Clamping Variables and Approximate Inference

Adrian Weller
University of Cambridge

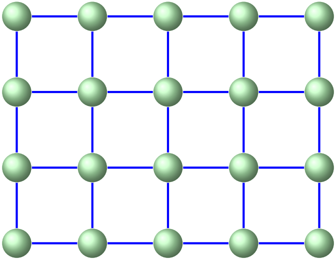
MSR Cambridge
Mar 18, 2016

Work with Tony Jebara and Justin Domke

For more information, see
http://mlg.eng.cam.ac.uk/adrian/

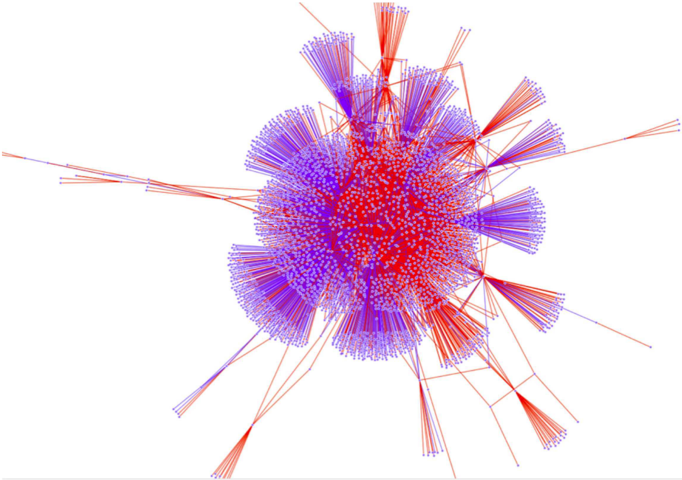
Motivation: *undirected graphical models*

- Powerful way to represent relationships across variables
- Many applications including: computer vision, social network analysis, deep belief networks, protein folding...
- In this talk, focus on binary pairwise (Ising) models



Example: Grid for computer vision (attractive)

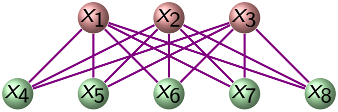
Motivation: *undirected graphical models*



Example: Part of epinions social network (mixed)

Figure courtesy of N. Ruozzi

Motivation: *undirected graphical models*



Example: Restricted Boltzmann machine (mixed)

A fundamental problem is *marginal inference*

- Estimate marginal probability distribution of one variable

$$p(x_1) = \sum_{x_2,\ldots,x_n} p(x_1, x_2, \ldots, x_n)$$

- Closely related to computing the *partition function*
- Computationally intractable, focus on approximate methods

- Our theme: combining approximate inference with *clamping* can be very fruitful as a **proof technique**, and in **practice**

Background: *Binary pairwise models*

- Binary variables $X_1, \ldots, X_n \in \{0, 1\}$
- Singleton and pairwise potentials θ
- Write $\theta \cdot x$ for the total score of a complete configuration
- Probability distribution given by

$$p(x) = \frac{1}{Z} \exp(\theta \cdot x)$$

- To ensure probabilities sum to 1, need normalizing constant

$$Z = \sum_x \exp(\theta \cdot x)$$

- Z is the *partition function*, a fundamental quantity we'd like to compute or approximate

Background: *A variational approximation*

$$\text{Recall} \quad p(x) = \frac{1}{Z} \exp(\theta \cdot x)$$

- Exact inference may be viewed as *optimization*,

$$\log Z = \max_{\mu \in \mathbb{M}} [\, \theta \cdot \mu + S(\mu) \,]$$

  $\mathbb{M}$ is the space of marginals that are *globally consistent*, S is the (Shannon) entropy

- Bethe makes two pairwise approximations,

$$\log Z_B = \max_{q \in \mathbb{L}} [\, \theta \cdot q + S_B(q) \,]$$

  $\mathbb{L}$ is the space of marginals that are *pairwise consistent*, S_B is the Bethe entropy approximation

- Loopy Belief Propagation finds stationary points of Bethe
- For models with no cycles (acyclic), Bethe is exact $Z_B = Z$

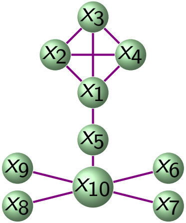
Background: When is Bethe a good approximation?

We know that Bethe is exact for acyclic models, $Z_B = Z$

When else does Bethe perform well?

- 'Tree-like models': models with long cycles or weak potentials
- Also: attractive models (all edges attractive)
- Sudderth, Wainwright and Willsky (NIPS 2007) used *loop series* to show that for a subclass of attractive binary pairwise models, $Z_B \leq Z$
- Conjectured $Z_B \leq Z$ for all attractive binary pairwise models
- Proved true by Ruozzi (NIPS 2012) using *graph covers*
- Here we provide a separate proof building from first principles, and also derive an upper bound for Z in terms of Z_B
- We use the idea of clamping variables

Background: *What is clamping?*

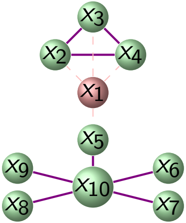


Example model

To compute the partition function Z, can enumerate all states and sum

$x_1 x_2 \ldots x_{10}$	*score*	exp(*score*)
0 0 ...0	1	2.7
0 0 ...1	2	7.4
...	...	...
0 1 ...1	1.3	3.7
1 0 ...0	-1	0.4
1 0 ...1	0.2	1.2
...	...	...
1 1 ...1	1.8	6.0
Total $Z =$		**47.1**

Background: *What is clamping?*



Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

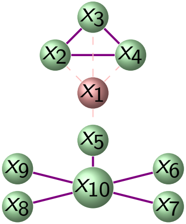
$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

After we clamp a variable, it may be removed

$x_1 x_2 \ldots x_{10}$	*score*	exp(*score*)	
0 0 . . . 0	1	2.7	
0 0 . . . 1	2	7.4	
. . .	. . .	. . .	
0 1 . . . 1	1.3	3.7	27.5
1 0 . . . 0	-1	0.4	
1 0 . . . 1	0.2	1.2	
. . .	. . .	. . .	
1 1 . . . 1	1.8	6.0	19.6
Total $Z =$		**47.1**	

$$p(X_1 = 1) = \frac{Z|_{X_1=1}}{Z}$$

Background: *What is clamping?*



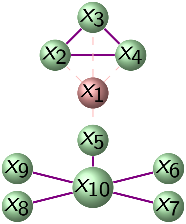
Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

After we clamp a variable, it may be removed

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)

Background: *What is clamping?*



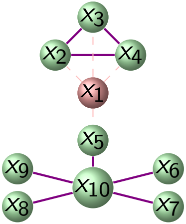
Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

After we clamp a variable, it may be removed

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)
- If not,
  - Can repeat: clamp and remove variables until acyclic, *or*
  - Settle for approximate inference on sub-models

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$



Can split Z in two: clamp variable X_1 to each of $\{0,1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

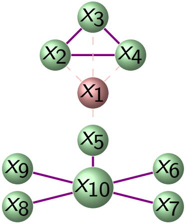
After we clamp a variable, it may be removed

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)
- If not,
  - Can repeat: clamp and remove variables until acyclic, *or*
  - Settle for approximate inference on sub-models

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Will this lead to a better estimate than approximate inference on the original model? Always?

Background: *What is clamping?*



Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

After we clamp a variable, it may be removed

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)
- If not,
    - Can repeat: clamp and remove variables until acyclic, *or*
    - Settle for approximate inference on sub-models

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Will this lead to a better estimate than approximate inference on the original model? Always? *Often but not always*

A variational perspective on clamping

- Bethe approximation

$$\log Z_B = \max_{q \in \mathbb{L}} [\, \theta \cdot q + S_B(q) \,]$$

- Observe that when X_i is clamped, we optimize over a subset

$$\log Z_B|_{X_i=0} = \max_{q \in \mathbb{L}: q_i = 0} [\, \theta \cdot q + S_B(q) \,]$$

$$\Rightarrow Z_B|_{X_i=0} \leq Z_B, \text{ similarly } Z_B|_{X_i=1} \leq Z_B$$

Recap of Notation

Z	true partition function
Z_B	Bethe optimum partition function
$Z_B^{(i)} := Z_B\|_{X_i=0} + Z_B\|_{X_i=1} \leq 2Z_B$	approximation obtained when *clamp and sum approximate* sub-partition functions

Clamping variables: *an upper bound on Z*

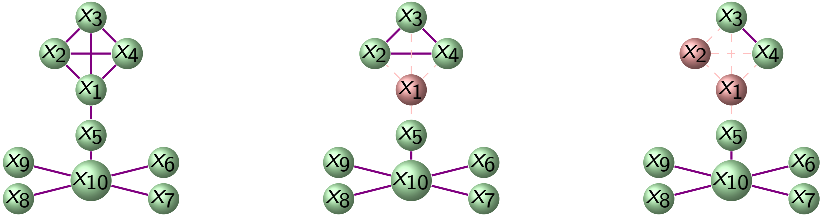
- From before,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1} \;\leq\; 2Z_B$$

- Repeat: clamp and remove variables, until remaining model is acyclic, where Bethe is exact
- For example, if must delete 2 variables X_i, X_j, obtain

$$Z_B^{(ij)} := \sum_{a,b\in\{0,1\}} Z_B|_{X_i=a,X_j=b} \;\leq\; 2^2 Z_B$$

But sub-partition functions are *exact*, hence LHS $= Z$



Clamping variables: *an upper bound on Z*

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1} \ \leq \ 2Z_B$$

- Repeat: clamp and remove variables, until remaining model is acyclic, where Bethe is exact
- Let $k(G)$ be the minimum size of a feedback vertex set

Theorem (result is tight in a sense)

$$Z \leq 2^k Z_B$$

Clamping variables: *an upper bound on Z*

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1} \ \leq \ 2Z_B$$

- Repeat: clamp and remove variables, until remaining model is acyclic, where Bethe is exact
- Let $k(G)$ be the minimum size of a feedback vertex set

Theorem (result is tight in a sense)

$$Z \leq 2^k Z_B$$

Attractive models: *a lower bound on Z*

- An *attractive* model is one with all edges attractive
- Recall definition,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Theorem (actually show a stronger result, ask if interested)

For an attractive binary pairwise model and any X_i, $Z_B \leq Z_B^{(i)}$

Repeat as before: $Z_B \leq Z_B^{(i)} \leq Z_B^{(ij)} \leq \cdots \leq Z$

Corollary (similar proof to earlier result; first proved Ruozzi, 2012)

For an attractive binary pairwise model, $Z_B \leq Z$

Attractive models: *a lower bound on Z*

- An *attractive* model is one with all edges attractive
- Recall definition,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Theorem (actually show a stronger result, ask if interested)

For an attractive binary pairwise model and any X_i, $Z_B \leq Z_B^{(i)}$

Repeat as before: $Z_B \leq Z_B^{(i)} \leq Z_B^{(ij)} \leq \cdots \leq Z$

Corollary (similar proof to earlier result; first proved Ruozzi, 2012)

For an attractive binary pairwise model, $Z_B \leq Z$

$\Rightarrow$ each clamp and sum can only *improve* Z_B

Recap of results so far

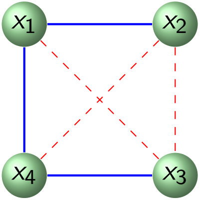
- We have used clamping as a proof technique
- Derived lower and upper bounds on Z for attractive models

$$\underbrace{Z_B \leq}_{\text{attractive only}} \quad Z \quad \underbrace{\leq 2^k Z_B}_{\text{attractive and mixed}}$$

$$\Leftrightarrow \quad \frac{Z}{2^k} \leq \quad Z_B \quad \underbrace{\leq Z}_{\text{attractive only}}$$

- We also proved that for attractive models, clamping and summing (optimum) Bethe sub-partition functions can only improve the estimate
- How about for mixed models?

Example: here clamping *any* variable *worsens* Z_B estimate



Blue edges are attractive with edge weight $+2$
Red edges are repulsive with edge weight -2
No singleton potentials

(performance is only slightly worse with clamping)

- In practice, if we pick a good variable to clamp, then clamping is usually helpful

New work: what does clamping do for MF and TRW?

- Mean field (MF) approximation assumes independent variables, yields a lower bound, $Z_M \leq Z$
- Tree-reweighted (TRW) is a pairwise approximation similar to Bethe but allows a convex optimization and yields an upper bound, $Z \leq Z_T$ $\qquad Z_M \leq Z \leq Z_T$
- Earlier, we showed that for Bethe, clamping always improves the approximation for attractive models; often but not always improves for mixed models
- How about for MF and TRW? $\qquad Z_M \leq Z_B \leq Z_T$

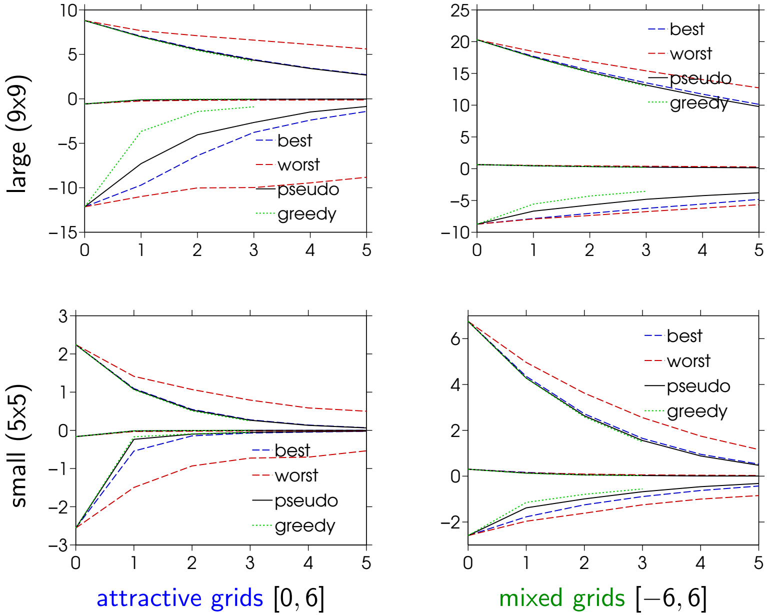
New work: what does clamping do for MF and TRW?

- Mean field (MF) approximation assumes independent variables, yields a lower bound, $Z_M \leq Z$
- Tree-reweighted (TRW) is a pairwise approximation similar to Bethe but allows a convex optimization and yields an upper bound, $Z \leq Z_T$ $\qquad Z_M \leq Z \leq Z_T$
- Earlier, we showed that for Bethe, clamping always improves the approximation for attractive models; often but not always improves for mixed models
- How about for MF and TRW? $\qquad Z_M \leq Z_B \leq Z_T$

Theorem

For both MF and TRW, for attractive and mixed models, clamping and summing approximate sub-partition functions can only improve the respective approximation and bound (any number of labels).

Error in log Z vs number of clamps: grids



Conclusions for practitioners

- Typically Bethe performs very well
- Clamping can be very helpful, more so for denser models with stronger edge weights, a setting where inference is often hard
- We provide fast methods to select a good variable to clamp
- MF and TRW provide useful bounds on Z and Z_B

Thank you

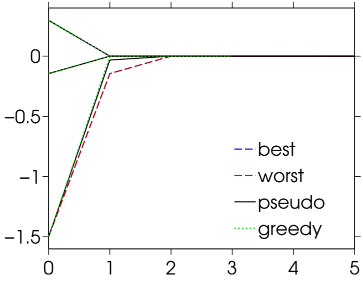
For more information, see
http://mlg.eng.cam.ac.uk/adrian/

References

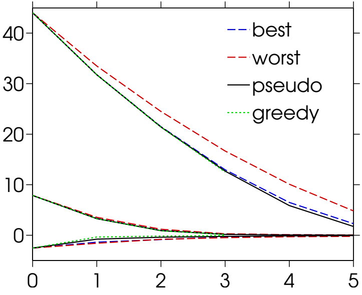
- N. Ruozzi. The Bethe partition function of log-supermodular graphical models. In *NIPS*, 2012.
- E. Sudderth, M. Wainwright, and A. Willsky. Loop series and Bethe variational bounds in attractive graphical models. In *NIPS*, 2007.
- A. Weller and J. Domke. Clamping improves TRW and mean field approximations. To appear in *AISTATS*, 2016.
- A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *AISTATS*, 2013.
- A. Weller and T. Jebara. Clamping variables and approximate inference. In *NIPS*, 2014.
- J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *IJCAI, Distinguished Lecture Track*, 2001.

Extra slides for questions or further explanation

Error in log Z vs number of clamps: complete graphs



attractive $K_{15}, [0, 6]$



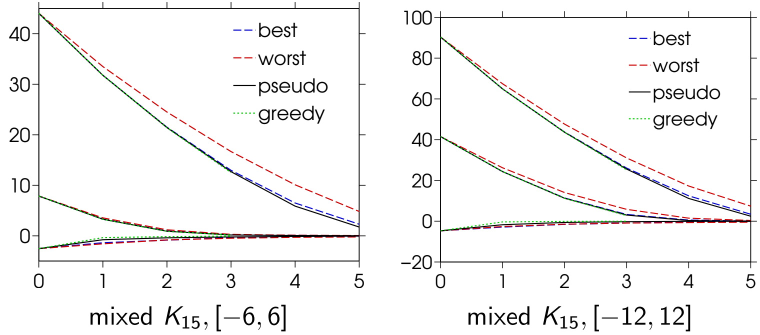
mixed $K_{15}, [-6, 6]$

- For dense mixed models (many edges),

  MF can be better than Bethe

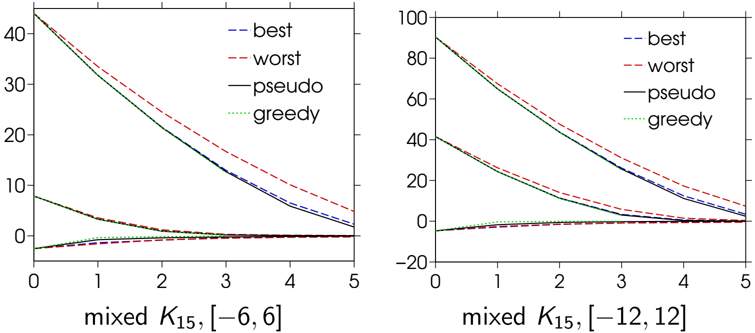
- What happens if we increase edge strength?

Error in log Z vs number of clamps: complete graphs



- With stronger edges, MF is much better than Bethe!
- But MF assumes variables are independent, what's going on?

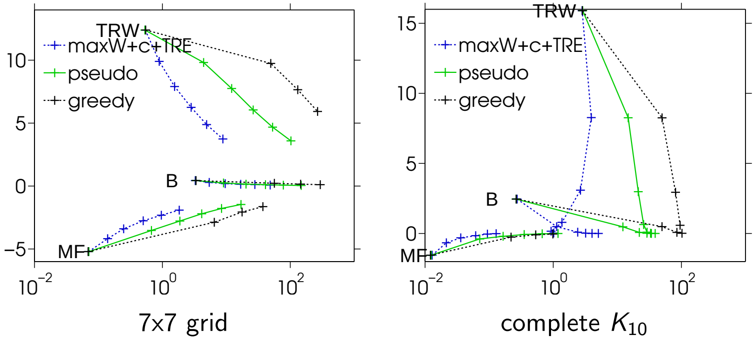
Error in log Z vs number of clamps: complete graphs



- With stronger edges, MF is much better than Bethe!
- But MF assumes variables are independent, what's going on?
  - Frustrated cycles cause Bethe to overestimate by a lot
    TRW is even worse
  - MF behaves much better (in marginal polytope)

Time (secs) vs error in log Z for various methods

Mixed models, $W_{ij} \sim U[-6, 6]$
Time shown on a log scale



- Clamping can make the subsequent optimization problems easier, hence sometimes total time with clamping is lower while also being more accurate

Clamping variables: strongest result for *attractive* models

$$\log Z_B = \max_{q \in \mathbb{L}} [\ \theta \cdot q + S_B(q)\ ]$$

- For any variable X_i and $x \in [0,1]$, let $q_i = q(X_i = 1)$ and

$$\log Z_{Bi}(x) = \max_{q \in \mathbb{L} : q_i = x} [\ \theta \cdot q + S_B(q)\ ]$$

- $Z_{Bi}(x)$ is '*Bethe partition function constrained to* $q_i = x$'

  Note: $Z_{Bi}(0) = Z_B|_{X_i=0},\ Z_{Bi}(x^*) = Z_B,\ Z_{Bi}(1) = Z_B|_{X_i=1}$

Clamping variables: strongest result for *attractive* models

$$\log Z_B = \max_{q \in \mathbb{L}} [\, \theta \cdot q + S_B(q) \,]$$

- For any variable X_i and $x \in [0,1]$, let $q_i = q(X_i = 1)$ and

$$\log Z_{Bi}(x) = \max_{q \in \mathbb{L}: q_i = x} [\, \theta \cdot q + S_B(q) \,]$$

- $Z_{Bi}(x)$ is '*Bethe partition function constrained to* $q_i = x$'
  Note: $Z_{Bi}(0) = Z_B|_{X_i=0}, \; Z_{Bi}(x^*) = Z_B, \; Z_{Bi}(1) = Z_B|_{X_i=1}$
- Define new function,

$$A_i(x) := \log Z_{Bi}(x) - S_i(x)$$

Theorem (implies all other results for attractive models)

For an attractive binary pairwise model, $A_i(x)$ is convex

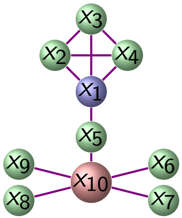
- Builds on derivatives of Bethe free energy from [WJ13]

Experiments: *Which variable to clamp?*

Compare error $|\log Z - \log Z_B^{(i)}|$ to original error $|\log Z - \log Z_B|$ for various ways to choose which variable X_i to clamp:

- best Clamp — best improvement in error of Z in hindsight
- worst Clamp — worst improvement in error of Z in hindsight
- avg Clamp — average performance

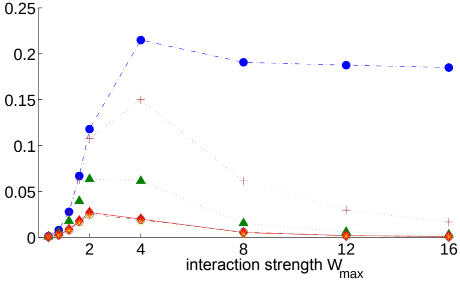
- maxW — max sum of incident edge weights $\sum_{j \in N(i)} |W_{ij}|$
- Mpower — more sophisticated, based on powers of related matrix



Experiments: *attractive random graph $n = 10, p = 0.5$*

unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

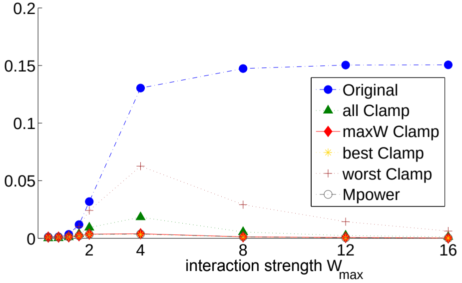
Error of estimate of $\log Z$



Observe

- Clamping any variable helps significantly
- Our selection methods perform well

Avg ℓ_1 error of singleton marginals

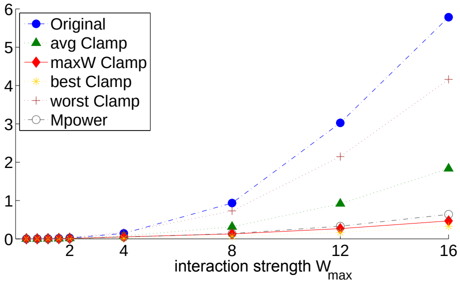


Using Frank-Wolfe to optimize Bethe free energy

Experiments: *mixed random graph $n = 10, p = 0.5$*

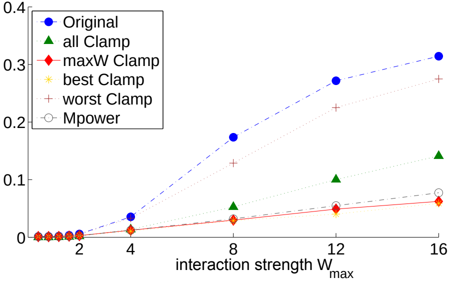
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

Error of estimate of log Z



Results remain promising for higher n

Avg ℓ_1 error of singleton marginals



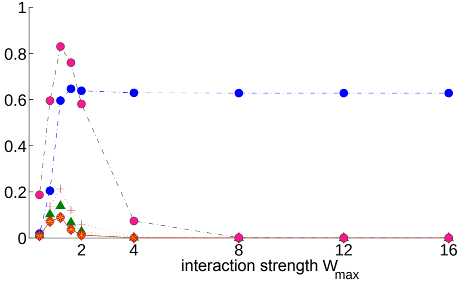
Using Frank-Wolfe to optimize Bethe free energy

Experiments: *attractive complete graph $n = 10$, TRW*

unary $\theta_i \sim U[-0.1, 0.1]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

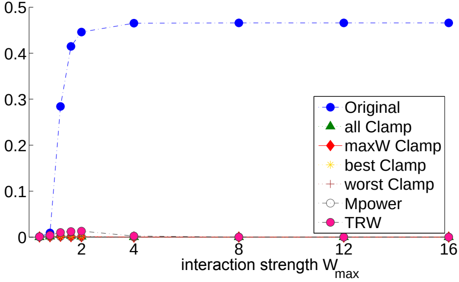
Error of estimate of $\log Z$

Note low unary potentials



Avg ℓ_1 error of singleton marginals

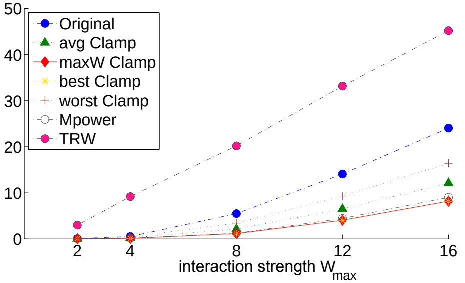
Clamping a variable ‘breaks symmetry’ and overcomes TRW advantage



Experiments: *mixed complete graph $n = 10$, TRW*

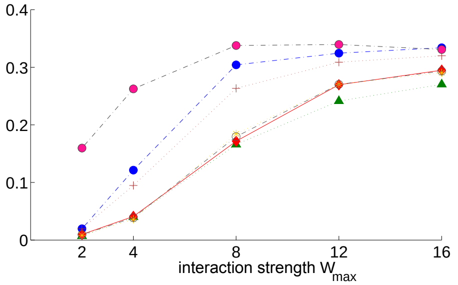
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

Error of estimate of log Z



Note regular singleton potentials

Avg ℓ_1 error of singleton marginals



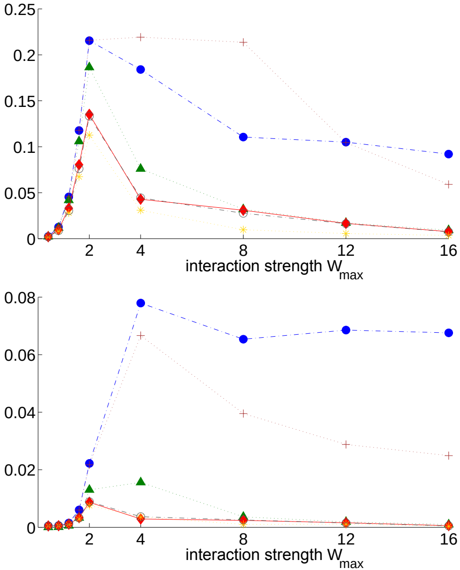
Experiments: *attractive random graph $n = 50, p = 0.1$*

unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

Error of estimate of $\log Z$

'worst Clamp' performs *worse* here due to suboptimal solutions found by Frank-Wolfe

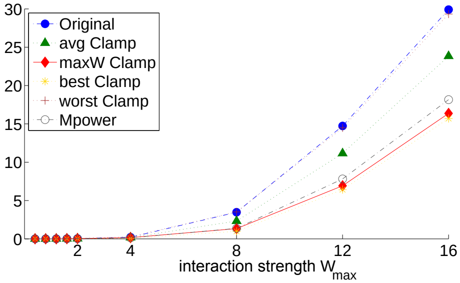
Avg ℓ_1 error of singleton marginals



Experiments: *mixed random graph $n = 50, p = 0.1$*

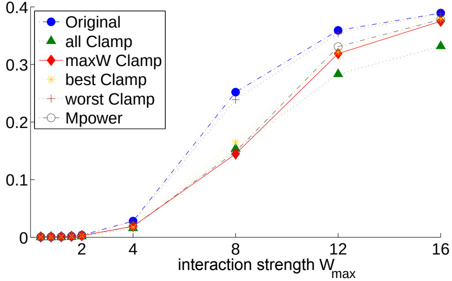
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

Error of estimate of $\log Z$



Performance still good for clamping just one variable

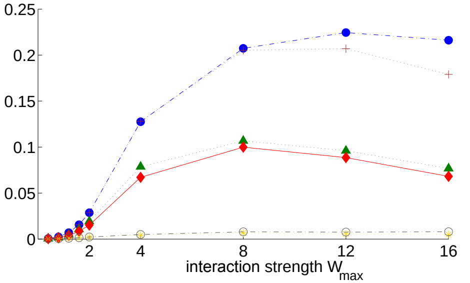
Avg ℓ_1 error of singleton marginals



Experiments: *attractive 'lamp' graph*

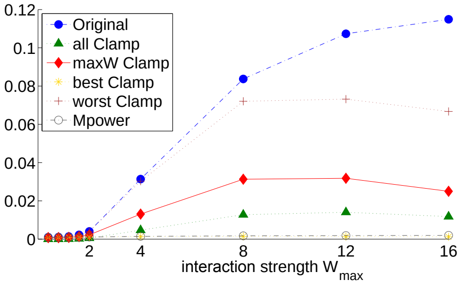
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

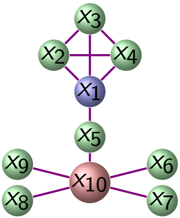
Error of estimate of $\log Z$



Mpower performs well,
significantly better than maxW

Avg ℓ_1 error of singleton marginals



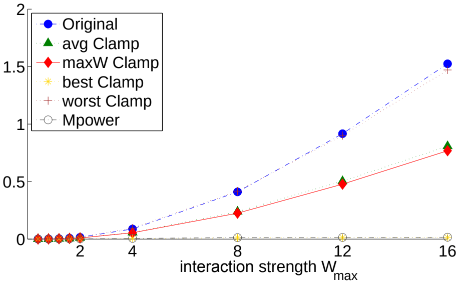


Experiments: *mixed 'lamp' graph*

unary $\theta_i \sim U[-2,2]$,

edge $W_{ij} \sim U[-W_{max}, W_{max}]$

Error of estimate of $\log Z$



Mpower performs well,

significantly better than maxW

Avg ℓ_1 error of singleton marginals