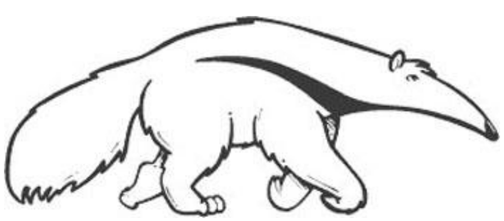
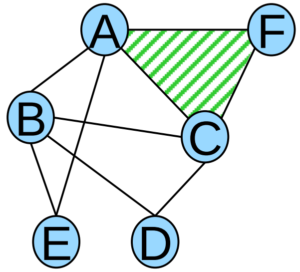
Anytime Probabilistic Reasoning

Rina Dechter

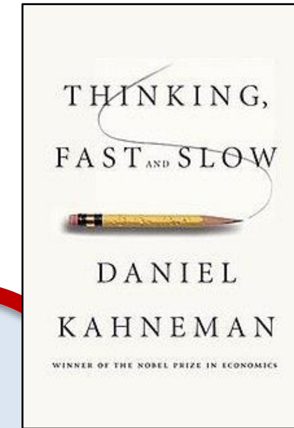
Bren School of Information and Computer Sciences

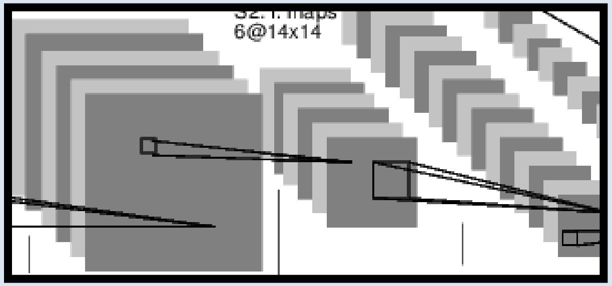




Main Collaborators:
Alexander Ihler
Kalev Kask
Radu Marinescu
Qi Lou
Junkyu Lee

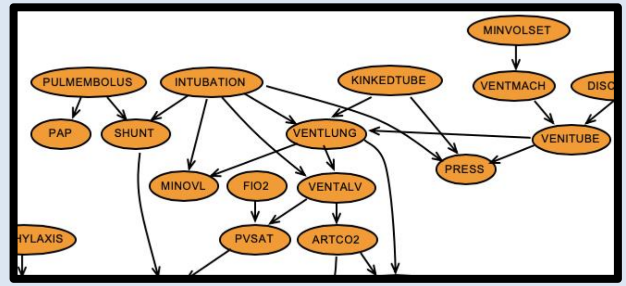
AI Renaissance





- Deep learning
  - Fast predictions
  - "Instinctive"

Tools:
Tensorflow, PyTorch, …

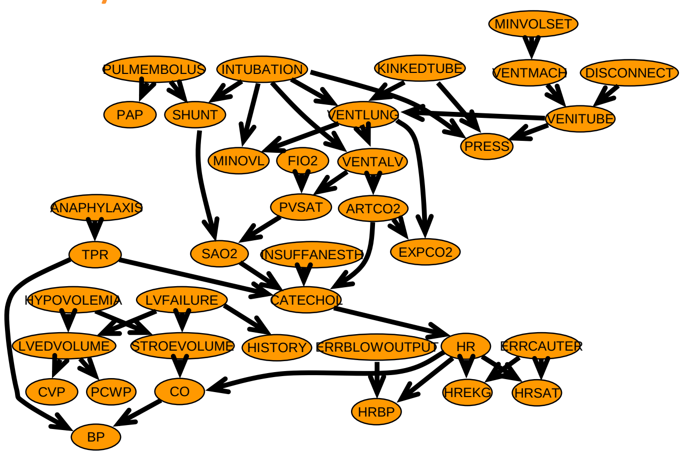


- Probabilistic models
  - Slow reasoning
  - "Logical / deliberative"

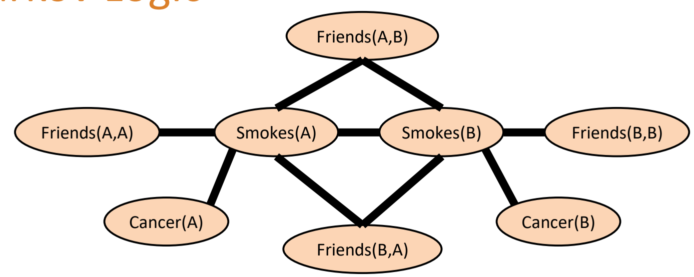
Tools:
Probabilistic programming, Markov Logic, …

Overview: Graphical Models

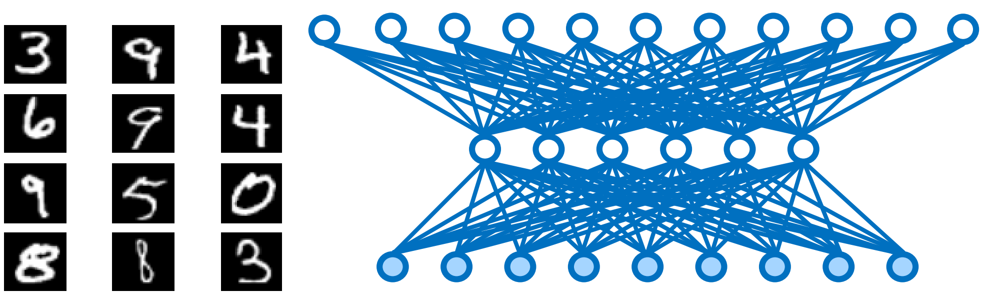
Bayesian Networks



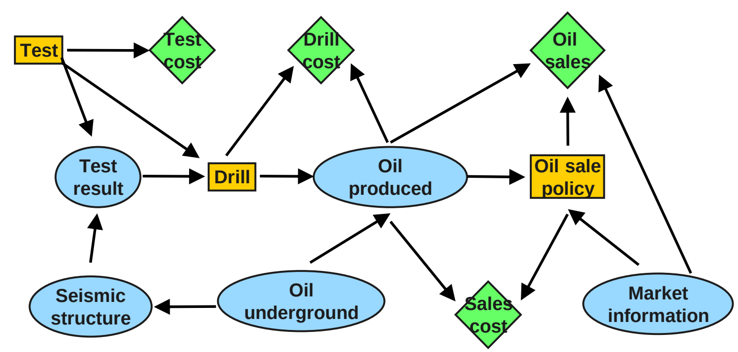
Markov Logic



Deep Boltzmann Machines



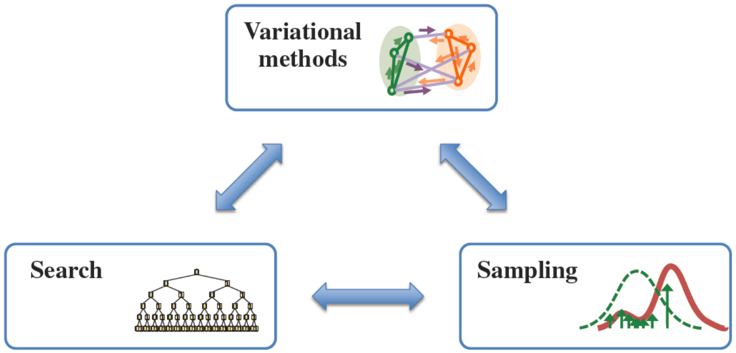
Influence Diagrams



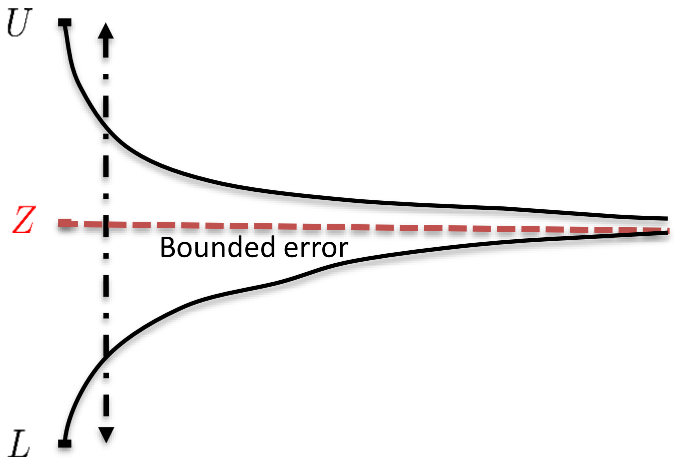
Outline

- Overview of problems and methodology
- Main paradigms of approximate reasoning:
  Variational, Search, Sampling
- Combining approaches
- Future challenges

Main paradigms



The anytime desirerata



Graphical Models

- Describe structure in large problems
  - Large complex system $f(X)$
  - Made of “smaller”, “local” interactions $f_\alpha(X_\alpha)$
  - Complexity emerges through interdependence
- More formally:

  A *graphical model* consists of:

  $X = \{X_1, \ldots, X_n\}$ -- variables (we’ll assume discrete)

  $D = \{D_1, \ldots, D_n\}$ -- domains

  $F = \{f_{\alpha_1}, \ldots, f_{\alpha_m}\}$ -- (non-negative) functions or “factors”

Example:

$A \in \{0,1\}$

$B \in \{0,1\}$

$C \in \{0,1\}$

$f_{AB}(A,B), \quad f_{BC}(B,C)$

- Example:

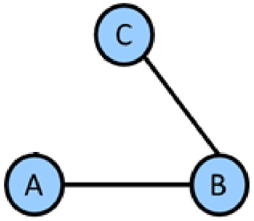
  $F(A,B,C) = f(A,B) \cdot f(B,C)$

A	B	f(A,B)
0	0	**0.24**
0	1	**0.56**
1	0	**1.1**
1	1	**1.2**

…

B	C	f(B,C)
0	0	**0.12**
0	1	**0.36**
1	0	**0.3**
1	1	**1.8**

Primal graph

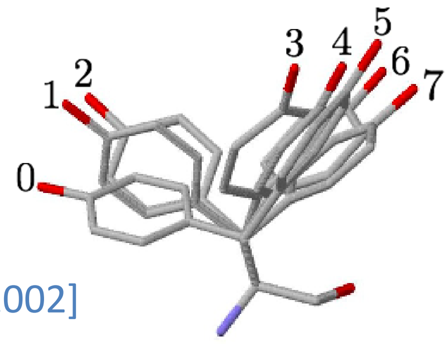


Graphical Models

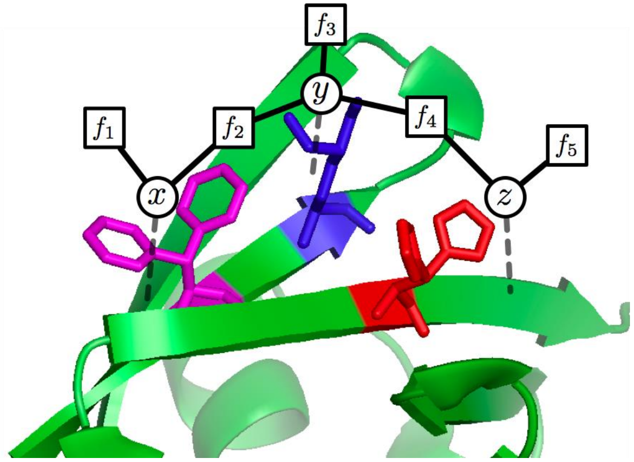
- Describe structure in large problems
  - Large complex system $f(X)$
  - Made of "smaller", "local" interactions $f_\alpha(X_\alpha)$
  - Complexity emerges through interdependence

- Examples & Tasks
  - Maximization (**MAP**): compute the most probable configuration

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \prod_\alpha f_\alpha(\mathbf{x}_\alpha) \qquad f(\mathbf{x}^*) = \max_{\mathbf{x}} \prod_\alpha f_\alpha(\mathbf{x}_\alpha)$$



Phenylalanine



[Yanover & Weiss 2002]

Graphical Models

- Describe structure in large problems
  - Large complex system $f(X)$
  - Made of “smaller”, “local” interactions $f_\alpha(X_\alpha)$
  - Complexity emerges through interdependence

- Examples & Tasks
  - Summation & marginalization

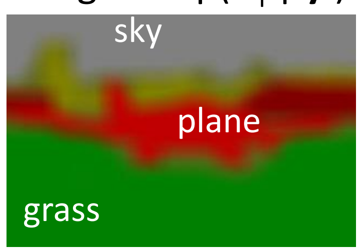
“partition function”

$$p(x_i) = \frac{1}{Z} \sum_{\mathbf{x} \setminus x_i} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha) \quad \text{and} \quad Z = \sum_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

Observation **y**

$\Rightarrow$

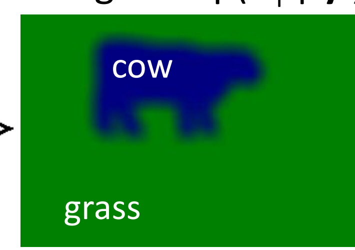
Marginals p(x_i | **y**)



Observation **y**

$\Rightarrow$

Marginals p(x_i | **y**)

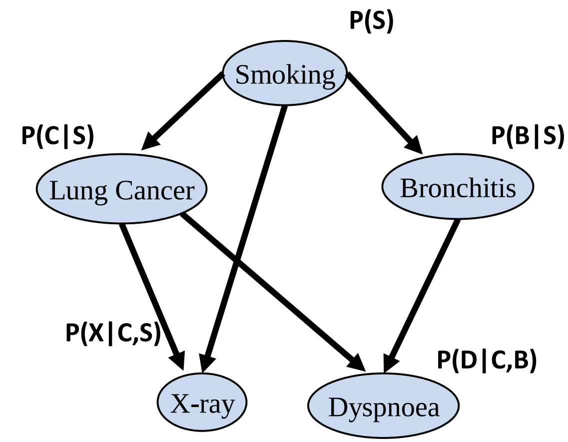


e.g., [Plath et al. 2009]

Graphical Models

- Describe structure in large problems
  - Large complex system $f(X)$
  - Made of "smaller", "local" interactions $f_\alpha(X_\alpha)$
  - Complexity emerges through interdependence
- Examples & Tasks
  - Mixed inference (marginal MAP, MEU, …)

$$f(\mathbf{x}_M^*) = \max_{\mathbf{x}_M} \sum_{\mathbf{x}_S} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

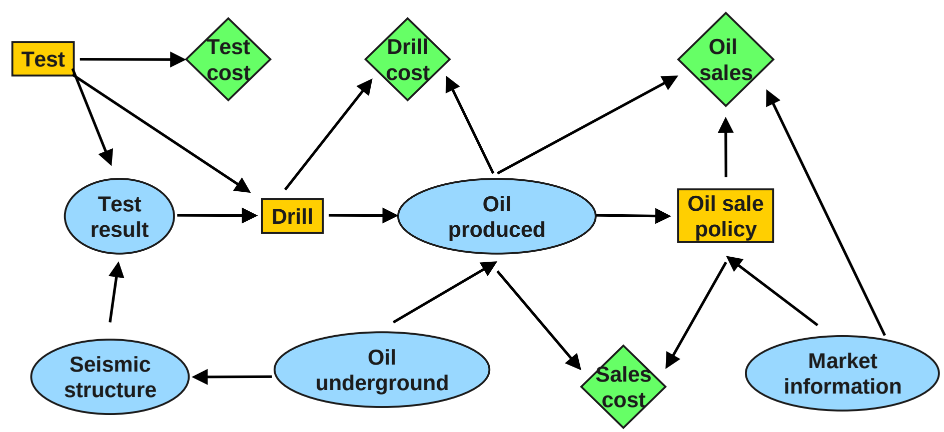


P(S,C,B,X,D) = P(S) · P(C|S) · P(B|S) · P(X|C,S) · P(D|C,B)

Influence diagrams & optimal decision-making

(the "oil wildcatter" problem)

e.g., [Raiffa 1968; Shachter 1986]



Graphical models

A ***graphical model*** consists of:

$X = \{X_1, \ldots, X_n\}$ -- variables

$D = \{D_1, \ldots, D_n\}$ -- domains

$F = \{f_{\alpha_1}, \ldots, f_{\alpha_m}\}$ -- functions

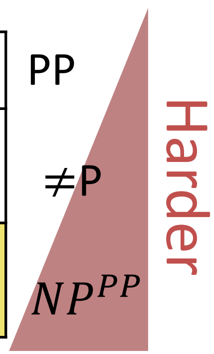
Operators:

combination operator (sum, product, join, ...)

elimination operator (projection, sum, max, min, ...)

Types of queries:

Max-Inference (MAP)	$f(\mathbf{x}^*) = \max_{\mathbf{x}} \prod_{\alpha} f_{\alpha}(\mathbf{x}_{\alpha})$
Sum-Inference (P(€)	$Z = \sum_{\mathbf{x}} \prod_{\alpha} f_{\alpha}(\mathbf{x}_{\alpha})$
Mixed-Inference	$f(\mathbf{x}_M^*) = \max_{\mathbf{x}_M} \sum_{\mathbf{x}_S} \prod_{\alpha} f_{\alpha}(\mathbf{x}_{\alpha})$

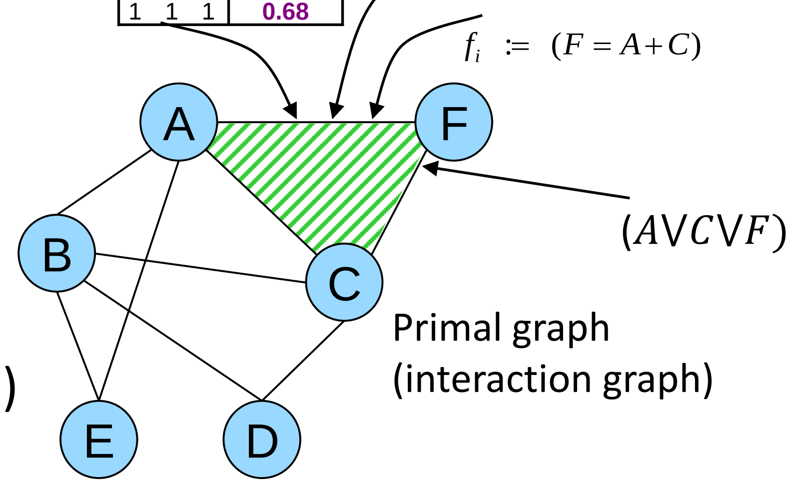


Conditional Probability Table (CPT)

A	C	F	P(F|A,C)
0	0	0	0.14
0	0	1	0.96
0	1	0	0.40
0	1	1	0.60
1	0	0	0.35
1	0	1	0.65
1	1	0	0.72
1	1	1	0.68

Relation

A	C	F
red	green	blue
blue	red	red
blue	blue	green
green	red	blue



Primal graph (interaction graph)

- All these tasks are NP-hard
  - exploit problem structure
  - identify special cases
  - approximate

Example Domains

- Natural Language processing
  - Information extraction, semantic parsing, translation, topic models, …
- Computer vision
  - Object recognition, scene analysis, segmentation, tracking, …
- Computational biology
  - Pedigree analysis, protein folding and binding, sequence matching, …
- Networks
  - Webpage link analysis, social networks, communications, citations, ….
- Robotics
  - Planning & decision making

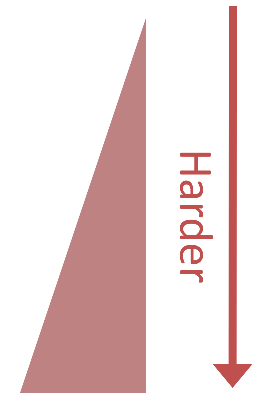
Anytime Bounds

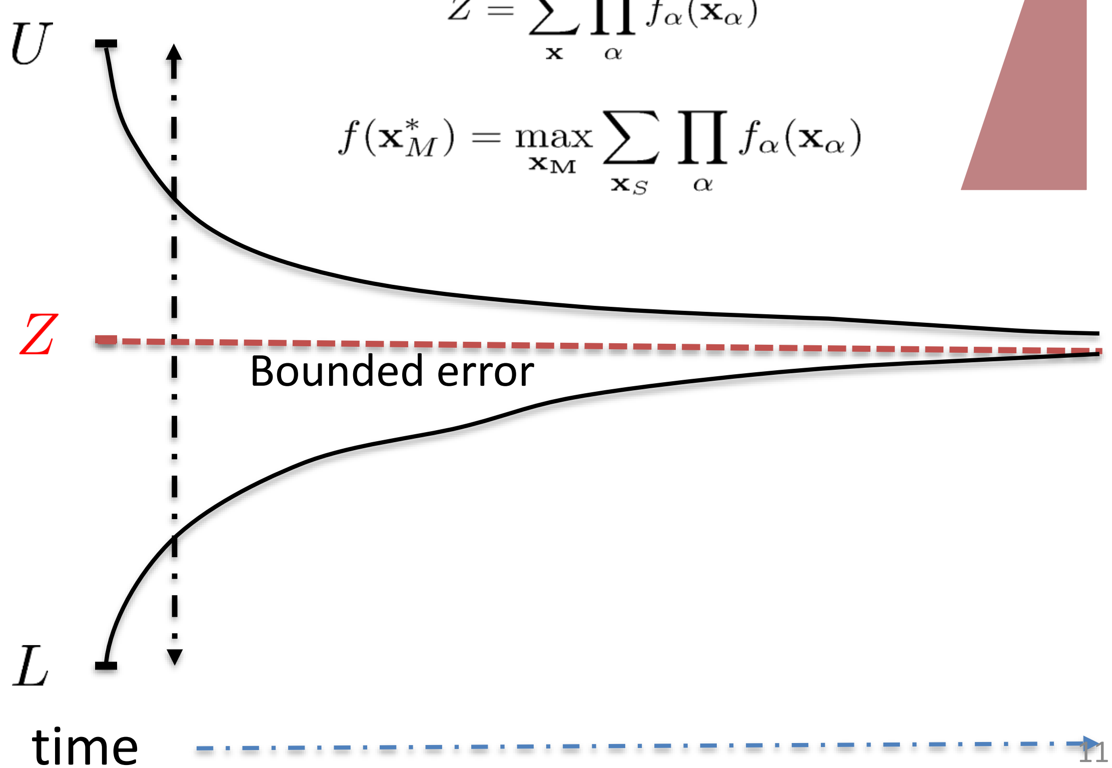
- Desiderata
  - Meaningful confidence interval
  - Responsive
  - Complete

$$f(\mathbf{x}^*) = \max_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

$$Z = \sum_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

$$f(\mathbf{x}_M^*) = \max_{\mathbf{x}_M} \sum_{\mathbf{x}_S} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$



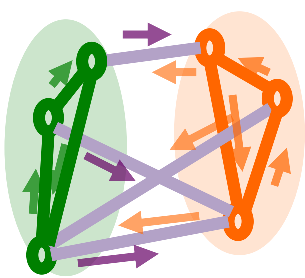


Approximate Inference

- Three major paradigms
  - Effective at different types of problems

Variational methods

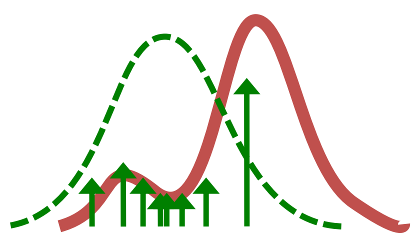
Reason over small subsets of variables at a time



- Bounds
- Responsive
- Complete

(Monte Carlo) Sampling

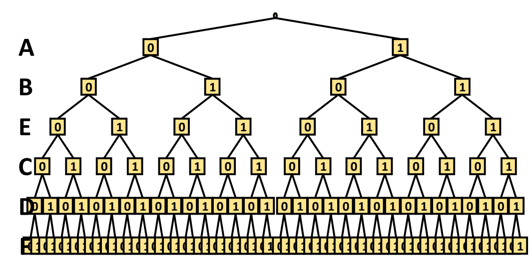
Use randomization to estimate averages over the state space



- Bounds
- Responsive
- Complete

(Heuristic) Search

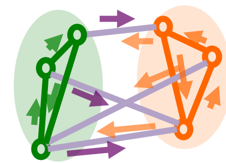
Structured enumeration over all possible states



- Bounds
- Responsive
- Complete

Combining Approaches

Inference, Variational methods



Bucket-elimination
weighted mini-bucket (WMB)

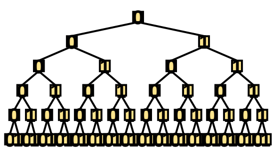
[Dechter 1999, Dechter and Rish, 2003
Liu and Ihler, ICML 2011]

provide heuristic

provide proposal

WMB-IS [Liu et al., NIPS 2015]

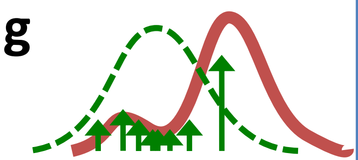
Search



refine proposal

help search

Sampling



AND/OR search (AODFS)

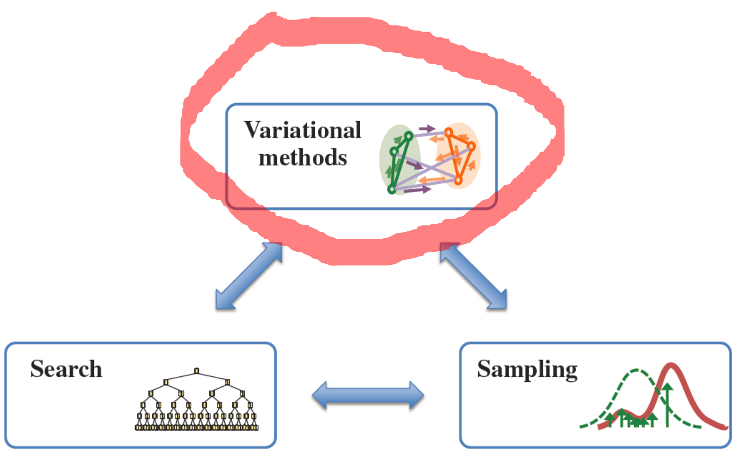
[Marinescu et al 2009, Lou et al., AAAI 2017]
Marinescu et al., IJCAI 2018]

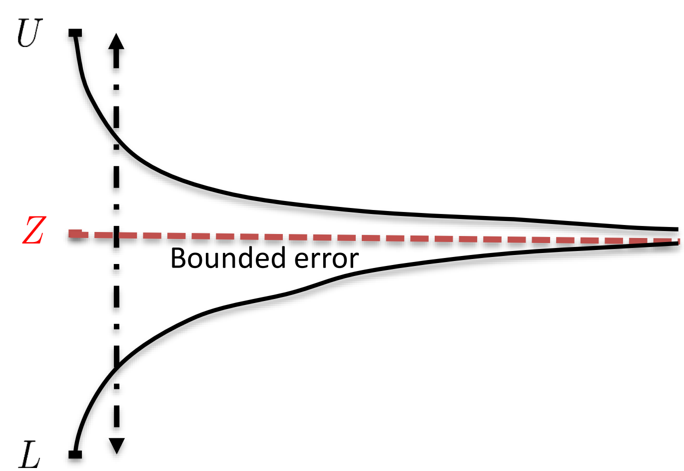
dynamic importance sampling (DIS)

[Lou et al., NIPS 2017]

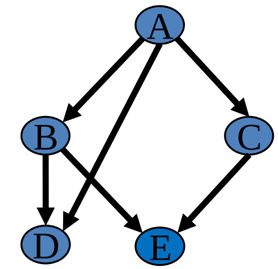
Outline

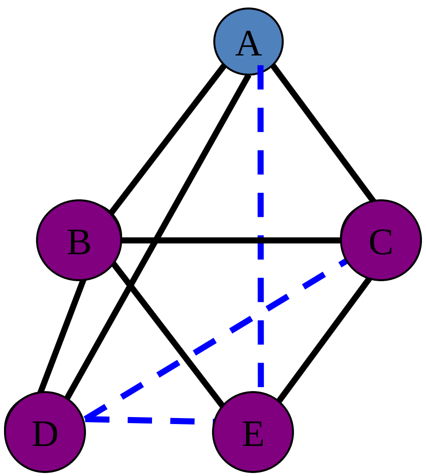
- Overview of problems and methodology
- Main paradigms of approximate reasoning:
  Variational, Search, Sampling
- Combining approaches
- Future challenges





Query 1: Belief updating: P(X|evidence)=?





"primal" graph

$$P(a|e=0) \propto P(a,e=0)=$$

$$\sum_{e=0,d,c,b} P(a)P(b|a)P(c|a)P(d|b,a)P(e|b,c)$$

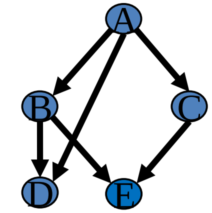
$$P(a)\sum_{e=0}\sum_{d}\sum_{c} P(c|a)\sum_{b} P(b|a)P(d|b,a)P(e|b,c)$$

Variable Elimination

$$h^B(a,d,c,e)$$

Marginals by Bucket Elimination

(Dechter 1999)



$$P(A \mid E=0) = \alpha \sum_{E=0,D,C,B} P(A)\cdot P(B \mid A)\cdot P(C \mid A)\cdot P(D \mid A,B)\cdot P(E \mid B,C)$$

$\sum_b \prod$ ← Elimination operator

bucket B: P(b|a) P(d|b,a) P(e|b,c)

bucket C: P(c|a) $\lambda_{B\to C}(a,d,c,e)$

bucket D: $\lambda_{C\to D}(a,d,e)$

bucket E: e=0 $\lambda_{D\to A}(a,e)$

bucket A: P(a) $\lambda_{E\to A}(a)$

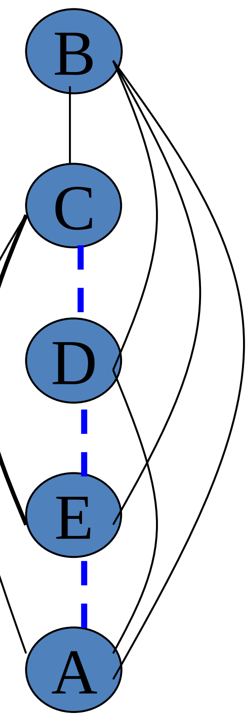
P(e=0)

P(a|e=0)

W*=4

"induced width"

(max clique size)



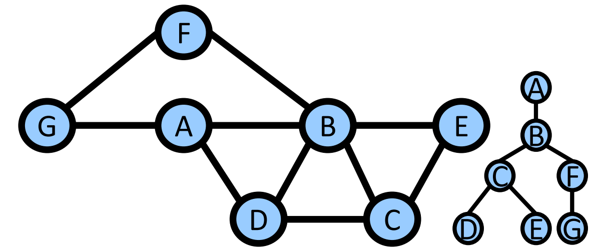
Complexity time and space $O(nk^{w*+1})$

Bucket and Mini-Bucket Elimination

$\sum_X F(X)$

X= {A,B,C,D,E,F,G}

$$F(\mathbf{X}) = f(A)f(A,B)f(A,D)f(A,G)f(B,C)f(B,D)$$
$$f(B,E)f(B,F)f(C,D)f(C,E)f(E,G)$$

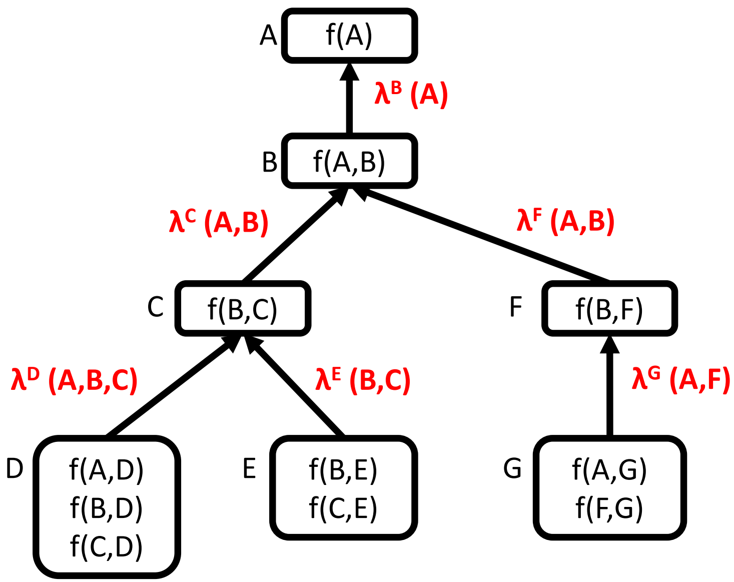


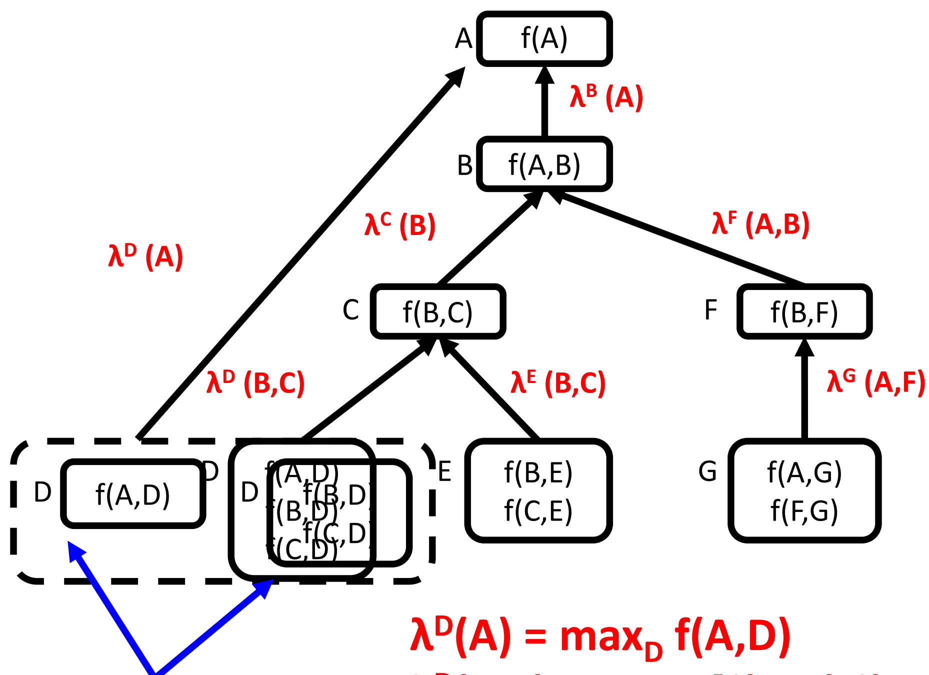
Bucket-Elimination (Dechter, 1999)

Exponential in tree-width O(nk^w)

Mini-Bucket Elimination (Dechter & Rish, 2003)

Exponential in i-bound O(nk^i)





mini-buckets

$\lambda^D(A,B,C) = \sum_D f(B,D)\, f(C,D) f(A,D)$

$\lambda^D(A) = \max_D f(A,D)$

$\lambda^D(B,C) = \max_D [f(B,D)\, f(C,D)]$

Bucket and Mini-Bucket Elimination

[Dechter 1999; Dechter & Rish, 2003]

A summation query; e.g., partition function

$$\sum_X F(X)$$

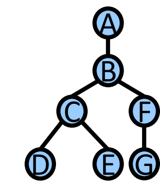
$$\{ f_1, f_2, \dots f_r, f_{r+1}, \dots f_n \}$$

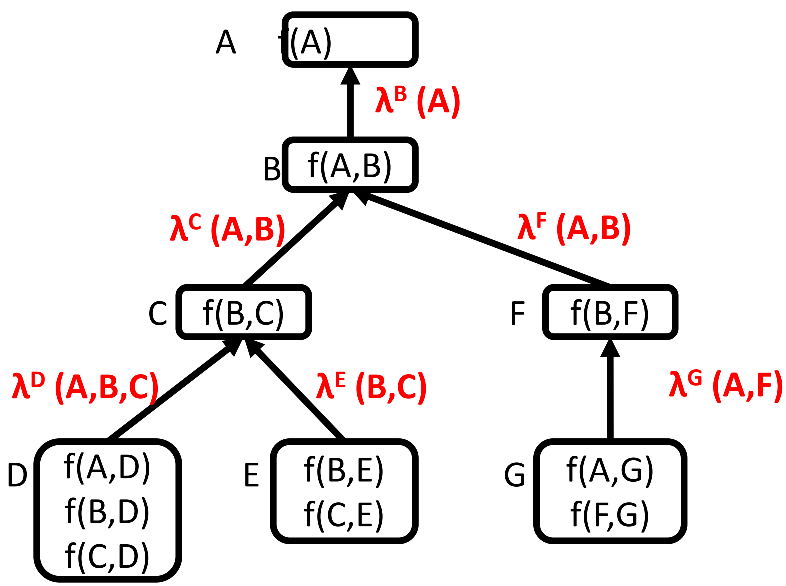
$$\lambda_X(\cdot) = \max_x \prod_{i=1}^{n} f_i(x, \dots)$$

$$\{ f_1, \dots f_r \} \qquad \{ f_{r+1}, \dots f_n \}$$

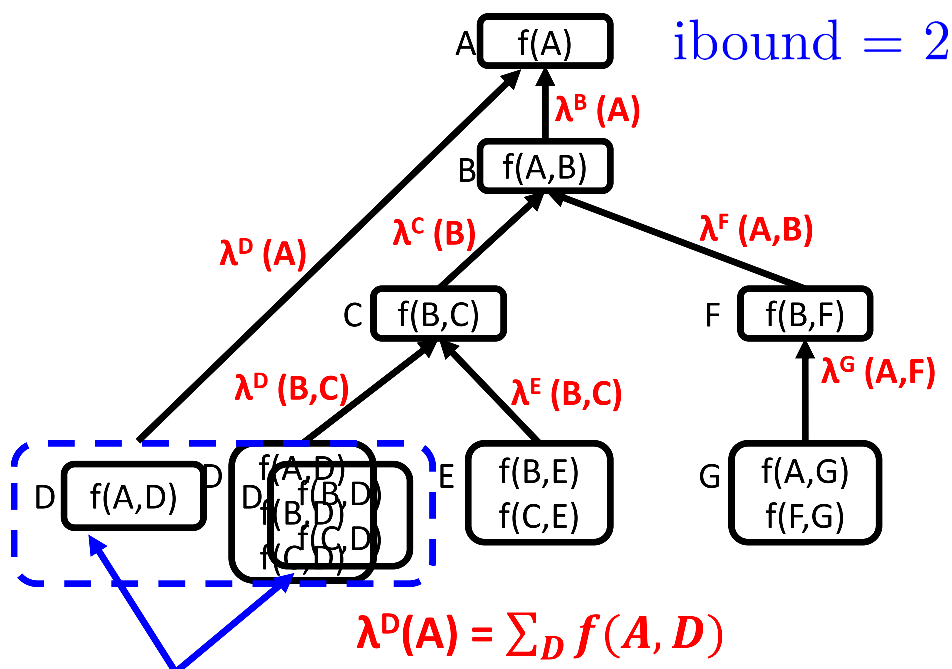
$$\lambda_{X,1}(\cdot) = \max_x \prod_{i=1}^{r} f_i(x, \dots) \qquad \lambda_{X,2}(\cdot) = \max_x \prod_{i=r+1}^{n} f_i(x, \dots)$$

$$\lambda_X(\cdot) \le \lambda_{X,1}(\cdot)\, \lambda_{X,2}(\cdot)$$





$\lambda^D(A,B,C) = \sum_D f(B,D)\, f(C,D) f(A,D)$



$\lambda^D(A) = \sum_D f(A,D)$

$\lambda^D(B,C) = \sum_D f(B,D)\, f(C,D)$

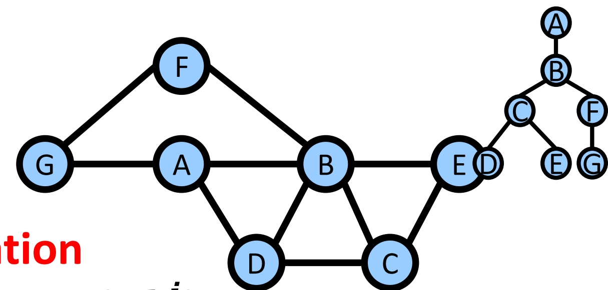
Bucket and Mini-Bucket Elimination

[Dechter 1999; Dechter & Rish, 2003]

A maximization query; e.g., MAP

$$\max_X F(X)$$

$$F(\mathbf{X}) = f(A)f(A,B)f(A,D)f(A,G)f(B,C)f(B,D)$$
$$f(B,E)f(B,F)f(C,D)f(C,E)f(E,G)$$

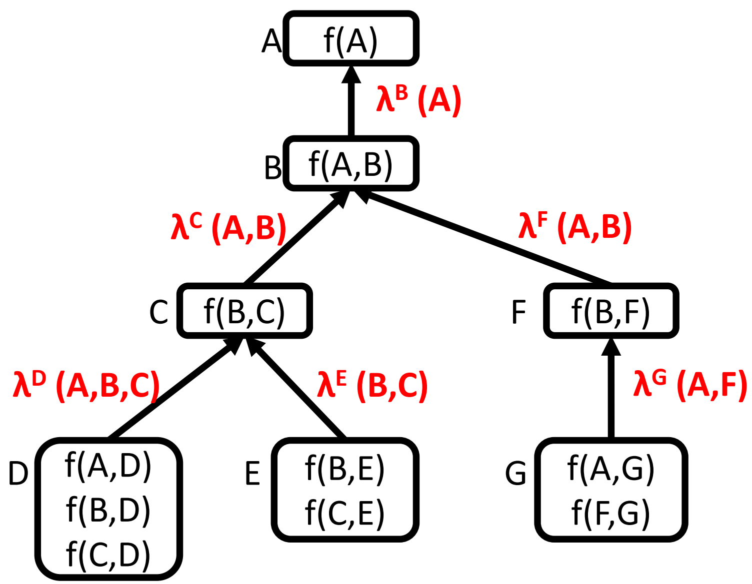


Bucket-Elimination

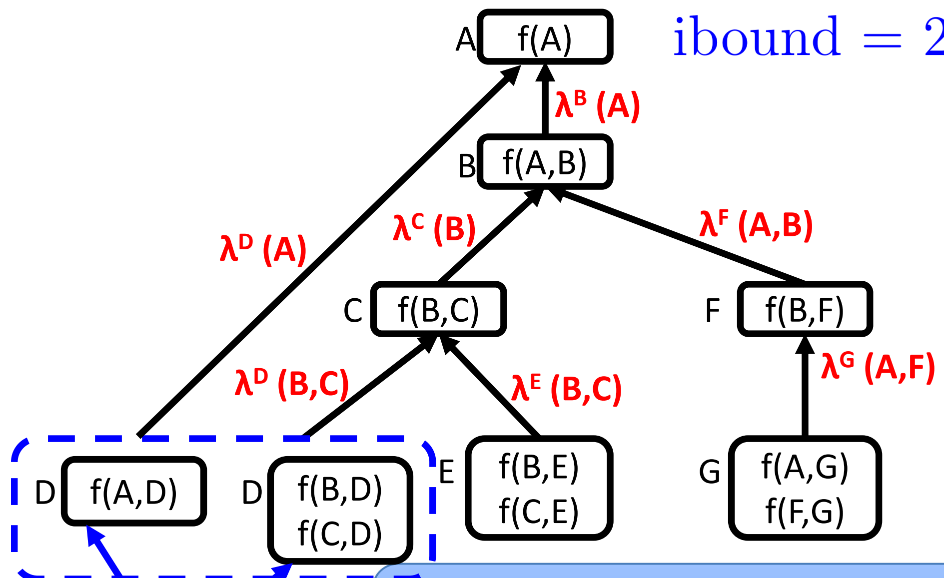
Exponential in tree-width O(nk^w)

Mini-Bucket Elimination

Exponential in i-bound O(nk^i)



λ^D(A,B,C) $= \max_D f(B,D)f(C,D)f(A,D)$



mini-buckets

λ^D(A) $= \max_D f(A,D)$

λ^D(B,C) $= \max_D f(B,D)f(C,D)$

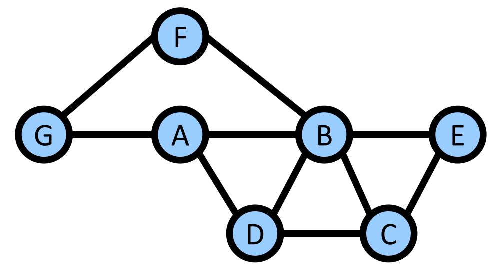
Bucket and Mini-Bucket Elimination

[Dechter 1999; Dechter & Rish, 2003, Liu & Ihler 2011]

A maximization query; e.g., MAP

$$\max_X F(X)$$

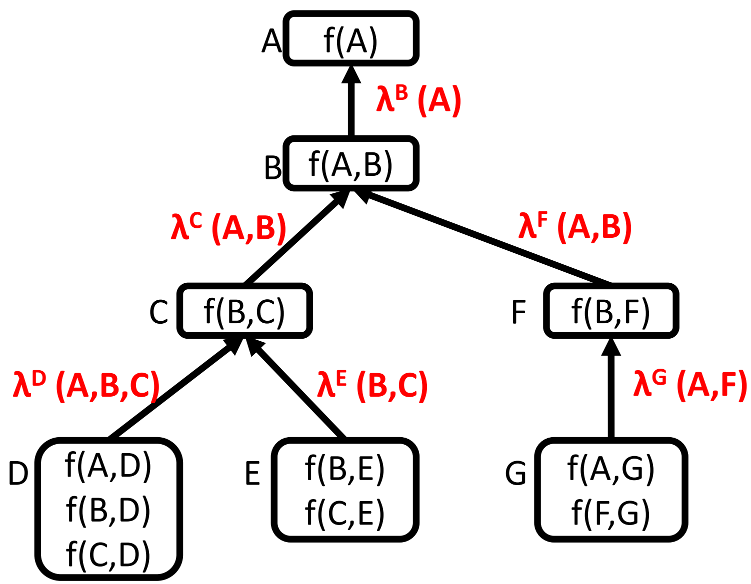
$$F(\mathbf{X}) = f(A)f(A,B)f(A,D)f(A,G)f(B,C)f(B,D)$$
$$f(B,E)f(B,F)f(C,D)f(C,E)f(E,G)$$



Bucket-Elimination

Exponential O(nk^w)

Assigning MAP value, greedily



λ^D(A,B,C) = $\max_D \boldsymbol{f(B,D)f(C,D)f(A,D)}$

$\boldsymbol{a}^* = argmax_a$ **f(a) λ^B (a)**

$\boldsymbol{b}^* = argmax_b$ **f(a,b) λ^C (a^*,b) λ^F (a^*,b)**

$\boldsymbol{c}^* = argmax_c$ **f(b,c) λ^D (a^*, b^*,C) λ^E (a^*,C)**

$\boldsymbol{f}^* = argmax_f$ **f(b,f) λ^G (a,f)**

$\boldsymbol{d}^* = argmax_c$ **f(a^*,d)f(b^*,d)f(c^*,d)**

$\boldsymbol{g}^* = argmax_f$.

$e^* = argmax_e$ **f(b^*,e)f(c^*,e)**

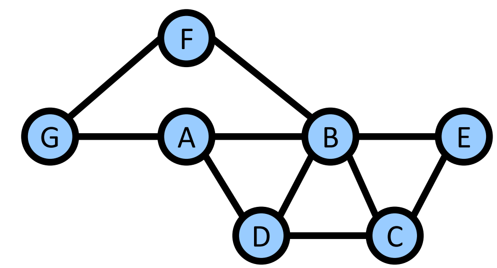
return ($\boldsymbol{a}^*$, $\boldsymbol{b}^*$, $\boldsymbol{c}^*$, $\boldsymbol{d}^*$, $\boldsymbol{e}^*$, $\boldsymbol{f}^*$, $\boldsymbol{g}^*$), exact

Bucket and Mini-Bucket Elimination

A maximization query; e.g., MAP

$$\max_{X} F(X)$$

$$F(\mathbf{X}) = f(A)f(A,B)f(A,D)f(A,G)f(B,C)f(B,D)$$
$$f(B,E)f(B,F)f(C,D)f(C,E)f(E,G)$$



$a^* = argmax_a$ **f(a) λ^B (a)λ^D(a)**

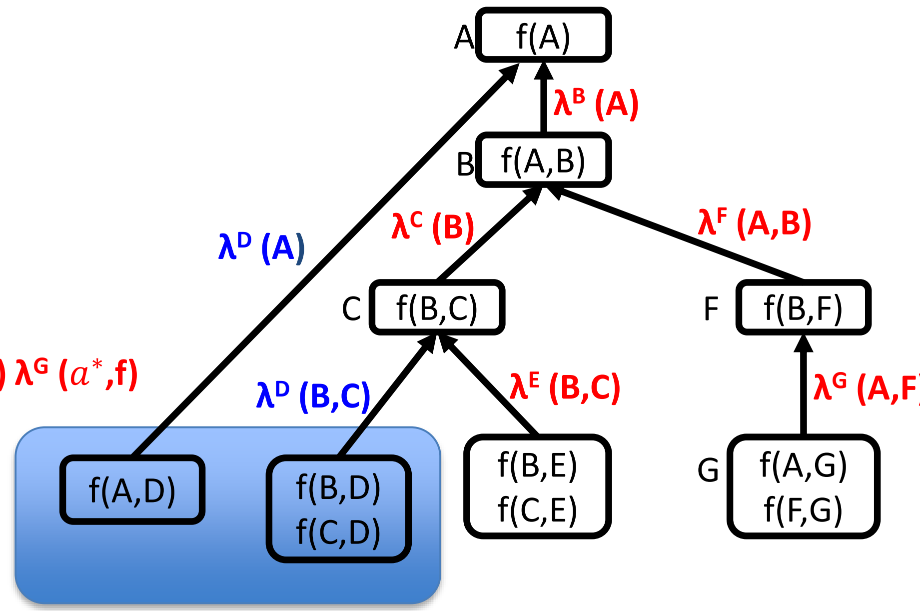
$b^* = argmax_b$ **f(a,b) λ^C (a^*,B) λ^C (a^*,B)**

$c^* = argmax_c$ **f(b,c) λ^D (b^*,C) λ^E (b^*,C)**

$f^* = argmax_f$ **f(b,f) λ^G (a^*,f)**

$d^* = argmax_c$ **f(a^*,d)f(b^*,d)f(c^*,d)**

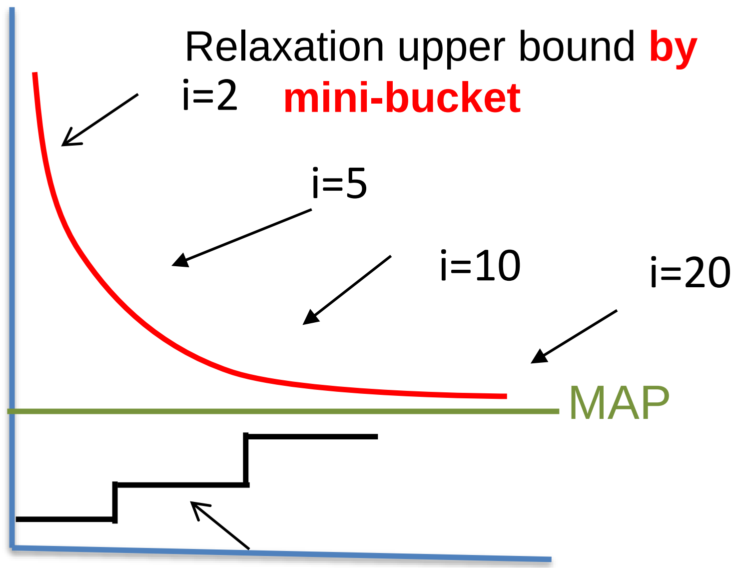
$e^* = argmax_e$ **f(b^*,e)f(c^*,e)**



returns F(a^*, b^*, c^*, d^*, e^*, f^*, g^*) a lower bound

Properties of Bucket Elimination and WMB

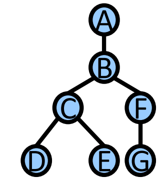
- Bounding from above and below



Consistent solutions (greedy search)

- Complexity: O(r exp(i)) time and O(exp(i)) space.
- Accuracy: determined by Upper/Lower bound.
- As i-bound increases, both accuracy and complexity increase.
- Message passing tightens bounds (next slides).

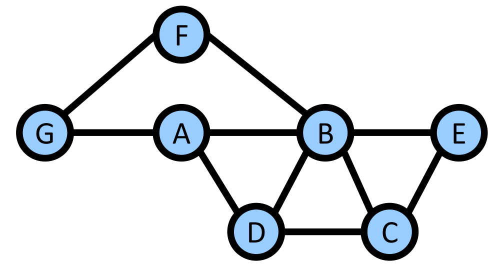
Tightening the Bound; Weighted Mini-Bucket (WMB)



[Dechter 2003, Liu & Ihler 2011]

Bounds can be tightened by optimizing weights.

$$F(\mathbf{X}) = f(A)f(A,B)f(A,D)f(A,G)f(B,C)f(B,D)$$
$$f(B,E)f(B,F)f(C,D)f(C,E)f(E,G)$$



- Holder inequality $w_1 + w_2 = 1$

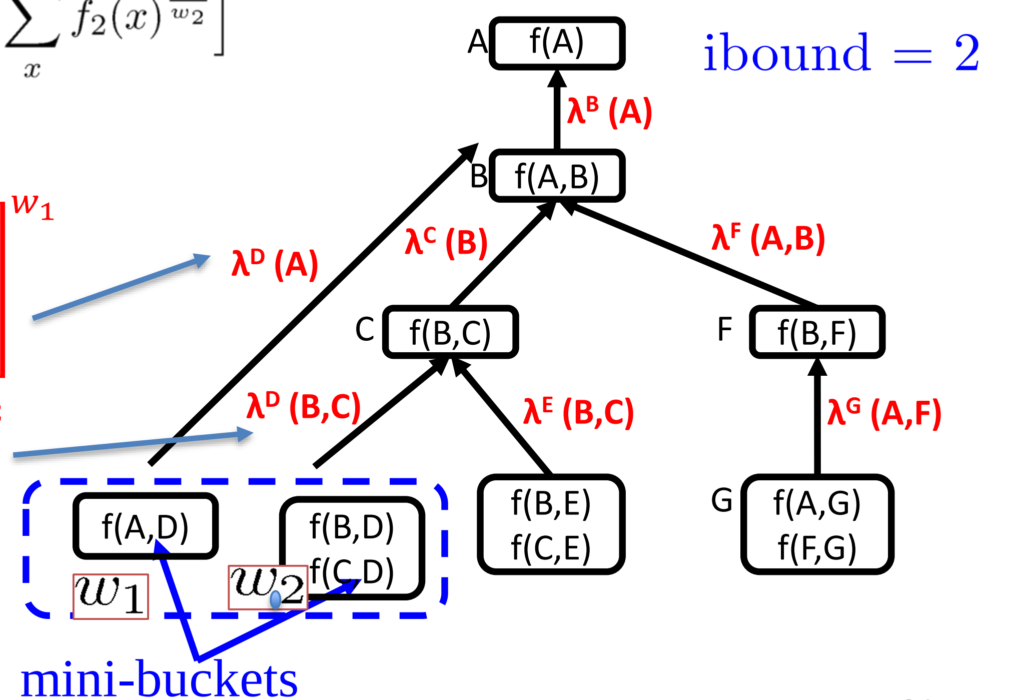
$$\sum_x f_1(x) \cdot f_2(x) \leq \left[\sum_x f_1(x)^{\frac{1}{w_1}}\right]^{w_1} \cdot \left[\sum_x f_2(x)^{\frac{1}{w_2}}\right]^{w_2}$$

$$\lambda^D(A) = \left[\sum_D f(A,D)^{\frac{1}{w}}\right]^{w_1}$$

$$\lambda^D(B,C) = \left[\sum_D f(B,D)f(C,D)^{\frac{1}{w}}\right]^{w_2}$$

($\lambda^D(A) = \sum_D f(A,D)$)

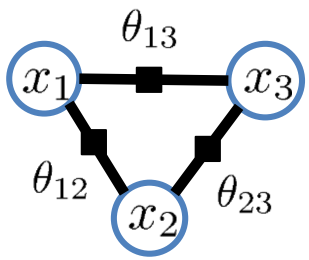
($\lambda^D(B,C) = \sum_D f(B,D)\, f(C,D)$)

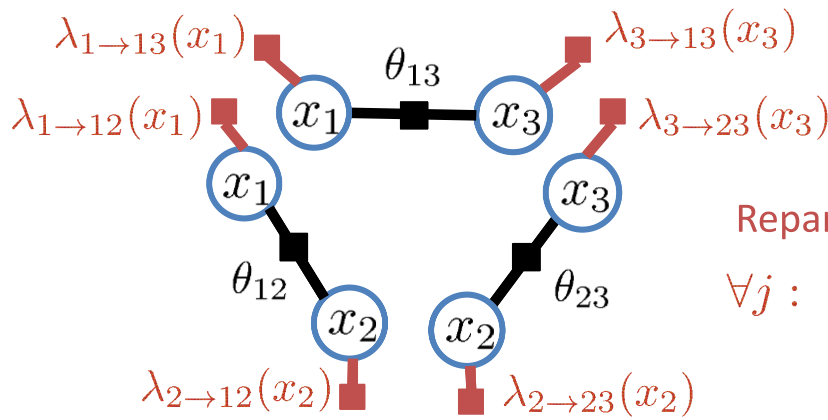


Tightening the Bound: Reparameterize fUNCTIONS

Add factors that "adjust" each local term, but cancel out in total

[Ihler,Flerova,Dechter, Otten 2012]





Reparameterization:

$$\forall j : \sum_{\alpha \ni j} \lambda_{j\to\alpha}(x_j) = 0$$

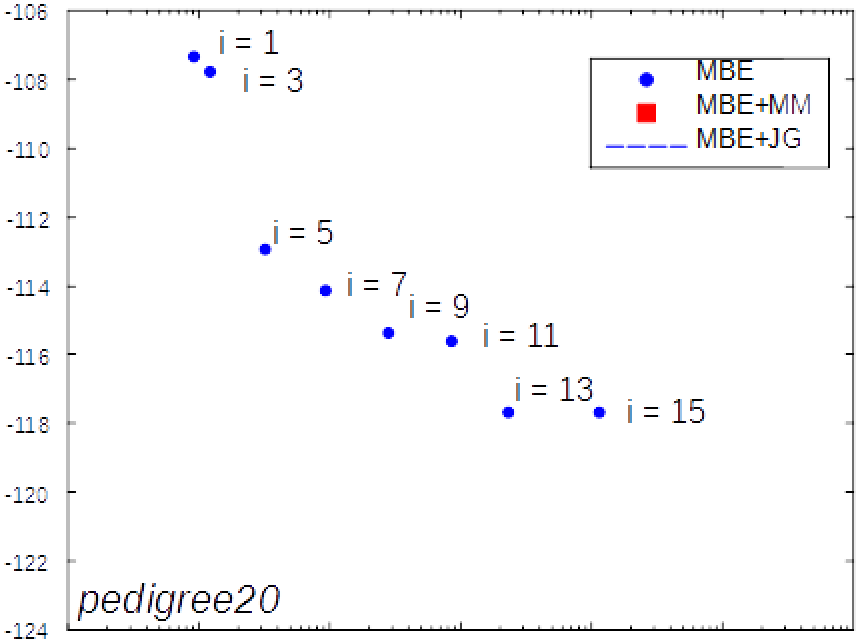
$$\log f(\mathbf{x}^*) = \max_{\mathbf{x}} \sum_{\alpha} \theta_\alpha(\mathbf{x}_\alpha) \quad \leq \min_{\{\lambda_{i\to\alpha}\}} \sum_{\alpha} \max_{\mathbf{x}_\alpha} \left[\theta_\alpha(\mathbf{x}_\alpha) + \sum_{i\in\alpha} \lambda_{i\to\alpha}(x_i) \right]$$

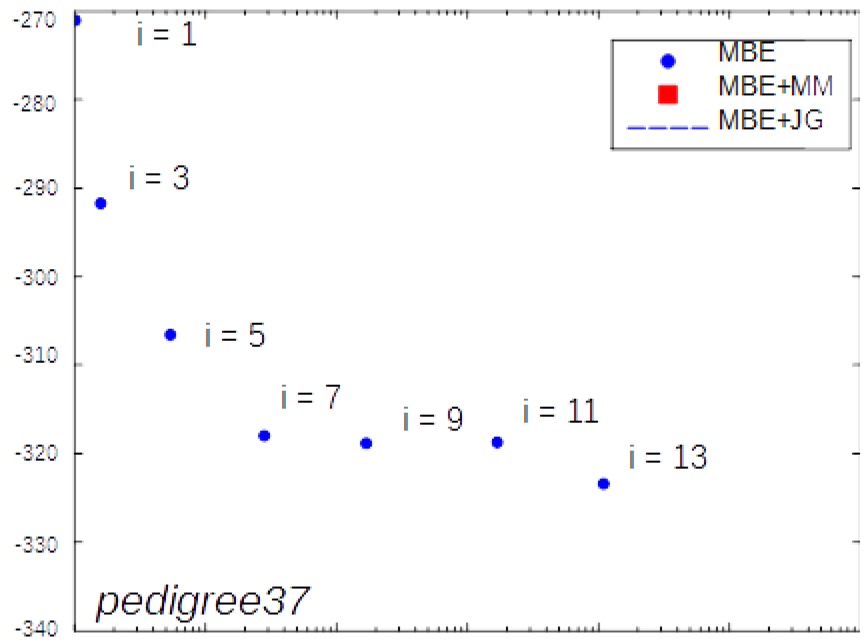
- Bound solution using decomposed optimization
- Solve independently: optimistic bound
- Tighten the bound by re-parameterization
  - Enforces lost equality constraints using Lagrange multipliers

Many names for the same class of bounds:

Dual decomposition [Komodakis et al. 2007], TRW, MPLP [Wainwright et al. 2005; Globerson & Jaakkola 2007, Soft arc consistency [Cooper & Schiex 2004] Max-sum diffusion [Warner 2007]

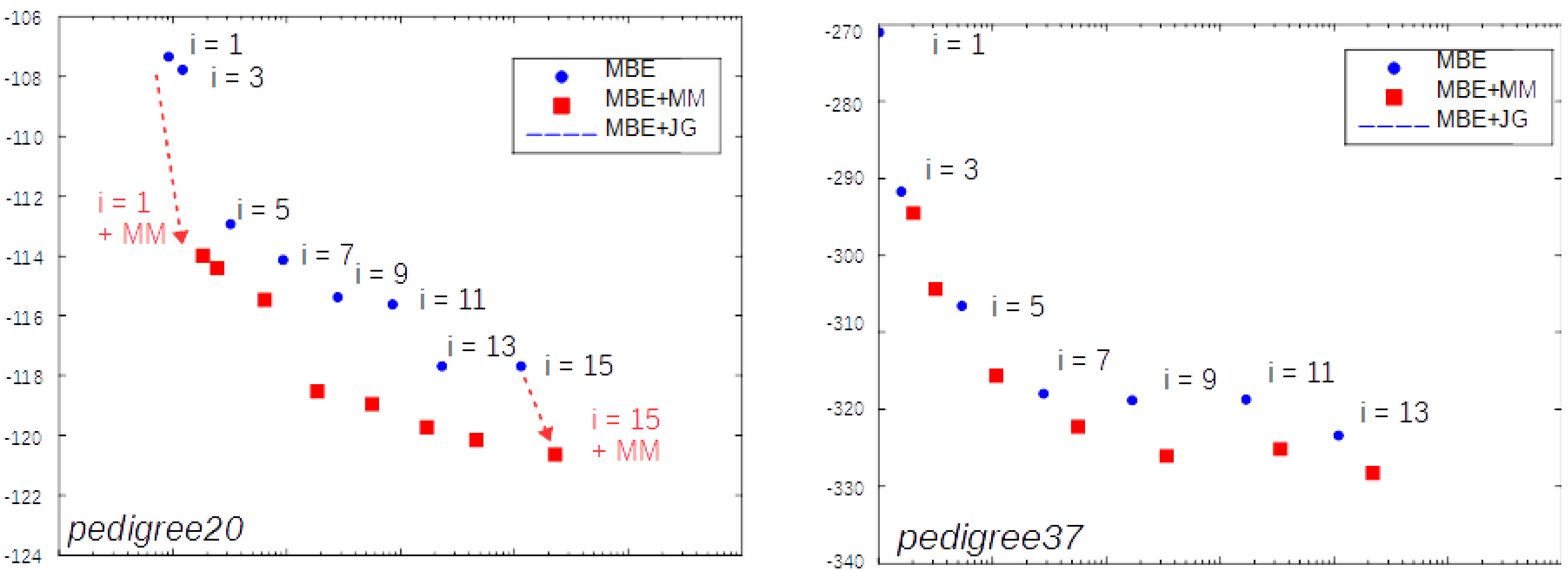
Anytime Approximation





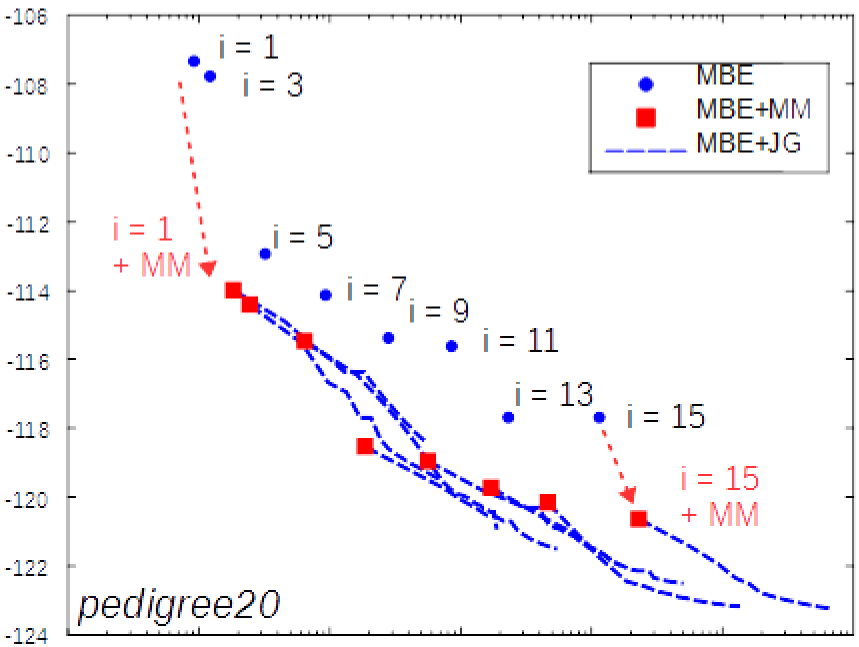
- Can tighten the bound in various ways
  - Cost-shifting (improve consistency between cliques)
  - Increase i-bound (higher order consistency)
- Simple moment-matching step improves bound significantly

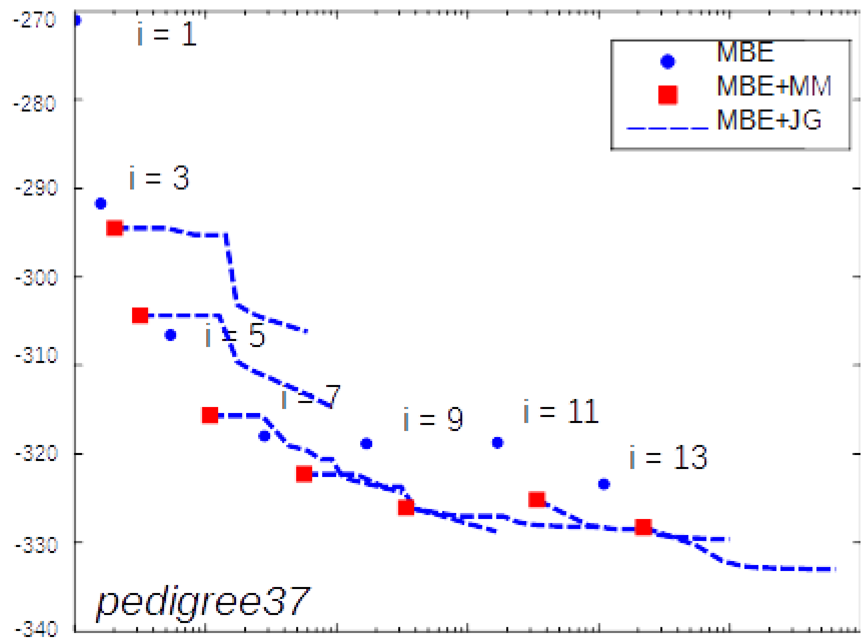
Anytime Approximation



- Can tighten the bound in various ways
  - Cost-shifting (improve consistency between cliques)
  - Increase i-bound (higher order consistency)
- Simple moment-matching step improves bound significantly

Anytime Approximation

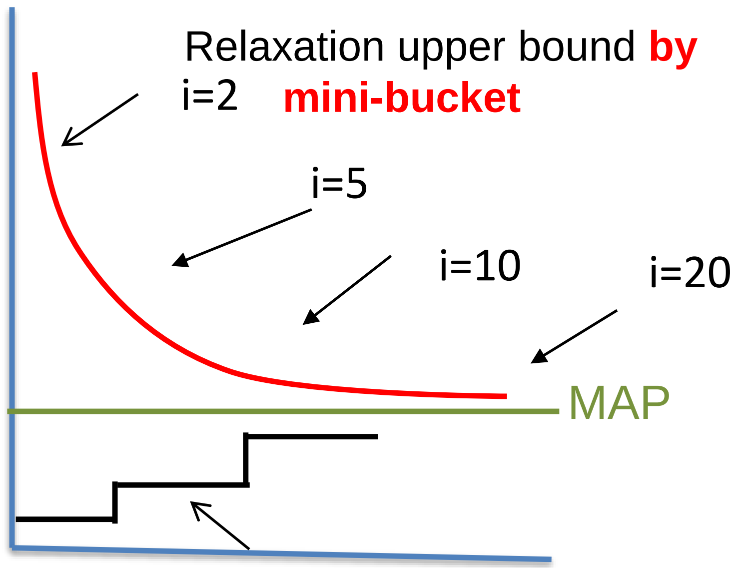




- Can tighten the bound in various ways
  - Cost-shifting (improve consistency between cliques)
  - Increase i-bound (higher order consistency)
- Simple moment-matching step improves bound significantly

Properties of Bucket Elimination and WMB

- Bounding from above and below



Consistent solutions (greedy search)

- Complexity: O(r exp(i)) time and O(exp(i)) space.
- Accuracy: determined by Upper/Lower bound.
- As i-bound increases, both accuracy and complexity increase.
- Message passing tightens bounds.

But:

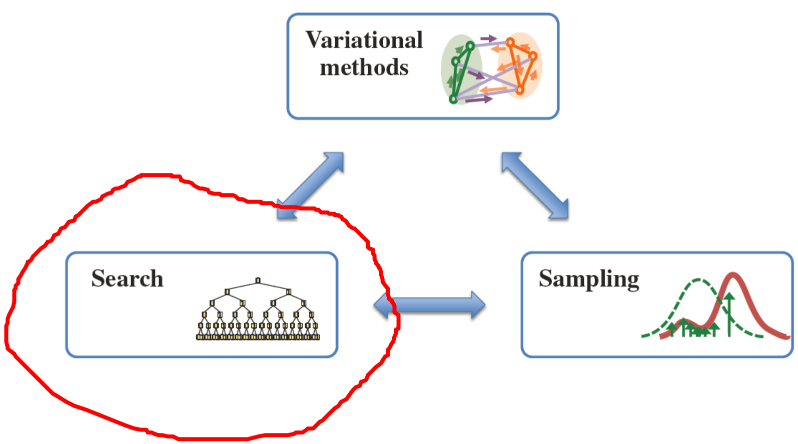
- Not anytime!

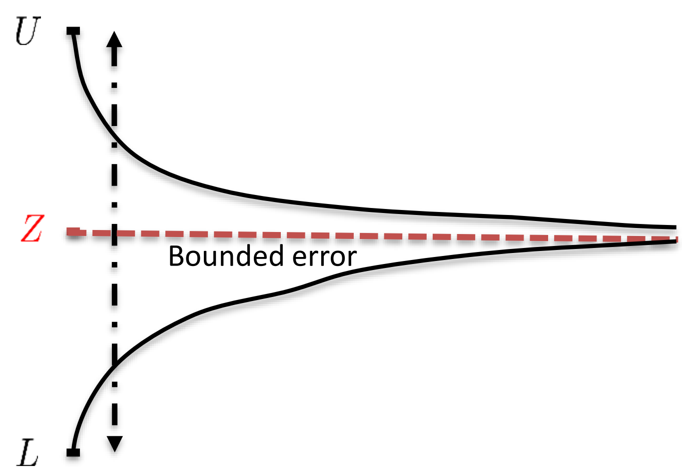
not asymp. tight w/o more memory

Outline

- Overview of problems and methodology
- Main paradigms of approximate reasoning:

  Variational, Search, Sampling
- Combining approaches
- Future challenges





Potential search spaces

A	B	f_1
0	0	2
0	1	0
1	0	1
1	1	4

A	C	f_2
0	0	3
0	1	0
1	0	0
1	1	1

A	E	f_3
0	0	0
0	1	3
1	0	2
1	1	0

A	F	f_4
0	0	2
0	1	0
1	0	0
1	1	2

B	C	f_5
0	0	0
0	1	1
1	0	2
1	1	4

B	D	f_6
0	0	4
0	1	2
1	0	1
1	1	0

B	E	f_7
0	0	3
0	1	2
1	0	1
1	1	0

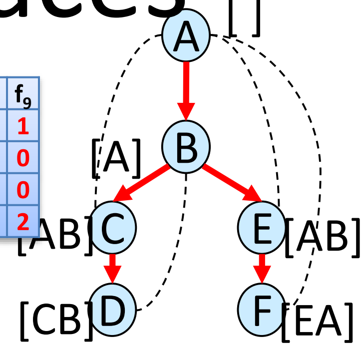
C	D	f_8
0	0	1
0	1	4
1	0	0
1	1	0

E	F	f_9
0	0	1
0	1	0
1	0	0
1	1	2

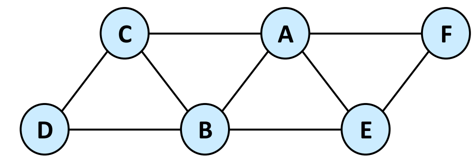
$$f(\mathbf{x}^*) = \max_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha) \qquad Z = \sum_{\mathbf{x}} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

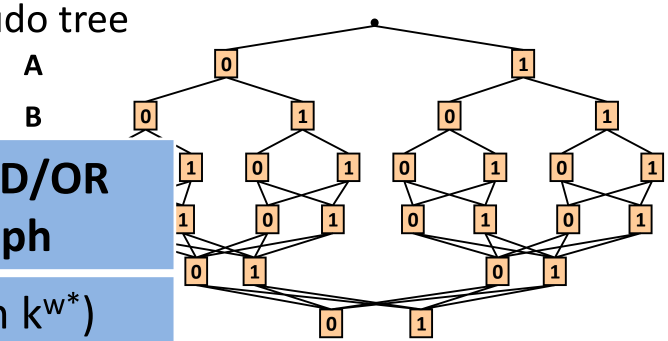
$$f(\mathbf{x}_M^*) = \max_{\mathbf{x}_M} \sum_{\mathbf{x}_S} \prod_{\alpha} f_\alpha(\mathbf{x}_\alpha)$$

Pseudo-tree



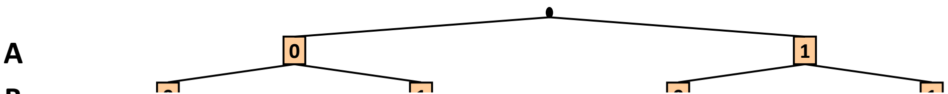
pseudo tree



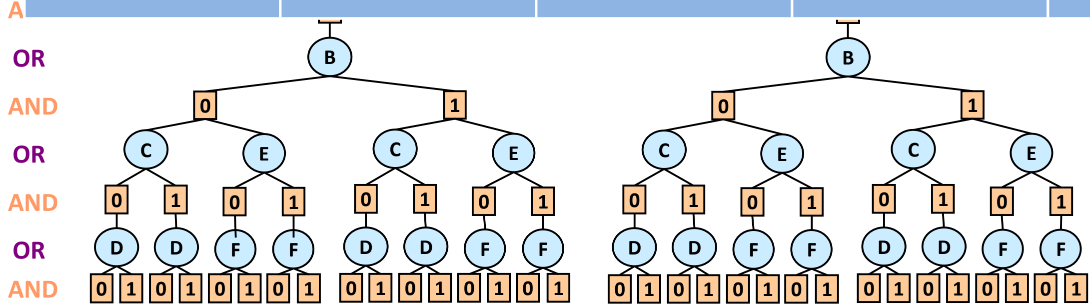


Context minimal OR search graph

28 nodes

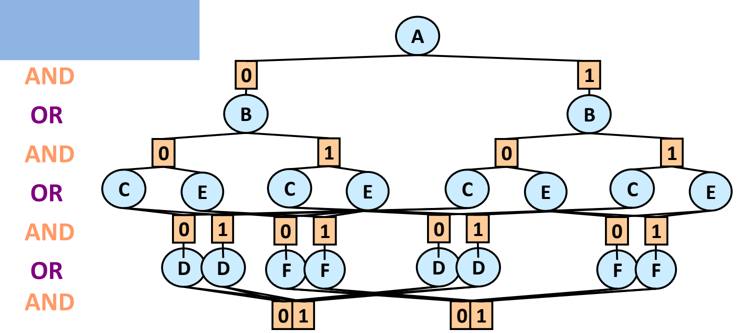


	OR tree	AND/OR tree	OR graph	AND/OR graph
time	$O(k^n)$	$O(nk^h)$	$O(n\,k^{pw^*})$	$O(n\,k^{w^*})$
memory	O(n)	O(n)	$O(n\,k^{pw^*})$	$O(n\,k^{w^*})$



Full AND/OR search tree

54 AND nodes



Context minimal AND/OR search graph

18 AND nodes

Any query can be computed over any of the search spaces

Cost of a Solution Tree

$P(E|A,B)$

A	B	E=0	E=1
0	0	.4	.6
0	1	.5	.5
1	0	.7	.3
1	1	.2	.8

Evidence: E=0

$P(B|A)$

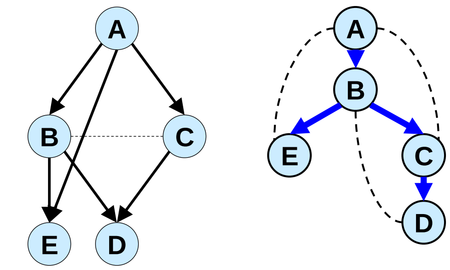
A	B=0	B=1
0	.4	.6
1	.1	.9

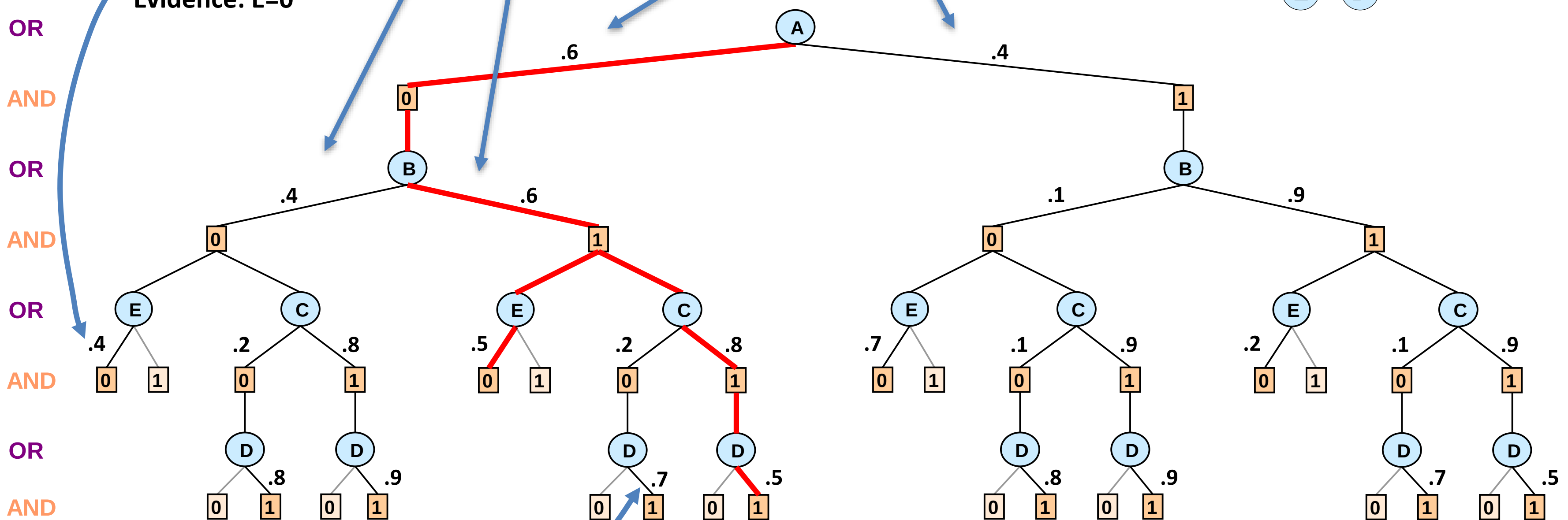
$P(C|A)$

A	C=0	C=1
0	.2	.8
1	.7	.3

$P(A)$

A	P(A)
0	.6
1	.4





$P(D|B,C)$

B	C	D=0	D=1
0	0	.2	.8
0	1	.1	.9
1	0	.3	.7
1	1	.5	.5

Evidence: D=1

Cost of the solution tree: the product of weights on its arcs

Cost of (A=0,B=1,C=1,D=1,E=0) = $0.6 \cdot 0.6 \cdot 0.5 \cdot 0.8 \cdot 0.5 = 0.0720$

Value of a Node (e.g., Probability of Evidence)

$P(E|A,B)$

A	B	E=0	E=1
0	0	.4	.6
0	1	.5	.5
1	0	.7	.3
1	1	.2	.8

Evidence: E=0

$P(B|A)$

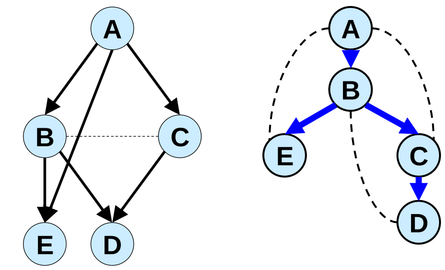
A	B=0	B=1
0	.4	.6
1	.1	.9

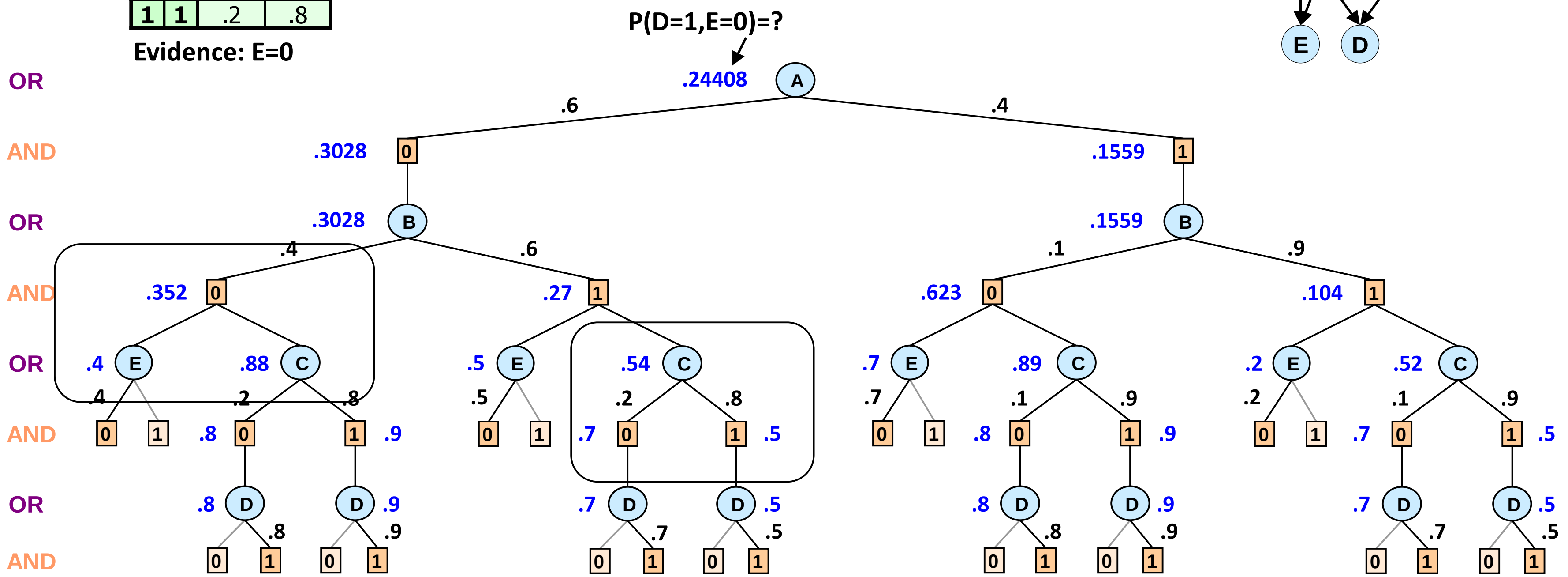
$P(C|A)$

A	C=0	C=1
0	.2	.8
1	.7	.3

$P(A)$

A	P(A)
0	.6
1	.4





$P(D|B,C)$

B	C	D=0	D=1
0	0	.2	.8
0	1	.1	.9
1	0	.3	.7
1	1	.5	.5

Evidence: D=1

Value of node = updated belief for sub-problem below

AND node: product $\prod_{n' \in children(n)} v(n')$

OR node: Marginalization by summation $\sum_{n' \in children(n)} w(n,n')\, v(n')$

Answering Queries: Sum-Product (Belief Updating)

$P(E|A,B)$

A	B	E=0	E=1
0	0	.4	.6
0	1	.5	.5
1	0	.7	.3
1	1	.2	.8

Evidence: E=0

$P(B|A)$

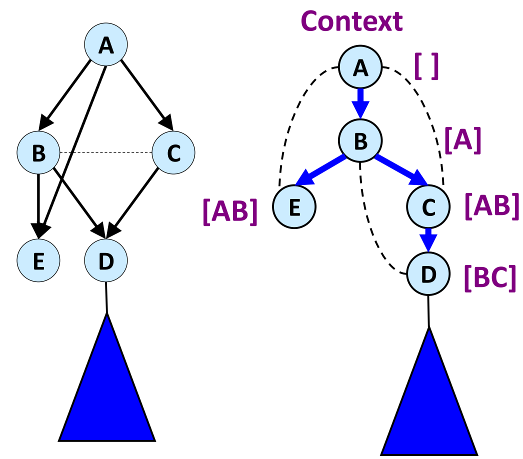
A	B=0	B=1
0	.4	.6
1	.1	.9

$P(C|A)$

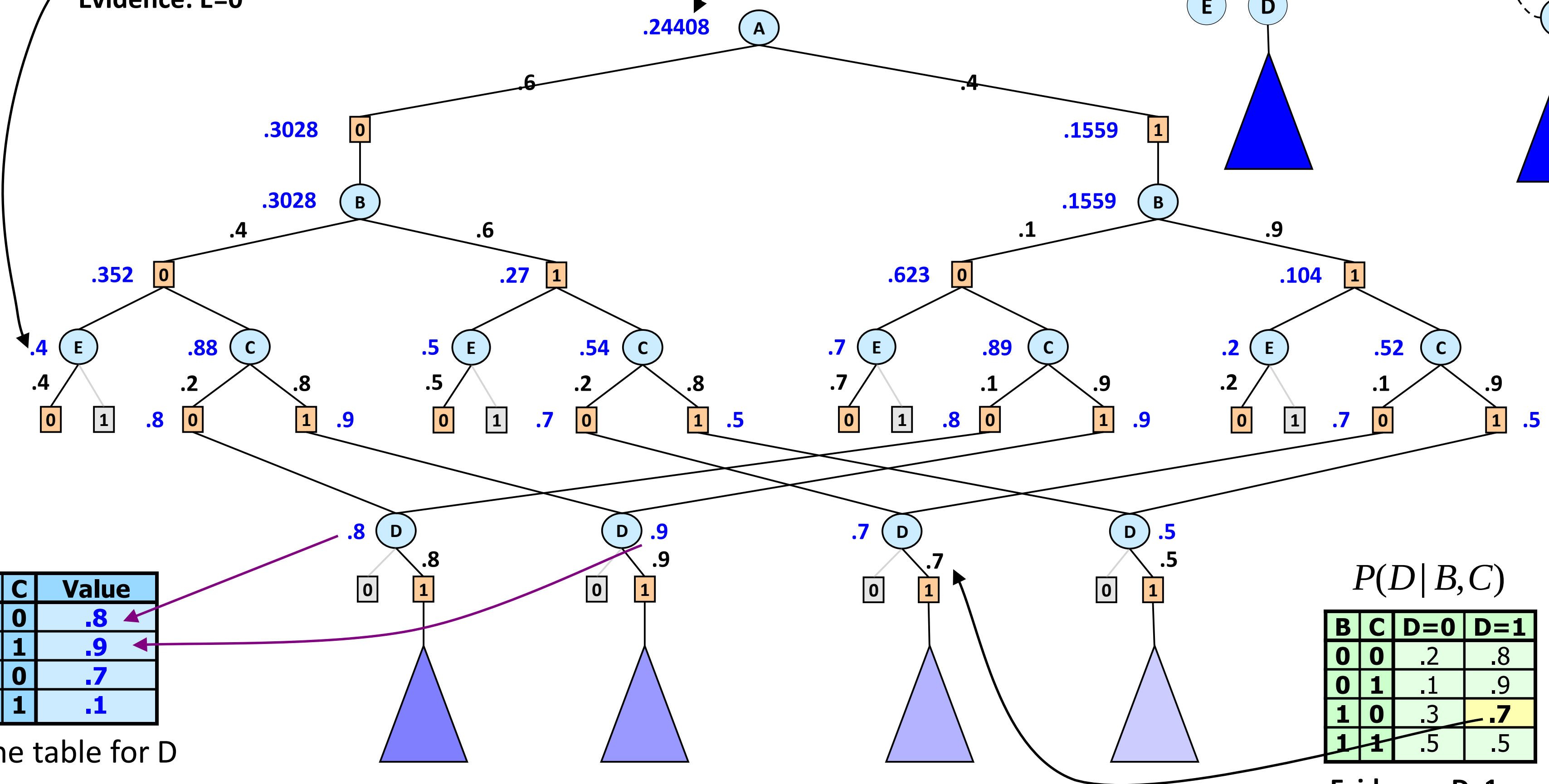
A	C=0	C=1
0	.2	.8
1	.7	.3

$P(A)$

A	P(A)
0	.6
1	.4



Result: P(D=1,E=0)



B	C	Value
0	0	.8
0	1	.9
1	0	.7
1	1	.1

Cache table for D

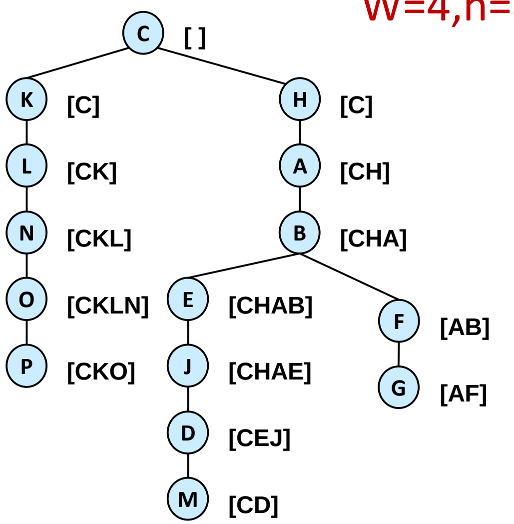
$P(D|B,C)$

B	C	D=0	D=1
0	0	.2	.8
0	1	.1	.9
1	0	.3	.7
1	1	.5	.5

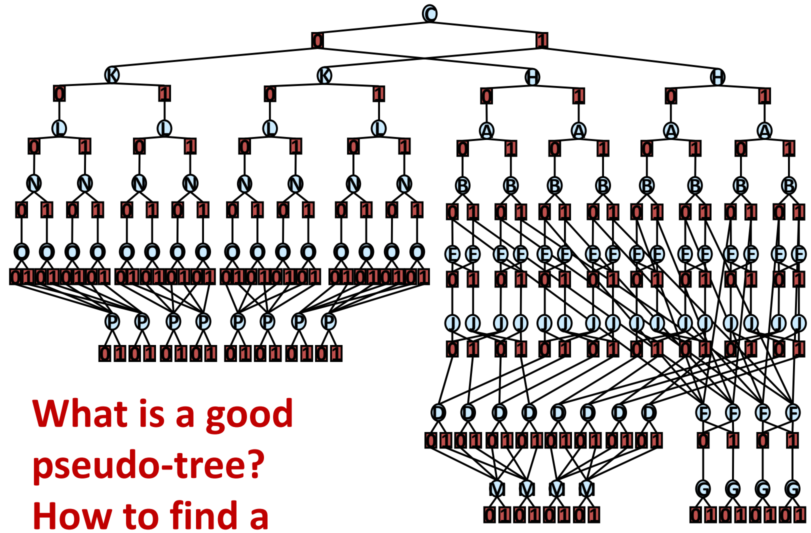
Evidence: D=1

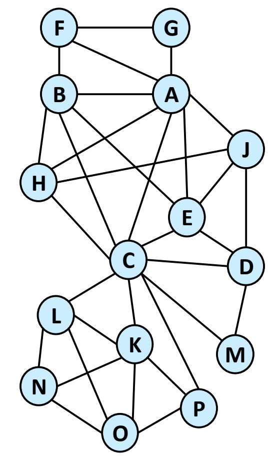
The Impact of the Pseudo-Tree

W=4,h=8



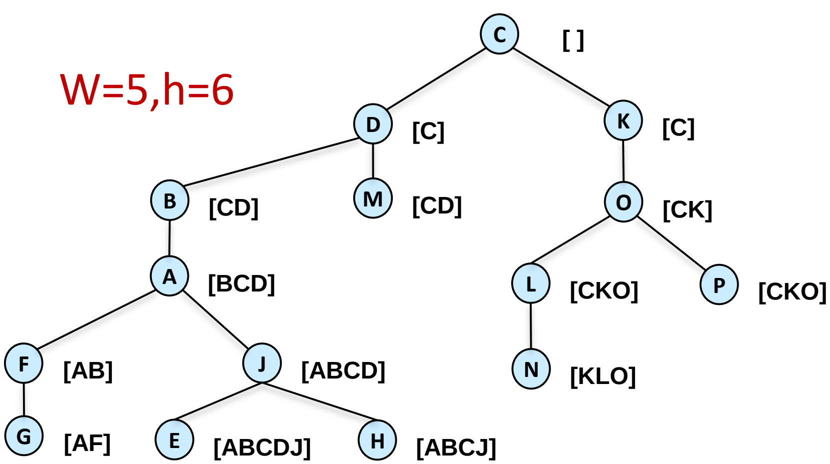
(C K H A B E J L N O D P M F G)

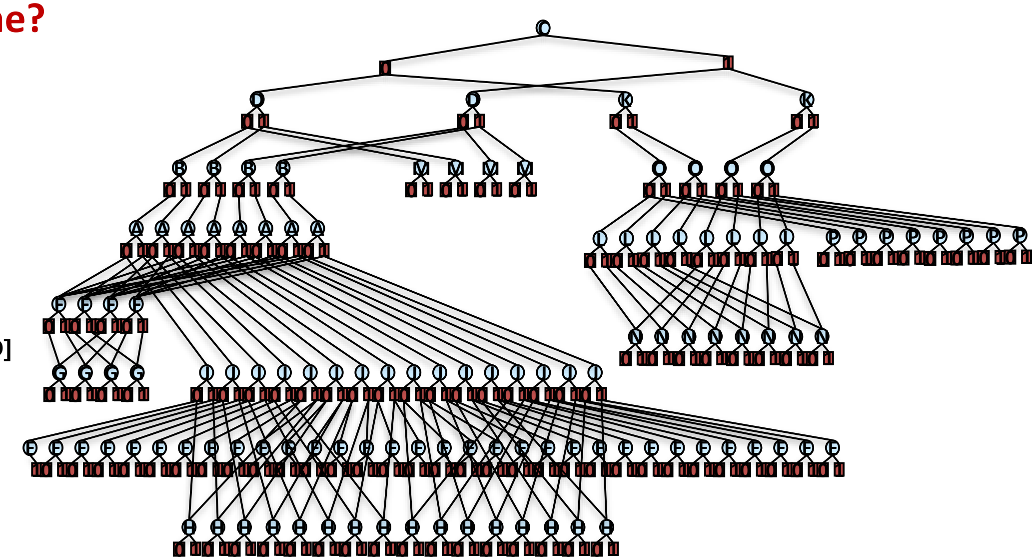




What is a good pseudo-tree?
How to find a good one?

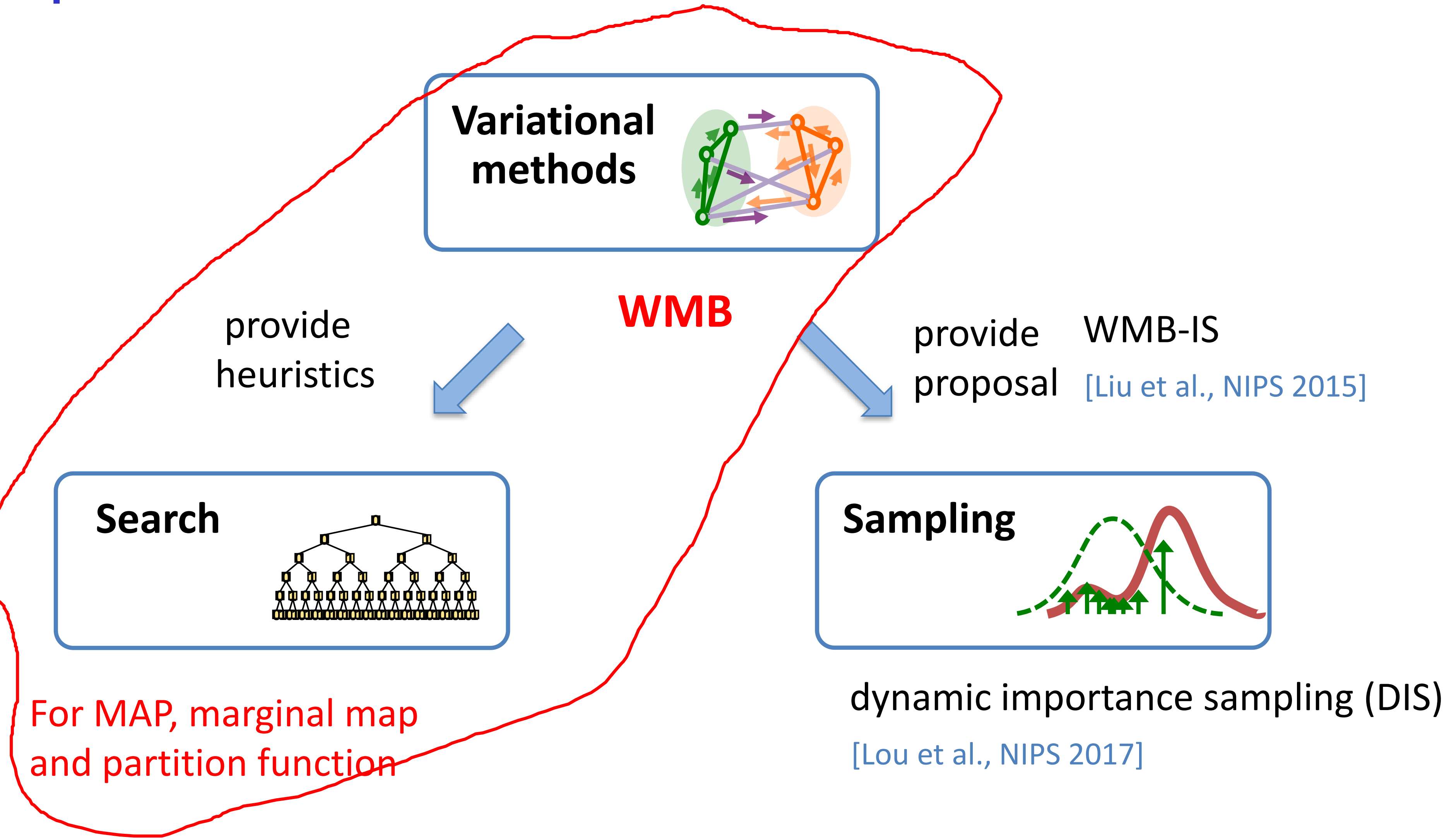
W=5,h=6





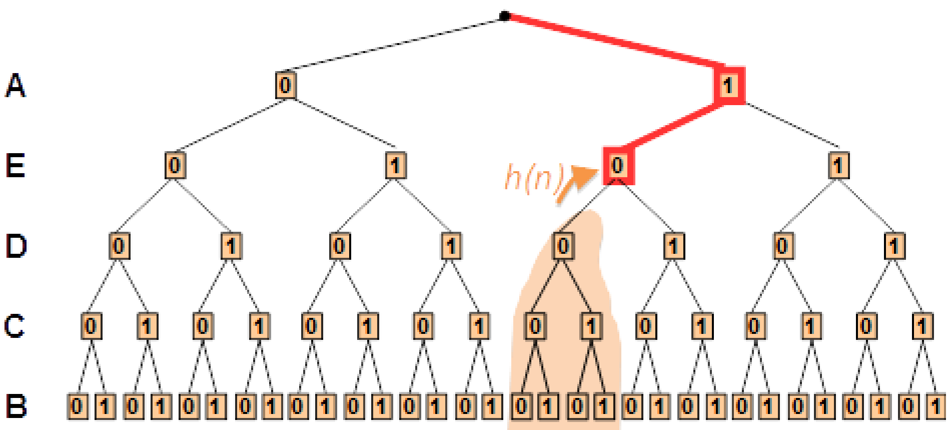
(C D K B A O M L N P J H E F G)

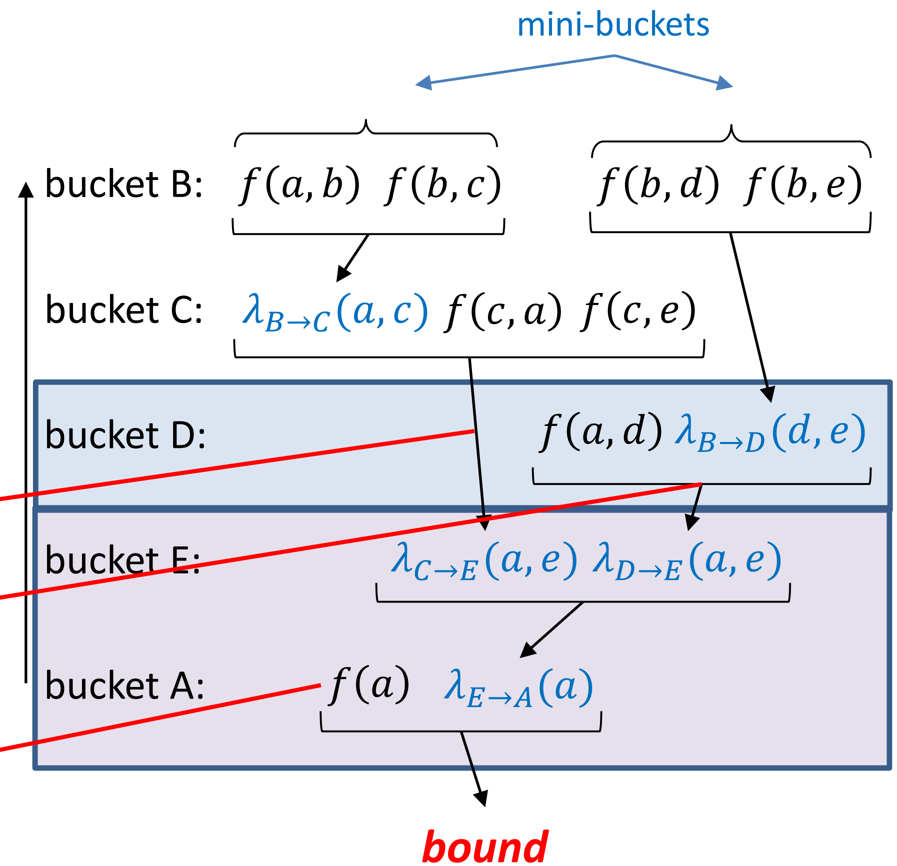
Combining Approaches: Search +Variational



WMBE Heuristic Guides OR Search

Given a partial assignment, $[\hat{a} = 1, \hat{e} = 0]$
(weighted) mini-bucket gives an admissible heuristic:





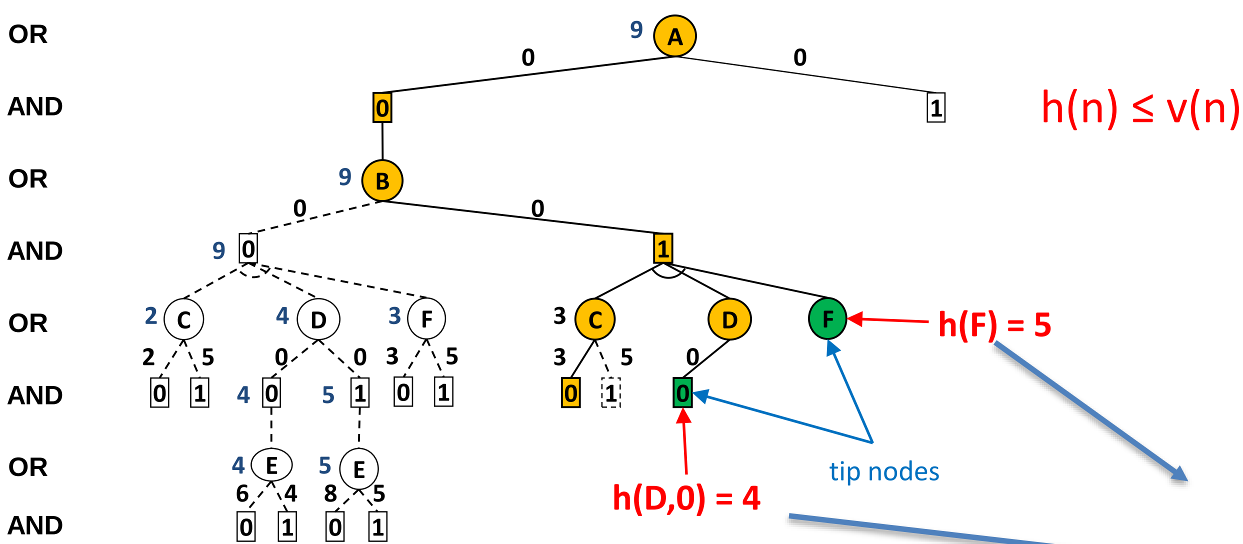
"cost to go":

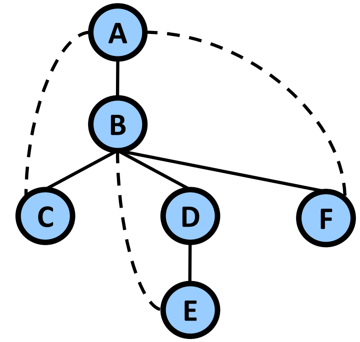
$$\tilde{h}(\hat{a},\hat{e},D) = \lambda_{C\to E}(\hat{a},\hat{e}) + f(\hat{a},D) + \lambda_{B\to D}(D,\hat{e})$$

"cost so far":

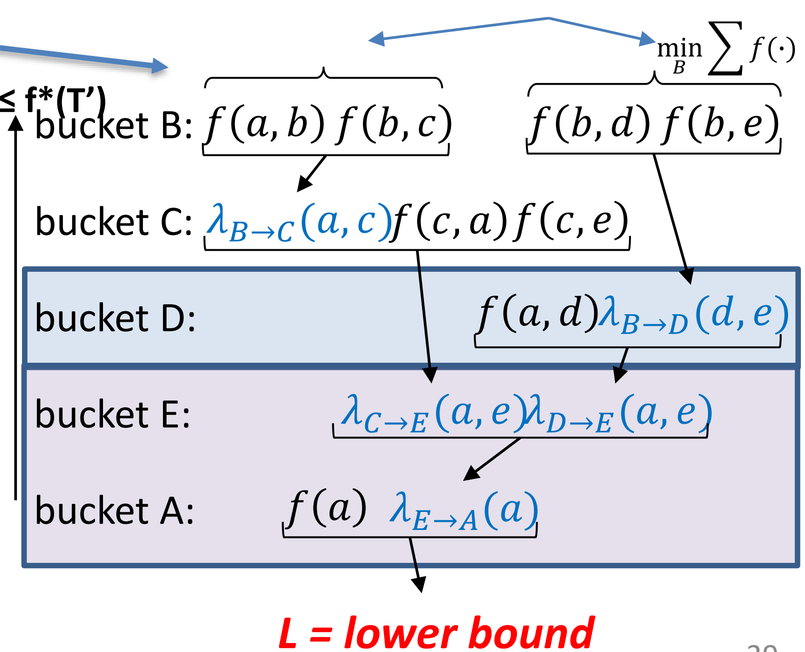
$$g(\hat{a},\hat{e},D) = f(A = \hat{a})$$

MBE Heuristic Guides AO Search





$f(T') = w(A,0) + w(B,1) + w(C,0) + w(D,0) + h(D,0) + h(F) = 12 \leq f^*(T')$



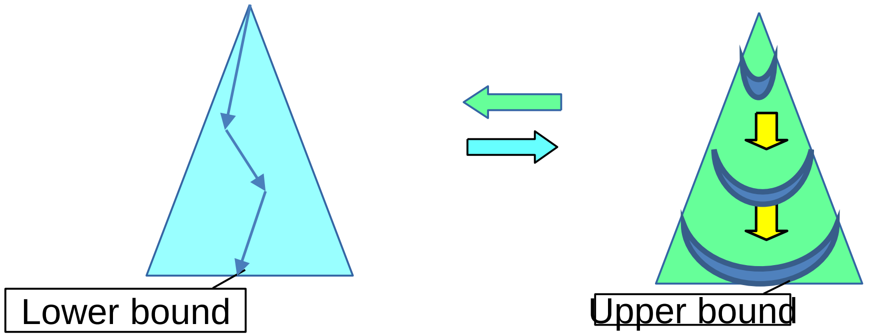
Exploiting Heuristic Search Principles

- **Weighted Heuristic:** [Pohl 1970]

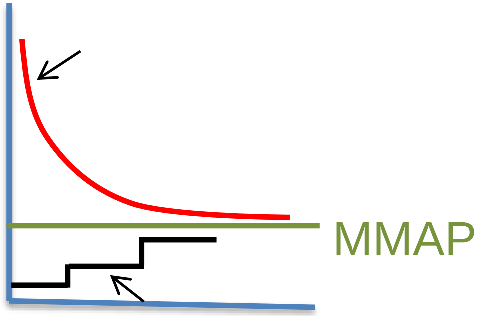
$$f(n)=g(n)+w \cdot h(n)$$

  - **Guaranteed w-optimal solution, cost $C \leq w \cdot C^*$**

- **Interleaving Best + Depth-First search**



Goal: anytime bounds
And anytime solution



Anytime Bounds for Optimization

(2001-2017)

- Kask and Dechter. Articial Intelligence, 2001,
- Marinescu and Dechter. Artif. Intell. (2009)
- Otten and Dechter: AI Commun. (2012)
- **Silberstein, et. Al (2013): Bioinformatics (2013)**
- Flerova, Marinescu and Dechter Articial Intelligence, 2016.
- Flerova, Marinescu and Dechter, (2016) JAIR", 2016
- Lam, Kask, Larrosa, and Dechter. " JAIR 2017.
- Otten and Dechter (JAIR 2017)
- Marinescu, Lee, Dechter and Ihler. JAIR, 2018.

- **2006** 
- **2008** 
- **2011** 
- **2014** 

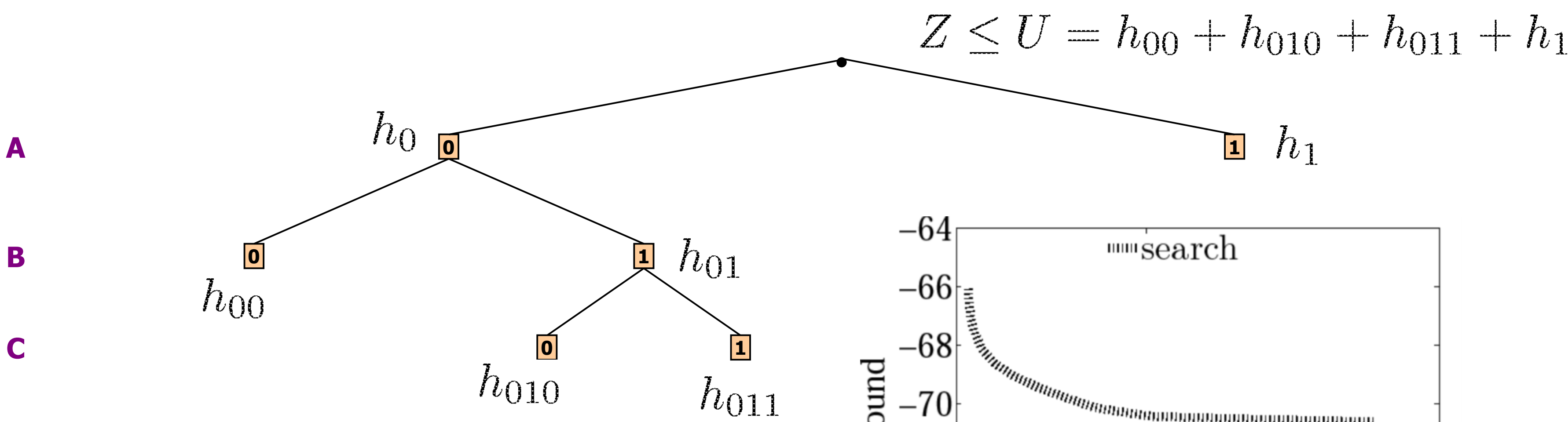
UAI Competition MPE/MAP

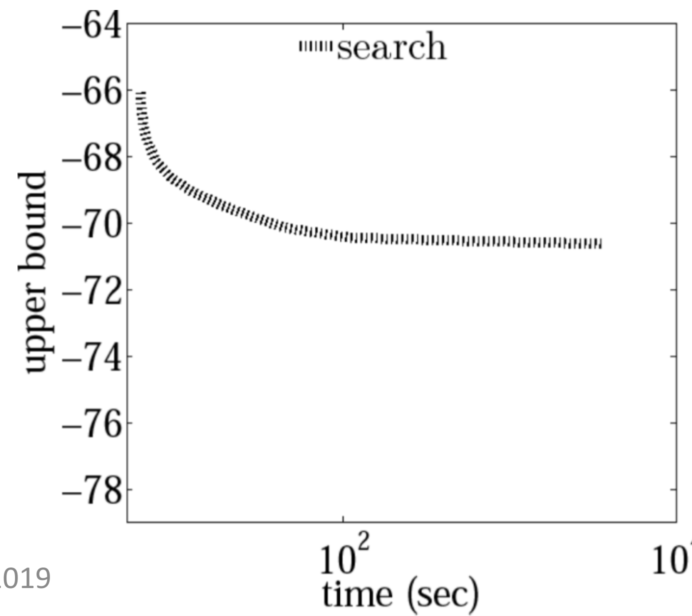
Winning first or second place UAI competitions, 2006, 2008, 2011, 2014, 2016

Anytime Bounds for Summation

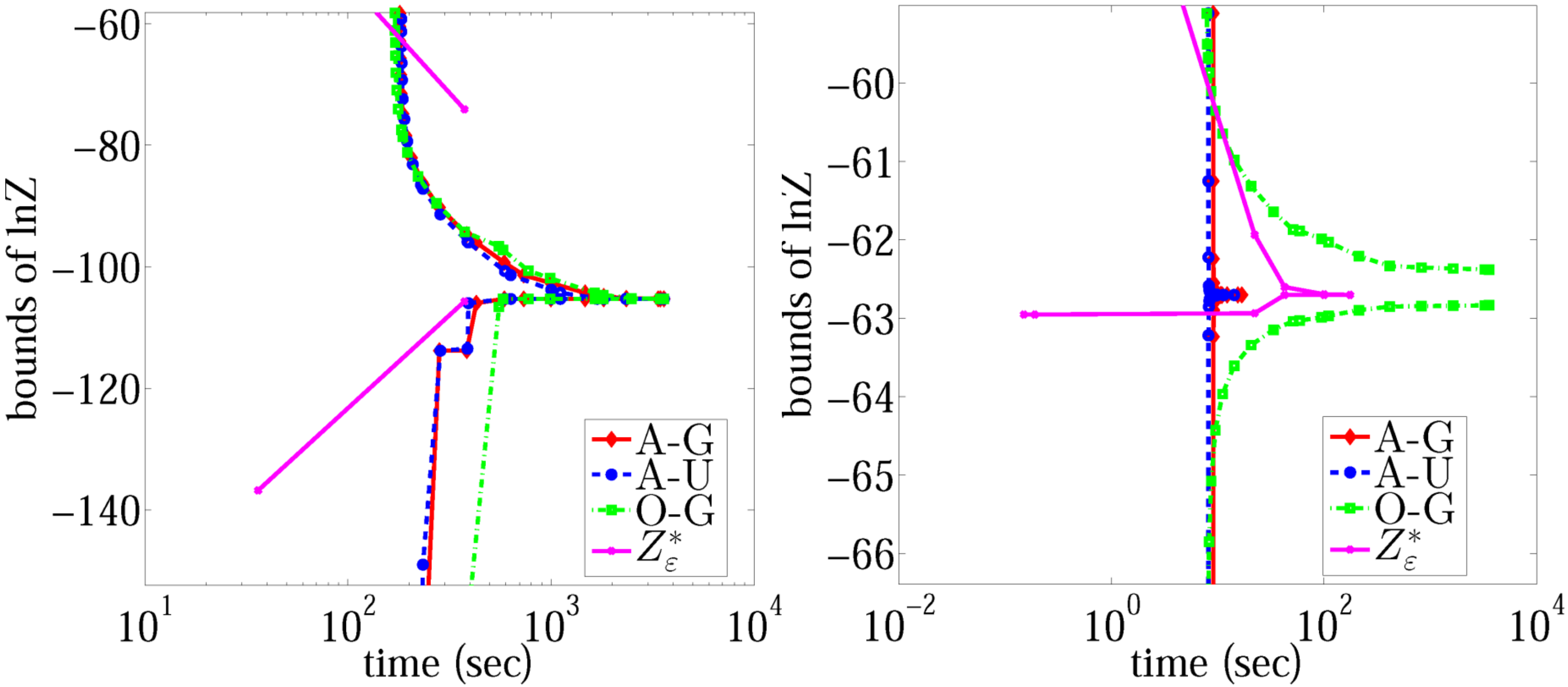
(AAAI'17, AAAI'18: Lou, Dechter and Ihler)

- Heuristic search for summation
  - Heuristic function upper bounds value (sum below) at any node
  - Expand tree and compute updated bounds, using a priority gap





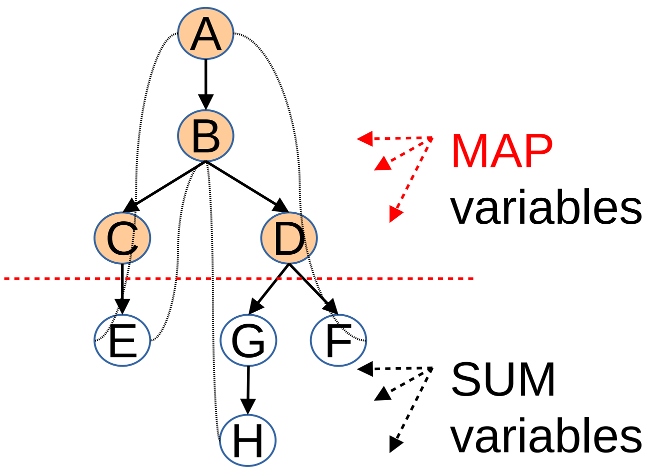
Anytime Behavior of AOBFS



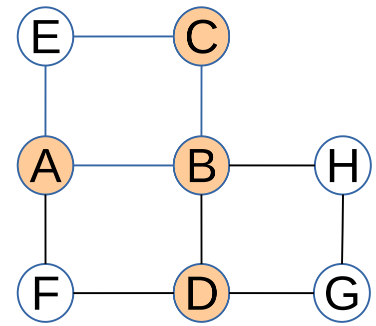
(a) PIC’11/queen5_5_4 (b) Protein/1g6x

Anytime Bounds for Marginal MAP

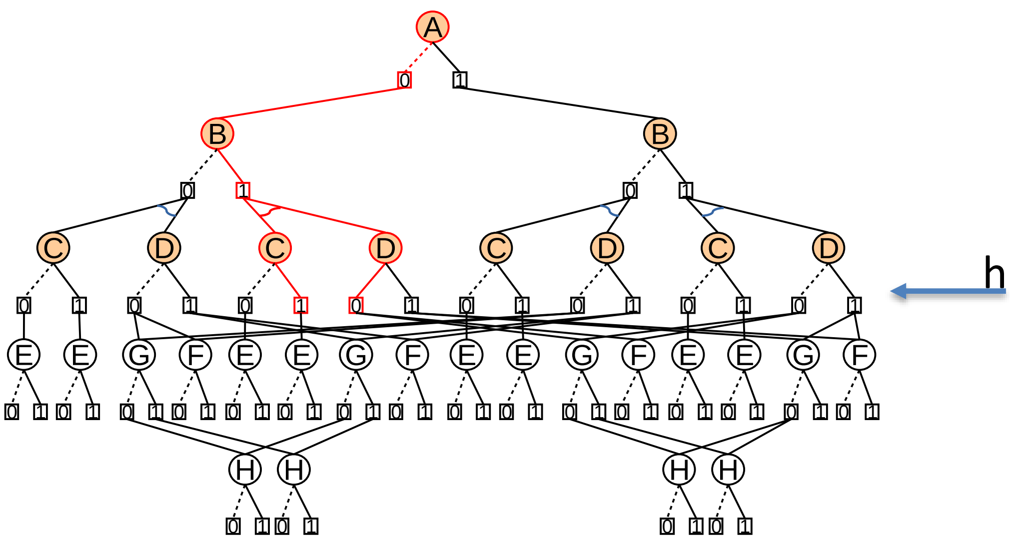
[Marinescu, Dechter and Ihler, 2014]

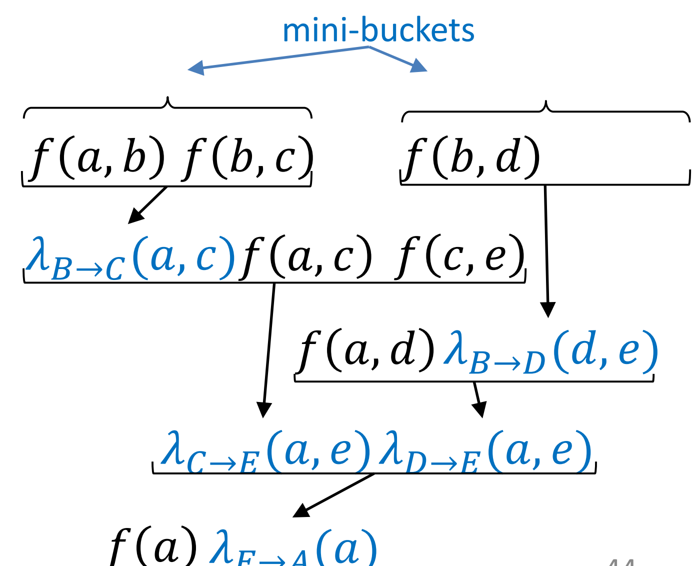


constrained
pseudo tree



- Complexity: NP^{pp} complete
- Not necessarily easy on trees





Anytime Solvers for Marginal MAP

- **Weighted Heuristic:** [Lee et. al. AAAI-2016, JAIR 2019]
  - Weighted Restarting AOBF (WAOBF)
  - Weighted Restarting RBFAOO (WRBFAOO)
  - Weighted Repairing AOBF (WRAOBF)

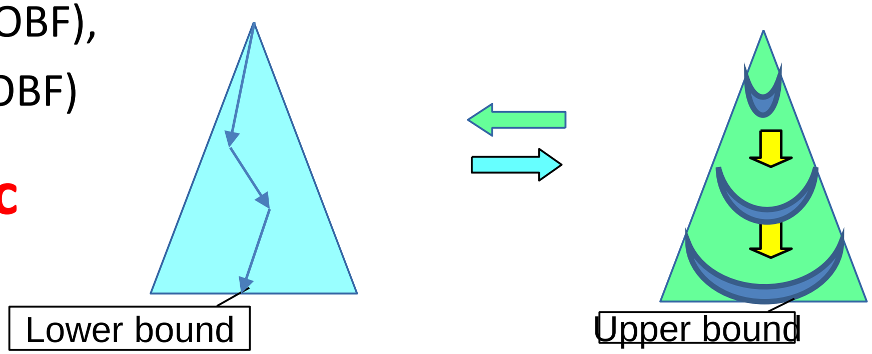
Weighted A* search [Pohl 1970]
- non-admissible heuristic
- Evaluation function:

  f(n)=g(n)+w·h(n)

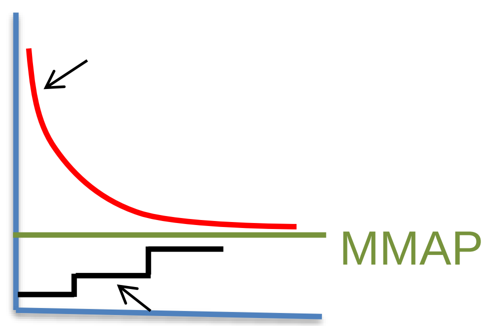
- **Guaranteed w-optimal solution, cost** $C \leq w \cdot C^*$

- **Interleaving Best and depth-first search**: (Marinescu et. al AAAI-2017)
  - Look-ahead (LAOBF),
  - alternating (AAOBF)

Exploiting heuristic search ideas



Goal: anytime bounds
And anytime solution



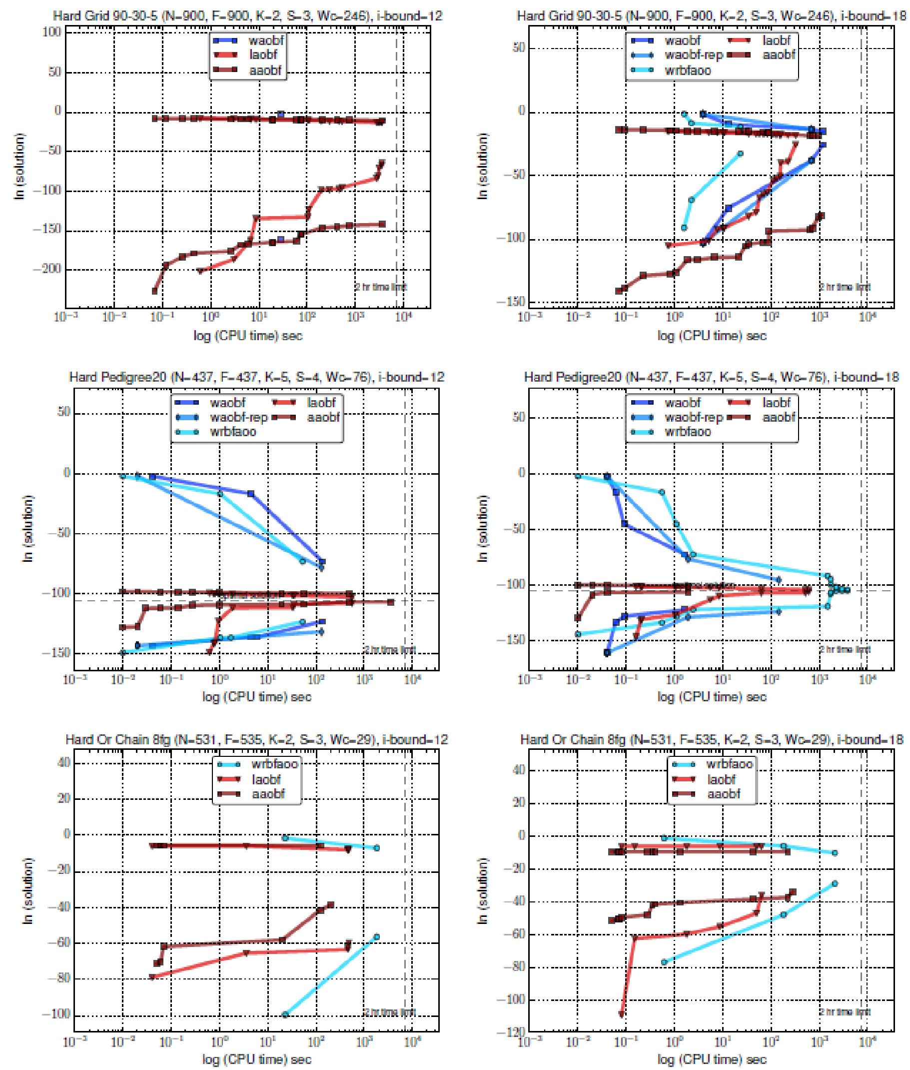
Anytime Bounds of Marginal MAP

(UAI’14, IJCAI’15, AAAI’16, AAAI’17, JAIR 2019 (Marinescu, Lee, Ihler, Dechter)

- Search: LAOBF, AAOBF, BRAOBB, WAOBF,WAOBF-rep
- heuristic: WMB-MM (20)
- memory: 24 GB

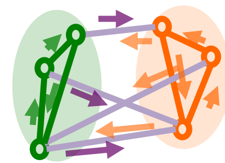
- Anytime lower and upper bounds from hard problem instances with **i-bound 12 (left) and 18 (right)**.

- The horizontal axis is the CPU time in log scale and the vertical axis is the value of marginal MAP in log scale.



Combining Approaches: Sampling+Search

Variational methods



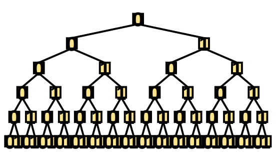
WMB

provide heuristics

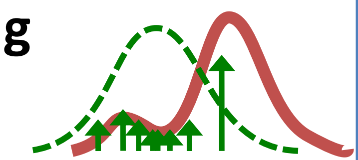
provide proposal

WMB-IS [Liu et al., NIPS 2015]

And/or Search



Sampling



dynamic importance sampling (DIS)

[Lou et al., NIPS 2017]

For MAP, marginal map and partition function

[Marinescu et al 2009, Lou et al., AAAI 2017, JAIR 2019]

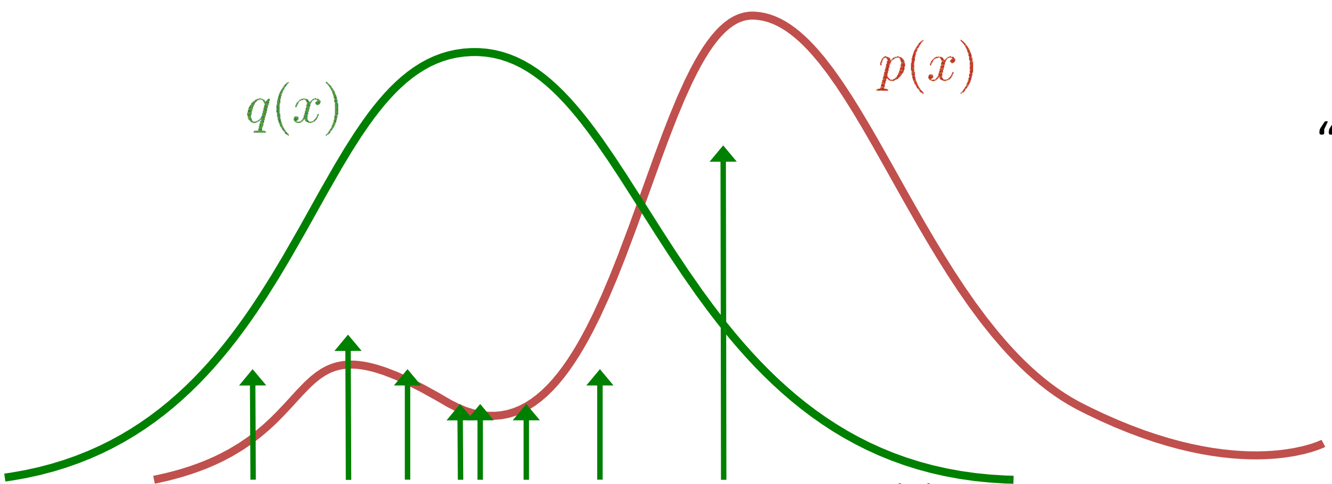
Importance Sampling

- Basic empirical estimate of probability:

$$\mathbb{E}[u(x)] = \int p(x)u(x) \approx \hat{u} = \frac{1}{m}\sum_i u(\tilde{x}^{(i)}) \qquad \tilde{x}^{(i)} \sim p(x)$$

- Importance sampling:

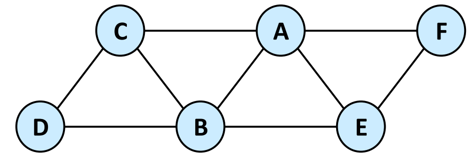
$$\int p(x)u(x) = \int q(x)\frac{p(x)}{q(x)}u(x) \approx \frac{1}{m}\sum_i \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})}u(\tilde{x}^{(i)}) \qquad \tilde{x}^{(i)} \sim q(x)$$



"importance weights"

$$w^{(i)} = \frac{p(\tilde{x}^{(i)})}{q(\tilde{x}^{(i)})}$$

IS on a Bayesian or Markov Network?



- Draw samples from P[A|E=e] directly?
  - Model defines un-normalized p(A,…,E=e)
  - Build (oriented) tree decomposition & sample

$$\tilde{\mathbf{b}} \sim f(\tilde{a}, b) \cdot f(b, \tilde{c}) \cdot f(b, \tilde{d}) \cdot f(b, \tilde{e}) / \lambda_{B \to C}$$

$$\tilde{\mathbf{c}} \sim f(c, \tilde{a}) \cdot f(c, \tilde{e}) \cdot \lambda_{B \to C}(\tilde{a}, c, \tilde{d}, \tilde{e}) / \lambda_{C \to D}$$

$$\tilde{\mathbf{d}} \sim f(\tilde{a}, d) \cdot \lambda_{B \to D}(d, \tilde{e}) / \lambda_{D \to E}(\tilde{a}, \tilde{e})$$

$$\tilde{\mathbf{e}} \sim \lambda_{D \to E}(\tilde{a}, e) / \lambda_{E \to A}(\tilde{a})$$

$$\tilde{\mathbf{a}} \sim p(A) = f(a) \cdot \lambda_{E \to A}(a) / Z$$

Downward message normalizes bucket; ratio is a conditional distribution

$$\sum_b \prod$$

B: $f(a,b)\ f(b,c)\ f(b,d)\ f(b,e)$

C: $f(c,a)\ f(c,e)\ \lambda_{B \to C}(a,c,d,e)$

D: $f(a,d)\ \lambda_{C \to D}(d,e,a)$

E: $\lambda_{D \to E}(a,e)$

A: $f(a)\ \lambda_{E \to A}(a)$

Z

Can use, WMB, Generalized belief propagation for proposal

Choose a Proposal Combine w Search

BuildIng blocks in current algorithms for Markov Logic Networks

- Probabilistic Theorem Proving: Gogate and Domingos, CACM 2016,
- Lifted Importance Sampling: Venugopal and Gogate, NeurIPS 2014.

- Cutset Sampling [Bidyuk and Dechter (JAIR, 2007)]
-]
- 012))]
- Sampling based lower bounds [Gogate, Dechter (Intelligenza Artificiale, 2011)]
- Dynamic Importance Sampling (DIS) [Lou, Dechter, and Ihler (NIPS 2017)]
- Abstraction Sampling [Broka, Dechter, Ihler and Kask (UAI, 2018)].
- Finite-sample Bounds for MMAP [Lou, Dechter, and Ihler. (UAI 2018)]
- WMB Importance Sampling (WMB-IS) [Liu, Fisher, Ihler (ICML 2015)]

Choosing a proposal- WMB-IS

[Liu, Fisher, Ihler 2015]

- Can use WMB upper bound to define a proposal $q_{\text{wmb}}(x)$

$$\tilde{\mathbf{b}} \sim w_1\, q_1(b|\tilde{a}, \tilde{c}) \;+\; w_2\, q_2(b|\tilde{d}, \tilde{e})$$

Weighted mixture:

use minibucket 1 with probability w_1

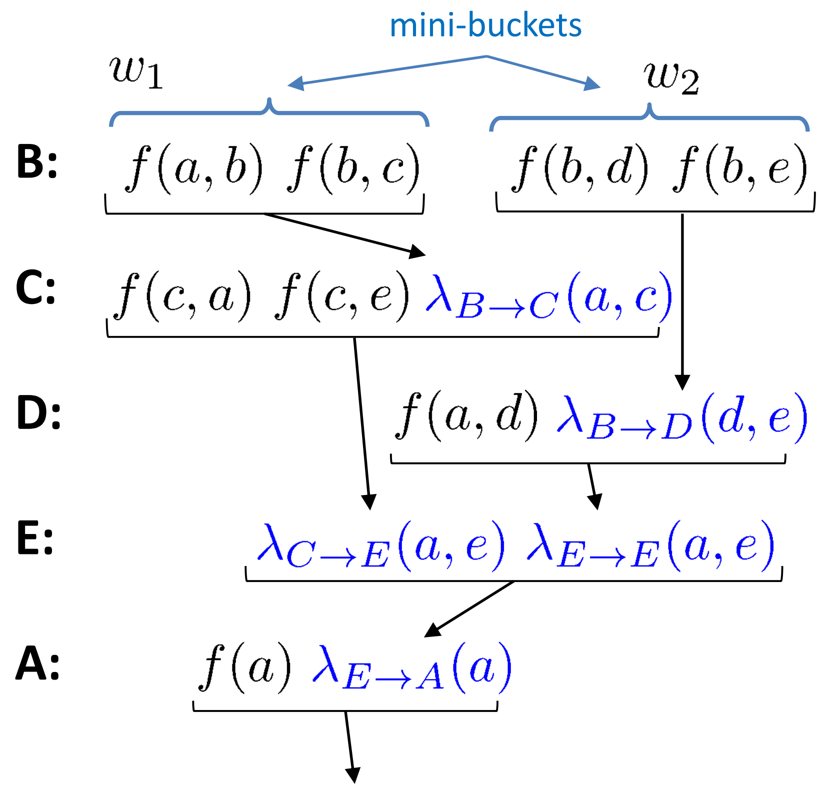
or, minibucket 2 with probability $w_2 = 1 - w_1$

where

$$q_1(b|a,c) = \left[\frac{f(a,b)\cdot f(b,c)}{\lambda_{B\to C}(a,c)}\right]^{\frac{1}{w_1}}$$

⋮

$$\tilde{\mathbf{a}} \sim q(A) = f(a) \cdot \lambda_{E\to A}(a)/U$$



***U* = upper bound**

Key insight: provides bounded importance weights!

$$0 \le f(x)/q_{\text{wmb}}(x) \le U \quad \forall x$$

WMB-IS Bounds

[Liu, Fisher, Ihler 2015]

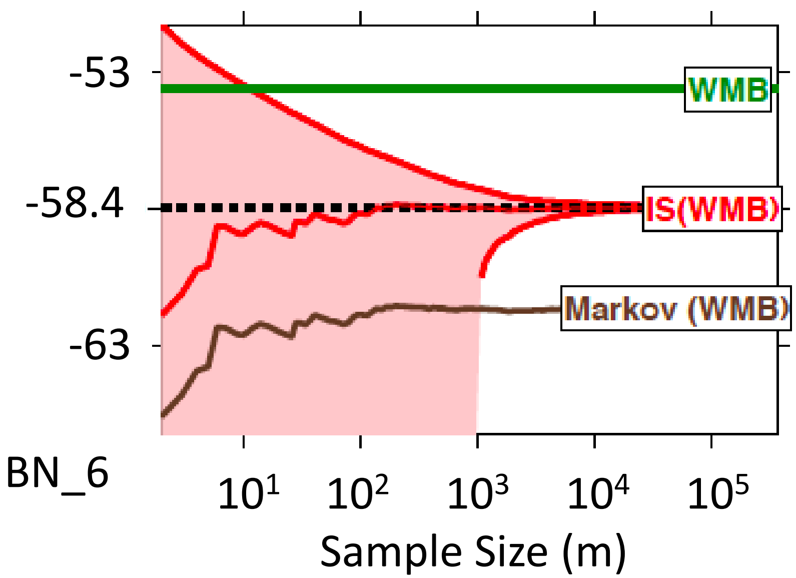
- Finite sample bounds on the average

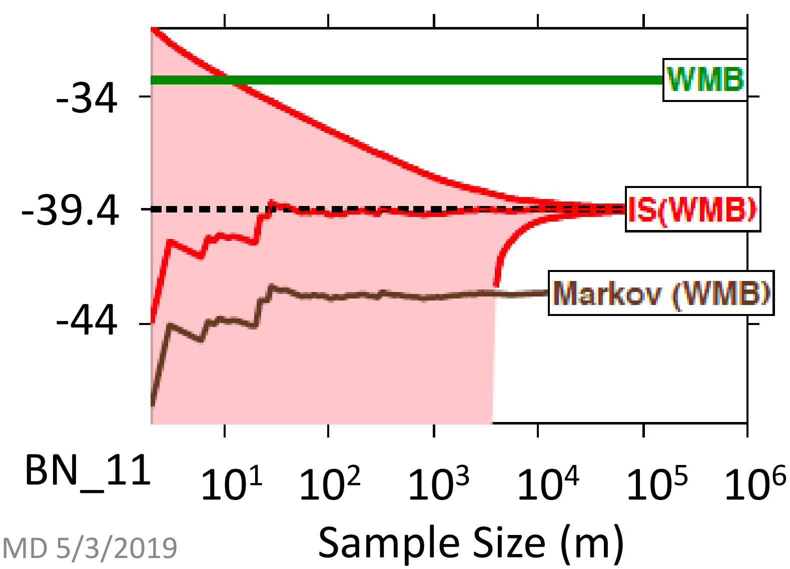
$$\Pr\Big[|\hat{Z} - Z| > \epsilon\Big] \le 1 - \delta$$

$$\epsilon = \sqrt{\frac{2\hat{V}\log(4/\delta)}{m}} + \frac{7\,U\;\log(4/\delta)}{3(m-1)}$$

"Empirical Bernstein" bounds

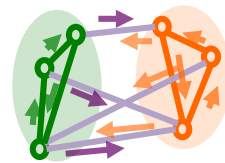
- Confidence interval depends on two parts
  - Empirical variance, decreasing as $1/\mathrm{m}^{1/2}$
  - Upper bound U, decreasing as $1/\mathrm{m}$





Combining Approaches

Variational methods



WMB

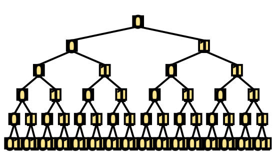
provide heuristics

provide proposal

WMB-IS

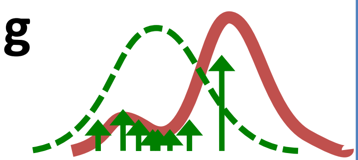
[Liu et al., NIPS 2015]

Search



refine proposal

Sampling



For MAP, marginal map and partition function

dynamic importance sampling (DIS)

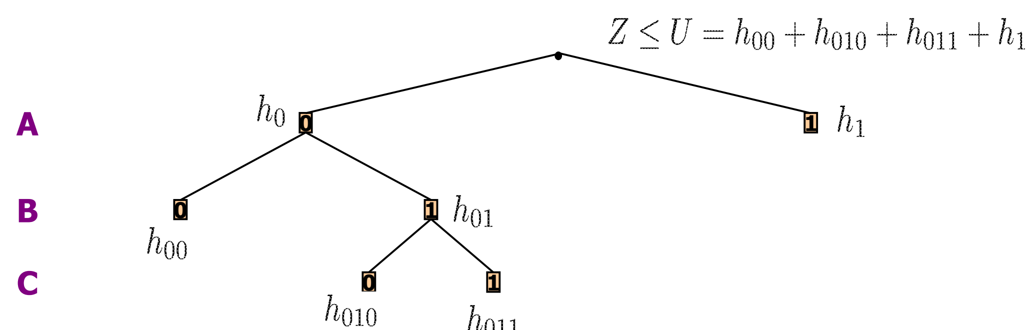
[Lou et al., NIPS 2017]

Dynamic Importance Sampling

[Lou, Dechter, Ihler, NIPS 2017, AAAI 2019]

- Interleave
  - Building search tree (expand Nd nodes)
  - Draw samples given search bound (Nl samples)

(For partition function)

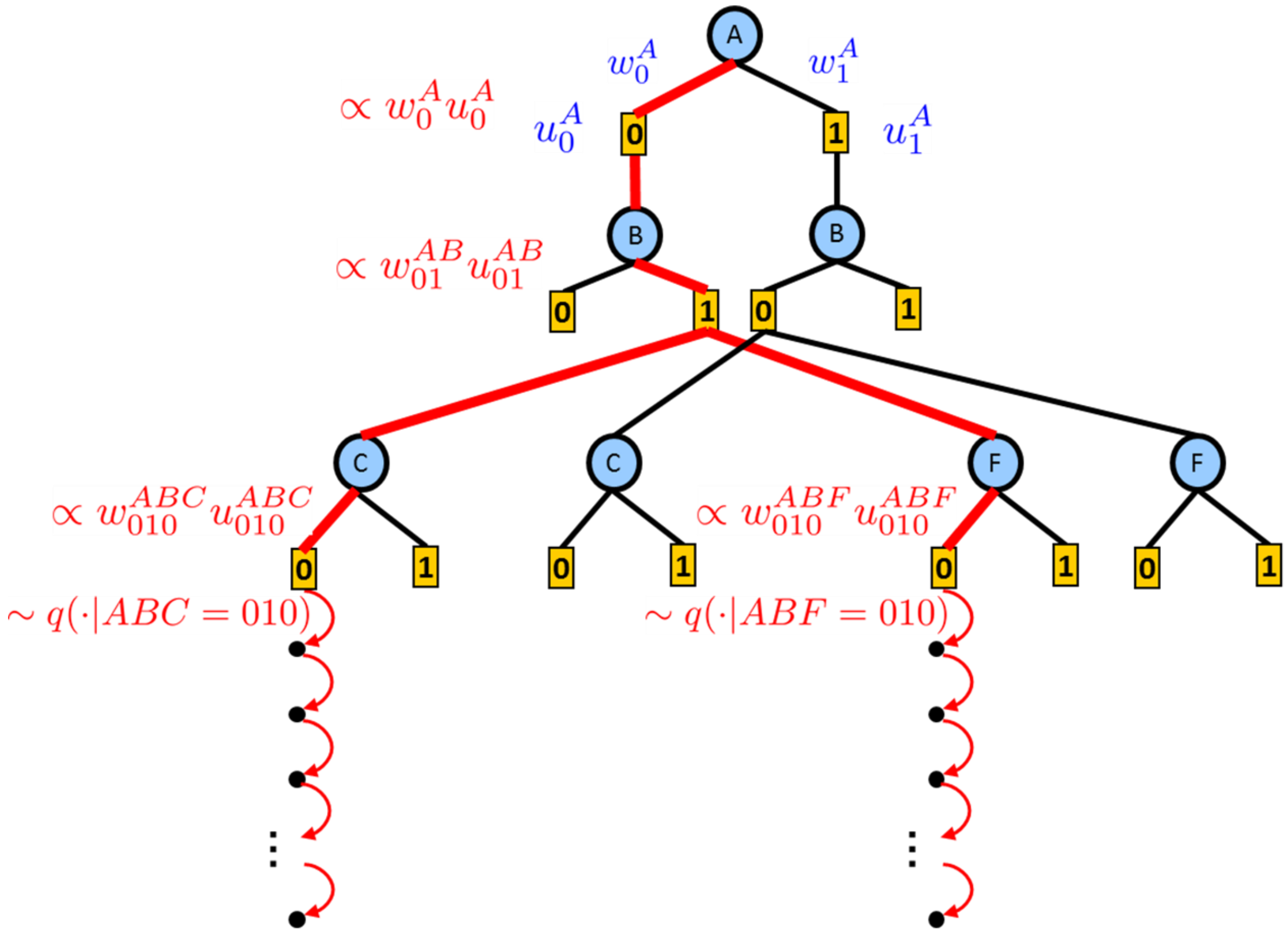


- Key insight: proposal changes (improves) with each step
  - Use weighted average: better samples get more weight

$$\widehat{Z} = \frac{\mathrm{HM}(\boldsymbol{U})}{N} \sum_{i=1}^{N} \frac{\widehat{Z}_i}{U_i}, \quad \mathrm{HM}(\boldsymbol{U}) = \Big[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{U_i}\Big]^{-1}$$

  - Derive corresponding concentration bound on Z

Two-step Sampling



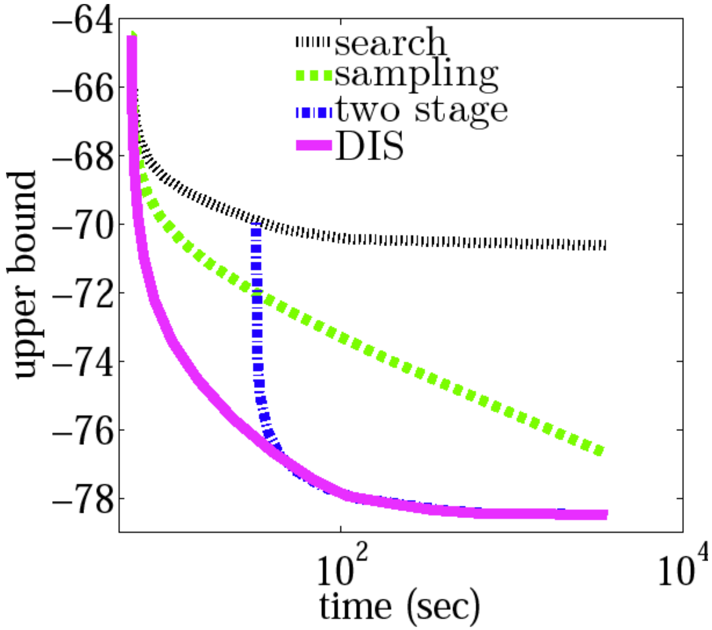
Finite-sample Bounds for DIS

Theorem: Define the deviation term

$$\Delta = \mathrm{HM}(\boldsymbol{U})\Big(\sqrt{\frac{2\widehat{\mathrm{Var}}(\{\widehat{Z}_i/U_i\}_{i=1}^{N})\ln(2/\delta)}{N}} + \frac{7\ln(2/\delta)}{3(N-1)}\Big)$$

then, $\Pr[Z \leq \widehat{Z} + \Delta] \geq 1 - \delta$ and $\Pr[Z \geq \widehat{Z} - \Delta] \geq 1 - \delta$.

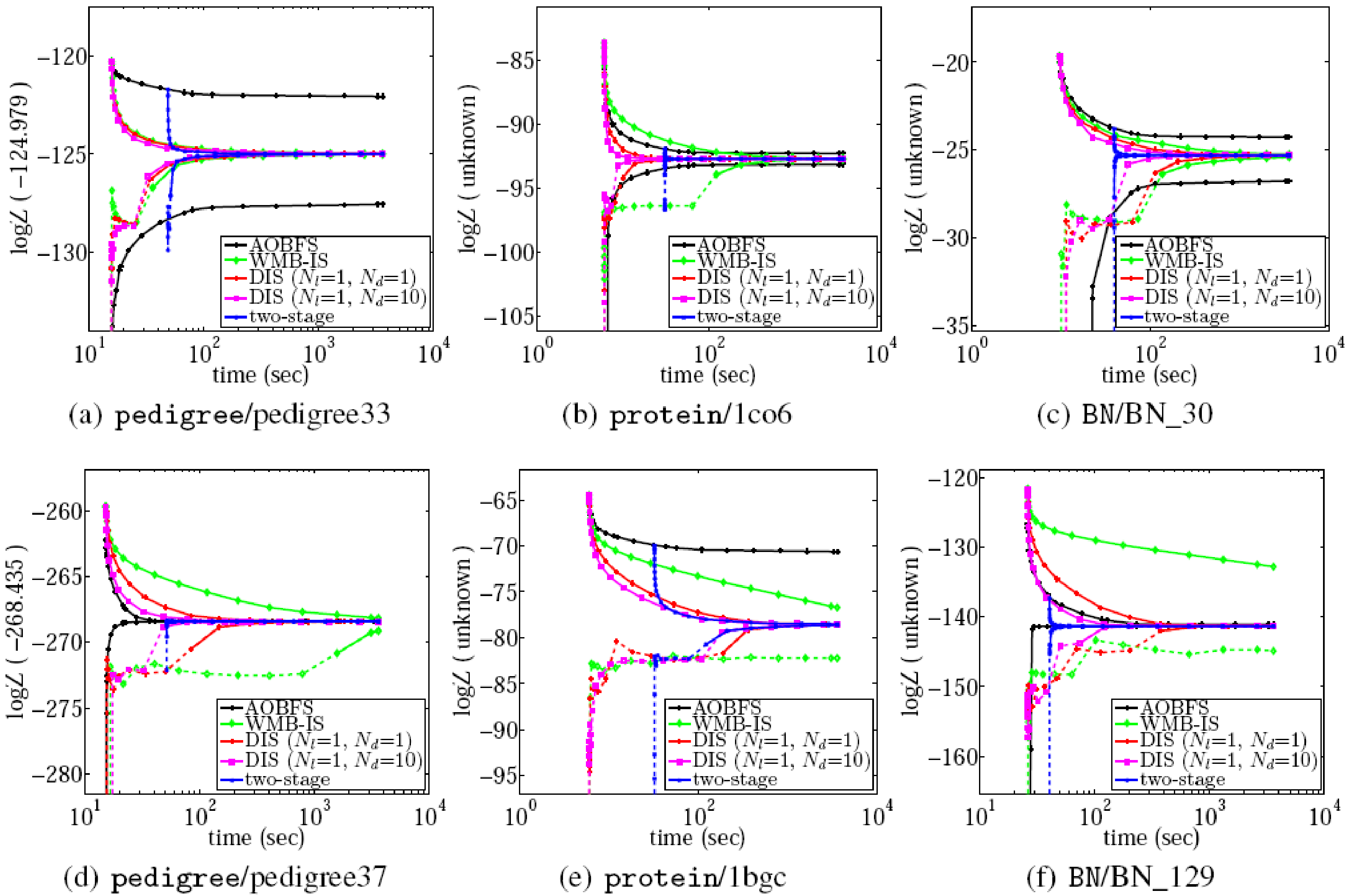
$\widehat{\mathrm{Var}}(\{\widehat{Z}_i/U_i\}_{i=1}^{N})$: empirical variance of $\{\widehat{Z}_i/U_i\}_{i=1}^{N}$.



Individual Results

(For partition function)

[Lou, Dechter, Ihler, NIPS 2017]

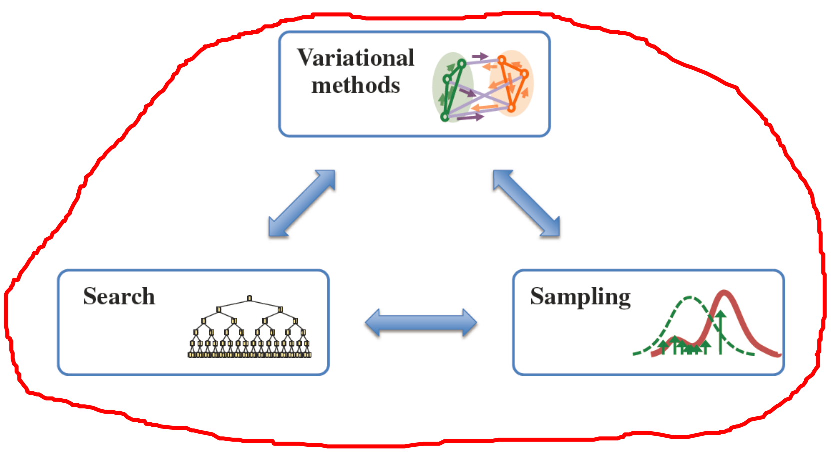


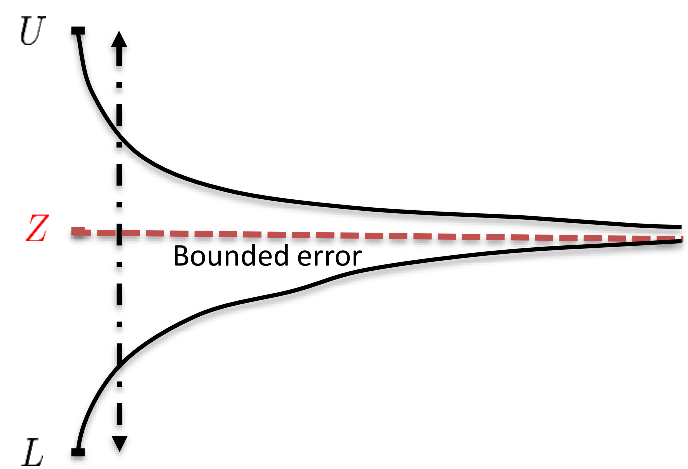
(a) pedigree/pedigree33 (b) protein/1co6 (c) BN/BN_30

(d) pedigree/pedigree37 (e) protein/1bgc (f) BN/BN_129

Outline and Challenges

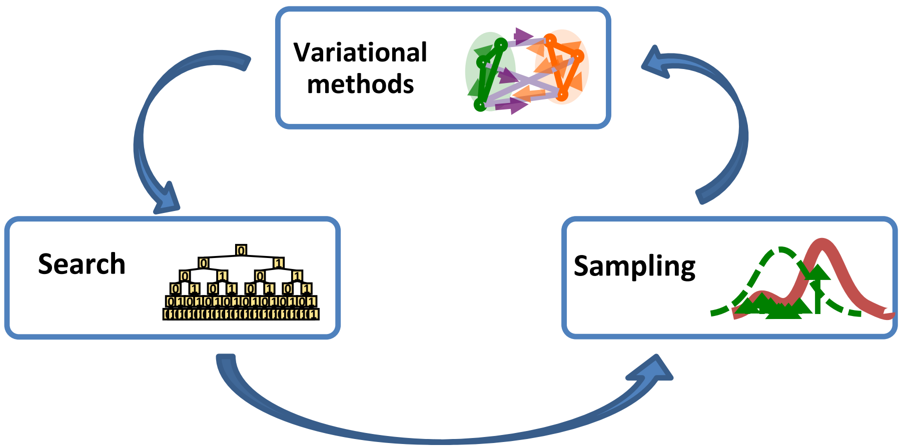
- Overview of problems and methodology
- Main paradigms of approximate reasoning:
  Variational, Search, Sampling
- Combining approaches
- Future challenges





Continuing Work

- Combining approaches:
  - Tune the hyper-parameters automatically
  - Extend to decision networks

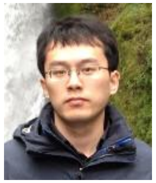


- Languages and Tools:
  - Relational languages
  - Handle constraints specification and continuous functions
  - Temporal domains; Planning, e.g., Influence diagrams, MDPs, POMDPs
  - Cross interaction of deep learning and graphical models → Reinforcement

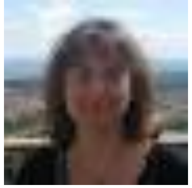
Thank You !

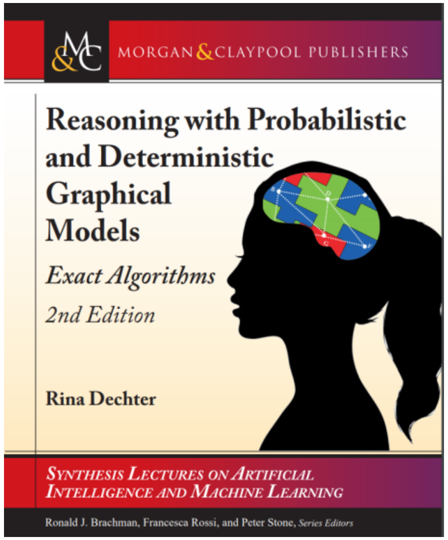
For publication see:
http://www.ics.uci.edu/~dechter/publications.html











Alex Ihler
Kalev Kask
Irina Rish
Bozhena Bidyuk
Robert Mateescu
Radu Marinescu
Vibhav Gogate
Emma Rollon
Lars Otten
Natalia Flerova
Andrew Gelfand
William Lam
Junkyu Lee
Qi Lou