Clamping Improves TRW and Mean Field Approximations
Adrian Weller and Justin Domke

SUMMARY
We examine the effect of clamping variables for approximate inference in undirected graphical models with pairwise relationships and discrete variables.

• For any number of variable labels, we demonstrate that clamping and summing approximate sub-partition functions can lead only to a decrease in the partition function estimate for TRW, and an increase for the naive mean field method, in each case guaranteeing an improvement in the approximation and bound.

• In contrast, Weller and Jebara (2014) showed that clamping with the Bethe approximation for attractive binary pairwise models always improves the global optimum, but for mixed models it can sometimes lead to a worse approximation.

• We introduce new ways to select variables to clamp, including: stripping to the core, identifying highly frustrated cycles, and checking singleton entropies.

BACKGROUND: VARIATIONAL INFERENCE
Computing the partition function may be formulated as an optimization problem,
\[
\log Z = \max_{\mu \in \mathcal{M}} \theta \cdot \mu + H(\mu).
\]
We consider three approximations:

• Naive mean field, restricts to fully factorized, \( \mu \in \mathcal{M}^\star \) where \( \mu(x) = \prod_i \mu_i(x_i) \)

\[
\log \hat{Z}_\mu = \max_{\mu \in \mathcal{M}^\star} \theta \cdot \mu + H(\mu).
\]

• Bethe, relaxes \( \mathcal{M} \) to the local polytope \( \mathcal{L} \) and uses Bethe entropy \( \tilde{H}_\theta \approx H \)

\[
\log \tilde{Z}_\theta = \max_{\mu \in \mathcal{L}} \theta \cdot \mu + \tilde{H}_\theta(\mu).
\]

• Tree-reweighted (TRW) also relaxes to \( \mathcal{L} \) but uses TRW entropy \( \tilde{H}_T \geq H \)

\[
\log \tilde{Z}_T = \max_{\mu \in \mathcal{L}} \theta \cdot \mu + \tilde{H}_T(\mu).
\]

The following inequalities hold, hence improving \( \hat{Z}_\mu \) and \( \tilde{Z}_T \) gives useful bounds:
\[
\hat{Z}_\mu \leq Z \leq \tilde{Z}_T \quad \text{and} \quad \hat{Z}_\mu \leq \hat{Z}_T \leq \tilde{Z}_T.
\]

CLAMPING TO ESTIMATE THE PARTITION FUNCTION
We consider an example for binary variables.

\[
Z = \sum_{x_1} \sum_{x_2} \mu(x_1, x_2) = \sum_{x_1} \mu_1(x_1) \sum_{x_2} \mu_2(x_2)
\]

After we clamp a variable \( x_1 \), it may be removed

• If the remaining sub-models are acyclic, then we can find the sub-partition functions efficiently

• If not,

  (a) Can repeat until acyclic, or

  (b) Settle for approximate inference on sub-models

\[
\tilde{Z}(i) = Z[x_1 \rightarrow 0] + Z[x_1 \rightarrow 1] \quad \text{(a new approx)}
\]

Theorems:
\[
\forall x_i, \quad \hat{Z}_\mu \leq \tilde{Z}_\mu(i) \leq Z \leq \tilde{Z}_T(i) \leq \tilde{Z}_T
\]

SELECTING A VARIABLE TO CLAMP, SINGLETON ENTROPY
We explore various methods to select a good variable, or sequence of variables, to clamp. See the paper for details.

• If the singleton entropy of a variable is already low, then it is already ‘almost’ clamped and we should avoid picking that variable for actual clamping.

• We use the TRW singleton entropy approximation to scale our heuristics when selecting a variable. We call this TRE (for TRW Entropy).

Example of a model where the earlier maxW and Mpower heuristics of Weller and Jebara (2014) perform poorly to select variables to clamp but our new methods perform well. Solid blue (dashed red) edges are strongly attractive (repulsive). maxW and Mpower both repeatedly select variables in \( X_1, \ldots, X_3 \); the first clamp is good but less so than picking from the frustrated cycle \( X_1, X_3, X_5 \); then repeat clamping in \( X_1, \ldots, X_3, \ldots, X_5 \) reap little benefit. It is best to pick \( X_3 \) then \( X_5 \).

REFERENCE

STRIPPING TO THE CORE
On the left is our example model. On the right is the core of the same model, obtained by iteratively removing variables with degree 1. Methods to select a variable to clamp might pick \( X_3 \) (which has highest degree) in the original model, whereas any of \( X_1, \ldots, X_4 \) would be better, and will be selected if the model is first stripped to its core.

FRUSTRATED CYCLES
In a binary pairwise model, each edge may be characterized as attractive (pulls variables toward the same value; shown in blue) or repulsive (pushes variables apart to different values; shown in red). A frustrated cycle contains an odd number of repulsive edges. These are challenging for many methods of inference.

Error in log Z vs number of clamps: grids

Error in log Z vs number of clamps: complete graphs

GUIDANCE FOR PRACTITIONERS
Based on our empirical results, we suggest:

• Typically Bethe performs best, provided convergence difficulties do not arise.

• However, for densely connected mixed models with strong edges, mean field can be much more accurate (optimizes within the marginal polytope, handles frustrated cycles much better).

• Clamping can be very helpful, more so for denser models with stronger edge weights, a setting where inference on the original model is hard.

• In some settings, the combined time to run inference on both clamped sub-models was faster than for the original model.

• No one clamping selection method dominated. The best was maxW+core+TRE. Guaranteed bounds from TRW and mean field can be very helpful.

Contact Adrian aw665@cam.ac.uk  Download the full paper at http://alg.eng.cam.ac.uk/adrian

AISTATS 2016, Cadiz