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*
Binary pairwise models

- Binary variables \( X_1, \ldots, X_n \in \{0, 1\} \)
- Singleton and pairwise potentials \( \theta \)
- 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*

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

- Exact inference may be viewed as *optimization*,
  \[
  \log Z = \max_{\mu \in \mathcal{M}} \left[ \theta \cdot \mu + S(\mu) \right]
  \]
  \( \mathcal{M} \) is the space of marginals that are *globally consistent*, \( S \) is the (Shannon) entropy

- Bethe makes two pairwise approximations,
  \[
  \log Z_B = \max_{q \in \mathcal{L}} \left[ \theta \cdot q + S_B(q) \right]
  \]
  \( \mathcal{L} \) 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 \( Z_B = Z \)
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
Example model

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

<table>
<thead>
<tr>
<th>$x_1x_2 \ldots x_{10}$</th>
<th>score</th>
<th>$\exp(score)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 \ldots 0</td>
<td>1</td>
<td>2.7</td>
</tr>
<tr>
<td>0 0 \ldots 1</td>
<td>2</td>
<td>7.4</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>0 1 \ldots 1</td>
<td>1.3</td>
<td>3.7</td>
</tr>
<tr>
<td>1 0 \ldots 0</td>
<td>-1</td>
<td>0.4</td>
</tr>
<tr>
<td>1 0 \ldots 1</td>
<td>0.2</td>
<td>1.2</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>1 1 \ldots 1</td>
<td>1.8</td>
<td>6.0</td>
</tr>
<tr>
<td><strong>Total $Z =$</strong></td>
<td></td>
<td><strong>47.1</strong></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$x_1x_2 \ldots x_{10}$</th>
<th>score</th>
<th>$\exp(score)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 \ldots 0</td>
<td>1</td>
<td>2.7</td>
</tr>
<tr>
<td>0 0 \ldots 1</td>
<td>2</td>
<td>7.4</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>0 1 \ldots 1</td>
<td>1.3</td>
<td>3.7 27.5</td>
</tr>
<tr>
<td>1 0 \ldots 0</td>
<td>-1</td>
<td>0.4</td>
</tr>
<tr>
<td>1 0 \ldots 1</td>
<td>0.2</td>
<td>1.2</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>1 1 \ldots 1</td>
<td>1.8</td>
<td>6.0 19.6</td>
</tr>
</tbody>
</table>

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

Total $Z = 47.1$
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)
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

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

*Note: The diagram represents a network with nodes $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ connected by edges.*
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

\[
\log Z_B = \max_{q \in \mathcal{L}} \left[ \theta \cdot q + S_B(q) \right]
\]

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

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

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

<table>
<thead>
<tr>
<th>Recap of Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>( Z )</strong></td>
</tr>
<tr>
<td><strong>( Z_B )</strong></td>
</tr>
<tr>
<td><strong>( Z_B^{(i)} )</strong> := ( Z_B</td>
</tr>
<tr>
<td>( \leq 2Z_B )</td>
</tr>
</tbody>
</table>

11 / 21
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: \textit{an upper bound on }$Z$

\[ Z^{(i)}_B := Z_B|_{X_i=0} + Z_B|_{X_i=1} \leq 2Z_B \]

- 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

\[ Z \leq 2^k Z_B \]
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, $Z_B \leq Z$*
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

\[
Z_B \leq Z \leq 2^k Z_B
\]

- **Attractive only**
- **Attractive and mixed**

\[
\Leftrightarrow \quad \frac{Z}{2^k} \leq Z_B \leq Z
\]

- **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, $Z \leq Z_T$  
  $Z_M \leq Z \leq Z_T$
- 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?  
  $Z_M \leq Z_B \leq Z_T$
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, $Z \leq Z_T$  $Z_M \leq Z \leq Z_T$
- 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? $Z_M \leq Z_B \leq Z_T$

**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

- Large grid (9x9)
  - Best
  - Worst
  - Pseudo
  - Greedy

- Small grid (5x5)
  - Best
  - Worst
  - Pseudo
  - Greedy

Attractive grids: $[0, 6]$

Mixed grids: $[-6, 6]$
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/


Extra slides for questions or further explanation
For dense mixed models (many edges),

**MF can be better than Bethe**

What happens if we increase edge strength?
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

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

- For any variable \( X_i \) and \( x \in [0, 1] \), let \( q_i = q(X_i = 1) \) and
  \[ \log Z_{Bi}(x) = \max_{q \in \mathbb{L}: q_i = x} \left[ \theta \cdot q + S_B(q) \right] \]

- \( Z_{Bi}(x) \) 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

- $\text{maxW}$: max sum of incident edge weights $\sum_{j \in N(i)} |W_{ij}|$
- $\text{Mpower}$: 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 \( W_{ij} \sim U[0, W_{\text{max}}] \)

**Error of estimate of \( \log Z \)**

**Observe**

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

**Avg \( \ell_1 \) 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 $W_{ij} \sim U[-W_{\text{max}}, W_{\text{max}}]$

Error of estimate of $\log Z$

Results remain promising for higher $n$

Avg $\ell_1$ 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 $W_{ij} \sim U[-W_{\text{max}}, W_{\text{max}}]$

Error of estimate of log $Z$

Note low unary potentials

Avg $\ell_1$ 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 \( W_{ij} \sim U[0, W_{\text{max}}] \)

Error of estimate of log \( Z \)

Note regular singleton potentials

Avg \( \ell_1 \) error of singleton marginals
Experiments: *attractive random graph* $n = 50, p = 0.1$

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

Error of estimate of $\log Z$

‘worst Clamp’ performs worse here due to suboptimal solutions found by Frank-Wolfe

Avg $\ell_1$ error of singleton marginals
Experiments: *mixed random graph* $n = 50, p = 0.1$

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

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

Error of estimate of log $Z$

Performance still good for clamping just one variable

Avg $\ell_1$ 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 \( \ell_1 \) error of singleton marginals
Experiments: *mixed ‘lamp’ graph*

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

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

Error of estimate of $\log Z$

$Mpower$ performs well, significantly better than $\max W$

Avg $\ell_1$ error of singleton marginals