$P_{err} \leq P_{err_B} \cdot 2^{n/2}$

$\frac{1}{n} \log_e 2 = \frac{1}{n}(0.69) = \eta$

$\eta_w = \frac{1}{n} \cancel{\log} 2$ $\qquad$ $\frac{1}{n} \cancel{\log} 2$ $\qquad$ $\frac{1}{n} \log_2$ $\qquad$ $\eta_r = \frac{1}{n} \cancel{\log} 2$

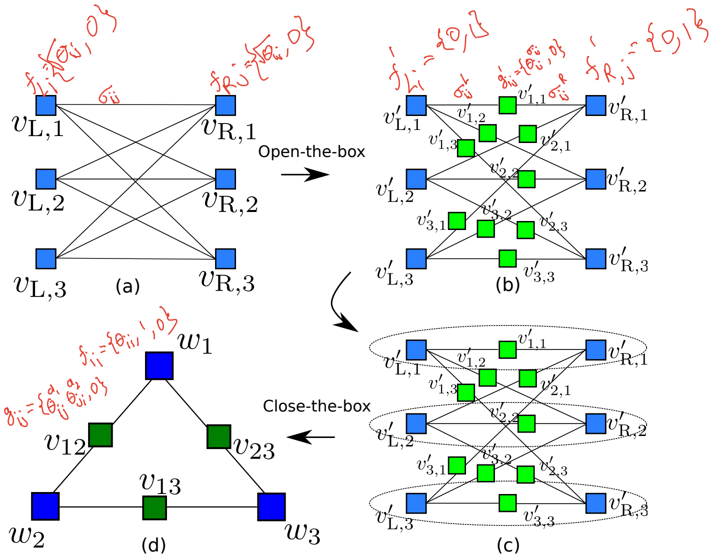


Fig. 1: From $\mathsf{N}_{\mathrm{B}}(\Theta)$ to $\mathsf{N}_{\mathrm{MB}}(\Theta)$ for $n = 3$. (See text for details.)

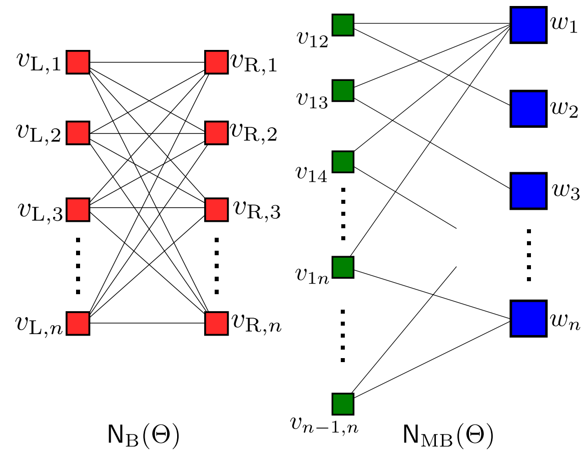


Fig. 2: $\mathsf{N}_{\mathrm{B}}(\Theta)$ and $\mathsf{N}_{\mathrm{MB}}(\Theta)$ for general n.

permanent will be characterized in terms of $\eta_{\mathrm{MB}}(\Theta) \coloneqq \frac{1}{n} \log \frac{\mathrm{perm}(\Theta)}{\mathrm{perm}_{\mathrm{MB}}(\Theta)}$. The modified NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$ strikes a good balance between removing cycles in the $\mathsf{N}_{\mathrm{B}}(\Theta)$ and not increasing the function node complexity by too much.[1]

In this paper, we study the NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$, in particular the associated Bethe free energy and the behavior of the SPA when operating on $\mathsf{N}_{\mathrm{MB}}(\Theta)$.

- For $n = 2$, the NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$ has no cycles an therefore $\eta_{\mathrm{MB}}(\Theta) = 0$. This case will not be discussed further in this paper.
- For $n = 3$, we show that the NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$ is very well behaved in the sense the Bethe free energy function is convex, that the SPA converges to the minimum of the Bethe free energy function, and that $0 \leqslant \eta_{\mathrm{MB}}(\Theta) \leqslant 1/3$, where both the lower and the upper bounds are tight.
- For $n > 3$, empirical results show that $\eta_{\mathrm{MB}}(\Theta)$ is always much less than $1/2$. Moreover, the SPA converges to an interior point of the local marginal polytope when none of the vertices are optimal. This is despite the fact that we can establish that the Bethe free energy function is not convex everywhere and despite the fact that there are vertices of the local marginal polytope where the directional derivative has infinite slope. This means that vertices with directional derivatives with infinite slope have also directional derivatives with finite slope, thereby allowing the SPA to escape a non-optimal vertex. (This behavior is in contrast to binary LDPC codes with bit node degree at least three, whose Bethe free energy function is such that at codeword vertices of the local marginal polytope all directional derivatives have infinite slope.)

Let us highlight two aspects which make $\mathsf{N}_{\mathrm{MB}}(\Theta)$ a worthwhile object of study:

- Because $\mathsf{N}_{\mathrm{MB}}(\Theta)$ breaks some symmetry in $\mathsf{N}_{\mathrm{B}}(\Theta)$, it is advisable to reorder the rows and columns of Θ (which leaves the permanent invariant) towards optimizing the properties of $\mathsf{N}_{\mathrm{B}}(\Theta)$. In particular, it is advisable to permute the rows and columns of Θ such that $\prod_{i=1}^{n} \theta_{i\sigma(i)}$ is maximized by the trivial permutation σ in S_n. The required row and

[1]Our approach at obtaining a better approximation can be seen as somewhat similar to region approximations [12]. For such region approximations it is clear that the larger the regions are, the better the approximation will be. However, the complexity will also grow exponentially with the region sizes.

column permutations can be found with the help of an algorithm that finds the maximum-weight matching. This can be accomplished efficiently, e.g., by running the max-product or min-sum algorithm on a suitable NFG (see [18], [19] for details). Such a low-complexity pre-processing of an NFG might also be interesting in other contexts.
- There is a tight connection between $\mathsf{N}_{\mathrm{MB}}(\Theta)$ and an NFG in [20] that was used for approximating a permanent-like quantity in quantum information processing. Similar NFGs have been used to represent quantum systems with multiple measurements [21], [22] and estimate rates of quantum channels with memory [23]. Insights gained for $\mathsf{N}_{\mathrm{MB}}(\Theta)$ are also relevant for these other NFGs.

Organisation of the paper: We assume that the reader is familiar with concepts of factor graphs and the SPA. For an introduction to these topics, we direct the reader to [12], [24], [25].

Sec. I-A introduces the notation used in this paper.

We describe the modified NFG and Bethe free energy function for the permanent in Sec. II. We will derive the modified NFG from that in [11] using open-the-box and close-the-box operations [24]. In Sec. II-A, we give an alternate derivation of the modified NFG starting from first principles.

We then describe the SPA update rules in Sec. III. Our results for the $n = 3$ case are described in Sec. IV, while results for $n > 3$ are presented in Sec. V. For easy reference, we list the main results and where they can be found in the paper.

- Proposition 1 shows that for $n = 3$, the modified BP algorithm initialized with random messages always converges, and that the modified Bethe permanent is equal to the largest eigenvalue of a certain matrix.
- Proposition 2 shows that for $n = 3$, the modified Bethe free energy function is convex.
- Proposition 3 gives tight upper and lower bounds on the modified Bethe permanent for $n = 3$ and we see that the approximation factor strictly improves upon $\eta_{\mathrm{B}}(\Theta)$.
- Proposition 4 shows that for $n > 3$, the Bethe free energy function is nonconvex. The proof can be found in Appendix B.
- In Sec. V-B, we discuss the structure of the local marginal polytope for the modified Bethe approximation, and prove in Lemma 1 that for $n > 3$, the local marginal polytope has fractional vertices.
- In Lemma 2, we give an exact expression for $\eta_{\mathrm{MB}}(\Theta)$ for any stationary point of the modified Bethe free energy function.
- Simulation results are presented in Sec. VI. Simulations suggest that for several ensembles of random matrices, the modified Bethe permanent approximates the true permanent better than the algorithm in [11].

A. Notation

The set of real numbers is denoted $\mathbb{R}$, while the set of integers is denoted $\mathbb{Z}$. For any positive integer n, we let $[n] := \{1, 2, \ldots, n\}$.

Ordered tuples are denoted within parentheses, e.g., $(a_i)_{i=1}^n := (a_1, a_2, \ldots, a_n)$.

Vectors are denoted by underlined characters, e.g., $\underline{a}, \underline{\sigma}$, etc. The dimension of a vector will usually be clear from the context.

To keep things simple, summations of the form $\sum_{j \in [n] \setminus i} a_j$ will be denoted by $\sum_{j \neq i} a_j$ and $\sum_{j=i+1}^{n} a_j$ will be denoted $\sum_{j>i} a_j$.

1) Notation for variables/configurations σ: In this paper, we will use special notation to refer to various variables of the factor graph and its configurations. Throughout, $\underline{\sigma}$ denotes the $n \times n$ matrix with the (i, j)th entry being equal to $\sigma_{ij} \in \{0, 1\}$. This is a special matrix that corresponds to a configuration of variables in the factor graph.

For any $1 \leqslant i, j \leqslant n$ with $i \neq j$, we define (with abuse of notation) $\underline{\sigma}_{ij}$ to be equal to the ordered pair $(\sigma_{ij}, \sigma_{ji})$ if $i < j$ and $(\sigma_{ji}, \sigma_{ij})$ otherwise. We also define, for every $i \in [n]$, $\underline{\sigma}_i := (\underline{\sigma}_{ij} : j \in [n] \setminus i)$.

II. The Modified NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$

In this section we introduce the modified NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$. It is most easily explained by first looking at the case $n = 3$. We do the following step for obtaining $\mathsf{N}_{\mathrm{MB}}(\Theta)$ from $\mathsf{N}_{\mathrm{B}}(\Theta)$ (see also Fig. 1):

1) Fig. 1(a): we define $\mathsf{N}_{\mathrm{B}}(\Theta)$ as in [11]. (See [11, Fig. 1 and Section II] for all the details.) The blue function nodes on the left and right encode the entries of the matrix Θ.

Specifically, each edge $(v_{\mathrm{L},i}, v_{\mathrm{R},j})$ is associated with a variable $\sigma_{ij} \in \{0, 1\}$. Corresponding to each left vertex $v_{\mathrm{L},i}$ is the associated local function

$$f_{\mathrm{L},i}((\sigma_{il})_{l=1}^n) = \begin{cases} \sqrt{\theta_{ij}} & \text{if } \sigma_{ij} = 1, \text{ and } \sigma_{il} = 0 \text{ for } l \neq j \\ 0 & \text{otherwise,} \end{cases}$$

and corresponding to each right vertex $v_{\mathrm{R},i}$ is

$$f_{\mathrm{R},j}((\sigma_{lj})_{l=1}^n) = \begin{cases} \sqrt{\theta_{ij}} & \text{if } \sigma_{ij} = 1, \text{ and } \sigma_{lj} = 0 \text{ for } l \neq i \\ 0 & \text{otherwise.} \end{cases}$$

If we define $\underline{\sigma} := (\sigma_{ij})_{i,j}$, then

$$\mathrm{perm}(\Theta) = \sum_{\underline{\sigma}} \left(\prod_{i=1}^{n} \prod_{j=1}^{n} f_{\mathrm{L},i}((\sigma_{il})_{l=1}^n) f_{\mathrm{R},j}((\sigma_{lj})_{l=1}^n) \right).$$

$\sigma^{(i)}$ and $\sigma^{(j)}$ are copies of same variable and should always match

2) From Fig. 1(a) to Fig. 1(b): we modify the NFG such that green function nodes encode the entries of the matrix Θ. The blue function nodes on the left and the right are now merely suitable indicator functions.
More precisely, the edge joining $v'_{L,i}$ and v'_{ij} corresponds to the variable σ^L_{ij}, while the edge joining v'_{ij} and $v'_{R,j}$ corresponds to the variable σ^R_{ij}. The local function associated with $v'_{L,i}$ is $f'_{L,i}((\sigma^L_{il})^n_{l=1}) = 1$ if exactly one of the σ^L_{il}'s is one, and $f'_{L,i}((\sigma^L_{il})^n_{l=1}) = 0$ otherwise. Similarly, $f'_{R,j}((\sigma^R_{lj})^n_{l=1}) = 1$ if exactly one of the σ^R_{lj}'s is one, and $f'_{L,i}((\sigma^R_{lj})^n_{l=1}) = 0$ otherwise. For every (i, j),

$$g'_{ij}(\sigma^L_{ij}, \sigma^R_{ij}) = \begin{cases} \theta_{ij}^{\sigma_{ij}} & \text{if } \sigma^L_{ij} = \sigma^R_{ij} = \sigma_{ij} \rightarrow \text{edge variable} \\ 0 & \text{otherwise} \end{cases}$$

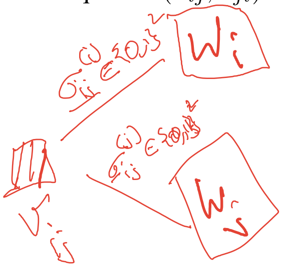
is associated with v'_{ij}.

3) From Fig. 1(c) to Fig. 1(d): for every $i \in [n]$, the function nodes within the ellipse containing $f'_{L,i}$, g'_{ii}, and $f'_{R,i}$ are replaced by a single blue function node f_i in Fig. 1(d). Moreover, for every $1 \leqslant i < j \leqslant n$, the function nodes g'_{ij} and g'_{ji} are merged to the green function node g_{ij} in Fig. 1(d). Finally pairs of parallel edges are replaced by a single edge and the corresponding variables concatenated.

For general n, the procedure is essentially the same. We start with the NFG $\mathsf{N}_{\mathrm{B}}(\Theta)$ in Fig. 2 (left) and obtain the NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$ in Fig. 2 (right). Note that the $\mathsf{N}_{\mathrm{MB}}(\Theta)$ has $\binom{n}{2}$ green function nodes of degree 2 on the left and n blue function nodes of degree n on the right.

Let us describe $\mathsf{N}_{\mathrm{MB}}(\Theta)$ in more detail. This has $\binom{n}{2}$ left vertices v_{ij} for $1 \leqslant i < j \leqslant n$ and n right vertices w_i for $1 \leqslant i \leqslant n$. Each edge (v_{ij}, w_i) (resp. (v_{ij}, w_j)) in $\mathsf{N}_{\mathrm{MB}}(\Theta)$ is associated with the variable $\underline{\sigma}^{(i)}_{ij} \in \{0,1\}^2$ (resp. $\underline{\sigma}^{(j)}_{ij}$). Here and in the rest of the paper, we will abuse notation and say that for any $1 \leqslant i, j \leqslant n$, $\underline{\sigma}_{ij}$ will be equal to $(\sigma_{ij}, \sigma_{ji})$ if $i < j$ and $(\sigma_{ji}, \sigma_{ij})$ if $j < i$.

The local function associated with each right function node w_i is



Contributes diagonal elements

$$f_i(\underset{(1\times n)\ \text{a permutation row}}{\underline{\sigma}^{(i)}_i}) := \begin{cases} \theta_{ii} & \text{if } \sum_{j_1 \neq i} \sigma^{(i)}_{ij_1} = \sum_{j_1 \neq i} \sigma^{(i)}_{j_1 i} = 0 \\ 1 & \text{if } \sum_{j_1 \neq i} \sigma^{(i)}_{ij_1} = \sum_{j_1 \neq i} \sigma^{(i)}_{j_1 i} = 1 \\ 0 & \text{otherwise} \end{cases}$$

while that associated with v_{ij} is

contributes diagonally opposite elements

$$g_{ij}(\underset{(1\times 2)}{\underline{\sigma}^{(i)}_{ij}}, \underset{(1\times 2)}{\underline{\sigma}^{(j)}_{ij}}) = \begin{cases} \theta_{ij}^{a_1} \theta_{ji}^{a_2} & \text{if } \underline{\sigma}^{(i)}_{ij} = \underline{\sigma}^{(j)}_{ij} = (a_1, a_2) \\ 0 & \text{otherwise.} \end{cases}$$

equal

Given $\mathsf{N}_{\mathrm{MB}}(\Theta)$, we use the standard approach [12] for formulating the Bethe free energy function F_{MB}. The modified Bethe permanent is then defined to be

$$\mathrm{perm}_{\mathrm{MB}}(\Theta) := \exp\left(- \min_{\underline{b}} F_{\mathrm{MB}}(\underline{b}) \right),$$

where the minimization is over the local marginal polytope associated with $\mathsf{N}_{\mathrm{MB}}(\Theta)$. The quality of the approximation will be measured by

$$\eta_{\mathrm{MB}}(\Theta) := \frac{1}{n} \log \frac{\mathrm{perm}(\Theta)}{\mathrm{perm}_{\mathrm{MB}}(\Theta)}.$$

A. *Derivation of* N_{MB} *from first principles*

We now give an alternate derivation of the modified factor graph N_{MB} starting from the definition of the permanent.

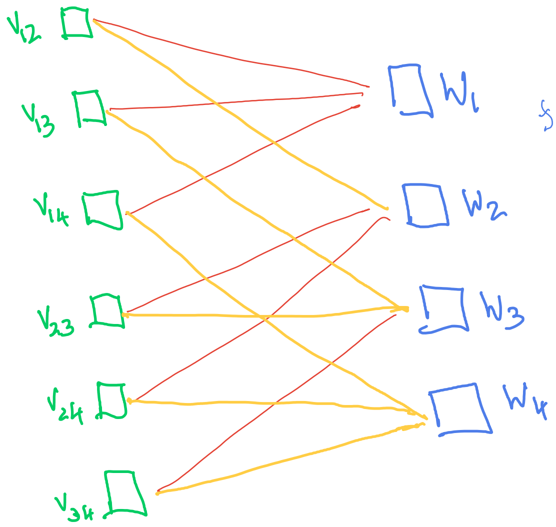
Let $\mathcal{P}_n$ denote the set of all $n \times n$ permutation matrices. For any set $\mathcal{S}$, define the indicator function $\chi_{\{S\}}(x)$ to be the function that takes value 1 if and only if $x \in \mathcal{S}$. Our algorithm for approximating the permanent stems from the following equivalent description of (1).

$$\begin{aligned} \mathrm{perm}(\Theta) &= \sum_{\sigma \in \mathcal{P}_n} \prod_{i=1}^n \prod_{j=1}^n \theta_{ij}^{\sigma_{ij}} \\ &= \sum_{\sigma \in \mathcal{P}_n} \prod_{i=1}^n \left(\theta_{ii}^{\chi_{\{1\}}(\sigma_{ii})} \prod_{j=i+1}^n \theta_{ij}^{\chi_{\{1\}}(\sigma_{ij})} \theta_{ji}^{\chi_{\{1\}}(\sigma_{ji})} \right). \end{aligned}$$

Since σ is a permutation matrix, $\sigma_{ii} = 1$ if and only if (iff) $\sum_{j \in [n] \setminus i} \sigma_{ij} = \sum_{j \in [n] \setminus i} \sigma_{ji} = 0$. Therefore,

$$\mathrm{perm}(\Theta) = \sum_{\sigma \in \mathcal{P}_n} \prod_{i=1}^n \left(\theta_{ii}^{\chi_{\{0\}}(\sum_{j_1 \neq i} \sigma_{ij_1}) \chi_{\{0\}}(\sum_{j_1 \neq i} \sigma_{j_1 i})} \prod_{j=i+1}^n \theta_{ij}^{\chi_{\{1\}}(\sigma_{ij})} \theta_{ji}^{\chi_{\{1\}}(\sigma_{ji})} \right). \tag{2}$$

$g(\sigma_{ij}, \sigma_{ji}) = \theta_{ij}^{\sigma_{ij}} \theta_{ji}^{\sigma_{ji}}$



$f(\underline{\sigma}_i) = \begin{cases} \theta_{ii}, & \sigma_i = w_i \\ 1, & \text{o/w} \end{cases}$

$$\underline{\sigma}^{(i)} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array} = \underline{\sigma}$$

$\underline{\sigma}^{(i)}$ = identical to $\underline{\sigma}^{(i)}$

$w_2 \; : \; \partial w_2 = \{v_{12}, v_{23}, v_{24}\}$

$\mu_{w_2 \to v_{12}}(11)$

$\mu_{v_{ij} \to w_1}(\sigma_{ij}) = \sum_{\sigma_{ij}} \cdots$ $\qquad$ $\mu_{w_i \to v_{ij}}$

Observe that a matrix $\sigma \in \{0,1\}^{n\times n}$ is permutation matrix iff $\sum_{j=1}^{n} \sigma_{ij} = \sum_{j=1}^{n} \sigma_{ji} = 1$ for all $i \in [n]$. Let us define $\mathcal{M}_n$ to be the set of all binary $n \times n$ matrices, and

$$f_i(\underline{\sigma}_i) := \begin{cases} \theta_{ii} & \text{if } \sum_{j_1 \neq i} \sigma_{ij_1} = \sum_{j_1 \neq i} \sigma_{j_1 i} = 0 \\ 1 & \text{if } \sum_{j_1 \neq i} \sigma_{ij_1} = \sum_{j_1 \neq i} \sigma_{j_1 i} = 1 \\ 0 & \text{otherwise} \end{cases}$$

contains both (σ_{ik}s and σ_{ki}s $\forall k$)

$$h_{ij}(\underline{\sigma}_{ij}) := \theta_{ij}^{\sigma_{ij}} \theta_{ji}^{\sigma_{ji}}$$

for all $1 \leqslant i < j \leqslant n$. To avoid cumbersome notation here and in the rest of the paper, we use $j \neq i$ to denote $j \in [n] \backslash \{i\}$. Also, the notation $j > i$ will typically mean $j \in \{i+1, \ldots, n\}$.

filters out cases where σ is a permutation

We can write (2) as follows:

$$\operatorname{perm}(\Theta) = \sum_{\sigma \in \mathcal{M}_n} \prod_{i=1}^{n} \left(f_i((\sigma_{ij_1}, \sigma_{j_1 i})_{j_1 \neq i}) \prod_{j=i+1}^{n} h_{ij}(\sigma_{ij}) \right). \tag{3}$$

The above sum-of-products expression can be described by the NFG in Fig. 2 (right). Note that in the figure,

$$g_{ij}(\underline{\sigma}_{ij}^{(i)}, \underline{\sigma}_{ij}^{(j)}) = h_{ij}(\underline{\sigma}_{ij}^{(i)}) \chi_{\{\underline{\sigma}_{ij}^{(i)}\}}(\underline{\sigma}_{ij}^{(j)}).$$

B. Bethe free energy function

The Bethe free energy function is a variational form that gives an analytical description of any quantity which can be expressed as a sum of products of local functions. We will use the Bethe approximation defined on the NFG $\mathsf{N}_{\mathrm{MB}}(\Theta)$ to approximate the permanent.

Let β be any distribution on $n \times n$ permutation matrices $\mathcal{P}_n$. Let $\mathcal{A}_n$ denote the simplex of all such distributions.

Define the *Gibbs average energy function*:

$$U_{\mathrm{G}}(\,\cdot\,;\Theta) : \mathcal{A}_n \to \mathbb{R}$$

$$U_{\mathrm{G}}(\beta;\Theta) := -\sum_{\sigma \in \mathcal{P}_n} \beta(\sigma) \sum_{i=1}^{n} \left(\log f_i(\underline{\sigma}_i) + \sum_{j>i} \log g_{ij}(\underline{\sigma}_{ij}) \right).$$

sum over all f's in the graph

If we define the marginals

$$\beta^{(i)}(\underline{\sigma}_i) := \sum_{\underline{\sigma}' \in \{0,1\}^n : \underline{\sigma}'_i = \underline{\sigma}_i} \beta(\underline{\sigma}')$$

and

$$\beta^{(ij)}(\underline{\sigma}_{ij}) := \sum_{\underline{\sigma}'_i \in \{0,1\}^n : \underline{\sigma}'_{ij} = \underline{\sigma}_{ij}} \beta(\underline{\sigma}'),$$

then we can write

$$U_{\mathrm{G}}(\beta;\Theta) = -\sum_{i=1}^{n} \sum_{\underline{\sigma}_i \in \{0,1\}^{2(n-1)}} \beta^{(i)}(\underline{\sigma}_i) \log f_i(\underline{\sigma}_i) - \sum_{i=1}^{n} \sum_{j>i} \sum_{\underline{\sigma}_{ij} \in \{0,1\}^2} \beta^{(ij)}(\underline{\sigma}_{ij}) \log g_{ij}(\underline{\sigma}_{ij}).$$

In a similar manner, we define the *Gibbs entropy function*:

$$H_{\mathrm{G}}(\,\cdot\,;\Theta) : \mathcal{A}_n \to \mathbb{R}$$

$$H_{\mathrm{G}}(\beta;\Theta) := -\sum_{\sigma \in \mathcal{P}_n} \beta(\sigma) \log \beta(\sigma)$$

and the *Gibbs free energy function*:

$$F_{\mathrm{G}}(\,\cdot\,;\Theta) : \mathcal{A}_n \to \mathbb{R}$$

$$F_{\mathrm{G}}(\beta;\Theta) := U_{\mathrm{G}}(\beta;\Theta) - H_{\mathrm{G}}(\beta;\Theta).$$

The Gibbs free energy function is convex, and its minimum is equal to the negative logarithm of the permanent [12].

$$-\log \operatorname{perm}(\Theta) = \min_{\beta \in \mathcal{A}_n} F_{\mathrm{G}}(\beta;\Theta).$$

This gives a different means to compute the permanent of a matrix: minimization of the Gibbs free energy function yields $-\log \operatorname{perm}(\Theta)$. However, minimizing F_{G} is at least as hard as computing the permanent. We therefore solve a relaxation of this minimization problem.

We define the following set, called the *local marginal polytope* or the *belief polytope*

$$
\begin{aligned}
\mathcal{B}_{\mathrm{MB}} := \Big\{ & \underline{b} := (b^{(i)}_{lm}, b^{(ij)}_{\underline{a}} : 1 \leqslant i < j \leqslant n; l, m \in [n]; \\
& \underline{a} \in \{0,1\}^2\}) \in \mathbb{R}^{k\times k\times k}_{\geqslant 0} \times \mathbb{R}^{k\times k\times 4}_{\geqslant 0} : \\
& b^{(i)}_{ij} = b^{(i)}_{ji} = 0, \qquad \forall (i,j) \in [n]\times[n],\ i \neq j \\
& \sum_{l,m} b^{(i)}_{lm} = 1, \qquad \forall i \\
& b^{(ij)}_{11} = b^{(i)}_{jj} = b^{(j)}_{ii} \qquad \forall\ 1 \leqslant i < j \leqslant n \\
& b^{(ij)}_{01} = \sum_{l\neq i,j} b^{(i)}_{lj} = \sum_{l\neq i,j} b^{(j)}_{il}, \qquad \forall\ 1 \leqslant i < j \leqslant n \\
& b^{(ij)}_{10} = \sum_{m\neq i,j} b^{(i)}_{jm} = \sum_{l\neq i,j} b^{(j)}_{li}, \qquad \forall\ 1 \leqslant i < j \leqslant n \\
& \sum_{\underline{a}\in\{0,1\}^2} b^{(ij)}_{\underline{a}} = 1, \qquad \forall\ 1 \leqslant i < j \leqslant n \Big\}.
\end{aligned} \tag{4}
$$

i=2

Summation=1

$b^{23}_{01} = b^{2}_{13} + b^{2}_{43}$

b^{i}_{10} . . .

The vector $\underline{b}$ is the vector of beliefs. For every $i \in [n]$, the quantity $b^{(i)}_{lm}$ is to be interpreted as the belief that the factor node w_i associates with $\sigma_{ij} = \sigma_{ji} = 1$. The quantity $b^{(ij)}_{00}$ is to be interpreted as the belief that the factor node g_{ij} associates with $\sigma_{ij} = \sigma_{ji} = 0$, whereas b^{ij}_{11} is the belief at factor node g_{ij} that $\sigma_{ij} = \sigma_{ji} = 1$. Similarly $b^{(ij)}_{10}$ is the belief that $\sigma_{ij} = 1, \sigma_{ji} = 0$ and $b^{(ij)}_{01}$ is the belief that $\sigma_{ij} = 0, \sigma_{ji} = 1$. We also define the *Bethe average energy function* (for convenience, we will omit the dependence on the matrix Θ in the notation):

$$
\begin{aligned}
& U_{\mathrm{MB}} : \mathcal{B}_{\mathrm{MB}} \to \mathbb{R} \\
& U_{\mathrm{MB}}(\underline{b}) := -\sum_{i=1}^{k} b^{(i)}_{ii} \log\theta_{ii} - \sum_i \sum_{j>i} \sum_{\underline{a}=(a_1 a_2)\in\{0,1\}^2} b^{(ij)}_{\underline{a}} \log(\theta^{a_1}_{ij}\theta^{a_2}_{ji}),
\end{aligned} \tag{5}
$$

and the *Bethe entropy function*:

$$
\begin{aligned}
& H_{\mathrm{MB}} : \mathcal{B}_{\mathrm{MB}} \to \mathbb{R} \\
& H_{\mathrm{MB}}(\underline{b}) := -\sum_i \sum_{l,m} b^{(i)}_{lm} \log b^{(i)}_{lm} + \sum_i \sum_{j>i} \sum_{\underline{a}\in\{0,1\}^2} b^{(ij)}_{\underline{a}} \log b^{(ij)}_{\underline{a}}.
\end{aligned} \tag{6}
$$

The Bethe average energy and entropy functions can be obtained by natural relaxations of U_{G} and H_{G} respectively [12]. The expressions for U_{MB} and H_{MB} above are obtained from these relaxations by a simple reparameterization along the lines of [11], and we omit the detailed calculations here. The *Bethe free energy function* is given by

$$
F_{\mathrm{MB}}(\underline{b}) \coloneqq U_{\mathrm{MB}}(\underline{b}) - H_{\mathrm{MB}}(\underline{b}). \tag{7}
$$

Just as minimizing the Gibbs free energy function gives the negative logarithm of the permanent of Θ, we define the *modified Bethe permanent* $\mathrm{perm}_{\mathrm{MB}}(\Theta)$ to be the quantity obtained from the following expression:

$$
\mathrm{perm}_{\mathrm{MB}}(\Theta) = \exp\left(-\min_{\underline{b}\in\mathcal{B}_{\mathrm{MB}}} F_{\mathrm{MB}}(\underline{b})\right).
$$

The modified Bethe permanent is an approximation of $\mathrm{perm}(\Theta)$. Indeed, we could have considered a different surrogate function to minimize in place of F_{MB} to approximate the permanent but the main advantage of using the Bethe approximation is its intimate relationship with the sum-product algorithm. The sum-product algorithm (SPA) is a local iterative algorithm performed on the normal factor graph. It is a well-known fact [12] that there is a one-to-one correspondence between the fixed points of the sum-product algorithm and the stationary points of the Bethe free energy function. We can therefore run the SPA to perform approximate minimization of the Bethe free energy function. Furthermore, if the Bethe free energy function is convex and the SPA converges, then it must converge to a unique fixed point. This was indeed the case in Vontobel's [11] formulation of the SPA for computing the permanent.

Let us now describe the sum-product algorithm for approximate computation of the permanent.

how is modified bethe different from the normal one? Why is it not convex?

III. The SPA update rules

The sum-product algorithm is an iterative algorithm run on the normal factor graph N_{MB}. For every edge in the normal factor graph, we assign two probability vectors called messages: a left-going message and a right-going message. These messages are updated iteratively as specified by the SPA until a suitable stopping point is reached.

In Fig. 2(right), every right function node w_i is connected to $n-1$ function nodes $\{v_{ij} : j \neq i\}$. Every left function node v_{ij} is connected to two right function nodes w_i and w_j.

We define $\mu^{(t)}_{ij\to i}$ be a vector in $\mathbb{R}^4$ that denotes the (right-going) message from the v_{ij} to w_i in the tth iteration. Likewise, $\mu^{(t)}_{i\to ij} \in \mathbb{R}^4$ will denote the (left-going) message from w_i to v_{ij} in the tth iteration. The initial messages $\mu^{(0)}_{ij\to j}$, $\mu^{(0)}_{i\to ij}$ are set randomly or according to some fixed rule. For all successive iterations t, the message $\mu^{(t)}_{i\to ij}$ is a function of the messages $\{\mu^{(t)}_{ij_1\to i} : j_1 \neq i, j\}$ and the local function f_i. The message $\mu^{(t)}_{ij\to i}$ is a function of $\mu^{(t-1)}_{j\to ij}$ and the local function g_{ij}.

A. Message update rules

Define $\mathcal{W} := \{w_i, i \in [n]\}$ and $\mathcal{V} := \{v_{ij}, 1 \leqslant i < j \leqslant n\}$. For $1 \leqslant i < j \leqslant n$, let ∂w_i denote the set of all neighbors v_{ij} of factor node w_i, i.e., let $\partial w_i := \{v_{ij} : i < j\} \cup \{v_{ji} : j < i\}$, and, likewise, $\partial v_{ij} := \{w_i, w_j\}$. For any neighbouring vertices v, w in the NFG, let $\mu^{(t)}_{v\to w}$ denote the message sent from v to w in iteration t.

The SPA message update equations are as follows: Let $g_{ij}(\underline{a}) := \theta_{ij}^{a_1}\theta_{ji}^{a_2}$. For $\underline{a} \in \{0,1\}^2$, $1 \leqslant i < j \leqslant n$, $w, w' \in \partial v_{ij}$ with $w \neq w'$, and $v := v_{ij}$, we have

$$\mu^{(t)}_{v\to w}(\underline{a}) \propto g_{ij}(\underline{a})\mu^{(t-1)}_{w'\to v}(\underline{a}),$$

} straight forward

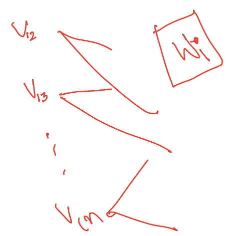
where the appropriate proportionality constant ensures that $\sum_{\underline{a}} \mu_{v\to w}(\underline{a}) = 1$. Moreover, for any $i \in [n]$, $w := w_i$, and $v \in \partial w$, we have

$$\mu^{(t)}_{w\to v}(11) \propto \prod_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v'\to w}(00),$$

$$\mu^{(t)}_{w\to v}(01) \propto \sum_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v'\to w}(10) \prod_{v''\in\partial w\setminus\{v,v'\}} \mu^{(t)}_{v''\to w}(00),$$

$$\mu^{(t)}_{w\to v}(10) \propto \sum_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v'\to w}(01) \prod_{v''\in\partial w\setminus\{v,v'\}} \mu^{(t)}_{v''\to w}(00),$$

$$\begin{aligned}\mu^{(t)}_{w\to v}(00) \propto \theta_{ii} &\prod_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v\to w}(00)\\ &+ \sum_{v'\in\partial w\setminus\{v\}} \Bigg[\mu^{(t)}_{v'\to w}(11) \prod_{v''\in\partial w\setminus\{v,v'\}} \mu^{(t)}_{v''\to w}(00)\\ &+ \mu^{(t)}_{v'\to w}(10) \sum_{v''\in\partial w\setminus\{v,v'\}} \Big(\mu^{(t)}_{v''\to w}(01) \prod_{v'''\in\partial w\setminus\{v,v',v''\}} \mu^{(t)}_{v'''\to w}(00)\Big)\Bigg].\end{aligned}$$



The proportionality constants enforce that $\sum_{\underline{a}\in\{0,1\}^2} \mu_{w\to v}(\underline{a}) = 1$. The messages are updated until convergence, or until some suitable stopping criterion is reached.

1) Computation of beliefs and $\mathrm{perm}_{\mathrm{MB}}$*:* Suppose that the algorithm is run for t_* steps. We use the final messages to compute an approximate minimizer of F_{MB}. For every $1 \leqslant i < j \leqslant n$ and $\underline{a} \in \{0,1\}^2$, we set

$$b^{(ij)}_{\underline{a}} \propto g_{ij}(\underline{a})\mu^{t*}_{i\to ij}(\underline{a})\mu^{t*}_{j\to ij}(\underline{a}). \tag{8}$$

where the proportionality constant ensures that $\sum_{\underline{a}} b^{(ij)}_{\underline{a}} = 1$. For every $i \in [n]$ and $l \neq i$,

$$b^{(i)}_{ll} \propto \mu^{(t*)}_{il\to i}(10)\mu^{(t*)}_{li\to i}(01) \prod_{j_1\neq i,l} \mu^{(t*)}_{ij_1}(00) \tag{9}$$

and for every $i \in [n]$ and $l, m \neq i$ such that $l \neq m$, we set

$$b^{(i)}_{lm} \propto \mu^{(t*)}_{il\to i}(10)\mu^{(t*)}_{mi\to i}(01) \prod_{j_1\neq i,l,m} \mu^{(t*)}_{ij_1}(00). \tag{10}$$

The proportionality constants ensure that $\sum_{l,m} b^{(i)}_{lm} = 1$ for all i. Our estimate of the Bethe permanent is

$$\mathrm{perm}_{\mathrm{MB}}(\Theta) = \exp(-F_{\mathrm{MB}}(\underline{b})).$$

Note that the computational complexity of updating each message $\mu_{i\to v}$ is $O(n^3)$, while that of updating each $\mu_{ij\to i}$ message is a constant. The overall computational complexity per iteration is therefore $O(n^4)$. However, this can be brought down by simplifying the message passing rules.

speed-up analogous to Huang Jebara

B. Simpler message passing rules

By reparameterizing the messages, we can reduce the computational complexity to $O(n^3)$. This reparameterization is also of relevance for the NFG in [20]. For $\underline{a} \in \{01, 10, 11\}$, and neighbours u, u' in the factor graph, define

$$V^{(t)}_{u \to u'}(\underline{a}) := \mu^{(t)}_{u \to u'}(\underline{a}) / \mu^{(t)}_{u \to u'}(00),$$

The components of the μ-messages sum to 1, hence there is a bijection between the μ messages and the V messages.

In terms of the new messages, we obtain the following message update rules. For every $1 \leqslant i < j \leqslant n$, $\underline{a} \in \{01, 10, 11\}$, $w, w' \in \partial v_{ij}$ and $v := v_{ij}$ we have

$$V^{(t)}_{v \to w}(\underline{a}) := g_{ij}(\underline{a}) V^{(t-1)}_{w' \to v}(\underline{a}).$$

Moreover, for each $i \in [n]$, $w := w_i$, and $v \in \partial w$, we have

$$\begin{aligned}
D^{(t)}_{w \to v} &:= \theta_{ii} + \sum_{v' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(11) \\
&\quad + \sum_{v', v'' \in \partial w \setminus \{v\}, v' \neq v''} V^{(t)}_{v' \to w}(10) \cdot V^{(t)}_{v'' \to w}(01), \\
V^{(t)}_{w \to v}(11) &:= \frac{1}{D^{(t)}_{w \to v}}, \\
V^{(t)}_{w \to v}(10) &:= \frac{\sum_{v' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(01)}{D^{(t)}_{w \to v}}, \\
V^{(t)}_{w \to v}(01) &:= \frac{\sum_{v' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(10)}{D^{(t)}_{w \to v}}.
\end{aligned}$$

We can obtain slightly simpler expressions through another reparameterization. For any edge $\overrightarrow{e}$ (either $w \to v$ or $v \to w$), let us define $W_{\overrightarrow{e}} := V_{\overrightarrow{e}}(11) - V_{\overrightarrow{e}}(10) V_{\overrightarrow{e}}(01)$. We can represent messages in terms of $(V_{\overrightarrow{e}}(01), V_{\overrightarrow{e}}(10), W_{\overrightarrow{e}})$ instead of $(V_{\overrightarrow{e}}(01), V_{\overrightarrow{e}}(10), V_{\overrightarrow{e}}(11))$ since there is a bijection between the two triples.

For every $1 \leqslant i < j \leqslant n$ and $\underline{a} \in \{01, 10\}$, the update equations for $(V_{\overrightarrow{e}}(01), V_{\overrightarrow{e}}(10), V_{\overrightarrow{e}}(11))$ are as follows (following the notation defined previously)

$$\begin{aligned}
V^{(t)}_{v \to w}(\underline{a}) &= g_{ij}(\underline{a}) V^{(t-1)}_{w' \to v}(\underline{a}) \\
W^{(t)}_{v \to w} &= g_{ij}(11) W^{(t-1)}_{w' \to v}.
\end{aligned}$$

and,

$$\begin{aligned}
W^{(t)}_{w \to v} &= \frac{\theta_{ii} + \sum_{v' \in \partial w \setminus \{v\}} W^{(t)}_{v' \to w}}{\theta_{ii} + \sum_{v' \in \partial w \setminus \{v\}} W^{(t)}_{v' \to w}(11) + \sum_{v', v'' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(10) V^{(t)}_{v'' \to w}(01)} \\
V^{(t)}_{w \to v}(10) &= \frac{\sum_{v' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to i}(01)}{\theta_{ii} + \sum_{v' \in \partial w \setminus \{v\}} W^{(t)}_{v' \to w}(11) + \sum_{v', v'' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(10) V^{(t)}_{v'' \to w}(01)} \\
V^{(t)}_{w \to v}(01) &= \frac{\sum_{v' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(01)}{\theta_{ii} + \sum_{v' \in \partial w \setminus \{v\}} W^{(t)}_{v' \to w}(11) + \sum_{v', v'' \in \partial w \setminus \{v\}} V^{(t)}_{v' \to w}(10) V^{(t)}_{v'' \to w}(01)}.
\end{aligned}$$

The final computation of the beliefs and the Bethe permanent is the same as in the previous subsection.

IV. The modified Bethe permanent of a 3×3 matrix

The $n = 3$ case is rather special as the Bethe free energy function and the SPA algorithm have some nice properties. The NFG is a cycle, illustrated in Fig. 1(d). In this special case, the SPA message update rules at each node can be represented as linear transformations followed by renormalization. Furthermore, the Bethe entropy function can be written as a sum of conditional entropy functions. We use this to show that the Bethe free energy function is convex.

Suppose that the matrix whose permanent we want to approximate is given by

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}.$$

$$\mu^{(t)}_{w\to v}(11) \propto \prod_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v'\to w}(00),$$

$$\mu^{(t)}_{w\to v}(01) \propto \sum_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v'\to w}(10) \prod_{v''\in\partial w\setminus\{v,v'\}} \mu^{(t)}_{v''\to w}(00),$$

$$\mu^{(t)}_{w\to v}(10) \propto \sum_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v'\to w}(01) \prod_{v''\in\partial w\setminus\{v,v'\}} \mu^{(t)}_{v''\to w}(00),$$

$$\begin{aligned}\mu^{(t)}_{w\to v}(00) \propto{}& \theta_{ii} \prod_{v'\in\partial w\setminus\{v\}} \mu^{(t)}_{v\to w}(00) \\ &+ \sum_{v'\in\partial w\setminus\{v\}} \Big[\mu^{(t)}_{v'\to w}(11) \prod_{v''\in\partial w\setminus\{v,v'\}} \mu^{(t)}_{v''\to w}(00) \\ &+ \mu^{(t)}_{v'\to w}(10) \sum_{v''\in\partial w\setminus\{v,v'\}} \Big(\mu^{(t)}_{v''\to w}(01) \prod_{v'''\in\partial w\setminus\{v,v',v''\}} \mu^{(t)}_{v'''\to w}(00)\Big)\Big].\end{aligned}$$

The message along any edge (u,u') can be written as a vector

$$\mu_{u\to u'} = \begin{pmatrix}\mu_{u\to u'}(00)\\ \mu_{u\to u'}(01)\\ \mu_{u\to u'}(10)\\ \mu_{u\to u'}(11)\end{pmatrix},$$

and we can write

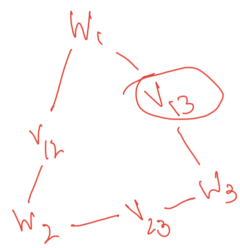
$$\mu^{(t)}_{w_i\to v_{ij}} \propto A_i \mu^{(t)}_{v_{ij'}\to w_i}$$

and

$$\mu^{(t)}_{v_{ij}\to w_i} \propto A_{ij} \mu^{(t-1)}_{w_j\to v_{ij}}.$$

Here, for $i = 1, 3$,

$$A_i := \begin{pmatrix}\theta_{ii} & 0 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\end{pmatrix},\ A_2 := \begin{pmatrix}\theta_{22} & 0 & 0 & 1\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 1 & 0 & 0 & 0\end{pmatrix},$$



whereas for $i < j$,

$$A_{ij} := \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & \theta_{ij} & 0 & 0\\ 0 & 0 & \theta_{ji} & 0\\ 0 & 0 & 0 & \theta_{ij}\theta_{ji}\end{pmatrix}.$$

1) Convergence: We can write the recursion

$$\mu^{(t)}_{w_2\to v_{23}} \propto A_{w_2\to v_{23}} \mu^{(t-1)}_{w_2\to v_{23}}$$

where $A_{w_2\to v_{23}} := A_{23}A_3A_{13}A_1A_{v_{12}}A_2 = (A_2A_{12}A_1A_{13}A_3A_{23})^T$. The message from v_{23} to w_2 can be written as

$$\mu^{(t)}_{v_{23}\to w_2} \propto A_{v_{23}\to w_2}\mu^{(t-1)}_{v_{23}\to w_2} \propto A^T_{w_2\to v_{23}}\mu^{(t-1)}_{v_{23}\to w_2}$$

Likewise, we can write $\mu^{(t)}_{w_i\to v_{ij}}$ in terms of $\mu^{(t-1)}_{w_i\to v_{ij}}$ as a linear transformation followed by renormalization. Since this factor graph has a single cycle, we can invoke the results of [26], which tell us that the SPA converges and the messages converge to the eigenvector corresponding to the largest eigenvalue of $A_{w_2\to v_{23}}$, and the partition function is equal to this eigenvalue. We cannot apply directly the Perron-Frobenius theorem to guarantee the existence of a unique largest eigenvalue because the matrices $A_{w_i\to v_{ij}}$ and $A_{v_{ij}\to w_i}$ need not be positive. However, with some minor modifications (permuting rows and columns), we can write $A_{w_i\to v_{ij}}$ as a block diagonal matrix, and try to apply the Perron-Frobenius theorem to each block. Yet, the existence of a unique largest eigenvalue for the matrix is not guaranteed. If the largest eigenvalue has multiplicity greater than one, then the SPA might oscillate. However, these scenarios can typically be circumvented by damping the messages.[2] The SPA is guaranteed to converge if the initial messages have nonzero projection onto the largest eigenvector. If we choose the initial vector randomly, then it will have a nonzero component in the direction of the eigenvector for the largest eigenvalue with high probability. We can prove the following.

Proposition 1. *For $n = 3$, if we initialize the messages randomly, then the sum-product algorithm converges with probability 1, and the Bethe permanent is equal to the largest eigenvalue of $A_{f_2\to g_{23}}$.*

2) Convexity:

Proposition 2. *The Bethe entropy function is a concave function of the beliefs. Hence, the Bethe free energy function is convex.*

[2]This is generally done by updating the messages using $\mu^{(t)} \leftarrow \alpha\mu^{(t)} + (1-\alpha)\mu^{(t-1)}$ at the end of each iteration, where α is the damping constant and is chosen heuristically.

Proof. For $n=3$, the Bethe entropy function can be written as a sum of conditional entropy functions. To see this, consider

$$\begin{aligned} H_{\mathrm{MB}}(\underline{b}) &= -\sum_{i=1}^{3}\sum_{l,m} b_{lm}^{(i)}\log b_{lm}^{(i)} + \sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{\underline{a}\in\{0,1\}^2} b_{\underline{a}}^{(ij)}\log b_{\underline{a}}^{(ij)} \\ &= -\sum_{l,m} b_{lm}^{(1)}\log b_{lm}^{(1)} + \sum_{\underline{a}\in\{0,1\}^2} b_{\underline{a}}^{(12)}\log b_{\underline{a}}^{(12)} \\ &\quad -\sum_{l,m} b_{lm}^{(2)}\log b_{lm}^{(2)} + \sum_{\underline{a}\in\{0,1\}^2} b_{\underline{a}}^{(23)}\log b_{\underline{a}}^{(23)} \\ &\quad -\sum_{l,m} b_{lm}^{(3)}\log b_{lm}^{(3)} + \sum_{\underline{a}\in\{0,1\}^2} b_{\underline{a}}^{(13)}\log b_{\underline{a}}^{(13)} \\ &= \left(H(\underline{b}^{(1)}) - H(\underline{b}^{(12)})\right) + \left(H(\underline{b}^{(2)}) - H(\underline{b}^{(23)})\right) + \left(H(\underline{b}^{(3)}) - H(\underline{b}^{(13)})\right). \end{aligned}$$

Every $b_{\underline{a}}^{(ij)}$ is obtained by marginalizing b^i (since $\underline{b}\in\mathcal{B}_{\mathrm{MB}}$). Specifically, $b_{11}^{(ij)} = b_{jj}^{(i)}$, $b_{10}^{(ij)} = b_{jj'}^{(i)}$, $b_{01}^{(ij)} = b_{j'j}^{(i)}$ and $b_{00}^{(ij)} = b_{ii}^{(i)} + b_{j'j'}^{(i)}$, where $j'\in[n]\backslash\{i,j\}$. Therefore, H_{MB} is a sum of conditional entropies. In Appendix A, we show that the conditional entropy $H(Y|X)$ is a concave function of the joint distribution p_{XY}. This completes the proof of Proposition 2. □

3) Correctness: For the remainder of this section, let us assume that $\theta_{11}=\theta_{22}=\theta_{33}=1$, and the identity permutation maximizes $\prod_{i=1}^{3}\theta_{i,\sigma(i)}$ (when viewed as a function of σ). We can guarantee this by permuting the columns, and then dividing each row (or each column) by θ_{ii}. The maximizing permutation can be found using a maximum weight matching algorithm. The permanent of the original matrix can be recovered from the permanent of the preprocessed matrix, and hence this preprocessing will not affect the following result.

Proposition 3. *For every nonnegative 3×3 matrix Θ with a nonzero permanent, we have*

$$0 \leqslant \eta_{\mathrm{MBP}}^{(3)}(\Theta) \leqslant \frac{1}{3}.$$

Furthermore, these bounds are tight.

Proof. From Prop. 1, we know that $\mathrm{perm}_{\mathrm{MB}}(\Theta)$ is equal to the largest eigenvalue of $A_{f_2\to g_{23}}$.

$$A_{f_2\to g_{23}} = \begin{pmatrix} \theta_{13}\theta_{22}\theta_{31} + \theta_{33}(\theta_{11}\theta_{22}+\theta_{12}\theta_{21}) & 0 & 0 & \theta_{11}\theta_{22}+\theta_{12}\theta_{21} \\ 0 & \theta_{12}\theta_{23}\theta_{31} & 0 & 0 \\ 0 & 0 & \theta_{13}\theta_{21}\theta_{32} & 0 \\ \theta_{11}\theta_{23}\theta_{32}\theta_{33} + \theta_{13}\theta_{23}\theta_{31}\theta_{32} & 0 & 0 & \theta_{11}\theta_{23}\theta_{32} \end{pmatrix}$$

We can explicitly compute the largest eigenvalue of $A_{f_2\to g_{23}}$, and find it to be[3]

$$\mathrm{perm}_B(\Theta) = \lambda_{\max} = \frac{1}{2}\left(\theta_{13}\theta_{22}\theta_{31} + \theta_{11}\theta_{23}\theta_{32} + \theta_{12}\theta_{21}\theta_{33} + \theta_{11}\theta_{22}\theta_{33} + \sqrt{f_1(\Theta)}\right) \tag{11}$$

where

$$\begin{aligned} f_1(\Theta) := {}& \theta_{13}^2\theta_{22}^2\theta_{31}^2 + \theta_{11}^2\theta_{23}^2\theta_{32}^2 + 2(2\theta_{12}\theta_{13}\theta_{21} + \theta_{11}\theta_{13}\theta_{22})\theta_{23}\theta_{31}\theta_{32} + (\theta_{12}^2\theta_{21}^2 + 2\theta_{11}\theta_{12}\theta_{21}\theta_{22} + \theta_{11}^2\theta_{22}^2)\theta_{33}^2 \\ & + 2\Big((\theta_{11}\theta_{12}\theta_{21} + \theta_{11}^2\theta_{22})\theta_{23}\theta_{32} + (\theta_{12}\theta_{13}\theta_{21}\theta_{22} + \theta_{11}\theta_{13}\theta_{22}^2)\theta_{31}\Big)\theta_{33}. \end{aligned} \tag{12}$$

Let us first show that $\eta_{\mathrm{MBP}}^{(3)}(\Theta)\geqslant 0$, or equivalently, that $\mathrm{perm}_{\mathrm{MB}}(\Theta)\leqslant\mathrm{perm}(\Theta)$. Since

$$\mathrm{perm}(\Theta) = \theta_{13}\theta_{22}\theta_{31} + \theta_{11}\theta_{23}\theta_{32} + \theta_{12}\theta_{21}\theta_{33} + \theta_{11}\theta_{22}\theta_{33} + \theta_{13}\theta_{21}\theta_{32} + \theta_{12}\theta_{23}\theta_{31},$$

it suffices to show that

$$f_1(\Theta) \leqslant (\theta_{13}\theta_{22}\theta_{31} + \theta_{11}\theta_{23}\theta_{32} + \theta_{12}\theta_{21}\theta_{33} + \theta_{11}\theta_{22}\theta_{33} + 2\theta_{13}\theta_{21}\theta_{32} + 2\theta_{12}\theta_{23}\theta_{31})^2.$$

The rest of the proof is a routine yet laborious calculation, expanding the square in the above, and comparing this with $f_1(\Theta)$ in (12). We omit the calculations here.

Let us now prove that $\eta_{\mathrm{MBP}}^{(3)}(\Theta)\leqslant 1/3$. This is equivalent to saying that $\mathrm{perm}_{\mathrm{MB}}(\Theta)\geqslant\mathrm{perm}(\Theta)/2$. The true permanent is equal to

$$\mathrm{perm}(\Theta) = \theta_{13}\theta_{22}\theta_{31} + \theta_{11}\theta_{23}\theta_{32} + \theta_{12}\theta_{21}\theta_{33} + \theta_{11}\theta_{22}\theta_{33} + \theta_{13}\theta_{21}\theta_{32} + \theta_{12}\theta_{23}\theta_{31}.$$

[3]This can be computed by hand or using symbolic computation software, see https://github.com/svatedka/bethepermanent_new for a Python implementation using the Sympy library [27].

Once again, we will use (11). Observe that the first three terms are contained in the expression for $\mathrm{perm}_{\mathrm{MB}}(\theta)$. It is therefore sufficient to show that the sum of the last two terms is at most $\sqrt{f_1(\theta)}$, or equivalently, to show that

$$\begin{aligned} f_1(\theta) &\geqslant (\theta_{13}\theta_{21}\theta_{32} + \theta_{12}\theta_{23}\theta_{31})^2 \\ &= \theta_{13}^2\theta_{21}^2\theta_{32}^2 + \theta_{12}^2\theta_{23}^2\theta_{31}^2 + 2\theta_{13}\theta_{21}\theta_{32}\theta_{12}\theta_{23}\theta_{31}. \end{aligned}$$

Without loss of generality, we can assume that $\theta_{13}\theta_{21}\theta_{32} \geqslant \theta_{12}\theta_{23}\theta_{31}$. If not, we can exchange the roles of the two terms. If $\theta_{12}\theta_{23}\theta_{31} = 0$, then using the fact that $\theta_{11}\theta_{22}\theta_{33} = 1 \geqslant \theta_{13}\theta_{21}\theta_{32}$, we have the above inequality being true. We can therefore assume that $\theta_{ij} > 0$ for all i, j. From our assumption, we have $\theta_{13}^2\theta_{21}^2\theta_{32}^2 \leqslant \theta_{11}^2\theta_{22}^2\theta_{33}^2 = 1$. We can then show that

$$\theta_{13}^2\theta_{22}^2\theta_{31}^2 + \theta_{11}^2\theta_{23}^2\theta_{32}^2 + \theta_{12}^2\theta_{21}^2\theta_{33}^2 \geqslant \theta_{12}^2\theta_{23}^2\theta_{31}^2$$

In fact, the following stronger result is true.

$$\max\left\{\theta_{13}^2\theta_{22}^2\theta_{31}^2, \theta_{11}^2\theta_{23}^2\theta_{32}^2, \theta_{12}^2\theta_{21}^2\theta_{33}^2\right\} \geqslant \theta_{12}^2\theta_{23}^2\theta_{31}^2$$

To see why this is true, let us assume otherwise and arrive at a contradiction. If

$$\max\left\{\theta_{13}^2\theta_{22}^2\theta_{31}^2, \theta_{11}^2\theta_{23}^2\theta_{32}^2, \theta_{12}^2\theta_{21}^2\theta_{33}^2\right\} < \theta_{12}^2\theta_{23}^2\theta_{31}^2,$$

then

$$\theta_{13}^2\theta_{22}^2\theta_{31}^2 \times \theta_{11}^2\theta_{23}^2\theta_{32}^2 \times \theta_{12}^2\theta_{21}^2\theta_{33}^2 < \theta_{12}^6\theta_{23}^6\theta_{31}^6,$$

or

$$\theta_{13}^2\theta_{22}^2\theta_{11}^2\theta_{32}^2\theta_{21}^2\theta_{33}^2 < \theta_{12}^4\theta_{23}^4\theta_{31}^4.$$

or (using the fact that $\theta_{11}\theta_{22}\theta_{33} = 1$),

$$\theta_{13}^2\theta_{32}^2\theta_{21}^2 < \theta_{12}^4\theta_{23}^4\theta_{31}^4.$$

This contradicts our original assumption that $1 \geqslant \theta_{13}\theta_{21}\theta_{32} \geqslant \theta_{12}\theta_{23}\theta_{31}$. Using this, and the fact that $\theta_{11}\theta_{22}\theta_{33} \geqslant \theta_{13}\theta_{21}\theta_{32}$, we see that

$$\theta_{13}^2\theta_{22}^2\theta_{31}^2 + \theta_{11}^2\theta_{23}^2\theta_{32}^2 + \theta_{12}^2\theta_{21}^2\theta_{33}^2 + \theta_{11}^2\theta_{22}^2\theta_{33}^2 \geqslant \theta_{13}^2\theta_{21}^2\theta_{32}^2 + \theta_{12}^2\theta_{23}^2\theta_{31}^2. \tag{13}$$

However, we also have that

$$f_1(\theta) \leqslant \theta_{13}^2\theta_{22}^2\theta_{31}^2 + \theta_{11}^2\theta_{23}^2\theta_{32}^2 + \theta_{12}^2\theta_{21}^2\theta_{33}^2 + \theta_{11}^2\theta_{22}^2\theta_{33}^2 + 2\theta_{13}\theta_{21}\theta_{32}\theta_{12}\theta_{23}\theta_{31}$$

which finally proves that $f_1(\theta) \geqslant (\theta_{13}\theta_{21}\theta_{32} + \theta_{12}\theta_{23}\theta_{31})^2$. This completes the second part of the proof.

We now show that the bounds are tight. The first part $\eta_{\mathrm{MBP}}^{(3)}(\Theta) = 0$ is achieved when Θ is equal to the identity matrix I_3. The second part $\eta_{\mathrm{MBP}}^{(3)}(\Theta) = 1/3$ is achieved when

$$\Theta = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

□

V. The Modified Bethe Permanent for $n > 3$

A. *Convexity of the Bethe free energy function*

We now proceed to study the case when $n > 3$. In this scenario, we observe that the Bethe entropy is not a concave function of the beliefs. In fact, we show in Appendix B that there exit vertices $\underline{b}_v \in \mathcal{B}_{\mathrm{MB}}$, direction $\underline{\xi}$ and a small $\epsilon > 0$ such that $H_{\mathrm{MB}}(\underline{b}_v + t\underline{\xi})$ is a convex function of t for $t \in [0, \epsilon]$.

Proposition 4. *For $n > 3$, the Bethe entropy is not a concave function of the beliefs.*

Proof. See Appendix B. The proof follows by computing expressions for the second directional derivative, and then finding a direction along which this is positive (in fact, equal to $+\infty$). □

Note: The sum-product algorithm can be thought of as an iterative algorithm akin to gradient descent that minimizes the Bethe free energy function. We have observed during simulations that the SPA never gets "stuck" at any vertex and even if we initialize at this vertex, the final beliefs lie within the interior of $\mathcal{B}_{\mathrm{MB}}$. This behaviour is unlike that for certain LDPC codes, where it has been observed that BP initialized at a vertex can get "stuck" because the free energy function is locally concave in all directions. Although there are directions along which F_{MB} is concave, we conjecture that for every point in the local marginal polytope, there are always directions along which F_{MB} is convex. We also conjecture that F_{MB} has a unique (global) minimum. Our conjectures are based on extensive simulations run for small n.

B. Vertices of the local marginal polytope

The structure of the local marginal polytope can give us insight into the behaviour of the sum product algorithm and the Bethe free energy function. Note that every belief $\underline{b}$ can be completely specified using the $b^{(i)}_{lm}$'s, and the beliefs $b^{(ij)}$'s corresponding to the factors g_{ij} are redundant.

Corresponding to every $i, l, m \in [n]$, we can define a variable $\sigma^{(i)}_{lm}$ which takes values from $\{0, 1\}$. Any assignment to $(\sigma^{(i)}_{lm})_{i,l,m}$ is called a *configuration*. A configuration is said to be *valid* if it corresponds to a permutation, or equivalently, satisfies the constraints

$$\sum_{l,m} \sigma^{(i)}_{lm} = 1$$

for all $i \in [n]$. Observe that every valid configuration is a vertex of $\mathcal{B}_{\text{MB}}$. Every valid configuration corresponds to a directed vertex disjoint cycle cover on n vertices.

Every belief vector can be represented as a set of weights on pairs of edges of a directed graph on n vertices. The belief $b^{(i)}_{lm}$ is a weight for the pair of directed edges $(l \to i, i \to m)$. For example, the belief vector (we only consider the $\underline{b}^{(i)}$'s)

$$\left[\underline{b}^{(1)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}, \underline{b}^{(2)} = \begin{pmatrix} 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \end{pmatrix}, \underline{b}^{(3)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \underline{b}^{(4)} = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\right]$$

corresponds to the graph in Fig. 3. It is easy to check that the above is in $\mathcal{B}_{\text{MB}}$. In the figure, weight on the dark blue pair of edges $(f_3 \to f_4, f_4 \to f_2)$ is the belief $b^{(4)}_{32}$, the weight on the dark green pair of edges $(3 \to 1, 1 \to 2)$ is equal to $b^{(1)}_{3,2}$, and so on.

In the original formulation of the Bethe permanent [11], the local marginal polytope $\mathcal{B}_{\text{B}}$ was found to be the same as the polytope of all $n \times n$ doubly stochastic matrices. Using the Birkhoff-von Neumann theorem, it was concluded that the vertices of $\mathcal{B}_{\text{B}}$ are precisely the set of all $n \times n$ permutation matrices $\mathcal{P}_n$. This fact was crucially used in obtaining upper and lower bounds for $\eta_{\text{B}}(\Theta)$.

We initally believed that a similar result holds for $\mathcal{B}_{\text{MB}}$. Let us call a point $\underline{b}$ in $\mathcal{B}_{\text{MB}}$ integral if all the entries of $\underline{b}$ are either 0 or 1. Every integral point corresponds to a valid configuration. Likewise, $\underline{b} \in \mathcal{B}_{\text{MB}}$ is called fractional if there is some entry of $\underline{b}$ which is strictly between 0 and 1. Clearly, every integral point in $\mathcal{B}_{\text{MB}}$ is a vertex. In fact, it is also easy to see that every integral point corresponds to a permutation. We initially suspected that $\mathcal{B}_{\text{MB}}$ contains no fractional vertices. However, using a numerical search, we found that $\mathcal{B}_{\text{MB}}$ contains fractional vertices.

1) Remarks: The notion of a doubly stochastic matrix can be generalized to tensors. An $n \times n \times n$ tensor A is said to be *plane stochastic* (these are called plane stochastic matrices in the literature) if for all $i \in [n]$,

$$\sum_{l,m} a_{ilm} = \sum_{l,m} a_{lim} = \sum_{l,m} a_{lmi} = 1.$$

A simple check verifies that the local marginal polytope is contained within the polytope of all plane stochastic tensors. Unlike the case of doubly stochastic matrices, the polytope of plane stochastic tensors has fractional vertices [28]. For large n, most of the vertices are fractional [29]. However, we cannot directly conclude that $\mathcal{B}_{\text{MB}}$ has fractional vertices (since it is a strict inclusion).

Let $\mathbb{B}$ denote the polytope of $n \times n$ doubly stochastic matrices. Since this is described by $2n$ linear constraints, $\mathbb{B}$ lies in a $n^2 - 2n$ dimensional subspace of $\mathbb{R}^{n^2}$. On the other hand, $\mathcal{B}_{\text{MB}}$ lies in a $n^3 - n - 7n(n-1)/2$ dimensional subspace of $\mathbb{R}^{n^3}$. For the special case of $n = 3$, we have $n^2 - 2n = n^3 - n - 7n(n-1)/2 = 3$. For $n = 3$, $\mathcal{B}_{\text{MB}}$ is equivalent to the polytope of 3×3 doubly stochastic matrices, and all vertices are integral.

2) The local marginal polytope has fractional vertices:

Lemma 1. *For $n > 3$, the local marginal polytope $\mathcal{B}_{\text{MB}}$ can have fractional vertices.*

We prove this lemma by providing an example. This was found by setting up an appropriate linear program and numerically finding the solution. The following is one such example

$$\underline{b} = \left[\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\right]. \tag{14}$$

The above can be represented graphically as in Fig. 3. This corresponds to a valid configuration (a vertex-disjoint cycle cover) on the double cover [30] of the original factor graph in Fig. 4. It is easy to show that the above cannot be expressed as a convex combination of integral vertices. Suppose that it were indeed possible to express $\underline{b}$ as $\underline{b} = \sum_{a=1}^{m} \xi(a)$, where $\xi(a)$ are integral beliefs and m is a positive integer. Then there must be some a such that $\xi^{(1)}_{44}(a) = 1$. However, this imposes the

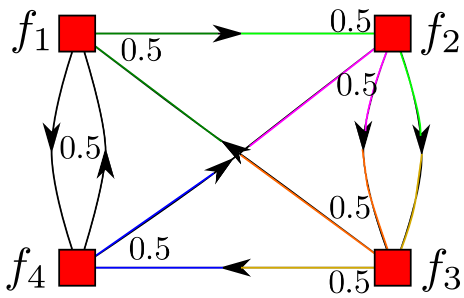


Fig. 3: The fractional vertex in (14). At each node, every monochromatic pair of incident edges represents a nonzero entry in (14).

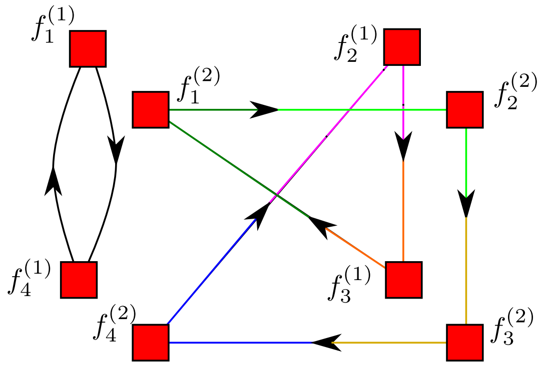


Fig. 4: Configuration in the double cover corresponding to example in Fig. 3.

constraint that either $\xi_{33}^{(2)}(a) = \xi_{22}^{(3)}(a) = 1$ or $\xi_{22}^{(2)}(a) = \xi_{33}^{(3)}(a) = 1$. Since $b_{33}^{(2)} = b_{22}^{(3)} = b_{22}^{(2)} = b_{33}^{(3)} = 0$, we have that $\underline{b}$ cannot be expressed as a convex combination of integral vertices.

We say that a point $\underline{b} \in \mathcal{B}_{\mathrm{MB}}$ is half integral if every entry belongs to $\{0, 0.5, 1\}$. For $n = 4$, we have not been able to find fractional vertices that are not half integral. However for larger n, we have observed vertices that are not half integral. For

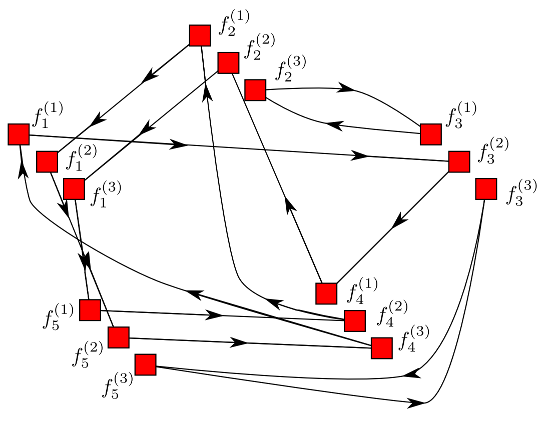


Fig. 5: Configuration in the triple cover which yields the fractional vertex in (15).

$n = 5$, the following is one such fractional vertex:

$$b = \left[\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2/3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 2/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 1/3 & 0 \\ 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/3 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2/3 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 2/3 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \right]. \tag{15}$$

This can be obtained from the pseudomarginal mapping of a triple cover, and the corresponding configuration is illustrated in Fig. 5

We initially wondered whether the fractional vertices were an artefact of the factor graph having odd cycles. However, this is not the case. If to $\mathcal{B}_{\mathrm{MB}}$ we further impose the constraint that $b_{lm}^{(i)} = 0$ for all $(i, l, m) \in [n/2] \times [n/2] \times [n]$ and all $(i, l, m) \in [n/2+1, n] \times [n/2+1, n] \times [n]$ — i.e., the factor graph is bipartite — we still have fractional vertices. The following is one example we obtained by numerical search (for $n = 6$).

$$b = \left[\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \right]. \tag{16}$$

C. An exact expression for η_{MB}

Watanabe and Chertkov [31] used the loop calculus technique [32] to show that for any stationary point $\underline{b}$ of the Bethe free energy function, the value of $\mathrm{perm}_{\mathrm{B}}(\Theta)$ is equal to the permanent of a certain matrix (which depends on $\underline{b}$). We can obtain a similar result:

Lemma 2. *For any stationary point of the Bethe free energy function, we have*

$$\eta_{\mathrm{MB}}(\Theta) = \frac{1}{n} \log \left(\frac{\sum_{\sigma \in S_n} \prod_i \varphi_i(\sigma)}{\prod_i \prod_{j \in [n] \setminus \{i\}} \sqrt{b_{00}^{(ij)}}} \right),$$

where for every $\underline{\sigma} \in S_n$,

$$\varphi_i(\sigma) = \begin{cases} b_{ii}^{(i)} & \text{if } \sigma_{ii} = 1 \\ \sqrt{b_{00}^{(ij)} b_{11}^{(ij)}} & \text{if } \sigma_{ij} = \sigma_{ji} = 1 \text{ for some } j \neq i \\ b_{lm}^{(i)} \sqrt{\frac{b_{00}^{(il)} b_{00}^{(im)}}{b_{10}^{(il)} b_{01}^{(im)}}} & \text{if } \sigma_{il} = \sigma_{mi} = 1 \text{ for some } l \neq m \neq i. \end{cases}$$

To prove the above lemma, we make use of the following result by Wainwright et al. [33]:

Lemma 3 ([33])**.** *Consider any factor graph (F, V) with V being the set of variable nodes and F being the set of factor nodes. Let Z denote the partition function of the factor graph. For any set of beliefs $(b_{\underline{a}}, b_i)_{\underline{a} \in F, i \in V}$, let $F_{\mathrm{MB}}(b)$ denote the corresponding Bethe free energy function, and $Z_{\mathrm{MB}}(b)$ denote $-\log(F_{\mathrm{MB}}(b))$. For any stationary point $(b_{\underline{a}}, b_i)_{\underline{a} \in F, i \in V}$ of the Bethe free energy function, we have*

$$\frac{Z}{Z_{\mathrm{MB}}(b)} = \sum_{\underline{x}} \left(\prod_{\underline{a} \in F} \frac{b_{\underline{a}}(\underline{x}_{\mathcal{N}_{\underline{a}}})}{\prod_{i \in \mathcal{N}_{\underline{a}}} b_i(x_i)} \right) \prod_{i \in V} b_i(x_i),$$

where $\mathcal{N}_{\underline{a}}$ denotes the set of variable nodes that form the neighbourhood of $\underline{a} \in F$.

$\mathrm{perm}(\Theta) \geqslant \mathrm{perm}_B(\Theta)$

1) Proof of Lemma 2: There is a one-to-one correspondence between the set of all valid configurations and the set of all $n \times n$ permutation matrices. Using Lemma 3, we have

$$
\begin{aligned}
\frac{\mathrm{perm}(\Theta)}{\mathrm{perm}_{\mathrm{MB}}(\Theta)} &= \sum_{\sigma \in S_n} \prod_{i=1}^{n} \left(\frac{b^{(i)}_{\sigma(i),\sigma^{-1}(i)}}{\prod_{l_1<i} b^{(l_1 i)}_{\sigma_{l_1 i}\sigma_{i l_1}} \prod_{l_2>i} b^{(i l_2)}_{\sigma_{i l_2}\sigma_{l_2 i}}} \right) \left(\prod_{i=1}^{k} \prod_{l>i} b^{(il)}_{\sigma_{il}\sigma_{li}} \right) \\
&= \sum_{\sigma \in S_n} \prod_{i=1}^{n} \left[\left(\frac{b^{(i)}_{\sigma(i),\sigma^{-1}(i)}}{\prod_{l_1<i} b^{(l_1 i)}_{\sigma_{l_1 i}\sigma_{i l_1}} \prod_{l_2>i} b^{(i l_2)}_{\sigma_{i l_2}\sigma_{l_2 i}}} \right) \left(\prod_{l_1<i} b^{(l_1 i)}_{\sigma_{l_1 i}\sigma_{i l_1}} \prod_{l_2>i} b^{(i l_2)}_{\sigma_{i l_2}\sigma_{l_2 i}} \right)^{1/2} \right] \qquad (17) \\
&= \sum_{\sigma \in S_n} \prod_{i=1}^{n} \left[\frac{b^{(i)}_{\sigma(i),\sigma^{-1}(i)}}{\left(\prod_{l_1<i} b^{(l_1 i)}_{\sigma_{l_1 i}\sigma_{i l_1}} \prod_{l_2>i} b^{(i l_2)}_{\sigma_{i l_2}\sigma_{l_2 i}} \right)^{1/2}} \right], \qquad (18)
\end{aligned}
$$

where we have used a slight abuse of notation by treating σ as both a (permutation) map and a (permutation) matrix. We can further simplify this by noticing that for each i, most of the σ_{ij} and σ_{ji}'s are zero. Let us define

$$
\phi_i(\sigma) := \frac{b^{(i)}_{\sigma(i),\sigma^{-1}(i)}}{\left(\prod_{l_1<i} b^{(l_1 i)}_{\sigma_{l_1 i}\sigma_{i l_1}} \prod_{l_2>i} b^{(i l_2)}_{\sigma_{i l_2}\sigma_{l_2 i}} \right)^{1/2}} \left(\prod_{l_1<i} b^{(l_1 i)}_{00} \prod_{l_2>i} b^{(i l_2)}_{00} \right)^{1/2}. \qquad (19)
$$

We now have to show that $\phi_i(\sigma) = \varphi_i(\sigma)$. First, if $\sigma(i) = i$, then we have $\sigma_{ij} = 0 = \sigma_{ji}$ for all $j \neq i$. Therefore, $\phi_i(\sigma) = b^{(i)}_{ii}$. If $\sigma(i) = j = \sigma^{-1}(i)$ for some $j \neq i$, then $\sigma_{ij} = \sigma_{ji} = 1$, while all other σ_{il}'s are zero. Moreover, $b^{(i)}_{jj} = b^{(ij)}_{11}$. Therefore, $\phi_i(\sigma) = \sqrt{b^{(ij)}_{11} b^{(ij)}_{00}}$. Similarly, if $\sigma(i) = l$ and $\sigma(m) = i$, we get $\phi_i(\sigma) = b^{(i)}_{lm} \sqrt{\frac{b^{(il)}_{00} b^{(im)}_{00}}{b^{(il)}_{10} b^{(im)}_{01}}}$, thus completing the proof. $\square$

VI. Simulation results

Simulations were performed using random matrices where each entry was chosen independently and uniformly at random from $[0, 1]$. The performance of the new algorithm (whose output we call the new Bethe permanent or $\mathrm{perm}_{\mathrm{MB}}(\Theta)$) was compared with [11] (whose output we refer to as the old Bethe permanent or $\mathrm{perm}_{\mathrm{B}}(\Theta)$). We can perform the comparison only for small values of n since it is computationally expensive to compute the true permanent. Before running the SPA, we preprocess the matrix by first running a maximum-weight matching algorithm and then permute the columns to obtain a Θ such that $\prod_{i=1}^{n} \theta_{i\sigma(i)}$ is maximized by the identity permutation σ. Fig. 6 illustrates that this preprocessing greatly improves the performance of the algorithm.

It was found that on average, the new Bethe permanent is a better approximation than the old Bethe permanent. We observed that even when $\mathrm{perm}_{\mathrm{MB}}(\Theta) < \mathrm{perm}_{\mathrm{B}}(\Theta)$, the two approximations are very close.

In all our simulations, we observed that the new algorithm converges (after introducing suitable damping).

We initially conjectured that $\eta_{\mathrm{MB}} \leqslant 1/3$. This is true for $n = 3$. However, there are matrices for $n > 3$ for which this was violated. We ran a numerical search for the worst matrix that maximizes the relative approximation error. This was done by iteratively finding a random direction along which η_{MB} is larger and perturbing Θ along that direction. The observed worst-case values were found to be roughly the same as in Table II.

Figure 6 shows empirical results compared with the algorithm of [11]. It is evident that the preprocessing step indeed improves the approximation factor of the modified Bethe permanent. Table II gives a comparison of the empirical average of the relative error. We see from Table II that our algorithm works very well for small n, and on average always performs better than [11]. However in all cases, η_{MB} is strictly below $1/2$.

The new algorithm significantly outperforms the old one when the matrix is close to being the tensor product of $1_{2\times 2}$ and $I_{n/2}$. As seen from Table IV, the improvement over [11] is pronounced for matrices that are close to $1_{k\times k} \otimes I_{n/k}$ for $k \ll n$. Here, $1_{k\times k}$ denotes the $k \times k$ all-1's matrix and I_m denotes the $m \times m$ identity matrix. Another ensemble where the new approximation outperforms the old one is Table V.

We also compared the two algorithms for the class of Toeplitz matrices

$$
\theta_{ij} = \begin{cases} 1 & \text{if } |i-j| \leqslant k \\ 0 & \text{otherwise,} \end{cases}
$$

where k is a fixed parameter. Estimates of the permanent of such matrices can help obtain bounds on the size of error correcting and covering codes over the metric space of permutations [34]. Table III illustrates a sample of our findings.

We also performed the simulations with initial messages corresponding to integral vertices of the local marginal polytope, and observed that the SPA converges to an internal point in the polytope even if convergence is slowed by damping. We saw in Sec. V-A that for certain integral vertices, there exist directions along which the Bethe entropy has infinite second derivative. This suggests that even though the Bethe free energy function has "steep hills" near certain vertices, there are also "valleys"

TABLE I: Simulation results for diagonally dominant matrices obtained by adding a diagonal matrix to the adjacency matrix of a random Erdos-Renyi graph. For each k, results were obtained by averaging over 1000 matrices.

k	$\mathbb{E}\eta_{\text{MB}}(\Theta)$	$\mathbb{E}\eta_{\text{B}}(\Theta)$	$\max \eta_{\text{MB}}(\Theta)$	$\max \eta_{\text{B}}(\Theta)$
12	0.014	0.35	0.08	0.10
13	0.02	0.04	0.08	0.10
15	0.03	0.06	0.08	0.10

along which the SPA can escape these hills to move towards the interior of the polytope. We conjecture that for every point in the local marginal polytope, there exist directions along which the Bethe free energy function is convex. We also conjecture that while F_{MB} can have multiple local maxima, it has no non-global local minima.

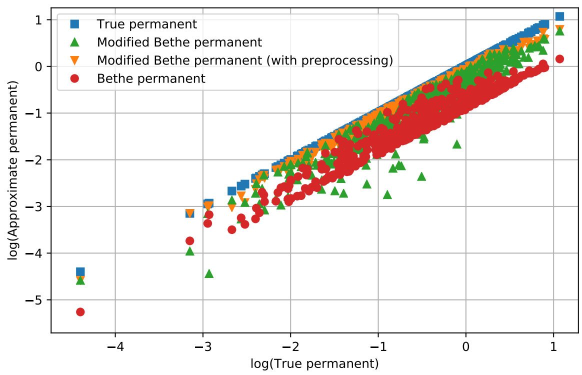


Fig. 6: Simulation results for $n = 3$. Entries of the random matrix were generated uniformly at random from $[0, 1]$. Here, preprocessing refers to the permutation of the rows/columns using the max-weight matching algorithm. The preprocessing significantly improves the performance of our algorithm.

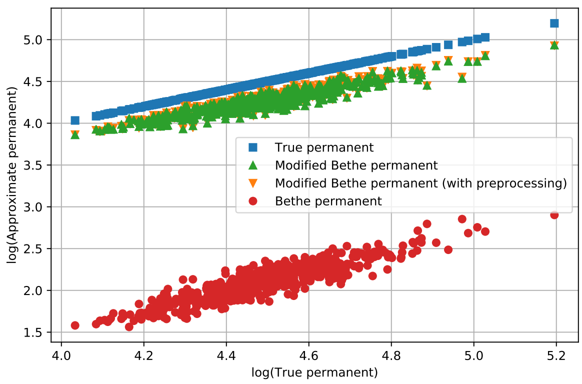


Fig. 7: Simulation results for $n = 12$. The random matrix Θ is obtained by adding random noise to the tensor product of $1_{2\times 2}$ and I_6. Here, preprocessing refers to the permutation of the rows/columns using the maximum-weight matching algorithm.

Appendix A
Concavity of the Bethe Entropy when $n = 3$

For $n = 3$, the Bethe entropy function can be written as a sum of conditional entropies. It then suffices to show that the conditional entropy

$$H_{Y|X}(p) = -\sum_{x,y} p_{xy} \log p_{xy} + \sum_{x} \left(\sum_{y} p_{xy} \right) \log \left(\sum_{y} p_{xy} \right)$$

is a concave function of p. For simplicity, we will assume that $\log$ denotes the natural logarithm for the rest of this section.

Proposition 5. *$H_{Y|X}(p)$ is a concave function of p.*

Proof. We can prove this by showing that the second-order directional derivative at every point and direction is less than or equal to zero. We will pick an arbitrary pmf p and a direction ξ that satisfies $\sum_{x,y} \xi_{xy} = 0$ and $p_{xy} + \xi_{xy} \geqslant 0$ for all (x, y). In

TABLE II: Simulation results for matrices of different dimensions. Matrices were generated randomly, with each entry drawn independently and uniformly from $[0, 1]$.

n	$\mathbb{E}\eta_{\mathrm{MB}}(\Theta)$	$\mathbb{E}\eta_{\mathrm{B}}(\Theta)$	$\max \eta_{\mathrm{MB}}(\Theta)$	$\max \eta_{\mathrm{B}}(\Theta)$
3	0.08	0.35	0.32	0.45
4	0.18	0.35	0.48	0.39
5	0.26	0.41	0.45	0.34
6	0.27	0.30	0.43	0.30
7	0.26	0.28	0.36	0.28
8	0.24	0.25	0.32	0.25
9	0.23	0.24	0.30	0.24
10	0.19	0.19	0.25	0.20

TABLE III: Simulation results for “banded” 12×12 matrices. In each case, $\theta_{ij} = 1$ if $|i - j| \leqslant k$ and $\theta_{ij} = 0$ otherwise.

k	$\eta_{\mathrm{MB}}(\Theta)$	$\eta_{\mathrm{B}}(\Theta)$
0	0.00	0.00
1	0.00	0.30
2	0.19	0.23
3	0.19	0.21
4	0.19	0.20
5	0.19	0.20

TABLE IV: Simulation results for 12×12 block-diagonal matrices. Matrices were generated by adding random i.i.d. noise to the tensor product $1_{k\times k} \otimes I_{12/k}$. For each k, results were obtained by averaging over 1000 matrices.

k	$\mathbb{E}\eta_{\mathrm{MB}}(\Theta)$	$\mathbb{E}\eta_{\mathrm{B}}(\Theta)$	$\max \eta_{\mathrm{MB}}(\Theta)$	$\max \eta_{\mathrm{B}}(\Theta)$
2	0.003	0.42	0.004	0.43
3	0.16	0.38	0.17	0.39
4	0.28	0.34	0.28	0.35
6	0.27	0.28	0.28	0.28
12	0.20	0.20	0.20	0.20

TABLE V: Simulation results for diagonally dominant matrices obtained by adding a diagonal matrix to the adjacency matrix of a random Erdős-Rényi graph. For each k, results were obtained by averaging over 1000 matrices.

k	$\mathbb{E}\eta_{\mathrm{MB}}(\Theta)$	$\mathbb{E}\eta_{\mathrm{B}}(\Theta)$	$\max \eta_{\mathrm{MB}}(\Theta)$	$\max \eta_{\mathrm{B}}(\Theta)$
12	0.01	0.35	0.08	0.10
13	0.02	0.04	0.08	0.10
15	0.03	0.06	0.08	0.10

other words, $p + \xi$ (and consequently $p + t\xi$ for any $0 \leqslant t \leqslant 1$) is a legitimate pmf. We then show that the second derivative of $H_{Y|X}(p + t\xi)$ for $t \to 0^+$ is nonpositive.

For every x, y, let us define $q_{xy}(t) := p_{xy} + t\xi_{xy}$.

We separately consider the following three cases: (a) p in the interior of the simplex, (b) p is a vertex (c) p is on an edge/face.

Case (a): We have $p_{xy} > 0$ for all x, y. Choose any valid direction ξ.

$$\begin{aligned}\frac{dH(q)}{dt} &= \frac{d}{dt}\left[-\sum_{x,y} q_{xy}(t)\log q_{xy}(t) + \sum_x (\sum_y q_{xy}(t))\log(\sum_y q_{xy}(t))\right]\\ &= -\sum_{x,y}[\xi_{xy}\log q_{xy}(t) + \xi_{xy}] + \sum_x\left[(\sum_y \xi_{xy})\log(\sum_y q_{xy}(t)) + \sum_y \xi_{xy}\right]\\ &= -\sum_{x,y}\xi_{xy}\log q_{xy}(t) + \sum_x(\sum_y \xi_{xy})\log(\sum_y q_{xy}(t))\end{aligned}$$

where in the last step, we have used the fact that $\sum_{x,y}\xi_{xy} = 0$.

Now compute the second derivative:

$$\begin{aligned}\frac{d^2H(q)}{dt^2} &= -\sum_{x,y}\frac{\xi_{xy}^2}{q_{xy}(t)} + \sum_x \frac{(\sum_y \xi_{xy})^2}{\sum_y q_{xy}(t)}\\ &= -\sum_{x,y}\frac{\xi_{xy}^2}{p_{xy} + t\xi_{xy}} + \sum_x \frac{(\sum_y \xi_{xy})^2}{\sum_y (p_{xy} + t\xi_{xy})}\end{aligned}$$

Note that $p_{xy} > 0$ for all (x, y). We have,

$$\left.\frac{d^2 H(q)}{dt^2}\right|_{t=0} = -\sum_{x,y} \frac{\xi_{xy}^2}{p_{xy}} + \sum_x \frac{(\sum_y \xi_{xy})^2}{\sum_y (p_{xy})}$$
$$= -\sum_{x,y} p_{xy} \left(\frac{\xi_{xy}}{p_{xy}}\right)^2 + \sum_x (\sum_y p_{xy}) \left(\frac{(\sum_y \xi_{xy})}{\sum_y (p_{xy})}\right)^2$$

Define $p_x := \sum_y p_{xy}$ and $p_{y|x} := \frac{p_{xy}}{\sum_{y'} p_{xy'}} = \frac{p_{xy}}{p_x}$. Using this in the above, and then using Jensen's inequality we get

$$\left.\frac{d^2 H(q)}{dt^2}\right|_{t=0} = -\sum_{x,y} p_x p_{y|x} \left(\frac{\xi_{xy}}{p_{xy}}\right)^2 + \sum_x p_x \left(\frac{(\sum_y \xi_{xy})}{p_x}\right)^2$$
$$\leqslant -\sum_x p_x \left(\sum_y p_{y|x} \frac{\xi_{xy}}{p_{xy}}\right)^2 + \sum_x p_x \left(\frac{(\sum_y \xi_{xy})}{p_x}\right)^2$$
$$= -\sum_x p_x \left(\sum_y \frac{\xi_{xy}}{p_x}\right)^2 + \sum_x p_x \left(\frac{(\sum_y \xi_{xy})}{p_x}\right)^2$$
$$= 0$$

Case (b): We have $p_{x^* y^*} = 1$ for some (x^*, y^*). This implies that $\xi_{x^* y^*} < 0$, while $\xi_{xy} \geqslant 0$ for all other (x, y). Without loss of generality, we can assume that for every x, there is at least one y such that $\xi_{xy} > 0$. Once again, we have (making use of the fact that $\xi_{xy} \geqslant 0$ for $(x, y) \neq (x^*, y^*)$)

$$\frac{d^2 H(q)}{dt^2} = -\sum_{(x,y):\xi_{xy}\neq 0} \frac{\xi_{xy}^2}{q_{xy}(t)} + \sum_x \frac{(\sum_y \xi_{xy})^2}{\sum_y q_{xy}(t)}$$
$$= -\frac{\xi_{x^* y^*}^2}{1 + t\xi_{x^* y^*}} - \sum_{(x,y)\neq(x^*,y^*)} \frac{\xi_{xy}}{t} + \frac{(\sum_y \xi_{x^* y})^2}{1 + t\sum_y \xi_{x^* y}} + \sum_{x\neq x^*} \frac{(\sum_y \xi_{xy})}{t}$$
$$= -\frac{\xi_{x^* y^*}^2}{1 + t\xi_{x^* y^*}} - \sum_{y\neq y^*} \frac{\xi_{x^* y}}{t} + \frac{(\sum_y \xi_{x^* y})^2}{1 + t\sum_y \xi_{x^* y}}$$

If $\xi_{x^* y} > 0$ for some $y \neq y^*$, then the above term tends to $-\infty$ as $t \to 0^+$. If $\xi_{x^* y} = 0$ for all $y \neq y^*$, then $\sum_y \xi_{x^* y} = \xi_{x^* y^*}$ and hence the second derivative is zero.

Case (c): In this case, there exists $\mathcal{A} \subset \mathcal{X} \times \mathcal{Y}$ with $|\mathcal{A}| < |\mathcal{X}||\mathcal{Y}| - 2$ such that $p_{xy} = 0$ for $(x, y) \in \mathcal{A}$ and $p_{xy} > 0$ otherwise. Therefore, $\xi_{xy} \geqslant 0$ for $(x, y) \in \mathcal{A}$. Let

$$\mathcal{A}_X := \{x : \exists y \text{ such that } p_{xy} = 0\},$$
$$\mathcal{A}_{X1} := \{x : p_{xy} = 0; \forall y\},$$

and

$$\mathcal{A}_{X2} := \{x : \exists y_1, y_2 \text{ such that } p_{xy_1} = 0, p_{xy_2} > 0\}.$$

Then,

$$\frac{d^2 H(q)}{dt^2} = -\sum_{(x,y):\xi_{xy}\neq 0} \frac{\xi_{xy}^2}{q_{xy}(t)} + \sum_x \frac{(\sum_y \xi_{xy})^2}{\sum_y q_{xy}(t)}$$
$$= -\sum_{(x,y)\in\mathcal{A}^c} \frac{\xi_{xy}^2}{p_{xy} + t\xi_{xy}} - \sum_{xy\in\mathcal{A}} \frac{\xi_{xy}}{t} + \sum_{x\in\mathcal{A}_X^c} \frac{(\sum_y \xi_{xy})^2}{\sum_y (p_{xy} + t\xi_{xy})} + \sum_{\mathcal{A}_{X1}} \sum_y \frac{\xi_{xy}}{t} + \sum_{x\in\mathcal{A}_{X2}} \frac{(\sum_y \xi_{xy})^2}{\sum_y (p_{xy} + t\xi_{xy})}$$

If $\mathcal{A}_{X2}$ is nonempty, then the above quantity tends to $-\infty$ as $t \to 0^+$. If $\mathcal{A}_{X2}$ is empty, then

$$\left.\frac{d^2 H(q)}{dt^2}\right|_{t=0} = -\sum_{(x,y)\in\mathcal{A}^c} \frac{\xi_{xy}^2}{p_{xy}} + \sum_{x\in\mathcal{A}_X^c} \frac{(\sum_y \xi_{xy})^2}{\sum_y (p_{xy})}$$

The remainder of the proof is identical to the last part of Case (a). Note that in the last part of Case (a), we did not require the assumption that $\sum_{xy} \xi_{xy} = 0$, and hence the above quantity is less than or equal to zero as a consequence of Jensen's inequality. $\square$

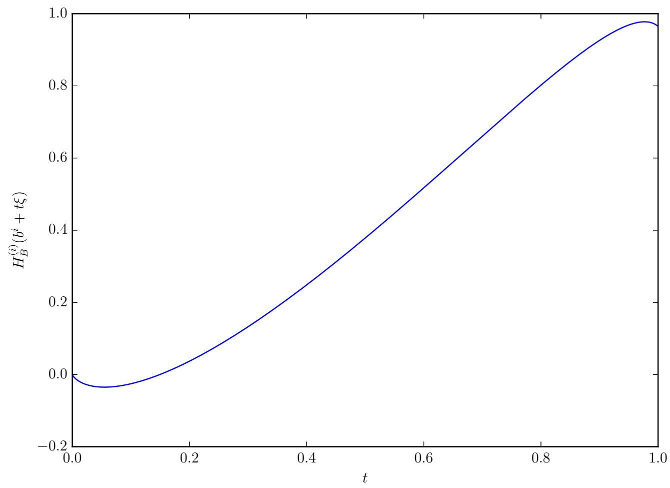


Fig. 8: Plot of $H_{\mathrm{MB}}^{(i)}(b^{(i)} + t\xi)$ as a function of t for a vertex b_i and a randomly generated ξ

Appendix B
H_{MB} is not concave for $n > 3$

Let

$$H_i(b^{(i)}) := -\sum_{l,m} b_{lm}^{(i)} \log b_{lm}^{(i)}$$

to be the entropy corresponding to f_i. Also, let

$$H_{lm}(b^{(l)}) := -\sum_{\underline{a}\in\{0,1\}^2} b_{\underline{a}}^{(lm)} \log b_{\underline{a}}^{(lm)}.$$

Note that

$$\begin{aligned} b_{11}^{(ij)} &= b_{jj}^{(i)} \\ b_{10}^{(ij)} &= \sum_{j'\neq i,j} b_{jj'}^{(i)} \\ b_{01}^{(ij)} &= \sum_{i'\neq i,j} b_{i'j}^{(i)} \\ b_{00}^{(ij)} &= \sum_{i',j'\neq i,j} b_{i'j'}^{(i)} \end{aligned}$$

Then, the modified Bethe entropy can be written as

$$H_{\mathrm{MB}}(\underline{b}) = \sum_i \left(H_i(b^{(i)}) - \frac{1}{2}\sum_{j\neq i} H_{ij}(b^{(i)}) \right)$$

We could hope that $H_{\mathrm{MB}}^{(i)}(b^{(i)}) := H_i(b^{(i)}) - \frac{1}{2}\sum_{j\neq i} H_{ij}(b^{(i)})$ is concave in $b^{(i)}$. However, it turns out that this is not the case. We observe that at certain vertices of the simplex, the directional derivative can be positive. Numerical computations seem to suggest that this function is concave as we go sufficiently into the interior of the simplex. Figure 8 illustrates this.

To see why this is the case, let us explicitly compute the second-order directional derivative of H_i at a vertex. We observe that at certain vertices, the directional derivative is always nonpositive. Specifically, if $b_{ii}^{(i)} = 1$, or if $b_{ll}^{(i)} = 1$ for some l, then the directional derivative is less than or equal to zero. This is not true for the remaining vertices.

Proposition 6. *$H_{\mathrm{MB}}^{(i)}(b^{(i)})$ is not a concave function of $b^{(i)}$.*

Proof. Let us consider an $(l*, m*)$ pair with $l* \neq m*$. Suppose that $b_{l*m*}^{(i)} = 1$. Consider any direction $\underline{\xi}$ having full support. In other words, there exists $\epsilon > 0$ such that $\underline{\xi}$ is such that $\underline{b} + t\underline{\xi}$ is a point in $\mathcal{B}_{\mathrm{MB}}$ for all $t \in [0, \epsilon]$ For ξ to be a valid direction, we must have $\xi_{l*m*} < 0$, while $\xi_{lm} > 0$ for all other (l, m). Also, we must have $\sum_{l,m} \xi_{lm} = 0$.

For notational convenience, we will drop the superscript i in the remainder of the proof. It is to be understood that indices in summations run over $[n]\backslash\{i\}$. As in the $n = 3$ case, we can compute

$$\frac{d^2 H_i}{dt^2} = -\sum_{l,m} \frac{(\xi_{lm})^2}{b_{lm} + t\xi_{lm}} - \frac{(\xi_i)^2}{b_i + t\xi_i}$$

and

$$\frac{d^2 H_{il}}{dt^2} = -\frac{\xi_{ll}^2}{b_{ll} + t\xi_{ll}} - \frac{(\sum_{l' \neq l} \xi_{l'l})^2}{\sum_{l' \neq l}(b_{l'l} + t\xi_{l'l})} - \frac{(\sum_{m' \neq l} \xi_{lm'})^2}{\sum_{m' \neq l}(b_{lm'} + t\xi_{lm'})} - \frac{(\xi_i + \sum_{l',m' \neq l} \xi_{l'm'})^2}{b_i + t\xi_i + \sum_{l',m' \neq l}(b_{l'm'} + t\xi_{l'm'})}$$

Substituting for b, we get

$$\frac{d^2 H_i}{dt^2} = -\frac{(\xi_{l*m*})^2}{1 + t\xi_{l*m*}} - \sum_{(l,m) \neq (l*,m*)} \frac{\xi_{lm}}{t} - \frac{\xi_i}{t}$$

For $l \notin \{l^*, m^*\}$, we have

$$\frac{d^2 H_{il}}{dt^2} = -\frac{\xi_{ll}}{t} - \frac{\sum_{l' \neq l} \xi_{l'l}}{t} - \frac{\sum_{m' \neq l} \xi_{lm'}}{t} - \frac{(\xi_i + \sum_{l',m' \neq l} \xi_{l'm'})^2}{1 + t\xi_i + t\sum_{l',m' \neq l} \xi_{l'm'}}$$

Also,

$$\frac{d^2 H_{il*}}{dt^2} = -\frac{\xi_{l*l*}}{t} - \frac{\sum_{l' \neq l*} \xi_{l'l*}}{t} - \frac{(\sum_{m' \neq l*} \xi_{lm'})^2}{1 + t\sum_{m' \neq l*} \xi_{l*m'}} - \frac{\xi_i + \sum_{l',m' \neq l*} \xi_{l'm'}}{t}$$

and

$$\frac{d^2 H_{im*}}{dt^2} = -\frac{\xi_{m*m*}}{t} - \frac{(\sum_{l' \neq m*} \xi_{l'm*})^2}{1 + t\sum_{l' \neq m*} \xi_{l'm*}} - \frac{\sum_{m' \neq m*} \xi_{m*m'}}{t} - \frac{\xi_i + \sum_{l',m' \neq m*} \xi_{l'm'}}{t}$$

With a little bit of simplification, we obtain

$$\begin{aligned}\frac{d^2 H^{(i)}_{\mathrm{MB}}}{dt^2} = &-\frac{(\xi_{l*m*})^2}{1 + t\xi_{l*m*}} - \frac{1}{2}\sum_{m' \neq l*} \frac{\xi_{l*m'}}{t} - \frac{1}{2}\sum_{l' \neq m*} \frac{\xi_{l'm*}}{t} + \frac{1}{2}\frac{(\sum_{m' \neq l*} \xi_{lm'})^2}{1 + t\sum_{m' \neq l*} \xi_{l*m'}} + \frac{1}{2}\frac{(\sum_{l' \neq m*} \xi_{l'm*})^2}{1 + t\sum_{l' \neq m*} \xi_{l'm*}} \\ &+ \frac{1}{2}\frac{\xi_i + \sum_{l',m' \neq l*} \xi_{l'm'}}{t} + \frac{1}{2}\frac{\xi_i + \sum_{l',m' \neq m*} \xi_{l'm'}}{t} + \frac{1}{2}\sum_{l \neq l*,m*} \frac{(\xi_i + \sum_{l',m' \neq l} \xi_{l'm'})^2}{1 + t\xi_i + t\sum_{l',m' \neq l} \xi_{l'm'}}\end{aligned}$$

If

$$\sum_{m' \neq l*} \xi_{l*m'} + \sum_{l' \neq m*} \xi_{l'm*} < \sum_{l',m' \neq l*} \xi_{l'm'} + \sum_{l',m' \neq m*} \xi_{l'm'},$$

then the second-order directional derivative tends to ∞ as $t \to 0$. Hence, $H^{(i)}_{\mathrm{MB}}$ is not concave. □

Corollary 1. *The Bethe entropy function is not concave*

Proof. Note that Proposition 6 in itself does not imply that the Bethe entropy function is nonconcave. However, we can easily extend the counterexample in the proof to show that H_{MB} is not concave.

Choose the permutation σ on $[n]$ such that $\sigma(i) = [i + 1] \bmod n$. For each i, fix

$$b^{(i)}_{lm} = \begin{cases} 1 & \text{if } l = \sigma(i) \text{ and } m = \sigma^{-1}(i) \\ 0 & \text{otherwise.} \end{cases}$$

and $b^{(i)}_i = 0$. Fix an $0 < \epsilon < 1$ and let

$$\xi^i_{lm} = \begin{cases} -\epsilon & \text{if } l = \sigma(i) \text{ and } m = \sigma^{-1}(i) \\ \epsilon/(k-1)^2 & \text{otherwise.} \end{cases}$$

and $\xi^i_i = \epsilon/(k-1)^2$. Then, for every $t \in [0, 1]$, $\{b^{(i)} + t\xi^i : 1 \leqslant i \leqslant n\}$ constitutes a valid set of beliefs. One can easily verify that all the consistency conditions are satisfied. The calculations in the proof of Prop. 6 say that for every i, $\frac{d^2 H^i_{\mathrm{MB}}}{dt^2}(b^{(i)} + t\xi^i)$ tends to infinity as $t \to 0$. Since $H_{\mathrm{MB}} = \sum_i H^i_{\mathrm{MB}}$, the Bethe entropy also tends to ∞ as $t \to 0$. This proves that it is nonconcave. □

REFERENCES

[1] S. Vatedka and P. O. Vontobel, “Modified Bethe approximation of a nonnegative matrix,” in *2020 Int. Conf. Sig. Proc. and Comm.*, Bengaluru, India, 2020.

[2] M. Chertkov, L. Kroc, and M. Vergassola, “Belief propagation and beyond for particle tracking,” *arXiv:0806.1199*, 2008.

[3] P. O. Vontobel, “The Bethe approximation of the pattern maximum likelihood distribution,” in *2012 IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, 2012.

[4] H. J. Ryser, *Combinatorial Mathematics*. Math. Assoc. America; dist. Wiley [New York], 1963, no. 14.

[5] L. G. Valiant, “The complexity of computing the permanent,” *Theoretical Comp. Sc.*, vol. 8, no. 2, pp. 189–201, 1979.

[6] N. Karmarkar, R. Karp, R. Lipton, L. Lovász, and M. Luby, “A Monte-Carlo algorithm for estimating the permanent,” *SIAM J. Computing*, vol. 22, no. 2, pp. 284–293, 1993.

[7] A. Barvinok, “Polynomial time algorithms to approximate permanents and mixed discriminants within a simply exponential factor,” *Random Structures & Algorithms*, vol. 14, no. 1, pp. 29–61, 1999.

[8] N. Linial, A. Samorodnitsky, and A. Wigderson, “A deterministic strongly polynomial algorithm for matrix scaling and approximate permanents,” *Combinatorica*, vol. 20, no. 4, pp. 545–568, 2000.

[9] M. Jerrum, A. Sinclair, and E. Vigoda, “A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries,” *J. ACM*, vol. 51, no. 4, pp. 671–697, 2004.

[10] B. Huang and T. Jebara, “Approximating the permanent with belief propagation,” *arXiv:0908.1769*, 2009.

[11] P. O. Vontobel, “The Bethe permanent of a nonnegative matrix,” *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1866–1901, Mar. 2013.

[12] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Constructing free-energy approximations and generalized belief propagation algorithms,” *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2288–2312, Jul. 2005.

[13] L. Gurvits, “Unleashing the power of Schrijver’s permanental inequality with the help of the Bethe approximation,” *arXiv preprint arXiv:1106.2844*, 2011.

[14] D. Straszak and N. K. Vishnoi, “Belief propagation, Bethe approximation and polynomials,” in *Proc. Allerton Conf. Comm., Cont., and Computing*, Monticello, IL, USA, Oct. 2017, pp. 666–671.

[15] L. Gurvits and A. Samorodnitsky, “Bounds on the permanent and some applications,” in *Proc. Ann. Symp. Found. Comp. Science*, PA, USA, Oct. 2014, pp. 90–99.

[16] N. Anari and A. Rezaei, “A tight analysis of bethe approximation for permanent,” in *Proc. Ann. Symp. Found. Comp. Science*, Baltimore, MD, USA, Nov. 2019, pp. 1434–1445.

[17] N. Anari, M. Charikar, K. Shiragur, and A. Sidford, “The bethe and sinkhorn permanents of low rank matrices and implications for profile maximum likelihood,” *arXiv preprint arXiv:2004.02425*, 2020.

[18] M. Bayati, D. Shah, and M. Sharma, “Max-product for maximum weight matching: Convergence, correctness, and LP duality,” *IEEE Trans. Inf. Theory*, vol. 54, no. 3, p. 1241–1251, Mar. 2008.

[19] S. Sanghavi, D. Malioutov, and A. Willsky, “Belief propagation and LP relaxation for weighted matching in general graphs,” *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2203–2212, Apr. 2011.

[20] M. X. Cao and P. O. Vontobel, “Double-edge factor graphs: definition, properties, and examples,” in *Proc. Inf. Theory Workshop*, Kaohsiung, Taiwan, Nov. 2017, pp. 136–140.

[21] H.-A. Loeliger and P. O. Vontobel, “Factor graphs for quantum probabilities,” *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5642–5665, Sep. 2017.

[22] ——, “Quantum measurement as marginalization and nested quantum systems,” *IEEE Trans. Inf. Theory, to appear*, 2020.

[23] M. X. Cao and P. O. Vontobel, “Bounding and estimating the classical information rate of quantum channels with memory,” *IEEE Trans. Inf. Theory, to appear*, 2020.

[24] H.-A. Loeliger, “An introduction to factor graphs,” *IEEE Sig. Proc. Magazine*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[25] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[26] Y. Weiss, “Correctness of local probability propagation in graphical models with loops,” *Neural Computation*, vol. 12, no. 1, pp. 1–41, 2000.

[27] A. Meurer et al., “SymPy: symbolic computing in Python,” *PeerJ Computer Science*, p. e103, Jan. 2017. [Online]. Available: https://doi.org/10.7717/peerj-cs.103

[28] J. Csima, “Multidimensional stochastic matrices and patterns,” *Journal of algebra*, vol. 14, p. 194–202, 1970.

[29] R. Brualdi and J. Csima, “Extremal plane stochastic matrices of dimension three,” *Linear Algebra and its Applications*, vol. 11, no. 2, p. 105–133, 1975.

[30] P. O. Vontobel, “Counting in graph covers: A combinatorial characterization of the Bethe entropy function,” *IEEE Trans. Inf. Theory*, vol. 59, no. 9, Nov.

[31] Y. Watanabe and M. Chertkov, “Belief propagation and loop calculus for the permanent of a non-negative matrix,” *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 24, p. 242002, 2010.

[32] M. Chertkov and V. Y. Chernyak, “Loop calculus in statistical physics and information science,” *Physical Review E*, vol. 73, no. 6, p. 065102, 2006.

[33] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, “Tree-based reparameterization framework for analysis of sum-product and related algorithms,” *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1120–1146, May 2003.

[34] M. Schwartz and P. O. Vontobel, “Improved lower bounds on the size of balls over permutations with the infinity metric,” *IEEE Trans. Inf. Theory*, vol. 63, no. 10, pp. 6227–6239, Oct. 2017.