Spectral Approximate Inference

Speaker: Sejun Park[1]

Joint work with Eunho Yang[1,2], Se-Young Yun[1] and Jinwoo Shin[1,2]

[1]Korea Advanced Institute of Science and Technology (KAIST)

[2]AITRICS

June 13th, 2019

Goal: Partition Function Approximation in GMs

Pairwise binary graphical model (GM) is a joint distribution, factorized by

- computer vision *[Freeman et al., 2000]*, social science *[Scott, 2017]* and deep learning *[Hinton et al., 2006]*

$$\mathbb{P}(\mathbf{x}) = \frac{1}{Z} \exp\left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle + \mathbf{x}^T A \mathbf{x}\right) \qquad \mathbf{x} \in \{-1, 1\}^n, \boldsymbol{\theta} \in \mathbb{R}^n, A \in \mathbb{S}^{n \times n}$$

Partition function Z is essential for inference, but it is NP-hard even to approximate *[Jerrum, 1993]*

$$Z = \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle + \mathbf{x}^T A \mathbf{x}\right)$$

Goal: Partition Function Approximation in GMs

Pairwise binary graphical model (GM) is a joint distribution, factorized by

- computer vision *[Freeman et al., 2000]*, social science *[Scott, 2017]* and deep learning *[Hinton et al., 2006]*

$$\mathbb{P}(\mathbf{x}) = \frac{1}{Z}\exp\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle + \mathbf{x}^T A \mathbf{x}\right) \qquad \mathbf{x} \in \{-1,1\}^n, \boldsymbol{\theta} \in \mathbb{R}^n, A \in \mathbb{S}^{n\times n}$$

Partition function Z is essential for inference, but it is NP-hard even to approximate *[Jerrum, 1993]*

$$Z = \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle + \mathbf{x}^T A \mathbf{x}\right)$$

In theory, Z of only a few restricted classes of GM can be approximated in polynomial time

1. Structured GMs: e.g., A is an adjacency matrix of tree/planar graphs *[Temperley et al., 1961; Pearl, 1982]*
2. GMs with homogeneous parameters: e.g., $A, \theta \geq 0$ *[Jerrum, 1993; Li et al., 2013; Liu, 2018]*
3. GMs under correlation decay/tree uniqueness: e.g., $\tanh|A_{ij}| \leq \frac{1}{\max_i |\{A_{ij} : A_{ij} \neq 0\}| - 1}$ *[Li et al., 2013]*

Goal: Partition Function Approximation in GMs

Pairwise binary graphical model (GM) is a joint distribution, factorized by

- computer vision *[Freeman et al., 2000]*, social science *[Scott, 2017]* and deep learning *[Hinton et al., 2006]*

$$\mathbb{P}(\mathbf{x}) = \frac{1}{Z} \exp\left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle + \mathbf{x}^T A \mathbf{x}\right) \qquad \mathbf{x} \in \{-1,1\}^n, \boldsymbol{\theta} \in \mathbb{R}^n, A \in \mathbb{S}^{n \times n}$$

Partition function Z is essential for inference, but it is NP-hard even to approximate *[Jerrum, 1993]*

$$Z = \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\langle \boldsymbol{\theta}, \mathbf{x} \rangle + \mathbf{x}^T A \mathbf{x}\right)$$

In practice, approximation algorithms based on certain local structures/consistency have been used

1. Markov chain Monte Carlo: e.g., annealed importance sampling *[Neal, 2001]*
2. Variational inference: e.g., belief propagation *[Pearl, 1982],* mean-field approximation *[Parisi, 1988]*
3. Variable elimination: e.g., minibucket *[Dechter et al., 2003]*, weighted minibucket *[Lie et al., 2011]*

However, due to their local nature, they often fails under large global correlation (i.e., large A)

Goal: Partition Function Approximation in GMs

Pairwise binary graphical model (GM) is a joint distribution, factorized by

- computer vision [Freeman et al., 2000], social science [Scott, 2017] and deep learning [Hinton et al., 2006]

Partition func... ...um, 1993]

We study the spectral properties of the parameter matrix A for more robust approximate inference

In practice, a... ...been used

1. Markov chain Monte Carlo: e.g., annealed importance sampling [Neal, 2001]
2. Variational inference: e.g., belief propagation [Pearl, 1982], mean-field approximation [Parisi, 1988]
3. Variable elimination: e.g., minibucket [Dechter et al., 2003], weighted minibucket [Liu et al., 2011]

However, due to their local nature, they often fails under large global correlation (i.e., large A)

Spectral Approximate Inference for Low-Rank GMs

Provable approximate inference algorithm for low-rank GMs (low-rank A)

Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A $(\boldsymbol{\theta} = 0, \mathrm{rank}(A) = 1, A = \lambda \mathbf{v}\mathbf{v}^T)$

$$Z = \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right)$$
$$= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\lambda \langle \mathbf{v}, \mathbf{x} \rangle^2\right)$$

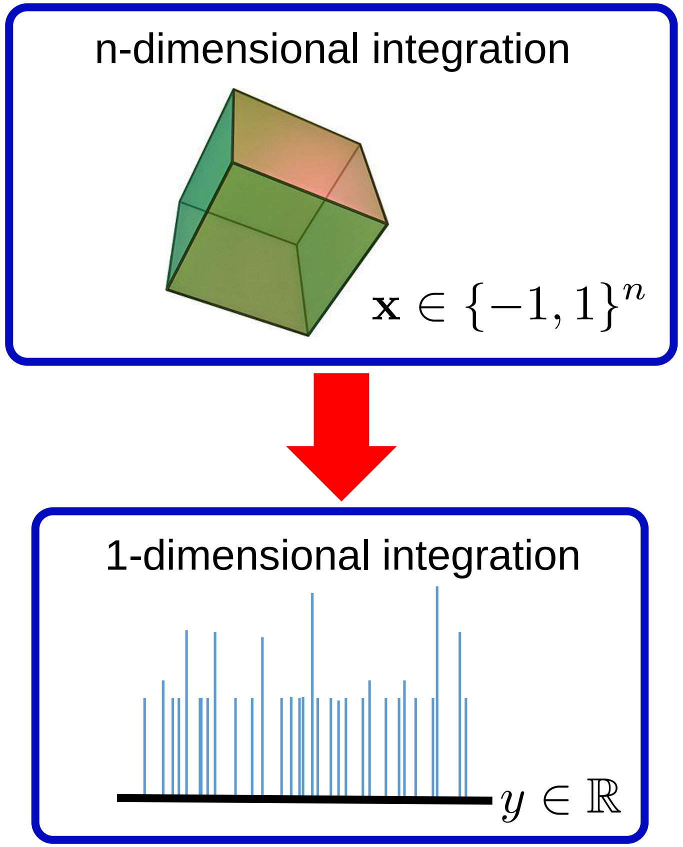
Eigenvalue decomposition

Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A $(\boldsymbol{\theta} = 0, \mathrm{rank}(A) = 1, A = \lambda \mathbf{v}\mathbf{v}^T)$

1. Transform the domain of integration from $\mathbf{x}$ to $\langle \mathbf{v}, \mathbf{x} \rangle$ using the identity $\mathbf{x}^T A \mathbf{x} = \lambda \langle \mathbf{v}, \mathbf{x} \rangle^2$

$$
\begin{aligned}
Z &= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\lambda \langle \mathbf{v}, \mathbf{x} \rangle^2\right) \\
&= \sum_{y = \langle \mathbf{v}, \mathbf{x} \rangle} \left| \{\mathbf{x} : y = \langle \mathbf{v}, \mathbf{x} \rangle\} \right| \exp\left(\lambda y^2\right)
\end{aligned}
$$

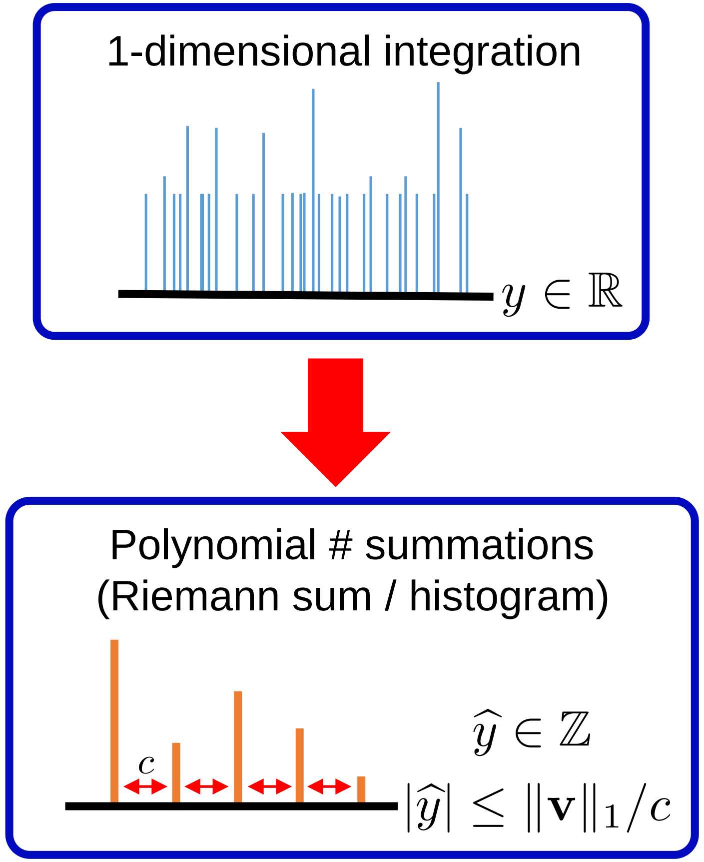


Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A $(\boldsymbol{\theta} = 0, \mathrm{rank}(A) = 1, A = \lambda \mathbf{v}\mathbf{v}^T)$

1. Transform the domain of integration from $\mathbf{x}$ to $\langle \mathbf{v}, \mathbf{x} \rangle$ using the identity $\mathbf{x}^T A \mathbf{x} = \lambda \langle \mathbf{v}, \mathbf{x} \rangle^2$
2. Approximate 1-dimensional integration into a polynomial number of summations using histogram

$$
\begin{aligned}
Z &= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\lambda \langle \mathbf{v}, \mathbf{x} \rangle^2\right) \\
&= \sum_{y = \langle \mathbf{v}, \mathbf{x} \rangle} \left| \{ \mathbf{x} : y = \langle \mathbf{v}, \mathbf{x} \rangle \} \right| \exp\left(\lambda y^2\right) \\
&\approx \sum_{\widehat{y} \in \mathbb{Z} : |\widehat{y}| \leq \|\mathbf{v}\|_1 / c} \left| \{ \mathbf{x} : c \cdot \widehat{y} \approx \langle \mathbf{v}, \mathbf{x} \rangle \} \right| \exp\left(\lambda \widehat{y}^2\right)
\end{aligned}
$$



Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A ($\boldsymbol{\theta} = 0, \mathrm{rank}(A) = 1, A = \lambda \mathbf{v}\mathbf{v}^T$)

1. Transform the domain of integration from $\mathbf{x}$ to $\langle \mathbf{v}, \mathbf{x} \rangle$ using the identity $\mathbf{x}^T A \mathbf{x} = \lambda \langle \mathbf{v}, \mathbf{x} \rangle^2$
2. Approximate 1-dimensional integration into a polynomial number of summations using histogram
3. Compute the weight of the histogram recursively

$$
\begin{aligned}
Z &= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\lambda \langle \mathbf{v}, \mathbf{x} \rangle^2\right) \\
&= \sum_{y = \langle \mathbf{v}, \mathbf{x} \rangle} \left|\{\mathbf{x} : y = \langle \mathbf{v}, \mathbf{x} \rangle\}\right| \exp\left(\lambda y^2\right) \\
&\approx \sum_{\widehat{y} \in \mathbb{Z} : |\widehat{y}| \le \|\mathbf{v}\|_1 / c} \underbrace{\left|\{\mathbf{x} : c \cdot \widehat{y} \approx \langle \mathbf{v}, \mathbf{x} \rangle\}\right|}_{\text{Weight of histogram}} \exp\left(\lambda \widehat{y}^2\right)
\end{aligned}
$$

Weight of histogram

For $\mathbf{x}, \ \mathbf{x}'$ differing only at $\mathbf{x}_i = 1, \ \mathbf{x}'_i = -1$

$$\langle \mathbf{v}, \mathbf{x} \rangle = \langle \mathbf{v}, \mathbf{x}' \rangle + 2v_i$$

$$t_i(\widehat{y}) = t_{i-1}(\widehat{y}) + t_{i-1}(\widehat{y} - \lfloor 2v_i / c \rfloor)$$

$$t_i(\widehat{y}) = \left|\{\mathbf{x} : c \cdot \widehat{y} \approx \langle \mathbf{v}, \mathbf{x} \rangle, x_j = -1 \ \forall j > i\}\right|$$

$$t_n(\widehat{y}) = \left|\{\mathbf{x} : c \cdot \widehat{y} \approx \langle \mathbf{v}, \mathbf{x} \rangle\}\right|$$

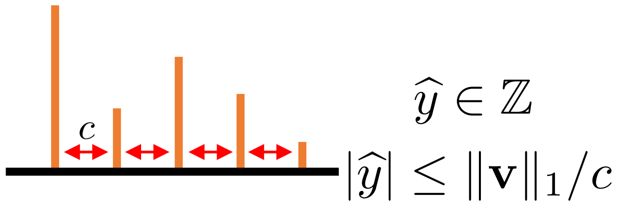
Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A ($\boldsymbol{\theta} = 0, \mathrm{rank}(A) = 1, A = \lambda \mathbf{v}\mathbf{v}^T$)

1. Transform the domain of integration from $\mathbf{x}$ to $\langle \mathbf{v}, \mathbf{x} \rangle$ using the identity $\mathbf{x}^T A \mathbf{x} = \lambda \langle \mathbf{v}, \mathbf{x} \rangle^2$
2. Approximate 1-dimensional integration into a polynomial number of summations using histogram
3. Compute the weight of the histogram recursively
4. Compute the approximated Z from using the histogram

$$
\begin{aligned}
Z &= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\lambda \langle \mathbf{v}, \mathbf{x} \rangle^2\right) \\
&= \sum_{y = \langle \mathbf{v}, \mathbf{x} \rangle} \left| \{\mathbf{x} : y = \langle \mathbf{v}, \mathbf{x} \rangle\} \right| \exp\left(\lambda y^2\right) \\
&\approx \sum_{\widehat{y} \in \mathbb{Z} : |\widehat{y}| \le \|\mathbf{v}\|_1 / c} \left| \{\mathbf{x} : c \cdot \widehat{y} \approx \langle \mathbf{v}, \mathbf{x} \rangle\} \right| \exp\left(\lambda \widehat{y}^2\right) = \widehat{Z}
\end{aligned}
$$

Polynomial # summations
(Riemann sum / histogram)



Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A

- The procedure for rank-1 GMs generalizes to arbitrary GMs by considering the histogram of $\mathrm{rank}(A)$-dimension

Theorem *[Park et al., 2019]*

For any $\varepsilon > 0$, the algorithm outputs $\widehat{Z}$ such that

$$(1-\varepsilon)Z \leq \widehat{Z} \leq (1+\varepsilon)Z$$

in $(n/\varepsilon)^{O(\mathrm{rank}(A))}$ time

Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A

- The procedure for rank-1 GMs generalizes to arbitrary GMs by considering the histogram of $\mathrm{rank}(A)$-dimension

Theorem *[Park et al., 2019]*

For any $\varepsilon > 0$, the algorithm outputs $\widehat{Z}$ such that

$$(1-\varepsilon)Z \leq \widehat{Z} \leq (1+\varepsilon)Z$$

in $(n/\varepsilon)^{O(\mathrm{rank}(A))}$ time

- The proposed algorithm is a fully polynomial-time approximation scheme (FPTAS) for $\mathrm{rank}(A) = O(1)$
- However, it is hard to use the algorithm for general GMs due to its complexity, exponential to $\mathrm{rank}(A)$

Spectral Approximate Inference for Low-Rank GMs

Proposed algorithm using spectral properties of A

- The procedure for rank-1 GMs generalizes to arbitrary GMs by considering the histogram of $\mathrm{rank}(A)$-dimension

Theorem *[Park et al., 2019]*

For any $\varepsilon > 0$, the algorithm outputs $\widehat{Z}$ such that

$$(1-\varepsilon)Z \le \widehat{Z} \le (1+\varepsilon)Z$$

in $(n/\varepsilon)^{O(\mathrm{rank}(A))}$ time

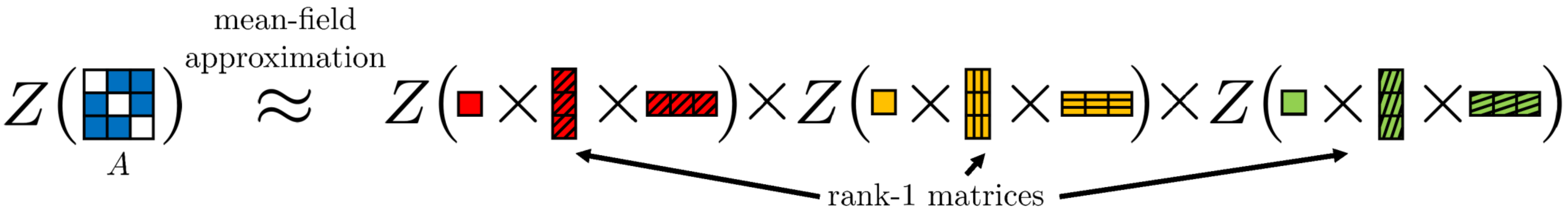
- The proposed algorithm is a fully polynomial-time approximation scheme (FPTAS) for $\mathrm{rank}(A) = O(1)$
- However, it is hard to use the algorithm for general GMs due to its complexity, exponential to $\mathrm{rank}(A)$

Next: we propose an algorithm for general high-rank GMs

Spectral Approximate Inference for High-Rank GMs

Approximation algorithm for general high-rank GMs

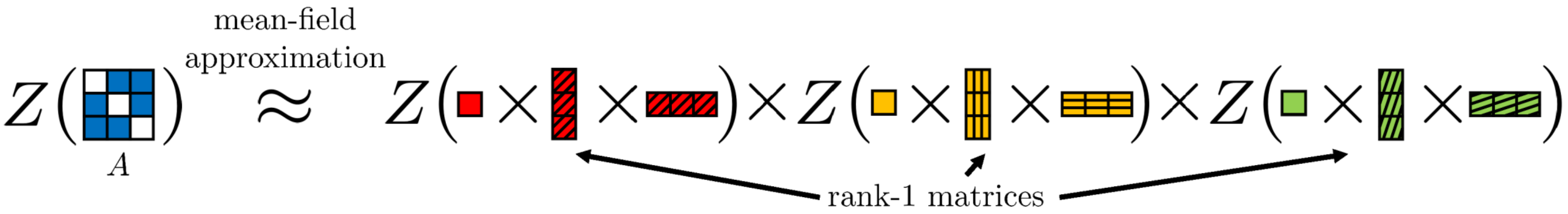
Mean-field approximation



Mean-field approximation

$$Z = \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right)$$
$$= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T A \mathbf{x}\right)$$

Transform summation into expectation

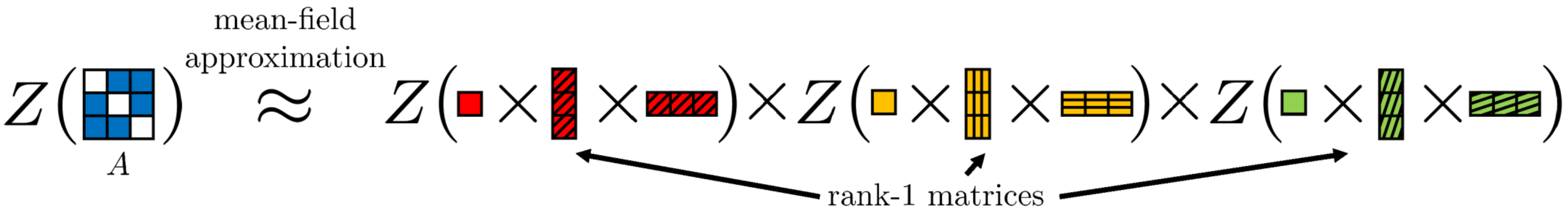


Spectral Approximate Inference for High-Rank GMs

Mean-field approximation

$$Z = \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\mathbf{x}^T A\mathbf{x}\right)$$
$$= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T A\mathbf{x}\right)$$
$$= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\sum_{i=1}^{n} \lambda_i \langle \mathbf{v}_i, \mathbf{x}\rangle^2\right)$$

Eigenvalue decomposition



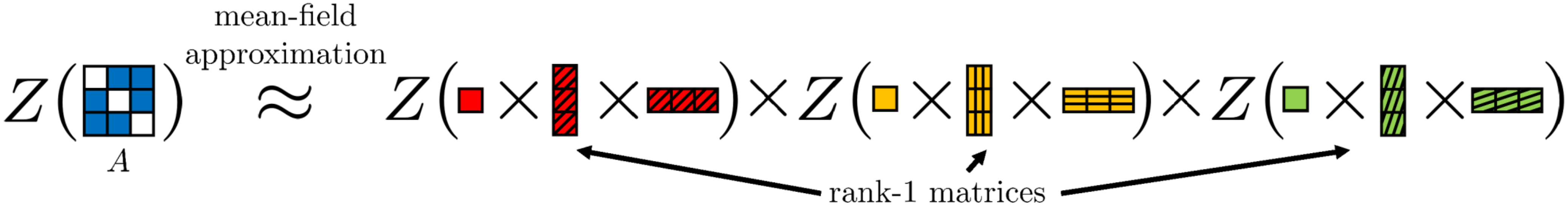
Spectral Approximate Inference for High-Rank GMs

Mean-field approximation

$$
\begin{aligned}
Z &= \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\sum_{i=1}^{n} \lambda_i \langle \mathbf{v}_i, \mathbf{x}\rangle^2\right) \\
&\approx 2^n \prod_{i=1}^{n} \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\lambda_i \langle \mathbf{v}_i, \mathbf{x}\rangle^2\right)
\end{aligned}
$$

Mean-field approximation

Product of rank-1 expectations



Spectral Approximate Inference for High-Rank GMs

Mean-field approximation

$$
\begin{aligned}
Z &= \sum_{\mathbf{x} \in \{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= 2^n \mathbb{E}_{\mathbf{x} \sim \text{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\
&= 2^n \mathbb{E}_{\mathbf{x} \sim \text{Uniform}(\{-1,1\}^n)} \exp\left(\sum_{i=1}^{n} \lambda_i \langle \mathbf{v}_i, \mathbf{x} \rangle^2\right) \\
&\approx 2^n \prod_{i=1}^{n} \mathbb{E}_{\mathbf{x} \sim \text{Uniform}(\{-1,1\}^n)} \exp\left(\lambda_i \langle \mathbf{v}_i, \mathbf{x} \rangle^2\right)
\end{aligned}
$$

Mean-field approximation

Product of rank-1 expectations

Controlling the mean-field approximation by varying the spectral property

Spectral Approximate Inference for High-Rank GMs

Mean-field approximation

$$\begin{aligned} Z &= \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\ &= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T A \mathbf{x}\right) \\ &= 2^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\sum_{i=1}^n \lambda_i \langle \mathbf{v}_i, \mathbf{x}\rangle^2\right) \\ &\approx 2^n \prod_{i=1}^n \mathbb{E}_{\mathbf{x}\sim\text{Uniform}(\{-1,1\}^n)} \exp\left(\lambda_i \langle \mathbf{v}_i, \mathbf{x}\rangle^2\right) \end{aligned}$$

Mean-field approximation

Product of rank-1 expectations

Controlling the mean-field approximation by varying the spectral property

Free parameter: diagonal matrix D

$$Z(A) = \exp\left(-\text{trace}(D)\right) Z(A + D)$$

Spectral Approximate Inference for High-Rank GMs

Mean-field approximation with a diagonal matrix D

$$Z = \exp(-\mathrm{trace}(D)) \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\mathbf{x}^T (A + D)\mathbf{x}\right)$$
$$= 2^n \exp(-\mathrm{trace}(D))\mathbb{E}_{\mathbf{x}\sim\mathrm{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T (A + D)\mathbf{x}\right)$$
$$= 2^n \exp(-\mathrm{trace}(D))\mathbb{E}_{\mathbf{x}\sim\mathrm{Uniform}(\{-1,1\}^n)} \exp\left(\sum_{i=1}^{n} \lambda_i^D \langle \mathbf{v}_i^D, \mathbf{x}\rangle^2\right)$$
$$\approx 2^n \exp(-\mathrm{trace}(D)) \prod_{i=1}^{n} \mathbb{E}_{\mathbf{x}\sim\mathrm{Uniform}(\{-1,1\}^n)} \exp\left(\lambda_i^D \langle \mathbf{v}_i^D, \mathbf{x}\rangle^2\right)$$

Goal: Reduce the error by choosing proper D

Controlling the mean-field approximation by varying the spectral property

Free parameter: diagonal matrix D

$$Z(A) = \exp\left(-\mathrm{trace}(D)\right) Z(A + D)$$

Spectral Approximate Inference for High-Rank GMs

Mean-field approximation with a diagonal matrix D

$$
\begin{aligned}
Z &= \exp(-\mathrm{trace}(D)) \sum_{\mathbf{x}\in\{-1,1\}^n} \exp\left(\mathbf{x}^T(A+D)\mathbf{x}\right) \\
&= 2^n \exp(-\mathrm{trace}(D)) \mathbb{E}_{\mathbf{x}\sim\mathrm{Uniform}(\{-1,1\}^n)} \exp\left(\mathbf{x}^T(A+D)\mathbf{x}\right) \\
&= 2^n \exp(-\mathrm{trace}(D)) \mathbb{E}_{\mathbf{x}\sim\mathrm{Uniform}(\{-1,1\}^n)} \exp\left(\sum_{i=1}^{n} \lambda_i^D \langle \mathbf{v}_i^D, \mathbf{x}\rangle^2\right) \\
&\approx 2^n \exp(-\mathrm{trace}(D)) \prod_{i=1}^{n} \mathbb{E}_{\mathbf{x}\sim\mathrm{Uniform}(\{-1,1\}^n)} \exp\left(\lambda_i^D \langle \mathbf{v}_i^D, \mathbf{x}\rangle^2\right)
\end{aligned}
$$

Goal: Reduce the error by choosing proper D

Optimizing diagonal matrix D for reducing the approximation error

Semi-definite programming

$$
\begin{aligned}
\mathrm{maximize}_D \quad & \mathrm{trace}(D) \\
\text{subject to} \quad & A + D \preceq 0
\end{aligned}
$$

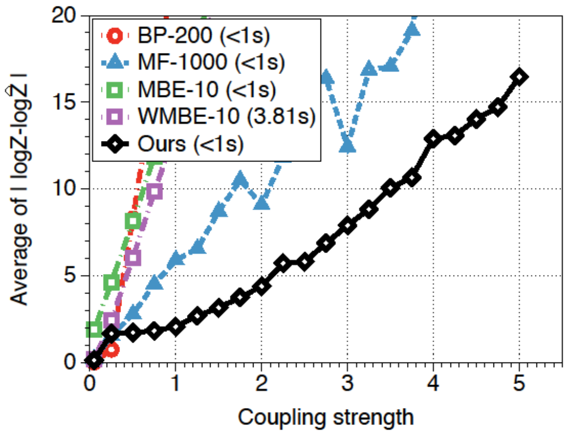
Experiments

Comparing our algorithm with popular approximate inference algorithms

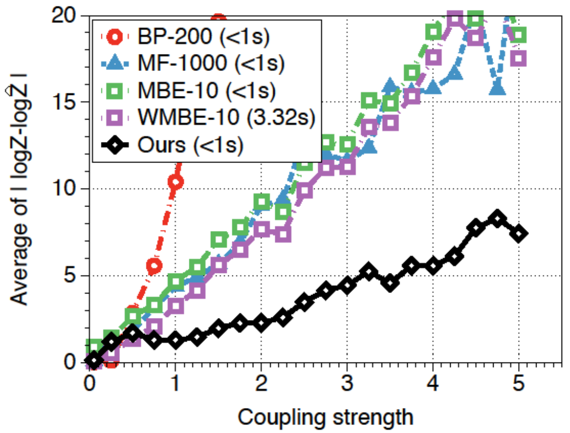
- Compared algorithms: Belief propagation *[Pearl, 1982]*, mean-field approximation *[Parisi, 1988]*, minibucket *[Dechter et al., 2003]*, weighted elimination *[Liu et al., 2011]*
- Synthetic dataset: Generated by varying the absolute magnitude of A (coupling strength)
- UAI grid dataset: Indices 1-4 are GMs on 10x10 grid graph and indices 5-8 are GMs on 20x20 grid graph

Our algorithm outperforms others even under large global correlation (i.e., large A)

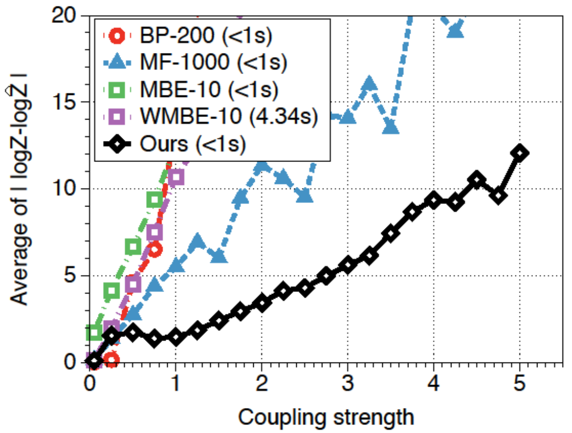
Complete graph on 20 vertices



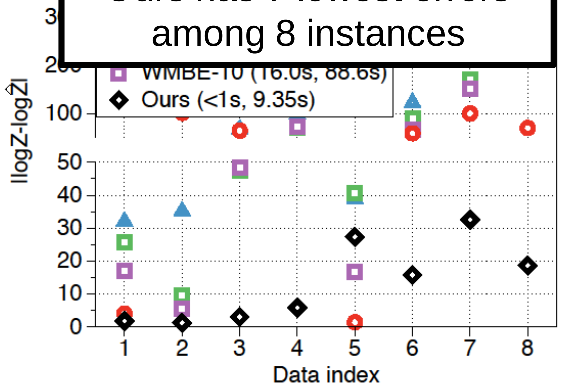
ER graph on 20 vertices: p=0.5



ER graph on 20 vertices: p=0.7



UAI grid dataset



Ours has 7 lowest errors among 8 instances

Conclusion

We develop partition function approximation algorithms using spectral properties of the parameter matrix

- For low-rank GMs, we propose a provable algorithm
- For high-rank GMs, we propose a mean-field type algorithm

Poster #213 Today @ Pacific Ballroom