Clamping Variables and Approximate Inference

Adrian Weller and Tony Jebara

Summary

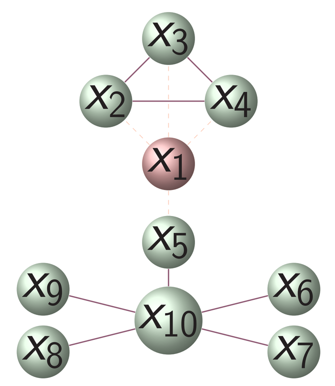
We address the problem of *marginal inference* for undirected graphical models - estimating the partition function Z and marginal probability distributions

We focus on *binary pairwise* (Ising) models, e.g. vision, RBMs, or social networks

Combining *clamping* of variables with *approximate variational inference* we obtain

- Strong theoretical results (middle column)
- Promising empirical results (right column)

Background: What is clamping?



$x_1x_2\ldots x_{10}$	*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**	

Z can be split into two parts: 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 clamping a variable, remove it from the graph

- If remaining sub-models are acyclic then can find sub-partition functions efficiently (Bethe approximation is exact on trees)
- If not,
  - (a) Can repeat until acyclic, *or*
  - (b) Settle for approximate inference on sub-models

    Will clamping and summing approximate sub-partition functions always lead to a better estimate of Z than approximate inference on the original model?

    Often but not always (see paper for example)

Variational inference

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

- Exact inference may be viewed as *optimization*,

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

- Bethe makes 2 pairwise approximations,

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

- Bethe is exact on trees
- Observe that when X_i is clamped, we optimize over a subset

$$\log Z_B|_{X_i=0} = \max_{q \in \mathbb{L}: q_i=0} [\,\theta \cdot q + S_B(q)\,], \qquad q_i = q(X_i = 1)$$

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

- **Notation:** If we *clamp* variable X_i *and sum approximate* sub-partition functions,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$
$$\le 2Z_B \quad \text{by above}$$

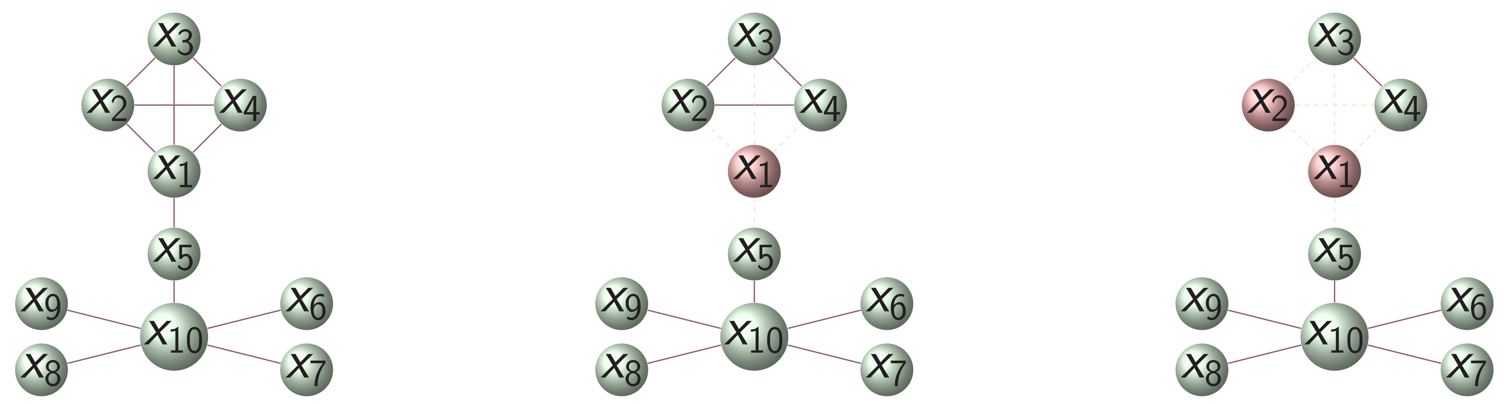
General models: upper bound on Z

Upper bound on $Z_B^{(i)}$ leads to upper bound on Z

- $$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1} \le 2Z_B$$
- Repeat, clamping variables until remaining model is acyclic, where Bethe is exact
- For example, if we must delete 2 variables X_i, X_j, obtain

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

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



Let $\nu(G)$ be minimum size of a feedback vertex set (set of vertices such that deleting them renders graph acyclic; $\nu \ge$ treewidth-1)

Theorem (result is tight)

$$Z \le 2^{\nu} Z_B$$

Attractive models: Master Theorem (strongest result)

An *attractive* model has all variables positively coupled

- 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 singleton $q_i = x$*'
- Define new function,

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

- By considering derivatives of the Bethe free energy, and how the optimum constrained to singleton q_i varies with q_i, we show

Theorem (strongest result for attractive models)

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

Attractive models: Consequences of Master Theorem

Lower bound on $Z_B^{(i)}$ leads to lower bound on Z

Theorem (clamp and sum can only increase Bethe)

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

Then with similar proof to result above for general models,

Corollary (lower bound on Z, first proved by Ruozzi, 2012)

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

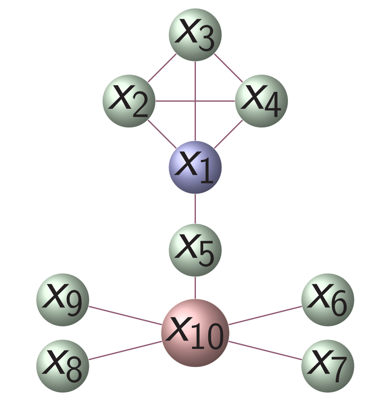
$\Rightarrow$ clamping can only *improve* the estimate of the partition function

$$Z_B \le Z_B^{(i)} \le Z_B^{(ij)} \le \cdots \le Z$$

Experiments

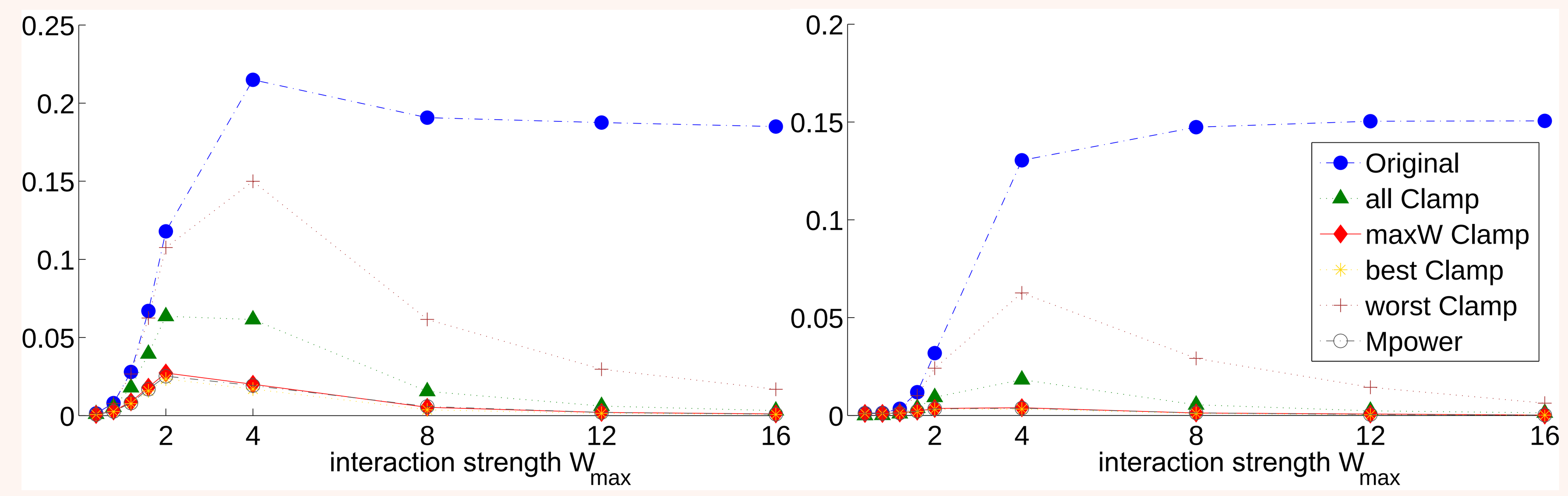
We investigate error of $Z_B^{(i)} = Z_B|_{X_i=0} + Z_B|_{X_i=1}$ compared to error of original Z_B, various methods of choosing variable to clamp X_i:

- 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, tries to break heavy cycles



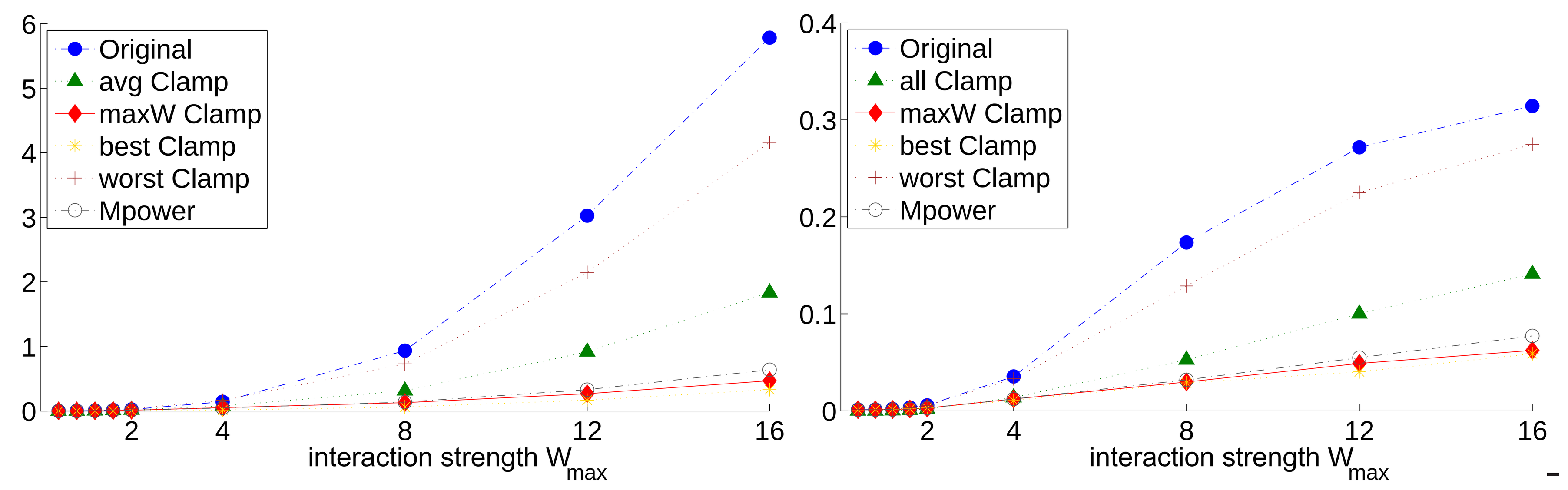
- Potentials drawn at random: unary $\theta_i \sim U[-2,2]$, edge $W_{ij} \sim U[-W_{max}, W_{max}]$ for *general*, $W_{ij} \sim U[0, W_{max}]$ for *attractive* models

Attractive random graph $n = 10, p = 0.5$



Error of $\log Z$ **Avg ℓ_1 error marginals**

General random graph $n = 10, p = 0.5$



As n grows, still helpful even just to clamp one variable (see paper)

General 'lamp' graph