Graph Covers and Quadratic Minimization

Nicholas Ruozzi
Computer Science
Yale University
New Haven, CT 06520, USA
nicholas.ruozzi@yale.edu

Justin Thaler*
Computer Science
Harvard University
Cambridge, MA 02138, USA
jthaler@fas.harvard.edu

Sekhar Tatikonda
Electrical Engineering
Yale University
New Haven, CT 06520, USA
sekhar.tatikonda@yale.edu

Abstract

We formulate a new approach to understanding the behavior of the min-sum algorithm by exploiting the properties of graph covers. First, we present a new, natural characterization of scaled diagonally dominant matrices in terms of graph covers; this result motivates our approach because scaled diagonal dominance is a known sufficient condition for the convergence of min-sum in the case of quadratic minimization. We use our understanding of graph covers to characterize the periodic behavior of the min-sum algorithm on a single cycle. Lastly, we explain how to extend the single cycle results to understand the 2-periodic behavior of min-sum for general pairwise MRFs. Some of our techniques apply more broadly, and we believe that by capturing the notion of indistinguishability, graph covers represent a valuable tool for understanding the abilities and limitations of general message-passing algorithms.

1. Introduction

Belief propagation (BP) and its variants perform empirically well in application areas including coding theory, statistical physics, and linear programming, but rigorously characterizing their behavior outside of a few well-structured instances has proved challenging. In general, BP is not even guaranteed to converge. However, in the absence of convergence we have empirically observed BP to exhibit only two "failure modes": either the messages diverge to $\pm\infty$ or the messages oscillate. In this paper, we lay the groundwork for a general understanding of the periodic behavior of belief propagation. Our primary tool for characterizing periodicity is the notion of a graph cover.

Graph covers have been used to understand the limits of message passing algorithms in distributed computation [1], but have received relatively little attention in the belief propagation community. The best known use of graph covers with respect to belief propagation appears in the recent work of Vontobel et al. In [2], the authors show that, when using max-product to solve linear programming problems, fractional solutions of the original linear program can be viewed as integer solutions on covers.

*Justin Thaler is supported in part by NSF grant CNS-0721491.

We believe that covers provide important insight into the periodic behavior of belief propagation. We demonstrate that for the specific case of minimizing a quadratic function, we can provably correct certain oscillatory behaviors to achieve the correct solution. We then use graph covers to suggest a tantalizing connection between periodicity and the behavior of min-sum on covers of the original problem.

In addition to providing insight into what are typically considered the failure modes of the min-sum algorithm, graph covers allow us to characterize when we should expect "good" behavior. To this end, we will consider the behavior of the min-sum algorithm for the unconstrained quadratic minimization problem:

$$\min_x \frac{1}{2}x^T\Gamma x - h^T x$$

where $\Gamma \in \mathbb{R}^{n\times n}$ is a symmetric positive definite with unit diagonal and $x \in \mathbb{R}$.

Quadratic minimization is equivalent to solving $\Gamma x = h$, and as a result, arises naturally in a variety of application areas. From our perspective, quadratic minimization is an important special case of the min-sum algorithm. Unlike many of the other applications of min-sum, for quadratic minimization over the reals we are able to provide closed form solutions for the message updates. This allows us to apply tools from the theory of differentiable functions to aid our understanding of the algorithm's behavior.

There are several known conditions that are sufficient to guarantee convergence of the min-sum algorithm for this problem:

Definition 1.1. $\Gamma \in \mathbb{R}^{n\times n}$ *is* ***walk-summable*** *if the spectral radius* $\rho(|I-\Gamma|) < 1$.

Definition 1.2. $\Gamma \in \mathbb{R}^{n\times n}$ *is* ***scaled diagonally dominant*** *if* $\exists w > 0 \in \mathbb{R}^n$ *such that* $|\Gamma_{ii}|w_i > \sum_{j\neq i}|\Gamma_{ij}|w_j$.

Here, we use $|A|$ to denote the matrix obtained from A by taking the absolute value of every entry.

Malioutov et al. [3] showed that walk-summability is a sufficient condition for convergence of min-sum. In [4] and [5], Moallemi et al. showed that scaled diagonal dominance is a sufficient condition for convergence for both quadratic and general convex functions. In this work, we will prove that the seemingly unnatural sufficient conditions above have a natural interpretation in the context of belief propagation via graph covers.

2. Preliminaries

Before we proceed to our main theorems, we briefly review the relevant background material pertaining to the min-sum algorithm and quadratic minimization.

2.1. The Min-Sum Algorithm and Quadratic Minimization

The min-sum algorithm attempts to compute the min-marginals of an objective function $p(x_1,...,x_n)$ that can be written as a sum of self-potentials and edge potentials,

$$p(x_1,...,x_n) = \sum_i \phi_i(x_i) + \sum_\alpha \psi_\alpha(x_\alpha)$$

where $\alpha \subseteq \{1,...,n\}$.

Every factorization of p has a corresponding graphical representation known as a factor graph. The factor graph consists of a node for each of the variables $x_1,...,x_n$ and each of the factors ψ_α with an edge joining ψ_α to x_i if $i \in \alpha$. In this paper we will only need the pair-wise case where $\forall\alpha, |\alpha| = 2$. In this special case, we can simplify the factor graph by eliminating the factor nodes and simply joining x_i and x_j by an edge if ψ_{ij} is a not identically zero. The min-sum algorithm is then a message passing algorithm on this reduced factor graph. On the t^{th} iteration of the algorithm, messages are passed along each edge of the factor graph as follows:

$$m^t_{i\to j}(x_j) = \min_{x_i} \phi_i(x_i) + \psi_{ij}(x_i,x_j) + \sum_{k\in\partial i - j} m^{t-1}_{k\to i}(x_i)$$

We will assume that the initial messages m^0 are zero. Given any vector of messages, m, we can construct a set of beliefs that are intended to estimate the min-marginals:

$$\begin{aligned} \tau^t_i(x_i) &= \phi_i(x_i) + \sum_{j\in\partial i} m^t_{k\to i}(x_i) \\ \tau^t_{ij}(x_i,x_j) &= \phi_j(x_j) + \phi_i(x_i) + \psi_{ij}(x_i,x_j) + \\ &\quad \sum_{k\in\partial i - j} m^t_{j\to i}(x_i) + \sum_{k\in\partial j - i} m^t_{k\to j}(x_j) \end{aligned}$$

We can then estimate the assignment that minimizes the objective function:

$$x^t_i = \arg\min_{x_i} \tau^t_i(x_i)$$

For the specific case of minimizing a quadratic function, equivalently finding the mean of a multivariate Gaussian probability distribution, we can write the objective function as:

$$\begin{aligned} p(x_1,...,x_n) &= \frac{1}{2}x^T\Gamma x - h^T x \\ &= \sum_i \frac{1}{2}\Gamma_{ii}x_i^2 - h_i x_i + \sum_{i>j}\Gamma_{ij}x_i x_j \end{aligned}$$

where $\Gamma \in \mathbb{R}^{n\times n}$ is symmetric positive definite. We will assume, without loss of generality, that Γ has been normalized to contain only ones along its diagonal.

Because the minimization is being performed over quadratic functions, we can explicitly compute the minimization required by the min-sum algorithm at each time step. In this way, the message update $m^t_{i\to j}$ can be parameterized as a quadratic function of the form $a^t_{ij}x_j^2 + b^t_{ij}x_j$ where the constants are given by:

$$\begin{aligned} a^t_{ij} &= \frac{-\frac{1}{2}\Gamma^2_{ij}}{\Gamma_{ii} + 2\sum_{k\in\partial i - j} a^{t-1}_{ki}} \\ b^t_{ij} &= \frac{(h_i - \sum_{k\in\partial i - j} b^{t-1}_{ki})\Gamma_{ij}}{\Gamma_{ii} + 2\sum_{k\in\partial i - j} a^{t-1}_{ki}} \end{aligned}$$

These updates are only valid for $\Gamma_{ii} + 2\sum_{k\in\partial i - j} a^{t-1}_{ki} > 0$. If this is not the case, then $a^t_{ij} = b^t_{ij} = -\infty$. For the initial constants, we take $a^0_{ij} = b^0_{ij} = 0$.

2.2. Definitions of Convergence

There are several notions of convergence that one may consider:

1. The beliefs converge to a fixed point.
2. The messages converge to a fixed point.
3. The estimates converge to a fixed point.

The standard notion of convergence is to consider convergence of the beliefs. We can see that if the messages converge then the beliefs must also converge, because the beliefs are defined as sums of the messages, self-potentials, and edge potentials. However, there are situations in which the messages and beliefs are converging, but the estimates are not. Consider the following objective function:

Example 2.1.

$$p(x_1,x_2,x_3) = \frac{1}{2}x^T \begin{pmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{pmatrix} x - \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^T x.$$

For this example, the min-sum algorithm oscillates between the all ones estimate and the all zeros estimate for any two consecutive time steps. The messages and the beliefs are converging, but will only reach their fixed point values in the limit. This situation is a special type of convergence: if we had infinite precision then the beliefs and the messages would continue to oscillate in such a way as to produce exactly the same series of alternating estimates. Because of this example, we will break with the standard definition of convergence in favor of a notion of periodic convergence. Let $T(m)$ denote the action of one step of the algorithm on the message m.

Definition 2.2. *The min-sum algorithm* ***converges to a solution of period*** k *if there exist messages* $\widetilde{m}^1, ..., \widetilde{m}^k$ *and estimates* $\widetilde{x}^1, ..., \widetilde{x}^k$ *for each* $s \in \{1, ..., k\}$

- $\forall i, j \lim_{t\to\infty} m_{i\to j}^{kt+s} = \widetilde{m}_{i\to j}^s$
- $\forall i \lim_{t\to\infty} x_i^{kt+s} = \widetilde{x}_i^s$

Recall that x_i^t is the *unique* value that minimizes the belief τ_i^t. If τ_i^t is degenerate then the limiting operation is ill-defined.

Denote the algorithm's one step action on a vector of messages m as $T(m)$. Notice that if the messages $\widetilde{m}^1, ..., \widetilde{m}^k$ are a fixed point of period k then they must satisfy $T(\widetilde{m}^i) = \widetilde{m}^{i+1 \mod k}$.

1-periodic convergence of the messages is not equivalent to obtaining a fixed point set of messages in the limit. This is because convergence toward a fixed point message requires convergence of the messages but not the estimates. However, in both 1-periodic convergence and convergence to a fixed point set of messages, we obtain fixed point beliefs. For a set of fixed point beliefs, we have the following result from Wainwright et al. [6]:

Theorem 2.3. *If the min-sum algorithm converges to a set of fixed point beliefs* τ^* *such that* $\forall i$*, there exists a unique* x_i^* *that minimizes* τ_i^**. Then* x^* *is a local optimum of the objective function.*

For the convex quadratic minimization problem, local optima correspond to global optima. So, if the beliefs converge and we can extract a unique estimate then this estimate is guaranteed to be the optimal solution.

2.3. Graph Covers

Definition 2.4. *A graph* H ***covers*** *a graph* G *if there exists a graph homomorphism* $\phi : H \to G$ *such that* ϕ *is an isomorphism on* ∂v *for all vertices* $v \in H$*. If* $\phi(v) = u$ *then we say that* $v \in H$ *is a copy of* $u \in G$*. We say that* H *is a k-cover of* G *if every vertex of* G *has* k *copies in* H*.*

If H covers the factor graph G then H has the same local properties as G. For any cover H of G and any set of initial messages on G, there exists a choice of initial messages on H such that the messages passed by min-sum are identical on both graphs: for every node $v \in G$ the messages received and sent by this node at time t are exactly the same as the messages sent and received at time t by any copy of v in H. As a result, if we use the min-sum algorithm to deduce an assignment for v, the algorithm run on the graph H must deduce the same assignment for each copy of v.

3. Quadratic Minimization and Covers

Let G be the factor graph for $p_G(x_1, ..., x_n) = \frac{1}{2}x^T\Gamma x - h^T x$. G has a node i for each variable x_i and an edge joining i to j if $\Gamma_{ij} \neq 0$.

Let H be a k-cover of G, and let $p_H(x_{11}, ..., x_{1k}, ...x_{nk}) = \frac{1}{2}x^T\widetilde{\Gamma}x - \widetilde{h}^T x$ be the corresponding objective function. Without loss of generality we can assume that p_H can be written with

$$\widetilde{\Gamma} = \begin{pmatrix} \Gamma_{11}P_{11} & \cdots & \Gamma_{1n}P_{1n} \\ \vdots & \ddots & \vdots \\ \Gamma_{n1}P_{n1} & \cdots & \Gamma_{nn}P_{nn} \end{pmatrix} \quad (1)$$

$$\widetilde{h}_i = h_{\lceil i/k \rceil} \quad (2)$$

where, for $i \neq j$, P_{ij} is a $k \times k$ permutation matrix and P_{ii} is the $k \times k$ identity matrix.

Definition 3.1. *Let* Γ_G *be the quadratic term of objective function* p_G *with factor graph* G*. We say that* Γ_H ***covers*** Γ_G *if* H *covers* G *and* Γ_H *is the quadratic term of the objective function* p_H*.*

For the quadratic minimization problem, the factor graphs for min-sum and their covers share many of the same properties. Most notably, we can transform critical points on covers to critical points of the original problem. Let H and G be as above. We have the following lemma:

Lemma 3.2. *Suppose* $\widetilde{\Gamma}x' = \widetilde{h}$ *for* $x' \in \mathbb{R}^{nk}$*. If* $x \in \mathbb{R}^n$ *is given by* $x_i = \frac{1}{k}\sum_{j=1}^k x'_{ki+j}$ *then* $\Gamma x = h$*. Conversely, suppose* $\Gamma x = h$*. If* x' *is given by* $x'_i = x_{\lceil i/k \rceil}$ *then* $\widetilde{\Gamma}x' = \widetilde{h}$*.*

Notice that these solutions correspond to critical points of the cover and the original problem. Similarly, we can transform eigenvectors of covers to either eigenvectors of the original problem or the zero vector.

Lemma 3.3. *Suppose* $\widetilde{\Gamma}x' = \lambda x'$*. If* $x \in \mathbb{R}^n$ *is given by* $x_i = \frac{1}{k}\sum_{j=1}^k x'_{ki+j}$ *then either* $\Gamma x = \lambda x$ *or* $\Gamma x = 0$*. Conversely, suppose* $\Gamma x = \lambda x$*. If* x' *is given by* $x'_i = x_{\lceil i/k \rceil}$ *then* $\widetilde{\Gamma}x' = \lambda x'$*.*

These lemmas demonstrate that we can scale critical points and eigenvectors of covers to critical points and eigenvectors (or the zero vector) of the original problem, and we can duplicate critical points and eigenvectors of the original problem to obtain critical points and eigenvectors of covers.

4. Graph Covers and Diagonal Dominance

Example 4.1. *Consider the following pair of matrices:*

$$\Gamma = \begin{pmatrix} 1 & .6 & .6 \\ .6 & 1 & .6 \\ .6 & .6 & 1 \end{pmatrix} \quad \widetilde{\Gamma} = \begin{pmatrix} 1 & 0 & .6 & 0 & 0 & .6 \\ 0 & 1 & 0 & .6 & .6 & 0 \\ .6 & 0 & 1 & 0 & .6 & 0 \\ 0 & .6 & 0 & 1 & 0 & .6 \\ 0 & .6 & .6 & 0 & 1 & 0 \\ .6 & 0 & 0 & .6 & 0 & 1 \end{pmatrix}$$

Γ *is positive definite, but* $\widetilde{\Gamma}$ *has negative eigenvalues.*

The example illustrates that there exist positive definite matrices that are covered by matrices which are not positive definite. This observation seems to be problematic for the convergence of the min-sum algorithm. The messages passed by the min-sum algorithm are exactly the same for each variable of Γ and their copies in $\widetilde{\Gamma}$.

One reasonable conjecture might be that if we are given a Γ such that Γ and all of its covers are positive definite then the min-sum algorithm converges to the correct solution. We show that this is indeed the case by demonstrating that this condition is equivalent to the known sufficient conditions for the convergence of min-sum described in [3] and [4].

Theorem 4.2. *Let Γ be a symmetric matrix with unit diagonal. The following are equivalent:*

1. *Γ is walk-summable.*
2. *Γ is scaled diagonally dominant.*
3. *All covers of Γ are positive definite.*
4. *All 2-covers of Γ are positive definite.*

Proof.
Without loss of generality, assume that the graph corresponding to Γ is connected. If not, the quadratic minimization breaks into several smaller quadratic minimizations and we can repeat this entire argument for each of the pieces separately.

- $(1 \Rightarrow 2)$ By assumption, $|I-\Gamma|$ is irreducible. Let λ be an eigenvalue of $|I-\Gamma|$ with eigenvector $x>0$ whose existence is guaranteed by the Perron-Frobenius theorem. For any row i, we have:

  $$x_i > \lambda x_i = \sum_{j\neq i} |\Gamma_{ij}| x_j$$

  Since $\Gamma_{ii} = 1$ this is the definition of scaled diagonal dominance with $w = x$.

- $(2 \Rightarrow 3)$ If Γ is scaled diagonally dominant then so is every one of its covers. Scaled diagonal dominance implies that a matrix is symmetric positive definite. Therefore, all covers must be symmetric positive definite.

- $(3 \Rightarrow 4)$ Trivial.

- $(4 \Rightarrow 1)$ Let $\widetilde{\Gamma}$ be any **connected** 2-cover of Γ. Recall that $\widetilde{\Gamma}$ has the form of equation 1.

  By assumption, $|I-\widetilde{\Gamma}|$ and $|I-\Gamma|$ are both irreducible.

  Observe that by the Perron-Frobenius theorem there exists an eigenvector $x > 0 \in \mathbb{R}^n$ of $|I-\Gamma|$ with eigenvalue $\rho(|I-\Gamma|)$. Let $y \in \mathbb{R}^{2n}$ be constructed by duplicating the values of x so that $y_{2i} = y_{2i+1} = x_i$ for each $i \in \{0...n\}$. By Lemma 3.3, y is an eigenvector of $|I-\widetilde{\Gamma}|$ with eigenvalue equal to $\rho(|I-\Gamma|)$. By the Perron-Frobenius theorem, $|I-\widetilde{\Gamma}|$ has a unique positive eigenvector, with eigenvalue equal to the spectral radius. Because $y > 0$, we must have $\rho(|I-\Gamma|) = \rho(|I-\widetilde{\Gamma}|)$.

  We will now construct a specific cover $\widetilde{\Gamma}$ such that $\widetilde{\Gamma}$ is positive definite iff Γ is walk-summable. To do this, we'll choose the P_{ij} as in Equation 1 such that $P_{ij} = I$ if $\Gamma_{ij} < 0$ and $P_{ij} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ otherwise. Now define $z \in \mathbb{R}^{2n}$ by setting $z_i = (-1)^i c y_i$, where the constant c ensures that $\|z\| = 1$.

  Consider the following:

  $$\begin{aligned} z^T \widetilde{\Gamma} z &= \sum_i \Gamma_{ii} z_i^2 + \sum_{i=1}^{n} \sum_{j\neq i} \Gamma_{ij} [z_{2i}, z_{2i+1}] P_{ij} \begin{bmatrix} z_{2j} \\ z_{2j+1} \end{bmatrix} \\ &= 1 - 2\sum_{i>j} |\Gamma_{ij}| c^2 y_i y_j \end{aligned}$$

  Recall that y is the eigenvector of $|I-\widetilde{\Gamma}|$ corresponding to the largest eigenvalue and $\|cy\| = 1$. By definition and the above,

  $$\begin{aligned} \rho(|I-\Gamma|) &= \rho(|I-\widetilde{\Gamma}|) \\ &= \frac{cy^T |I-\widetilde{\Gamma}| cy}{c^2 y^T y} \\ &= 2\sum_{i>j} |\Gamma_{ij}| c^2 y_i y_j \end{aligned}$$

  Combining all of the above we see that $z^T \widetilde{\Gamma} z = 1 - \rho(|I-\Gamma|)$. Now, $\widetilde{\Gamma}$ positive definite implies that $z^T \widetilde{\Gamma} z > 0$, so $1 - \rho(|I-\Gamma|) > 0$. In other words, Γ is walk-summable.

□

From the theorem, we can immediately infer the following:

1. If Γ is scaled diagonally dominant then all covers of Γ are also positive definite. In this case it does not matter that min-sum cannot distinguish between Γ and its covers; for by converging to a global minimum for Γ, the algorithm also converges to a global minimum for any cover of Γ.
2. If Γ is not scaled diagonally dominant then at least one of its 2-covers has an eigenvalue that is less than or equal to zero. In this case Γ has a finite minimum, but the 2-cover does not. Therefore if the the min-sum algorithm converges, it must produce an incorrect answer on either Γ or one of its covers because min-sum cannot distinguish between the two.

This theorem has several important consequences. First, it provides us with a new characterization of scaled diagonal dominance and walk-summability, which offers an intuitive explanation for why these appear as sufficient conditions for the convergence of the min-sum algorithm. Second, although this characterization is specific to the case of quadratic minimization, we might expect similar requirements on covers to be sufficient for convergence of the min-sum algorithm in other problem domains such as convex minimization, linear programming, etc. Lastly, as scaled diagonal dominance is a well-studied property of matrices, the theorem may also be of independent interest. For example, this intuition may explain why scaled diagonal dominance is also a sufficient condition for convergence and correctness of the related Gauss-Seidel and Jacobi algorithms.

5. Weak Scaled Diagonal Dominance

We can weaken the strict inequalities in our previous definitions to obtain a weaker version of Theorem 4.2. Consider the following definitions:

Definition 5.1. *$\Gamma \in \mathbb{R}^{n\times n}$ is **weakly walk-summable** if the spectral radius $\rho(|I-\Gamma|) \leq 1$.*

Definition 5.2. *Γ is **weakly scaled diagonally dominant** if $\exists w \in \mathbb{R}^{n\times n} > 0$ such that $|\Gamma_{ii}|w_i \geq \sum_{j\neq i}|\Gamma_{ij}|w_j$.*

Theorem 5.3. *Let Γ be a symmetric matrix with unit diagonal. The following are equivalent:*

1. *Γ is weakly walk-summable.*
2. *Γ is weakly scaled diagonally dominant.*
3. *All covers of Γ are positive semi-definite.*
4. *All 2-covers of Γ are positive semi-definite.*

The proof of this theorem is almost identical to the proof of Theorem 4.2 and those details will not be repeated here.

Weak scaled diagonal dominance is a somewhat special case. If Γ is weakly scaled diagonally dominant but not scaled diagonally dominant then we know that the objective function is covered by quadratic objective functions that have an infinite number of critical points. To see this, note that the minima of the quadratic equation must satisfy $\Gamma x = h$. The two-cover $\widetilde{\Gamma}$ constructed in the proof of Theorem 4.2 has 0 as an eigenvalue, and therefore there are infinitely many vectors in the kernel of $\widetilde{\Gamma}$. Because $\Gamma x = h$ has at least one solution, $\widetilde{\Gamma} y = \widetilde{h}$ must have infinitely many.

Theorem 2.3 implies that if the min-sum algorithm converges, it must converge to a critical point. If Γ is weakly scaled diagonally dominant, we may suspect that the presence of multiple critical points on covers could present difficulties for convergence. Specifically, observe that Example 2.1 is weakly scaled diagonally dominant. In that example, both the messages and beliefs are converging, but the beliefs are converging to the constant zero function. This means that we cannot easily recover the minimum from the fixed point beliefs using Theorem 2.3. However, as we will demonstrate below, we can still extract meaningful estimates from the min-sum algorithm in this case.

6. Periodic Behavior

As mentioned earlier, we have empirically observed that when min-sum fails to converge, it exhibits only two failure modes: either the messages diverge to $\pm\infty$, or they oscillate with some period. When periodicity is encountered in the estimates, one of several approaches is typically used in an attempt to recover the correct behavior. The first approach is to damp the messages. The second approach is to average the estimates (or the beliefs). For the quadratic minimization problem, we will show that this second approach provides a correct solution. First, we need the following lemma:

Lemma 6.1. *$a^t_{ij} \leq a^{t-1}_{ij}$ for all $t \geq 1$.*

Proof. This result follows trivially by induction. Note that $\Gamma_{ii} + 2\sum_{k\in\partial i-j} a^{t-1}_{ki} > 0$ if the update rule is applied. Otherwise $a^t_{ij} = -\infty$. □

Theorem 6.2. *If the min-sum algorithm converges k-periodically to a set of messages $\widetilde{m}^1,\ldots,\widetilde{m}^k$ and a set of estimates $\widetilde{x}^1,\ldots,\widetilde{x}^k$ then the vector $x^* \in \mathbb{R}^n$ given by*

$$x^*_i = \lim_{t\to\infty}\frac{1}{k}\sum_{s=1}^{k} x^{t+s}_i = \frac{1}{k}\sum_{s=1}^{k}\widetilde{x}^s_i$$

is the unique minimizer of the objective function.

Proof. Define $m^* \equiv \frac{1}{k}\sum_{s=1}^{k}\widetilde{m}^s$. Observe that, because the a^t_{ij} are monotonic decreasing, we can infer that $T(m^*) = m^*$. So, m^* is a fixed point of the message updates. Let τ^* denote the corresponding fixed point beliefs.

Notice that $\tau^*_i(x_i)$ need not have a unique minimum. Even so, if we can demonstrate that x^* minimizes τ^*_i for each i and τ^*_{ij} for each i and j then we can guarantee that x^* is locally optimal in the sense of [6]. Let m^t be the messages at time t of the algorithm. Observe that $m^* = \lim_{t\to\infty}\frac{1}{k}\sum_{s=1}^{k} m^{kt+s}$. Similarly, all of the fixed point beliefs are (possibly degenerate) quadratic functions, and they can be written as the limit of quadratic functions with unique minima:

- $\tau^*_i(x_i) = \lim_{t\to\infty}\frac{1}{k}\sum_{s=1}^{k}\tau^{kt+s}_i(x_i)$
- $\tau^*_{ij}(x_i,x_j) = \lim_{t\to\infty}\frac{1}{k}\sum_{s=1}^{k}\tau^{kt+s}_{ij}(x_i,x_j)$

In general, one cannot exchange limits and derivatives, but for quadratic functions, we can (for a discussion of this, see Appendix A). Now, either τ^*_i is a constant function in which case x^*_i trivially minimizes it, or τ^*_i is a parabola with a unique minimum. In the latter case, we have

$$\begin{aligned}
\arg\min_{x_i}\lim_{t\to\infty}\frac{1}{k}\sum_{s=1}^{k}\tau^{kt+s}_i(x_i) &= \arg\min_{x_i}\frac{1}{k}\sum_{s=1}^{k}\lim_{t\to\infty}\tau^{kt+s}_i(x_i)\\
&= \frac{1}{k}\sum_{s=1}^{k}\arg\min_{x_i}\lim_{t\to\infty}\tau^{kt+s}_i(x_i)\\
&= \frac{1}{k}\sum_{s=1}^{k}\lim_{t\to\infty}\arg\min_{x_i}\tau^{kt+s}_i(x_i)\\
&= \frac{1}{k}\sum_{s=1}^{k}\lim_{t\to\infty}x^{kt+s}_i\\
&= \frac{1}{k}\sum_{s=1}^{k}\widetilde{x}^s_i\\
&= x^*_i
\end{aligned}$$

where the second equality follows by monotonicity of the a_{ij} and the third equality follows from Appendix A.

From the above, we can infer that x_i^* is a critical point of τ_i^*. Now, consider minimizing $\tau_{ij}^*(x_i, x_j)$. Again, either τ_{ij}^* is a constant function in which case (x_i^*, x_j^*) trivially minimizes it, or τ_{ij}^* is a quadratic function with at least one minima. Using an argument similar to the one presented above, we can show that (x_i^*, x_j^*) minimizes $\tau_i j^*$.

□

This theorem is strictly stronger than the result of Wainwright et al. [6] even if the messages approach a fixed point. Recall that in Example 2.1 the messages are converging to a set of fixed point messages, but the corresponding fixed point beliefs τ^* have $\tau_i^*(x_i) = 0$ for all i. Because the τ_i^* do not have unique minimum, we cannot apply Theorem 2.3 to obtain a local optima. However, the estimates in this instance oscillate between 0 and 1 allowing us to apply Theorem 6.2 to obtain a globally optimal solution.

We should note that Theorem 6.2 cannot be used to fix periodicity for arbitrary objective functions. There are many application areas where the averaging procedure does not result in a correct set of fixed point messages (for example see [7], [8], and [9]).

6.1. The Single Cycle Case

Because Theorem 6.2 uses a scaling operation similar to Lemma 3.2 we might suspect that there is a relationship between graph covers and periodicity. For quadratic minimization, this is indeed the case when the factor graph consists of a single cycle.

Lemma 6.3. *For the single cycle case, if the min-sum algorithm converges to a set of k-periodic messages then there is a corresponding 1-periodic solution on some k-cover of the original graph.*

Proof. Let H be a connected k-cover of our original cycle. H must be a single cycle. Starting at some node $i \in H$ and proceeding clockwise around H, place the fixed point messages in order from 1 to k, restarting the process at 1 every k nodes. Repeat the same placement technique in the counter clockwise direction. We can see that this corresponds to a 1-periodic set of messages and estimates. □

If the beliefs are not converging to a degenerate solution, Lemma 6.3 provides an alternate proof of Theorem 6.2 in the single-cycle case that does make any explicit reference to the messages themselves. The one-periodic solution on the k-cover H constructed in Lemma 6.3 defines a fixed point set of beliefs on H. As long as there is a unique x_i minimizing each belief τ_i^*, the corresponding estimates must be globally optimal on H by Theorem 2.3 and convexity. By Lemma 3.2, we can scale the solution on H down to a solution on the original graph.

6.2. 2-Periodic Solutions

As observed in the previous section, we can use graph covers to understand the behavior of min-sum in the single cycle case. For the quadratic minimization problem on general graphs the algorithm may still converge to the correct solution even if the original matrix is not weakly scaled diagonally dominant. However, even in the more general case, we can apply our observations from the single cycle case to show that 2-periodic solutions always correspond to a 1-periodic solution on some 2-cover of the original problem.

Lemma 6.4. *If the min-sum algorithm converges to a* 2*-periodic solution then there is a corresponding 1-periodic solution on some* 2*-cover of the original graph.*

Proof. Every graph G has a bipartite 2-cover (A, B). Consider messages m^1 and m^2 corresponding to consecutive time steps in the min-sum algorithm. Define the messages m for the 2-cover (A, B) as follows:

$$m_{ij} = \begin{cases} m_{ij}^1 & \text{if } i \in A \text{ and } j \in B \\ m_{ij}^2 & \text{if } i \in B \text{ and } j \in A \end{cases}$$

m is a 1-periodic solution on (A, B). □

Observe that the construction of this lemma and Lemma 6.3 apply more generally: for the min-sum algorithm on any pairwise MRF we still have that every 2-periodic solution corresponds to a 1-periodic solution on a 2-cover. For arbitrary pairwise MRFs, we have no guarantee that fixed point beliefs on the 2-cover are globally optimal. However, they remain locally optimal in the sense of [6]. In this sense, we conjecture that periodicity can always be explained by looking at covers of the original MRF.

References

[1] D. Angluin, "Local and global properties in networks of processors (extended abstract)," in *STOC '80: Proceedings of the Twelfth Annual ACM symposium on Theory of Computing*. New York, NY, USA: ACM, 1980, pp. 82–93.

[2] P. O. Vontobel and R. Koetter, "Graph-cover decoding and finite-length analysis of message-passing iterative decoding of ldpc codes," *CoRR*, vol. abs/cs/0512078, 2005.

[3] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-sums and belief propagation in gaussian graphical models," *J. Mach. Learn. Res.*, vol. 7, pp. 2031–2064, 2006.

[4] C. C. Moallemi and B. Van Roy, "Convergence of the min-sum message passing algorithm for quadratic optimization," May 2008. [Online]. Available: http://arxiv.org/abs/cs/0603058

[5] ——, "Convergence of the min-sum algorithm for convex optimization," *Forty-Fifth Annual Allerton Conference on Communication, Control, and Computing*, 2007.

[6] M. Wainwright, T. Jaakkola, and A. Willsky, "Tree consistency and bounds on the performance of the max-product algorithm and its generalizations," *Statistics and Computing*, vol. 14, no. 2, pp. 143–166, 2004.

[7] K. P. Murphy, Y. Weiss, and M. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Uncertainty in Artificial Intelligence*, 1999, pp. 467–475.

[8] S. Sanghavi, D. Shah, and A. Willsky, "Message-passing for maximum weight independent set," Jul 2008. [Online]. Available: http://arxiv.org/abs/0807.5091
[9] S. Sanghavi and D. Shah, "Tightness of lp via max-product belief propagation," 2005. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0508097

A. Uniform Convergence and Quadratics

In this section we discuss the mathematical results relating to uniform convergence. Most notably, when a sequence of functions f_n converges uniformly to a function f then f must have minima at the same place as the limit of the f_n. We make use of this observation without much discussion in Theorem 6.2.

Definition A.1. *A sequence of functions $f_n : D \to R$* ***converges uniformly*** *to $f : D \to R$ if $\forall \varepsilon > 0$ there exists an N such that for all $n > N$ and all $x \in D$, $|f_n(x) - f(x)| < \varepsilon$.*

We state the following standard results without proof:

Lemma A.2. *A sequence of functions $f_n : D \to R$ converges uniformly to $f : D \to R$ if and only if*

$$\lim_{n\to\infty} \sup_{x\in D} |f_n(x) - f(x)| = 0$$

Lemma A.3. *If a sequence of functions $f_n : D \to R$ converges uniformly to $f : D \to R$ then the*

$$\lim_{n\to\infty} \arg\min_x f_n(x) \subseteq \arg\min_x f(x)$$

For any bounded region in $D \subseteq \mathbb{R}$, if the coefficients of the messages m^t_{ij} are converging pointwise to m^*_{ij}, then m^t_{ij} converges uniformly to m^*_{ij} on D. Similarly, the derivatives of m^*_{ij} are converging uniformly to the derivatives of m^*_{ij} on D. As a result, we can use Lemma A.3 to exchange argmins and limits.