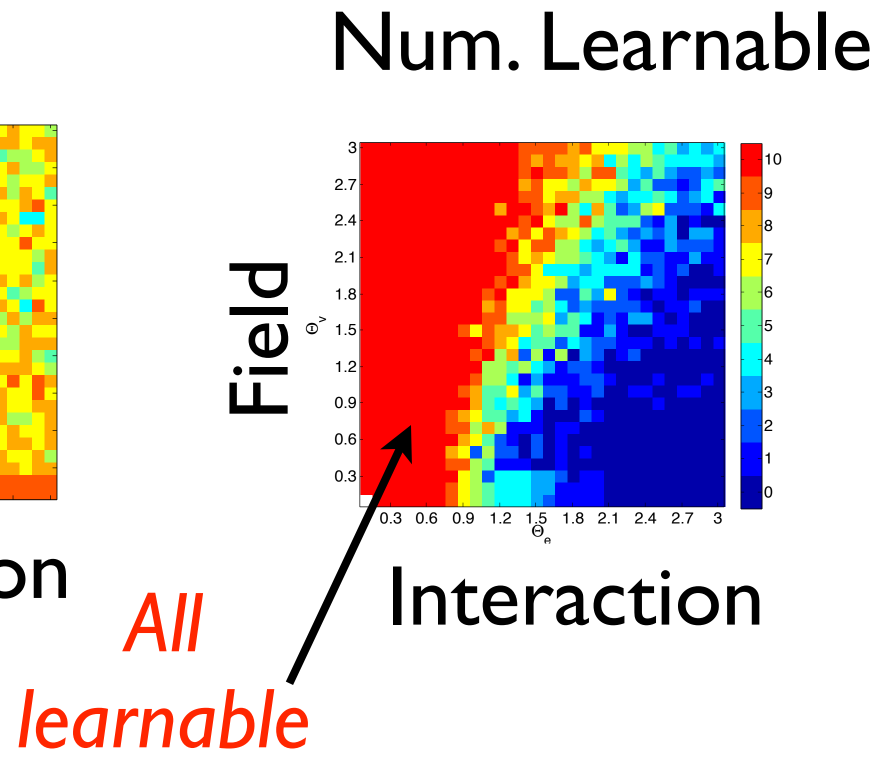
Learnability and Performance



Learnability and Performance

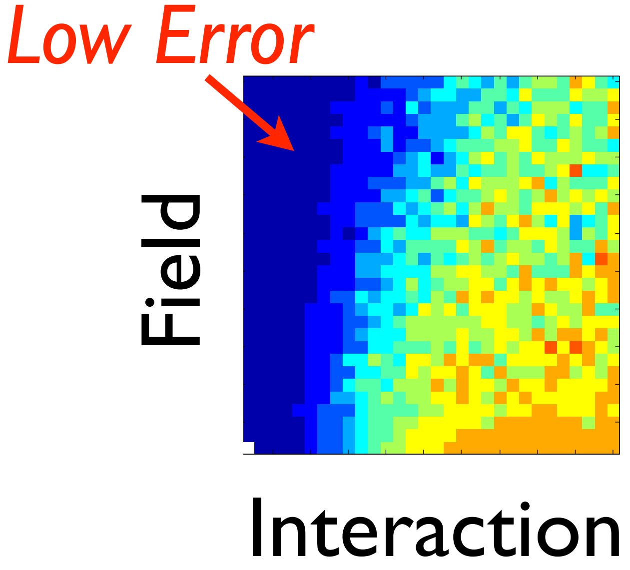




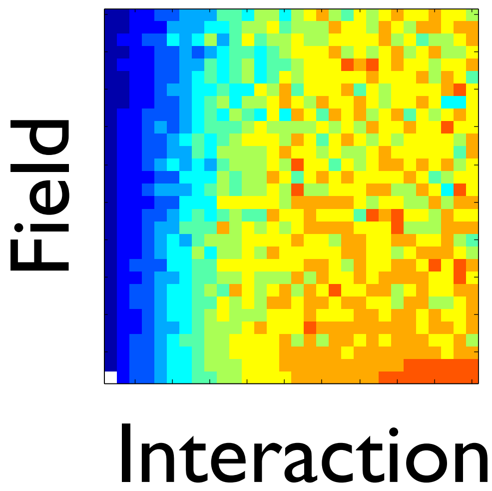


Learnability and Performance

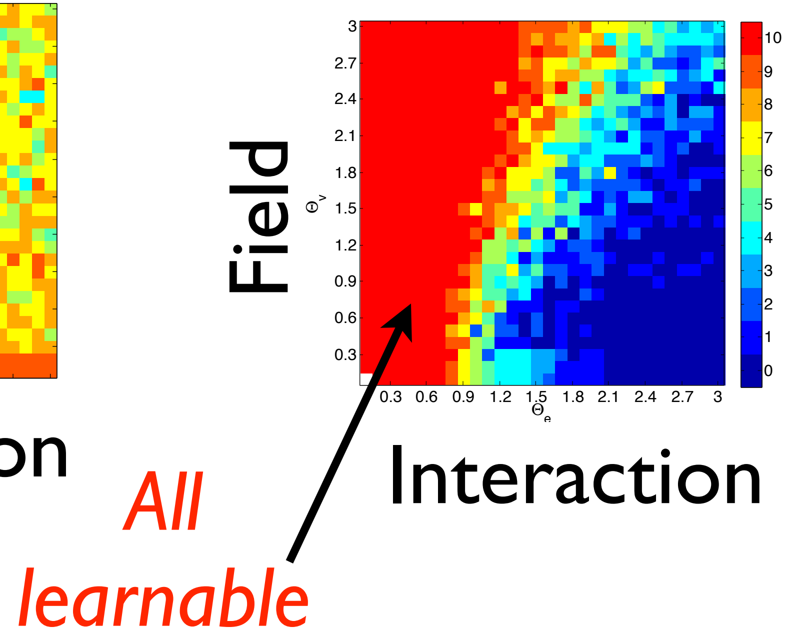
BP



TRW



Num. Learnable



- Learnability is well correlated with performance!

Take Home Messages

- Some marginals cannot be obtained with BP!
- These can be analytically characterized.
- Learning with BP will “often” not even achieve moment matching.
- Cannot recover marginals of the data.
- No reason to use BP in these cases.
- For learnable marginals BP performs well.

Future Work

- Tighter characterization
- Use BP on models where it works.
- Workarounds: Maybe ML is not the right criterion. Try to match moment directly.
- Use higher order approximations (Kikuchi). Could improve learnability (provably does it for sufficiently tight approximations).