October 20, 2000

An Analysis of Belief Propagation on the Turbo Decoding Graph with Gaussian Densities

Paat Rusmevichientong and Benjamin Van Roy
Stanford University
{paatrus,bvr}@stanford.edu

ABSTRACT

Motivated by its success in decoding turbo codes, we provide an analysis of the belief propagation algorithm on the turbo decoding graph with Gaussian densities. In this context, we are able to show that, under certain conditions, the algorithm converges and that – somewhat surprisingly – though the density generated by belief propagation may differ significantly from the desired posterior density, the means of these two densities coincide.

Since computation of posterior distributions is tractable when densities are Gaussian, use of belief propagation in such a setting may appear unwarranted. Indeed, our primary motivation for studying belief propagation in this context stems from a desire to enhance our understanding of the algorithm's dynamics in non-Gaussian setting, and to gain insights into its excellent performance in turbo codes. Nevertheless, even when the densities are Gaussian, belief propagation may sometimes provide a more efficient alternative to traditional inference methods.

Key words: approximate inference, belief network, belief propagation, Gaussian densities, and turbo decoding.

1 Introduction

Probability distributions provide a tool for characterizing beliefs about unobserved quantities and relationships among them. As observations are made, beliefs change and posterior distributions evolve to reflect improved understanding. Unfortunately, the process of *inference* – that of computing posterior distributions – often entails integration over high–dimensional spaces and is typically intractable. One exception arises when densities are Gaussian. In this case, posterior distributions – which are also Gaussian – can be computed efficiently and represented compactly in terms of means and covariances.

Another case that admits efficient computation arises when conditional independencies among random variables form a convenient pattern. Belief networks and Markov random fields offer two approaches to characterizing such conditional independencies in terms of directed and undirected graphs, respectively. In either case, when the graph is singly connected (i.e., when there are no cycles), belief propagation – an efficient inference algorithm – becomes applicable [13, 22].

Many distributions of interest are not Gaussian and do not accommodate singly–connected graphs. In such cases, exact inference is typically intractable, and approximations are called for. Surprisingly, although belief propagation was developed for singly-connected graphs, it has been shown to deliver impressive performance in many applications involving graphs with cycles. A notable example of this is the turbo decoding algorithm used in turbo codes.

The turbo decoding algorithm is an approximation method that has delivered impressive performance in certain coding applications [5, 6]. The inference task originally addressed by the turbo decoding algorithm involves computing a distribution over the underlying message after receiving an encoded transmission across a noisy communication channel. The structure of the encoding scheme – which makes use of "turbo codes" – leads to efficient transmission rates, but leaves the decoder with the job of solving an intractable inference problem. The turbo decoding algorithm has proven to be an effective approximation method for this task. Because its initial development was not supported by mathematical theory, spectacular empirical success was received with surprise, excitement, and intrigue.

It turns out that the turbo decoding algorithm is equivalent to belief propagation. This connection was first noted by Frey and Kschischang [11] and McEliece [20]. In particular, McEliece, MacKay, and Cheng [19] presented an interpretation of the turbo decoding algorithm as an application of belief propagation in a graph with cycles. Since belief propagation was developed for singly–connected graphs, application in the presence of cycles – as is done in turbo decoding – was not supported by pre–existing principles.

With the excitement spawned by success of the turbo decoding algorithm came a reexamination of iterative decoding algorithms for codes on graphs [31, 32] and message passing algorithms [12]. Message passing algorithms were proposed decades earlier in the coding literature and bear similarities with the turbo decoding algorithm. Designed for decoding of low density parity check codes, message passing algorithms turned out also to correspond to belief propagation in graphs with cycles. Furthermore, a recent empirical study establishes that message–passing algorithms share the impressive performance demonstrated by turbo decoding [18, 24].

Indeed, Kschischang and Frey [14, 15] have shown that iterative decoding algorithms, belief propagation, and various message passing algorithms are unified by a single framework involving a distributed marginalization algorithm for functions characterized by factor graphs. They also show that many algorithms in artificial intelligence, signal processing, and digital communications, which were each developed independently, fit naturally into this framework. A similar unifying framework was also proposed in [3].

There are also signs of promise for belief propagation (in graphs with cycles) in inference problems beyond those arising in coding. Positive results have been generated – for example – in empirical case studies motivated by applications in image processing and medical decision making [21]. However, in some case studies, the algorithm fails, and factors influencing performance are not well–understood. Analytical work has focused on identifying suitable classes of problems and understanding why their properties foster success.

Recent analyses focusing on the context of coding [17, 23, 25] extend early work by Gallager [12] to shed light on the success of turbo decoding and message passing algorithms. In very rough terms, the thrust of this line of research involves establishing that cycles arising in relevant coding applications are "generally very long" and showing that this allows belief propagation to work "almost as well as in singly–connected graphs." Additional work specialized to the context of low density parity check codes further strengthens these results [24].

Another line of analytical work has aimed at understanding the behavior of belief propagation in general graphs with cycles. As a starting point, several researchers have studied the case involving a graph with a single cycle [2, 8, 28]. This case is not useful in its own right, since exact inference is tractable in the presence of a single cycle. However, the study of this case has lead to concise results that enhance our state of understanding. In particular, results pertaining to the case of a single cycle include:

1. Belief propagation converges to a unique stationary point.

2. If all random variables are binary–valued, the component–wise maximum likelihood estimates offered by the resulting approximation concur with true maximum likelihood values.

Unfortunately, the line of analysis employed for the case with a single cycle does not immediately extend to graphs with multiple cycles.

In this paper, we study belief propagation from a new angle by analyzing its dynamics in a restrictive setting where densities are Gaussian. We focus our attention on the case where the dependence structure of the random variables is similar to the one that appeared in the original turbo decoding application. The graph that captures this dependency will be referred to as the turbo decoding graph. In this case, exact inference is tractable and use of belief propagation is not entirely necessary. Nevertheless, belief propagation may sometimes provide a more efficient method for solving certain inference problems in this context. Our primary motivation for studying the Gaussian case, however, is to provide a setting amenable to a streamlined analysis. A clear understanding here may offer insights into behavior of belief propagation in more general settings, and possibly shed light on its success in turbo codes.

Contributions of our analysis include certain concise results concerning use of belief propagation when densities are Gaussian:

1. If belief propagation is initialized with Gaussian densities, each iterate is also Gaussian (Lemma 2).

2. The associated sequence of covariance matrices converges to a unique stationary point (Theorem 1).

3. Under certain conditions, the sequence of mean vectors also converges to a unique stationary point (Theorem 2, Proposition 1, 2, 3, 4, and 5).

4. When belief propagation converges, the mean of the resulting approximation coincides with that of the true posterior density (Theorem 3). (Note that, since the distribution is Gaussian, the mean corresponds to the maximum likelihood value, so this result parallels an aforementioned result concerning the case of a graph with a single cycle and binary variables.)

While preparing this paper, we became aware of two related initiatives, both involving analysis of belief propagation when densities are Gaussian and graphs possess cycles. Weiss and Freeman [30] were studying the case of 2-dimensional lattice. Here, they were able to show that, if belief propagation converges, the mean of the resulting approximation coincides with that of the true posterior distribution. Weiss and Freeman also derived equations characterizing dynamics of means and covariance matrices generated by belief propagation. At the same time, Frey [10] studied a case involving graphical structures that generalize those employed in turbo decoding. He derived an equation satisfied by stationary points and provided an analysis relating convergence of means to the spectral radius of a particular matrix (we will present and analyze a related matrix in Section 5.2). He also conducted an empirical study. Coincidentally, short papers describing the work of Weiss and Freeman [29] and Frey [9], as well as one summarizing results in this paper [26], were simultaneously submitted to the same conference.

The paper is organized as follow. In the next section, we provide our working definition of the belief propagation algorithm. To lend concreteness to this definition, we present an example in Section 3 of a situation where belief propagation might be more efficient than traditional inference methods. In Section 4, we discuss specialization of belief propagation to the Gaussian case. A convergence analysis is then presented in Section 5. In section 6, we prove that the mean of the approximation generated by belief propagation coincides with that of the desired posterior distribution. After presenting some experimental results, we close with a concluding section.

2 Belief Propagation on the Turbo Decoding Graph

Consider a random variable x that takes on values in $\Re^n$ and has independent components. Let p_0 denote the prior density of x. Also, let y_1 and y_2 be two random variables that are conditionally independent given x. For example, y_1 and y_2 might represent outcomes of two independent transmissions of the signal x over a memoryless communication channel. The turbo decoding graph depicting the dependence among the random variables (both in Bayesian network and factor graph representation) is given in Figure 1. Our definition of belief propagation will exploit the dependence structure of these random variables.

If y_1 and y_2 are observed, one might want to infer a posterior density f of x conditioned on y_1 and y_2. This can be obtained by first computing densities p_1^* and p_2^*, where the first

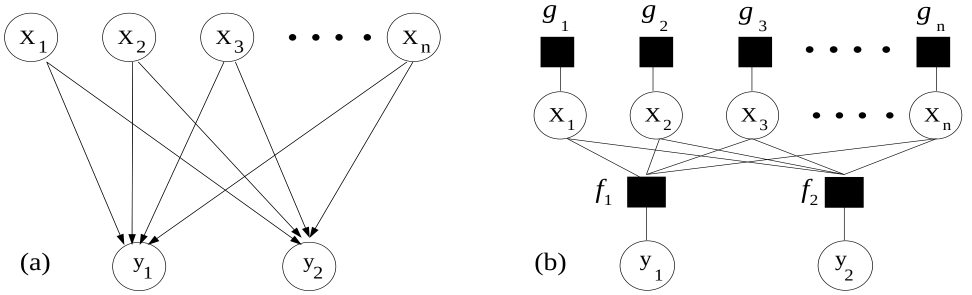

Figure 1: Turbo Decoding Graph (a) Bayesian network representation, (b) factor graph representation. In (b), the function f_1 (resp. f_2) corresponds to the conditional density of y_1 (resp. y_2) given x. The function g_i corresponds to the prior density of x_i.

is conditioned on y_1 and the second is conditioned on y_2. Then,

$$f = \alpha \left(\frac{p_1^* p_2^*}{p_0} \right),$$

where α is a "normalizing operator" defined by

$$\alpha g = \frac{g}{\int g(\overline{x}) d\overline{x}},$$

and multiplication and division are carried out pointwise.

Unfortunately, even when p_1^* and p_2^* are known, computation of f can be intractable. The burden associated with storing and manipulating high–dimensional densities appears to be the primary obstacle. This motivates the idea of limiting attention to densities that factor. In this context, it is convenient to define an operator π that generates a density that factors while possessing the same marginals as another density. In particular, this operator is defined by

$$(\pi g)(a) = \prod_{i=1}^{n} \int_{\{\bar{x} \in \Re^n | \bar{x}_i = a_i\}} g(\bar{x}) d\bar{x} \wedge d\bar{x}_i,$$

for any density g and any $a \in \Re^n$, where $d\bar{x} \wedge d\bar{x}_i = d\bar{x}_1 \cdots d\bar{x}_{i-1} d\bar{x}_{i+1} \cdots d\bar{x}_n$. One might aim at computing πf as a proxy for f. Unfortunately, even this problem can be intractable. Belief propagation can be viewed as an iterative algorithm for approximating πf.

Let operators T_1 and T_2 be defined by

$$T_1 g = \alpha \left(\left(\pi \frac{p_1^* g}{p_0} \right) \frac{p_0}{g} \right),$$

and

$$T_2 g = \alpha \left(\left(\pi \frac{g p_2^*}{p_0} \right) \frac{p_0}{g} \right),$$

for any density g. Belief propagation is applicable in cases where computation of these two operations is tractable. The algorithm generates sequences $q_1^{(k)}$ and $q_2^{(k)}$ according to

$$q_1^{(k+1)} = T_1 q_2^{(k)} \quad \text{and} \quad q_2^{(k+1)} = T_2 q_1^{(k)}.$$

initialized with densities $q_1^{(0)}$ and $q_2^{(0)}$ that factor. The hope is that $\alpha(q_1^{(k)} q_2^{(k)} / p_0)$ converges to a useful approximation of πf.

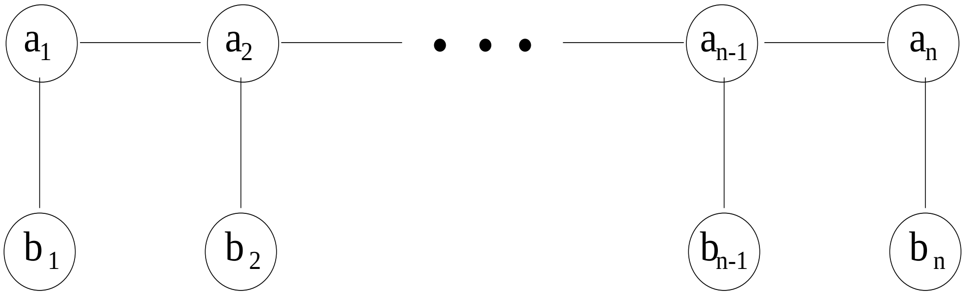

Figure 2: An example of a hidden Markov model.

3 An Example

The preceding abstract definition relied on use of operators T_1 and T_2 as subroutines. For the sake of concreteness, we will discuss in this section certain situations where computation of T_1 and T_2 is tractable. It is in such situations that belief propagation may constitute a legitimate approximation scheme.

We will describe an example in terms of Markov random fields, so let us begin by reviewing the semantics of this graphical modeling framework. A Markov random field is an undirected graph with each node corresponding to a random variable. The arcs convey information about conditional independencies. In particular, if $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$, are mutually exclusive sets of nodes and $\mathcal{C}$ separates $\mathcal{A}$ from $\mathcal{B}$, then the random variables corresponding to $\mathcal{A}$ are conditionally independent from those corresponding to $\mathcal{B}$ conditioned on those corresponding to $\mathcal{C}$. The term *separates* refers to the fact that every path from a node in $\mathcal{A}$ to a node in $\mathcal{B}$ visits at least one node in $\mathcal{C}$.

When a Markov random field is singly connected, belief propagation offers an efficient approach to inference. In particular, when some of the variables are observed, a posterior distribution over the remaining variables, and furthermore, marginal distributions over individual variables, can be efficiently computed.

One common class of Markov random fields that accommodates efficient inference is the class of hidden Markov models. Figure 2 depicts the Markov random field associated with a simple hidden Markov model. The nodes are labeled with corresponding random variables. It is easy to see that the graph is singly connected, and the common inference problem of computing a posterior distribution over $a_1, \ldots, a_n$ conditioned on $b_1, \ldots, b_n$ is efficiently solved by belief propagation.

In the presence of cycles, inference becomes more complicated and often intractable. We will now describe one class of problems for which belief propagation may constitute a useful approximation scheme. Consider two singly connected Markov random fields – $\mathcal{M}_1$ and $\mathcal{M}_2$ – each with $2n$ nodes. The nodes of $\mathcal{M}_1$ correspond to the components of two n–dimensional random vectors y_1 and z_1, while those of $\mathcal{M}_2$ correspond to y_2 and z_2. In either graph, belief propagation offers efficient inference when y_1 or y_2 is observed.

Consider now an augmented Markov random field $\mathcal{M}$ containing $5n$ nodes, corresponding to components of y_1, y_2, z_1, z_2, and another random vector x. The arcs include those connecting components of y_1 and z_1 in $\mathcal{M}_1$, as well as those connecting components of

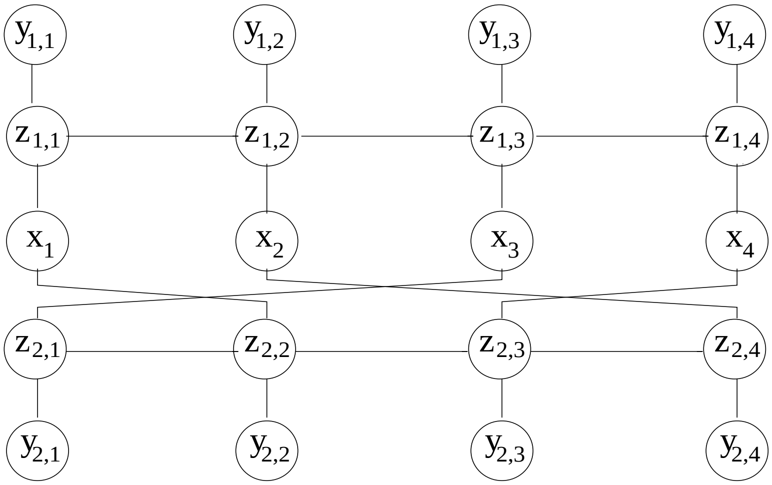

Figure 3: An example of the Markov random field $\mathcal{M}$ when $n = 4$. Note the presence of cycles.

y_2 and z_2 in $\mathcal{M}_2$. Furthermore, $2n$ additional arcs connect each component of x with components of z_1 and z_2. As illustrated by an example in Figure 3, $\mathcal{M}$ can possess cycles. In the presence of cycles, traditional exact inference method [16] requires construction of a junction tree, where nodes in the tree correspond to cliques in a triangulated graph. The resulting clique is generally very large due to the presence of cycles. Since the running time of these algorithms is exponential in the clique size, exact inference is typically infeasible in these problems.

Though the presence of cycles can render many inference tasks intractable, there are at least some forms of inference in $\mathcal{M}$ that can be performed efficiently. For example, upon observation of y_1, the posterior distribution p_1^* over x can be efficiently computed by belief propagation. This is possible because the nodes corresponding to z_2 and y_2 can be ignored, and the remaining nodes form a singly connected graph. Similarly, if y_2 is observed, the posterior distribution p_2^* over x can be efficiently inferred. However, if we observe both y_1 and y_2, inference becomes complex. In this context, belief propagation may provide a suitable approximation algorithm. Ideally, the algorithm should generate marginal distributions over individual components of x, conditioned on simultaneous observation of y_1 and y_2.

We assume that the prior distribution p_0 over x factors (i.e., $p_0 = \pi p_0$, or equivalently, the components of x are initially independent). For any density g over x, $p_1^* g/p_0$ would be the posterior density over x conditioned on y_1 if the prior density over x were g, rather than p_0. Consequently, for any density g that factors, by appropriately altering the priors on x (while keeping fixed conditional probabilities of y_1 and z_1, conditioned on x) and applying belief propagation, we can efficiently compute $\pi(p_1^* g/p_0)$. This in turn enables efficient computation of

$$T_1 g = \alpha \left(\pi \left(\frac{p_1^* g}{p_0} \right) \frac{p_0}{g} \right),$$

since pointwise multiplication and normalization are tractable for functions that factor. The operator T_2 similarly accommodates efficient computation.

In conclusion, for the Markov random field $\mathcal{M}$, there are tractable implementations of T_1 and T_2, and application of belief propagation is therefore feasible. Whether or not belief propagation will generate useful approximations, however, is a separate issue.

4 The Gaussian Case

In the remainder of this paper, we will focus on a setting in which the joint distribution of x, y_1, and y_2, is Gaussian. In this context, application of belief propagation may appear to be unwarranted – there are tractable algorithms for computing conditional distributions when priors are Gaussian. Indeed, our primary motivation is to provide a setting amenable to a streamlined analysis and concise results. It is worth noting, nevertheless, that belief propagation may provide a more efficient means than traditional algorithms for solving certain Gaussian inference problems. We will further discuss this possibility in the concluding section.

Let us define some notation that will facilitate our exposition. Let $\mathcal{D}$ denote the set of covariance matrices that are diagonal and positive definite. Let $\mathcal{G}$ denote the set of Gaussian densities with covariance matrices in $\mathcal{D}$. We will write $g \sim N(\mu_g, \Sigma_g)$ to denote a Gaussian density g with mean vector μ_g and covariance matrix Σ_g. For any matrix A, let $\delta(A)$ denote a diagonal matrix with entries equal to the diagonal elements of A. Hence,

$$\pi g \sim N(\mu_g, \delta(\Sigma_g)),$$

for any Gaussian density $g \sim N(\mu_g, \Sigma_g)$. For any diagonal matrices D and $\overline{D}$, we write $D \leq \overline{D}$ if $D_{ii} \leq \overline{D}_{ii}$ for all i and $D < \overline{D}$ if $D_{ii} < \overline{D}_{ii}$ for all i. For any two pairs of diagonal matrices (C, D) and $(\overline{C}, \overline{D})$, we write $(C, D) \leq (\overline{C}, \overline{D})$ if $C \leq \overline{C}$ and $D \leq \overline{D}$. Similarly, we write $(C, D) < (\overline{C}, \overline{D})$ if $C < \overline{C}$ and $D < \overline{D}$. For any matrices Σ_u, Σ_v for which $\Sigma_u^{-1} + \Sigma_v^{-1} - I$ is nonsingular, we define a matrix A_{Σ_u,Σ_v} by

$$A_{\Sigma_u,\Sigma_v} = (\Sigma_u^{-1} + \Sigma_v^{-1} - I)^{-1}.$$

To abbreviate, we will sometimes denote this matrix by A_{uv}. Finally, all vectors are assumed to be column vectors unless explicitly stated otherwise.

When the random variables x, y_1, and y_2, are jointly Gaussian, the densities p_1^*, p_2^*, f, and p_0, are also Gaussian. We define μ, μ_1, μ_2, Σ, Σ_1, and Σ_2, to be means and covariance matrices satisfying

$$p_1^* \sim N(\mu_1, \Sigma_1), \quad p_2^* \sim N(\mu_2, \Sigma_2), \quad f \sim N(\mu, \Sigma).$$

We make the following assumptions concerning these parameters.

Assumption 1
(a) *$p_0 = N(0, I)$, where I is the identity matrix.*
(b) *$\Sigma_1^{-1} - I$ and $\Sigma_2^{-1} - I$ are positive definite.*
(c) *Σ_1 and Σ_2 are positive definite.*

The first assumption simplifies the exposition at no sacrifice of generality. Any problem with a nondegenerate Gaussian prior on x can be transformed to meet this requirement by appropriate translation and scaling of the coordinate system. The second assumption implies that the observations y_1 and y_2 each provide at least some information pertinent to every component of x. The final assumption, on the other hand, requires that neither observation rules out possible outcomes – every outcome for x is possible both before and after an observation, though the prior and posterior probabilities may differ substantially.

Since $f = \alpha(p_1^* p_2^* / p_0)$, its mean μ and covariance matrix Σ are determined by those of p_1^*, p_2^*, and p_0. The nature of this dependence is identified by the following lemma, which will be reused for various purposes in subsequent sections.

Lemma 1 *Let $u \sim N(\mu_u, \Sigma_u)$ and $v \sim N(\mu_v, \Sigma_v)$, where Σ_u and Σ_v are positive definite. If $\Sigma_u^{-1} + \Sigma_v^{-1} - I$ is positive definite, then*

$$\alpha\left(\frac{uv}{p_0}\right) \sim N\left(A_{uv}\left(\Sigma_u^{-1}\mu_u + \Sigma_v^{-1}\mu_v\right), A_{uv}\right).$$

This result follows from simple algebra, and we omit the proof. The implications with respect to μ and Σ are, of course, that

$$\mu = A_{\Sigma_1,\Sigma_2}\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right) \quad \text{and} \quad \Sigma = A_{\Sigma_1,\Sigma_2}.$$

It turns out that, if initialized with Gaussian densities $q_1^{(0)}, q_2^{(0)} \in \mathcal{G}$, all iterates $q_1^{(k)}$ and $q_2^{(k)}$ generated by belief propagation are also in $\mathcal{G}$. This fact simplifies analysis of the algorithm's dynamics – we need only attend to sequences of means and covariance matrices. In particular, we can define sequences $m_1^{(k)}$, $m_2^{(k)}$, $C_1^{(k)}$, and $C_2^{(k)}$, such that

$$q_1^{(k)} \sim N\left(m_1^{(k)}, C_1^{(k)}\right) \quad \textit{and} \quad q_2^{(k)} \sim N\left(m_2^{(k)}, C_2^{(k)}\right).$$

The fact that iterates remain Gaussian is a consequence of the following lemma, the proof of which is provided in Appendix B.

Lemma 2 *The set $\mathcal{G}$ is closed under T_1 and T_2.*

It follows from this lemma that, over the domain $\mathcal{G}$, the mappings T_1 and T_2, which act on densities, can be represented in terms of operations on mean vectors and covariance matrices. We will provide characterizations of these operations in the form of a lemma. For a concise statement of the lemma, let us define some notation. For any $D \in \mathcal{D}$, let functions $\mathcal{F}_1$ and $\mathcal{F}_2$ be defined by

$$\mathcal{F}_1(D) = \left(\left(\delta\left(A_{\Sigma_1,D}\right)\right)^{-1} + I - D^{-1}\right)^{-1},$$

and

$$\mathcal{F}_2(D) = \left(\left(\delta\left(A_{D,\Sigma_2}\right)\right)^{-1} + I - D^{-1}\right)^{-1}.$$

Furthermore, for any $m \in \Re^n$ and any $D \in \mathcal{D}$, let functions $\mathcal{H}_1$ and $\mathcal{H}_2$ be defined by

$$\mathcal{H}_1(m, D) = \mathcal{F}_1(D)\left(A_{\mathcal{F}_1(D),D}^{-1} A_{\Sigma_1,D} - I\right) D^{-1} m + \mathcal{F}_1(D) A_{\mathcal{F}_1(D),D}^{-1} A_{\Sigma_1,D} \Sigma_1^{-1}\mu_1,$$

and

$$\mathcal{H}_2(m, D) = \mathcal{F}_2(D)\left(A_{D,\mathcal{F}_2(D)}^{-1} A_{D,\Sigma_2} - I\right) D^{-1} m + \mathcal{F}_2(D) A_{D,\mathcal{F}_2(D)}^{-1} A_{D,\Sigma_2} \Sigma_2^{-1}\mu_2.$$

The lemma follows.

Lemma 3 *For all $g \in \mathcal{G}$, if $g \sim N(\mu_g, D_g)$ then*

$$T_1 g \sim N(\mathcal{H}_1(\mu_g, D_g), \mathcal{F}_1(D_g)) \quad \text{and} \quad T_2 g \sim N(\mathcal{H}_2(\mu_g, D_g), \mathcal{F}_2(D_g)).$$

Given this lemma, dynamics of belief propagation can be characterized by

$$C_1^{(k+1)} = \mathcal{F}_1\left(C_2^{(k)}\right) \quad \text{and} \quad C_2^{(k+1)} = \mathcal{F}_2\left(C_1^{(k)}\right),$$

and

$$m_1^{(k+1)} = \mathcal{H}_1\left(m_2^{(k)}, C_2^{(k)}\right) \quad \text{and} \quad m_2^{(k+1)} = \mathcal{H}_2\left(m_1^{(k)}, C_1^{(k)}\right).$$

Once again, we postpone the proof of this lemma to Appendix B.

5 Convergence Analysis

Two immediate consequences of Lemma 3 guide the general structure of our convergence analysis. The first is that covariance matrices generated by belief propagation evolve independently from mean vectors. This fact leads us to begin by studying the dynamics of covariance matrices without paying any attention to that of the mean vectors.

We will show that each sequence of covariance matrices converges to a unique stationary point. Denoting the stationary points by C_1^* and C_2^*, this allows us to approximate the dynamics of the means for large k by

$$m_1^{(k+1)} = \mathcal{H}_1\left(m_2^{(k)}, C_2^*\right) \quad \text{and} \quad m_2^{(k+1)} = \mathcal{H}_2\left(m_1^{(k)}, C_1^*\right).$$

A second consequence of Lemma 3 – that the functions $\mathcal{H}_1$ and $\mathcal{H}_2$ are affine in their first arguments – then renders the convergence analysis for $m_1^{(k)}$ and $m_2^{(k)}$ amenable to the tools of linear systems. Unfortunately, unlike the sequences of covariance matrices, the sequences of means do not always converge. We will, however, provide conditions under which convergence is guaranteed.

Stability of a particular matrix constitutes a sufficient condition for global convergence of the mean vectors. We will show that the set of Σ_1 and Σ_2 that lead to stability of this matrix is invariant under a certain type of transformation. In addition, to facilitate understanding, we will provide simpler conditions under which the matrix is stable. As a preview, let us state – in rough terms – three such conditions, each of which ensures convergence:

1. Σ_1 and Σ_2 are "complementary." (Proposition 2)

2. Either Σ_1 or Σ_2 is diagonal or "nearly diagonal." (Proposition 3 and 4)

3. Σ_1 and Σ_2 are "well-conditioned." In other words, for each matrix, the ratio of the largest to the smallest eigenvalue is not large. (Proposition 5)

In analyzing each of the above conditions, we will use a customized argument. A unified approach that offers interpretable means to distinguishing convergent cases from those that are not would be desirable, but finding such an approach remains an open problem.

Let us now move on to formal statements of our results and the corresponding analyses. The following subsection addresses convergence of the sequences of covariance matrices, while dynamics of the mean vectors are treated in Section 5.2.

5.1 Convergence of the Covariance Matrices

Defining $\mathcal{F}$ by

$$\mathcal{F}(D_1, D_2) = (\mathcal{F}_1(D_2), \mathcal{F}_2(D_1)),$$

for all $D_1, D_2 \in \mathcal{D}$, it is clear from Lemma 3 that

$$\left(C_1^{(k)}, C_2^{(k)}\right) = \mathcal{F}^k\left(C_1^{(0)}, C_2^{(0)}\right).$$

The following theorem establishes that such a sequence converges to a point that is independent of the initial iterate.

Theorem 1 *The operator $\mathcal{F}$ possesses a unique fixed point in $\mathcal{D}\times\mathcal{D}$. Furthermore, denoting this fixed point by (C_1^*, C_2^*),*

$$\lim_{k\to\infty} \mathcal{F}^k(D_1, D_2) = (C_1^*, C_2^*),$$

for all $D_1, D_2 \in \mathcal{D}$.

Since the operator $\mathcal{F}$ is uniquely determined by Σ_1 and Σ_2, it follows from Theorem 1 that the unique fixed point (C_1^*, C_2^*) is completely determined by Σ_1 and Σ_2, which are the covariance matrices of the conditional densities p_1^* and p_2^*, respectively. For ease of exposition, we do not make explicit the dependence of C_1^* and C_2^* on Σ_1 and Σ_2. The proof of Theorem 1 relies on the following lemma. The first lemma captures the essential properties of the operator $\mathcal{F}$. The proof of this result is given in Appendix C

Lemma 4
(a) *Continuity: The function $\mathcal{F}$ is continuous on $\mathcal{D}\times\mathcal{D}$.*
(b) *Monotonicity: For all $X_1, X_2, Y_1, Y_2 \in \mathcal{D}$, if $(X_1, X_2) \le (Y_1, Y_2)$, then*

$$\mathcal{F}(X_1, X_2) \le \mathcal{F}(Y_1, Y_2).$$

(c) *Boundedness: There exist matrices $\overline{D}_1, \overline{D}_2 \in \mathcal{D}$ such that for all $D_1, D_2 \in \mathcal{D}$,*

$$\left(\overline{D}_1, \overline{D}_2\right) \le \mathcal{F}(D_1, D_2) < (I, I).$$

(d) *Scaling: For all $\beta \in (0,1)$ and $D_1, D_2 \in \mathcal{D}$,*

$$\beta\mathcal{F}(D_1, D_2) < \mathcal{F}(\beta D_1, \beta D_2).$$

The following lemma establishes convergence when the sequence of covariance matrices is initialized with the identity matrix.

Lemma 5 *The sequence $\mathcal{F}^k(I, I)$ converges in $\mathcal{D}\times\mathcal{D}$ to a fixed point of $\mathcal{F}$.*

Proof: By Lemma 4(c), $\mathcal{F}(I,I) < (I,I)$. It then follows from monotonicity (Lemma 4(b)) that $\mathcal{F}^{k+1}(I,I) \le \mathcal{F}^k(I,I)$. Because $\mathcal{F}^k(I,I)$ is bounded below by a pair of matrices in $\mathcal{D}$ (Lemma 4(c)), the sequence must converge in $\mathcal{D}\times\mathcal{D}$. Furthermore, because $\mathcal{F}$ is continuous on $\mathcal{D}\times\mathcal{D}$ (Lemma 4(a)), the limit $\lim_{k\to\infty}\mathcal{F}^k(I,I)$ must be a fixed point of $\mathcal{F}$. ■

Let $(C_1^*, C_2^*) = \lim_{k\to\infty}\mathcal{F}^k(I,I)$. (By Lemma 5, the limit exists and $C_1^*, C_2^* \in \mathcal{D}$.) The following lemma establishes this as the unique fixed point in $\mathcal{D}\times\mathcal{D}$.

Lemma 6 *(C_1^*, C_2^*) is the unique fixed point in $\mathcal{D} \times \mathcal{D}$ of $\mathcal{F}$.*

Proof: By Lemma 5, (C_1^*, C_2^*) is a fixed point of $\mathcal{F}$. Let $(D_1, D_2) \in \mathcal{D} \times \mathcal{D}$ be a different fixed point. It follows from Lemma 4(c) that $(D_1, D_2) \leq (I, I)$. By monotonicity (Lemma 4(b)),

$$(D_1, D_2) = \mathcal{F}^k (D_1, D_2) \leq \mathcal{F}^k (I, I),$$

for all k. Hence, $(D_1, D_2) \leq (C_1^*, C_2^*)$.

Let

$$\beta = \sup \left\{ \gamma \in (0, 1] \middle| (\gamma C_1^*, \gamma C_2^*) \leq (D_1, D_2) \right\}.$$

Note that β is well–defined because D_1 and D_2 are positive definite. Furthermore, since $(D_1, D_2) \neq (C_1^*, C_2^*)$, we have $\beta < 1$. It follows from Lemma 4(d) that

$$\beta (C_1^*, C_2^*) = \beta \mathcal{F} (C_1^*, C_2^*) < \mathcal{F} (\beta C_1^*, \beta C_2^*).$$

In addition, due to monotonicity of $\mathcal{F}$ (Lemma 4(b)),

$$\mathcal{F} (\beta C_1^*, \beta C_2^*) \leq \mathcal{F} (D_1, D_2) = (D_1, D_2).$$

Hence,

$$(\beta C_1^*, \beta C_2^*) < (D_1, D_2),$$

which implies existence of some $\alpha > 0$ such that

$$(\alpha + \beta) (C_1^*, C_2^*) \leq (D_1, D_2).$$

However, this contradicts the definition of β. It follows that (C_1^*, C_2^*) is the unique fixed point of $\mathcal{F}$ in $\mathcal{D} \times \mathcal{D}$. ■

5.1.1 Proof of Theorem 1

Lemma 6 established uniqueness of a fixed point (C_1^*, C_2^*) and Lemma 5 asserts that $\mathcal{F}^k(I, I)$ converges to this fixed point. To complete the proof of Theorem 1, we need to show that $\mathcal{F}^k(D_1, D_2)$ converges for all $D_1, D_2 \in \mathcal{D}$, not only $D_1 = D_2 = I$.

If $(C_1^*, C_2^*) \leq (D_1, D_2) \leq (I, I)$ convergence to (C_1^*, C_2^*) follows from monotonicity (Lemma 4(b)). In particular, $(C_1^*, C_2^*) = \mathcal{F}^k(C_1^*, C_2^*) \leq \mathcal{F}^k(D_1, D_2) \leq \mathcal{F}^k(I, I)$, and since $\mathcal{F}^k(I, I)$ converges to (C_1^*, C_2^*), so must $\mathcal{F}^k(D_1, D_2)$.

For the more general case of $(D_1, D_2) \geq (C_1^*, C_2^*)$, convergence follows from the fact that $(C_1^*, C_2^*) = \mathcal{F}(C_1^*, C_2^*) \leq \mathcal{F}(D_1, D_2) < (I, I)$ (a consequence of Lemmas 4(b) and 4(c)). Considering $\mathcal{F}(D_1, D_2)$ as a starting point for the sequence leads to the preceding case for which we have already established convergence.

Let us now address the case of $(D_1, D_2) \leq (C_1^*, C_2^*)$. Let

$$\beta = \sup \left\{ \gamma \in (0, 1] \middle| (\gamma C_1^*, \gamma C_2^*) \leq (D_1, D_2) \right\}.$$

By Lemma 4(d),

$$(\beta C_1^*, \beta C_2^*) \leq \mathcal{F} (\beta C_1^*, \beta C_2^*).$$

It follows from monotonicity (Lemma 4(b)) that,

$$\mathcal{F}^k (\beta C_1^*, \beta C_2^*) \leq \mathcal{F}^{k+1} (\beta C_1^*, \beta C_2^*) \leq \mathcal{F}^{k+1} (D_1, D_2) \leq (C_1^*, C_2^*)$$

for all k. Hence, $\mathcal{F}^k(\beta C_1^*, \beta C_2^*)$ converges in $\mathcal{D} \times \mathcal{D}$, and since $\mathcal{F}$ is continuous, the limit must be a fixed point. Uniqueness of the fixed point (C_1^*, C_2^*) makes it the only viable limit. Since

$$(\beta C_1^*, \beta C_2^*) \leq (D_1, D_2) \leq (C_1^*, C_2^*),$$

$\mathcal{F}^k(D_1, D_2)$ must also converge to (C_1^*, C_2^*) by monotonicity.

To complete the proof, we consider the case of an arbitrary pair $D_1, D_2 \in \mathcal{D}$. For this case, there exist matrices $\underline{D}, \overline{D} \in \mathcal{D}$ such that $\underline{D} \leq C_i^* \leq \overline{D}$ and $\underline{D} \leq D_i \leq \overline{D}$ for $i = 1, 2$. By monotonicity,

$$\mathcal{F}^k(\underline{D}, \underline{D}) \leq \mathcal{F}^k(D_1, D_2) \leq \mathcal{F}^k\left(\overline{D}, \overline{D}\right).$$

Our previous arguments establish that $\mathcal{F}^k(\underline{D}, \underline{D})$ and $\mathcal{F}^k(\overline{D}, \overline{D})$ both converge to (C_1^*, C_2^*), and consequently $\mathcal{F}^k(D_1, D_2)$ must also converge to (C_1^*, C_2^*). ■

5.2 Convergence of the Mean Vectors

Unlike the sequences of covariance matrices, the sequences of mean vectors do not always converge. In this section, we establish sufficient conditions that ensure convergence. We will first show that convergence is guaranteed by the stability of a certain matrix $\mathbf{T}_{\Sigma_1,\Sigma_2}$, defined by

$$\mathbf{T}_{\Sigma_1,\Sigma_2} = \begin{pmatrix} 0 & A^{-1}_{C_1^*,C_2^*} A_{\Sigma_1,C_2^*} - I \\ A^{-1}_{C_1^*,C_2^*} A_{C_1^*,\Sigma_2} - I & 0 \end{pmatrix}.$$

Unfortunately, this matrix and the factors influencing its stability are difficult to interpret. Consequently, the remainder of this section will be devoted to understanding properties of those Σ_1 and Σ_2 that give rise to a stable $\mathbf{T}_{\Sigma_1,\Sigma_2}$ and to establishing interpretable conditions that ensure stability, and thus, convergence of the mean vectors.

Let us begin by stating and proving the result linking convergence to stability of $\mathbf{T}_{\Sigma_1,\Sigma_2}$. For the purpose of this theorem as well as the associated analysis, we will denote the spectral radius of any matrix A by $\rho(A)$.

Theorem 2 *If $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$, then there exist vectors m_1^* and m_2^* such that, for any $m_1^{(0)}, m_2^{(0)}$ and any $C_1^{(0)}, C_2^{(0)} \in \mathcal{D}$, the sequence $(m_1^{(k)}, m_2^{(k)})$ converges to (m_1^*, m_2^*).*

This theorem provides a sufficient and "almost necessary" condition for convergence. However, because the matrix $\mathbf{T}_{\Sigma_1,\Sigma_2}$ is difficult to interpret, this condition offers little insight into factors influencing convergence. After proving Theorem 2, we will provide in subsequent subsections more interpretable conditions under which $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$. Let us now move on to prove Theorem 2. We will rely on a lemma that is somewhat standard in flavor. We state the result here and provide its proof in Appendix D.

Lemma 7 *Let $\{A_k\}$ be a sequence of matrices that converges to A, and let $\{b_k\}$ be a sequence of vectors that converges to b. Consider a sequence of vectors $\{x_k\}$ with*

$$x_{k+1} = A_k x_k + b_k,$$

for all $k \geq 0$. If $\rho(A) < 1$, then there exists a vector x^* such that the sequence $\{x_k\}$ converges to x^* for any x_0.

Proof of Theorem 2

Recall from Lemma 3 that the mean vectors evolve according to

$$m_1^{(k+1)} = \mathcal{H}_1\left(m_2^{(k)}, C_2^{(k)}\right) \quad \text{and} \quad m_2^{(k+1)} = \mathcal{H}_2\left(m_1^{(k)}, C_1^{(k)}\right),$$

which we can rewrite as

$$m_1^{(k+1)} = C_1^{(k+1)}\left(A^{-1}_{C_1^{(k+1)},C_2^{(k)}} A_{\Sigma_1,C_2^{(k)}} - I\right)\left(C_2^{(k)}\right)^{-1} m_2^{(k)} + C_1^{(k+1)} A^{-1}_{C_1^{(k+1)},C_2^{(k)}} A_{\Sigma_1,C_2^{(k)}} \Sigma_1^{-1}\mu_1,$$

and

$$m_2^{(k+1)} = C_2^{(k+1)}\left(A^{-1}_{C_1^{(k)},C_2^{(k+1)}} A_{C_1^{(k)},\Sigma_2} - I\right)\left(C_1^{(k)}\right)^{-1} m_1^{(k)} + C_2^{(k+1)} A^{-1}_{C_1^{(k)},C_2^{(k+1)}} A_{C_1^{(k)},\Sigma_2} \Sigma_2^{-1}\mu_2.$$

To highlight the relation between these dynamics and those addressed by Lemma 7, let us introduce some additional notation. For each k, let $\mathbf{C_k}$, $\mathbf{R_k}$, and $\mathbf{T_k}$ be defined by

$$\mathbf{C_k} = \left(\begin{array}{cc} C_1^{(k)} & 0 \\ 0 & C_2^{(k)} \end{array}\right), \quad \mathbf{R_k} = \left(\begin{array}{cc} A^{-1}_{C_1^{(k+1)},C_2^{(k)}} A_{\Sigma_1,C_2^{(k)}} & 0 \\ 0 & A^{-1}_{C_1^{(k)},C_2^{(k+1)}} A_{C_1^{(k)},\Sigma_2} \end{array}\right),$$

and

$$\mathbf{T_k} = \left(\begin{array}{cc} 0 & A^{-1}_{C_1^{(k+1)},C_2^{(k)}} A_{\Sigma_1,C_2^{(k)}} - I \\ A^{-1}_{C_1^{(k)},C_2^{(k+1)}} A_{C_1^{(k)},\Sigma_2} - I & 0 \end{array}\right).$$

We then have

$$\left(\begin{array}{c} m_1^{(k+1)} \\ m_2^{(k+1)} \end{array}\right) = \mathbf{C_{k+1}}\,\mathbf{T_k}\,\mathbf{C_k}^{-1}\left(\begin{array}{c} m_1^{(k)} \\ m_2^{(k)} \end{array}\right) + \mathbf{C_{k+1}}\,\mathbf{R_k}\left(\begin{array}{c} \Sigma_1^{-1}\mu_1 \\ \Sigma_2^{-1}\mu_2 \end{array}\right).$$

Theorem 1, asserts that $(C_1^{(k)}, C_2^{(k)})$ converges to (C_1^*, C_2^*). It follows that the matrices $\mathbf{C_{k+1}}$, $\mathbf{T_k}$, $\mathbf{C_k}^{-1}$, and $\mathbf{R_k}$, converge. Furthermore, the limit of convergence of $\mathbf{C_{k+1}}\,\mathbf{T_k}\,\mathbf{C_k}^{-1}$ is given by

$$\left(\begin{array}{cc} C_1^* & 0 \\ 0 & C_2^* \end{array}\right)\mathbf{T}_{\Sigma_1,\Sigma_2}\left(\begin{array}{cc} C_1^* & 0 \\ 0 & C_2^* \end{array}\right)^{-1}.$$

Since $\rho(A) = \rho(MAM^{-1})$ for any matrix A and nonsingular matrix M, we have

$$\rho\left(\left(\begin{array}{cc} C_1^* & 0 \\ 0 & C_2^* \end{array}\right)\mathbf{T}_{\Sigma_1,\Sigma_2}\left(\begin{array}{cc} C_1^* & 0 \\ 0 & C_2^* \end{array}\right)^{-1}\right) = \rho(\mathbf{T}_{\Sigma_1,\Sigma_2}).$$

The result therefore follows from Lemma 7. ■

5.2.1 Region of Convergence

We know from Theorem 2 that a sufficient (and almost necessary) condition for convergence of the mean vectors is $\rho\left(\mathbf{T}_{\Sigma_1,\Sigma_2}\right) < 1$. Let $\mathcal{C}$ denote the set of (Σ_1, Σ_2) satisfying Assumption 1 such that $\rho\left(\mathbf{T}_{\Sigma_1,\Sigma_2}\right) < 1$. Thus, $\mathcal{C}$ can be interpreted as the region in the space of symmetric positive definite matrices where belief propagation converges. In this section, we will show that $\mathcal{C}$ is invariant under a certain type of transformation. This result will provide us with some information on the shape of $\mathcal{C}$. In the next section, we will demonstrate that certain classes of "well-behaved" symmetric positive definite matrices belong to $\mathcal{C}$.

Before we proceed to the main result of this section, let us introduce some notation. For any symmetric matrix A, let $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the smallest and largest eigenvalues of A, respectively.

Proposition 1 *Let*

$$\Sigma_1^\beta = \left(\beta\Sigma_1^{-1} + (1-\beta)\frac{I}{2}\right)^{-1}, \quad \text{and} \quad \Sigma_2^\beta = \left(\beta\Sigma_2^{-1} + (1-\beta)\frac{I}{2}\right)^{-1}, \quad \beta \geq 1.$$

If $(\Sigma_1, \Sigma_2) \in \mathcal{C}$*, then* $\left(\Sigma_1^\beta, \Sigma_2^\beta\right) \in \mathcal{C}$ *for all* $\beta \geq 1$*.*

Proof: It is not hard to verify that Σ_1^β and Σ_2^β satisfy Assumption 1. Let (C_1^*, C_2^*) denote the unique fixed point of the sequence of covariance matrices generated by belief propagation when the covariance matrices of p_1^* and p_2^* are Σ_1 and Σ_2, respectively. Also, let

$$C_1^\beta = \left(\beta\left(C_1^*\right)^{-1} + (1-\beta)\frac{I}{2}\right)^{-1}, \quad \text{and} \quad C_2^\beta = \left(\beta\left(C_2^*\right)^{-1} + (1-\beta)\frac{I}{2}\right)^{-1}.$$

It follows from the definition of $\left(\Sigma_1^\beta, \Sigma_2^\beta\right)$ and $\left(C_1^\beta, C_2^\beta\right)$ that

$$A_{\Sigma_1^\beta, C_2^\beta} = \frac{1}{\beta}A_{\Sigma_1, C_2^*}, \quad A_{C_1^\beta, \Sigma_2^\beta} = \frac{1}{\beta}A_{C_1^*, \Sigma_2}, \quad \text{and} \quad A_{C_1^\beta, C_2^\beta} = \frac{1}{\beta}A_{C_1^*, C_2^*}.$$

Since (C_1^*, C_2^*) is the unique fixed point of $\mathcal{F}$ (Theorem 1), it follows that

$$C_1^* = \left(\left(\delta\left(A_{\Sigma_1, C_2^*}\right)\right)^{-1} + I - \left(C_2^*\right)^{-1}\right)^{-1},$$

or equivalently

$$A_{C_1^*, C_2^*} = \left(\left(C_1^*\right)^{-1} + \left(C_2^*\right)^{-1} - I\right)^{-1} = \delta\left(A_{\Sigma_1, C_2^*}\right).$$

Thus,

$$A_{C_1^\beta, C_2^\beta} = \frac{1}{\beta}A_{C_1^*, C_2^*} = \frac{1}{\beta}\delta\left(A_{\Sigma_1, C_2^*}\right) = \frac{1}{\beta}\delta\left(\beta A_{\Sigma_1^\beta, C_2^\beta}\right) = \delta\left(A_{\Sigma_1^\beta, C_2^\beta}\right).$$

Hence,

$$C_1^\beta = \left(\left(\delta\left(A_{\Sigma_1^\beta, C_2^\beta}\right)\right)^{-1} + I - \left(C_2^\beta\right)^{-1}\right)^{-1}.$$

A similar argument shows that

$$C_2^\beta = \left(\left(\delta\left(A_{C_1^\beta, \Sigma_2^\beta}\right)\right)^{-1} + I - \left(C_1^\beta\right)^{-1}\right)^{-1}.$$

It follows from Theorem 1 that $\left(C_1^\beta, C_2^\beta\right)$ is the unique fixed point of the sequence of covariance matrices generated by belief propagation when the covariance matrices of p_1^* and

p_2^* are Σ_1^β and Σ_2^β, respectively. Therefore,

$$\begin{aligned}\mathbf{T}_{\Sigma_1^\beta,\Sigma_2^\beta} &= \begin{pmatrix} 0 & A^{-1}_{C_1^\beta,C_2^\beta}A_{\Sigma_1^\beta,C_2^\beta} - I \\ A^{-1}_{C_1^\beta,C_2^\beta}A_{C_1^\beta,\Sigma_2^\beta} - I & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*} - I \\ A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2} - I & 0 \end{pmatrix} \\ &= \mathbf{T}_{\Sigma_1,\Sigma_2}\end{aligned}$$

The desired result follows. ■

The previous proposition provides us with some information on the shape of $\mathcal{C}$. If we let $\mathcal{C}^{-1} = \left\{\left(\Sigma_1^{-1},\Sigma_2^{-1}\right) : (\Sigma_1,\Sigma_2) \in \mathcal{C}\right\}$ be a collection of the inverses of covariance matrices in $\mathcal{C}$, the previous proposition suggests that there is an open set centered at $\left(\frac{I}{2},\frac{I}{2}\right)$ such that $\mathcal{C}^{-1}$ consists of rays emanating from the boundary of this set. Consequently, we conjecture that the region of convergence $\mathcal{C}$ should be star-shaped with a center at the origin $(0,0)$. Currently, we do not have a formal proof of this result, but we plan to pursue this in our future work. Also, this result appears to resemble a result reported in [23] (Theorem 6.2) on stability of fixed points of general turbo decoding.

5.2.2 Sufficient Conditions for $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$

In the previous section, we showed that covariance matrices Σ_1 and Σ_2 that lead to convergence of the mean vectors are invariant under a certain type of transformation. This result provides us with information on the shape of the region of convergence. Unfortunately, it does not help us in determining if a particular pair of covariance matrices (Σ_1,Σ_2) will lead to a convergent sequence of mean vectors. In this section, we offer four sufficient conditions that ensure stability of the matrix $\mathbf{T}_{\Sigma_1,\Sigma_2}$, and thus, convergence of the mean vectors. Since the proofs of these conditions are quite complicated, we defer them to the appendices. We will instead focus on the insights derived from each of these conditions. The first condition is expressed in the following proposition, whose proof is given in Appendix E.

Proposition 2 *For any symmetric positive definite matrix Σ such that $\Sigma^{-1} - I$ is positive definite, if*

$$\Sigma_1 = \left(\Sigma^{-1} + \gamma I\right)^{-1}, \quad \text{and} \quad \Sigma_2 = \left(\Sigma^{-1} - \gamma I\right)^{-1},$$

then $(\Sigma_1,\Sigma_2) \in \mathcal{C}$ for all $-\frac{1-\lambda_{max}(\Sigma)}{\lambda_{max}(\Sigma)} < \gamma < \frac{1-\lambda_{max}(\Sigma)}{\lambda_{max}(\Sigma)}$.

We should note that since $\Sigma^{-1} - I$ is positive definite, all eigenvalues of Σ are less than one. This implies that the range of allowable γ's in Proposition 2 includes zero. So, if $\mathcal{S} = \{(\Sigma_1,\Sigma_2) : \Sigma_1 = \Sigma_2\}$, it follows that there is an open set $\mathcal{U}$ containing $\mathcal{S}$ such that $\rho\left(\mathbf{T}_{\Sigma_1,\Sigma_2}\right) < 1$ for all $(\Sigma_1,\Sigma_2) \in \mathcal{U}$. Thus, whenever Σ_1 and Σ_2 are equal or "close", the mean vectors converge.

In general, Proposition 2 shows that the mean vectors will converge if the covariance matrices Σ_1 and Σ_2 are "complementary" in the sense that the total variance, as measured by $\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}$, is not too large. The degree of "complementarity" between Σ_1 and Σ_2 is captured by the parameter γ. As γ increases, the variance of Σ_1 decreases (relative to

Σ) while that of Σ_2 increases. This might correspond to the situation in which additional errors are introduced, resulting in greater uncertainty over the expected value of x given y_2 (thus, the increase in the variance of Σ_2). Proposition 2 tells us that belief propagation still converges, provided that there is a corresponding increase in the precision associated with the estimate of the expected value of x given y_1 (i.e., a decrease in the variance of Σ_1). Furthermore, if we start with a fairly certain estimate of x ($\lambda_{max}(\Sigma) \approx 0$), we see that belief propagation would still converge despite a wide range of variation in the covariance matrices, since the range of γ is inverse proportional to the largest eigenvalue of Σ.

Whether or not we are dealing with Gaussians, when the components of x conditioned on y_1 are independent – or equivalently, p_1^* factors – it is easy to see that belief propagation converges to πf. Since p_1^* factors, it follows that for any density q,

$$T_1 q = \alpha\left(\left(\pi \frac{p_1^* q}{p_0}\right)\frac{p_0}{q}\right) = p_1^*,$$

which implies that $q_1^{(k)} = p_1^*$ for $k \geq 1$. Furthermore, note that

$$T_2 p_1^* = \alpha\left(\left(\pi \frac{p_1^* p_2^*}{p_0}\right)\frac{p_0}{p_1^*}\right) = \alpha\left(\frac{(\pi f) p_0}{p_1^*}\right).$$

Therefore, we have

$$\alpha\left(\frac{q_1^{(k)} q_2^{(k)}}{p_0}\right) = \pi f,$$

for $k \geq 2$. Independence of components of x conditioned on y_2 leads to an analogous outcome.

In the Gaussian case, independence corresponds to the fact that a covariance matrix is diagonal. The argument we have discussed in the context of general distributions implies that belief propagation converges when either Σ_1 or Σ_2 is diagonal. However, a stronger result, formalized in the following proposition, establishes that convergence holds for a range of matrices that are "nearly diagonal." The proof of this proposition is given in Appendix F.

Proposition 3 *For $i = 1, 2$, let L_i and U_i be defined by*

$$L_i = 1 - \lambda_{min}\left(\Sigma_i \left(\delta\left(\Sigma_i\right)\right)^{-1}\right) \quad \text{and} \quad U_i = \lambda_{max}\left(\Sigma_i \delta\left(\Sigma_i^{-1}\right)\right) - 1.$$

If

$$\prod_{i=1}^{2} \left(L_i \vee U_i\right) < 1,$$

then $\rho\left(\mathbf{T}_{\Sigma_1,\Sigma_2}\right) < 1$.

Let us discuss a certain interpretation of the proposition. When Σ_1 is diagonal, $L_1 = U_1 = 0$ and $\prod_{i=1}^{2}(L_i \vee U_i) = 0$. This is an extreme case that leaves much leeway in the requirement that $\prod_{i=1}^{2}(L_i \vee U_i) < 1$. An analogous extreme case arises when Σ_2 is diagonal.

As Σ_1 becomes "less diagonal," L_1 and U_1 grow – the former is bounded by 1 but the latter can become arbitrarily large. In any event, $L_1 \vee U_1$ can be viewed as a measure of how far Σ_1 is from being diagonal, or alternatively, how correlated the components of x become upon observation of y_1. Furthermore, the product $\prod_{i=1}^{2}(L_i \vee U_i)$ combines this measure for Σ_1 and Σ_2, and the requirement for this product to be less than 1 allows for one covariance matrix to become more diagonal as the other becomes less so.

Proposition 3 places constraints on Σ_1 and Σ_2 under which convergence is guaranteed, and we discussed how covariance matrices that are "nearly diagonal" should satisfy such constraints. The next proposition extends this result further by showing that if the off-diagonal elements of Σ_1 and Σ_2 are small relative to the diagonal elements, then belief propagation converges. The proof of this proposition follows directly from Proposition 3, and we refer the reader to Appendix G.

Proposition 4 *For any Σ_1 and Σ_2 satisfying Assumption 1, let*

$$\Sigma_1^\beta = \left(\Sigma_1^{-1} + (\beta - 1)I\right)^{-1}, \quad \text{and} \quad \Sigma_2^\beta = \left(\Sigma_2^{-1} + (\beta - 1)I\right)^{-1}.$$

Then, there exist $U_{\Sigma_1,\Sigma_2} > 1$ such that $\rho\left(\mathbf{T}_{\Sigma_1^\beta,\Sigma_2^\beta}\right) < 1$ for all $\beta > U_{\Sigma_1,\Sigma_2}$.

Let us discuss an interpretation of the above result in coding context. The covariance matrices Σ_1^β and Σ_2^β can be written as

$$\Sigma_1^\beta = \left(\Sigma_1^{-1} + \left(\frac{1}{\beta}I\right)^{-1} - I\right)^{-1}, \quad \text{and} \quad \Sigma_2^\beta = \left(\Sigma_2^{-1} + \left(\frac{1}{\beta}I\right)^{-1} - I\right)^{-1}.$$

It follows from Lemma 1 that Σ_1^β can be interpreted as the covariance matrix of the conditional density of x given y_1, when the prior density p_0 of x has variance $\frac{1}{\beta}I$ instead of I. A similar interpretation applies to Σ_2^β. As β increases, the uncertainty over the prior estimate of the random variable x decreases. Thus, β can be thought of as the signal-to-noise ratio of x. Moreover, recall that Σ_1 and Σ_2 represent the covariances of the conditional density of x given y_1 and y_2, respectively. These covariances encapsulate the correlations among information bits conditioned on the observed transmissions. These correlations are determined by the encoding scheme and the channel characteristics. Viewing from this perspective, the result of Proposition 4 implies that, for a given encoding scheme and channel characteristics, there is a threshold U_{Σ_1,Σ_2} such that if the signal-to-noise ratio exceeds this threshold, then belief propagation converges. This result appears to be related to results reported in [1, 7, 25].

It was observed by Agrawal and Vardy [1] that for codes with finite length, there are two thresholds L and U such that when the signal-to-noise ratio is higher than U, turbo decoding converges, but when the signal-to-noise ratio is below L, the algorithm diverges. We expect that a similar result should hold in our context of Gaussian densities. Unfortunately, we currently do not have a formal proof this result. We plan to pursue this in our future work.

The last two propositions show that if the covariance matrices are "close" to diagonal, then the mean vectors converge. Here, we identify an additional situation where convergence occurs, which involves covariance matrices that are "well–conditioned." This result is stated in the following proposition, whose proof is given in Appendix H.

Proposition 5 *If*

$$\frac{\lambda_{min}(\Sigma_1)}{\lambda_{max}(\Sigma_1)} + \frac{\lambda_{min}(\Sigma_2)}{\lambda_{max}(\Sigma_2)} > 1,$$

then $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$.

As an immediate corollary of this proposition, we have $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$ if

$$\frac{\lambda_{max}(\Sigma_i)}{\lambda_{min}(\Sigma_i)} < 2$$

for $i = 1, 2$. Hence, the means converge if the covariance matrices are "well–conditioned."

5.2.3 Example of Divergence

In this section, we provide an example that demonstrates the possibility of a divergent sequence of mean vectors. We should note that when Σ_i is a 2×2 matrix, the variable U_i in Proposition 3 is bounded by 1. This implies that belief propagation will converge if Σ_1 and Σ_2 are 2×2 matrices. So, consider the following 3×3 matrices

$$\Sigma_1 = \begin{pmatrix} 0.2158720 & 0.3135334 & 0.1844028 \\ 0.3135334 & 0.5006464 & 0.3364129 \\ 0.1844028 & 0.3364129 & 0.2653747 \end{pmatrix},$$

and

$$\Sigma_2 = \begin{pmatrix} 0.0346151 & -0.0211820 & 0.0274089 \\ -0.0211820 & 0.0157915 & -0.0175615 \\ 0.0274089 & -0.0175615 & 0.0250863 \end{pmatrix}.$$

For this particular choice of Σ_1 and Σ_2, it turns out that

$$C_1^* = \begin{pmatrix} 0.0095633 & 0 & 0 \\ 0 & 0.0157210 & 0 \\ 0 & 0 & 0.0255243 \end{pmatrix},$$

and

$$C_2^* = \begin{pmatrix} 0.0145617 & 0 & 0 \\ 0 & 0.0050802 & 0 \\ 0 & 0 & 0.0072025 \end{pmatrix}.$$

With this information, it is not hard to verify that $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) = 1.0132513$. Thus, the mean vectors diverge.

6 Analysis of the Fixed Point

We have established that the covariance matrices generated by belief propagation converge, and under certain conditions, so do the means. In this section, we will show that the limits of convergence may provide useful information relating to the desired posterior density $f \sim N(\mu, \Sigma)$. In particular, it turns out – somewhat surprisingly – that the mean of the approximation resulting from belief propagation coincides with that of f. To formalize this result, let the limiting means and covariance matrices be denoted by m_1^*, m_2^*, C_1^*, and C_2^*. Furthermore, let q_1^* and q_2^* be the limiting densities with $q_1^* \sim N(m_1^*, C_1^*)$ and $q_2^* \sim N(m_2^*, C_2^*)$. The following theorem establishes the main result of this section: the mean of the density $\alpha\,(q_1^* q_2^*/p_0)$ generated by belief propagation coincides with that of the desired posterior density f.

Theorem 3 *Let (C_1^*, C_2^*) denote the limit of the sequence of covariance matrices $\left(C_1^{(k)}, C_2^{(k)}\right)$. Suppose that sequences of the mean vectors $m_1^{(k)}$ and $m_2^{(k)}$ converge to m_1^* and m_2^*, respectively. If $q_1^* \sim N(m_1^*, C_1^*)$ and $q_2^* \sim N(m_2^*, C_2^*)$, then*

$$\alpha(q_1^* q_2^*/p_0) \sim N(\mu, A_{C_1^*, C_2^*}).$$

Our proof of this theorem relies on the following lemma, which provides an equation relating means associated with the fixed points. It is not hard to show that $A_{C_1^*, C_2^*}$, A_{Σ_1, C_2^*}, and $A_{C_1^*, \Sigma_2}$, which are used in the statement, are well–defined.

Lemma 8 *Let (C_1^*, C_2^*) denote the limit of the sequence of covariance matrices $\left(C_1^{(k)}, C_2^{(k)}\right)$. Suppose that sequences of the mean vectors $m_1^{(k)}$ and $m_2^{(k)}$ converge to m_1^* and m_2^*, respectively. If $q_1^* \sim N(m_1^*, C_1^*)$ and $q_2^* \sim N(m_2^*, C_2^*)$, then*

$$A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = A_{\Sigma_1, C_2^*}\left(\Sigma_1^{-1}\mu_1 + C_2^{*-1} m_2^*\right) = A_{C_1^*, \Sigma_2}\left(C_1^{*-1} m_1^* + \Sigma_2^{-1}\mu_2\right).$$

Proof: Let $q_1^* \sim N(m_1^*, C_1^*)$ and $q_2^* \sim N(m_2^*, C_2^*)$. Since q_1^* and q_2^* denote the fixed points of belief propagation, we have

$$q_1^* = T_1 q_2^* \quad \text{and} \quad q_2^* = T_2 q_1^*.$$

It follows that

$$\alpha \frac{q_1^* q_2^*}{p_0} = \alpha\pi \frac{p_1^* q_2^*}{p_0} = \alpha\pi \frac{q_1^* p_2^*}{p_0}.$$

The result is then a consequence of Lemma 1 and the fact that π does not alter the mean of a density. ■

Proof of Theorem 3

By Lemma 1, $\mu = A_{\Sigma_1, \Sigma_2}\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right)$, while the mean of $\alpha(q_1^* q_2^*/p_0)$ is $A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right)$. We will show that these two expressions are equal.

Multiplying the equations from Lemma 8 by appropriate matrices, we obtain

$$A^{-1}_{\Sigma_1, C_2^*} A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = \Sigma_1^{-1}\mu_1 + C_2^{*-1} m_2^*,$$

and

$$A^{-1}_{C_1^*, \Sigma_2} A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = C_1^{*-1} m_1^* + \Sigma_2^{-1}\mu_2.$$

It follows that

$$\left(A^{-1}_{\Sigma_1, C_2^*} + A^{-1}_{C_1^*, \Sigma_2}\right) A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2 + C_1^{*-1} m_1^* + C_2^{*-1} m_2^*,$$

which implies that

$$\left((A^{-1}_{\Sigma_1, C_2^*} + A^{-1}_{C_1^*, \Sigma_2}) A_{C_1^*, C_2^*} - I\right)\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2.$$

Therefore,

$$\left(A^{-1}_{\Sigma_1, C_2^*} + A^{-1}_{C_1^*, \Sigma_2} - A^{-1}_{C_1^*, C_2^*}\right) A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = \Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2.$$

Note that $A^{-1}_{\Sigma_1, C_2^*} + A^{-1}_{C_1^*, \Sigma_2} - A^{-1}_{C_1^*, C_2^*} = A^{-1}_{\Sigma_1, \Sigma_2}$. It follows that

$$A_{C_1^*, C_2^*}\left(C_1^{*-1} m_1^* + C_2^{*-1} m_2^*\right) = A_{\Sigma_1, \Sigma_2}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) = \mu.$$

■

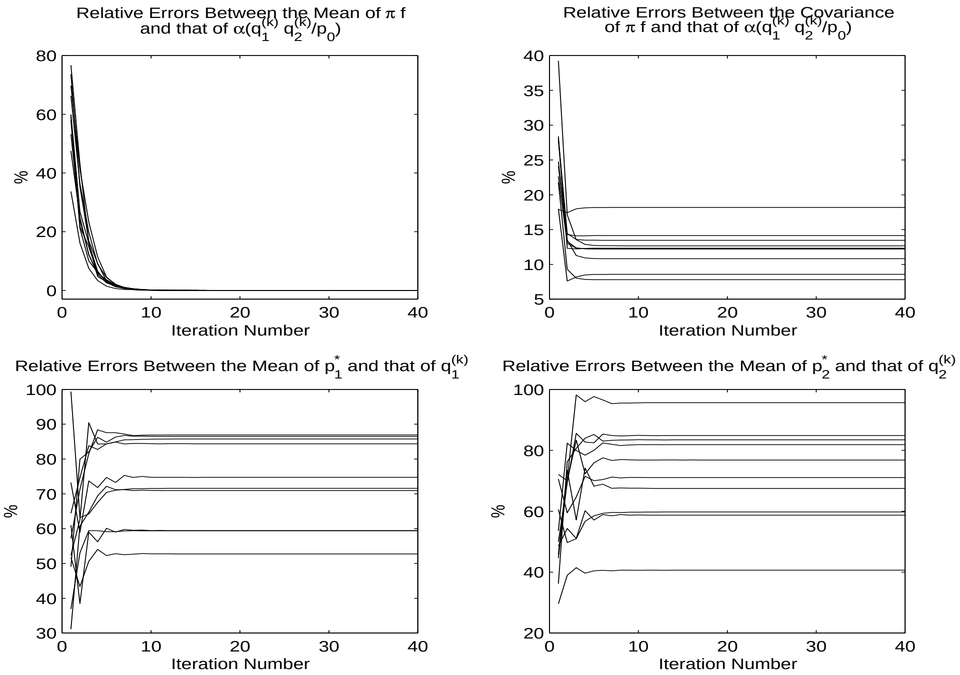

Figure 4: Evolution of errors during 10 representative runs of belief propagation.

7 Experimental Results

The limits of convergence q_1^* and q_2^* of belief propagation provide an approximation $\alpha(q_1^* q_2^*/p_0)$ to πf. We have established that the mean of this approximation coincides with that of the desired posterior density. One might further expect that the covariance matrix of $\alpha(q_1^* q_2^*/p_0)$ approximates that of πf, and even more so, that q_1^* and q_2^* bear some relation to p_1^* and p_2^*. Unfortunately, as will be illustrated by experimental results in this section, such expectations appear to be inaccurate.

We performed experiments involving 20 and 50 dimensional Gaussian densities (i.e., x was either 20 or 50 dimensional in each instance). The covariance matrices Σ_1 and Σ_2 were generated according to

$$\Sigma_1 = \left(I + U_1 U_1'\right)^{-1} \quad \text{and} \quad \Sigma_2 = \left(I + U_2 U_2'\right)^{-1},$$

where U_1 and U_2 are independent random matrices with elements drawn from a uniform distribution over $(-1, 1)$. The means μ_1 and μ_2 were generated as independent random vectors with elements drawn from a uniform distribution over $(-20, 20)$.

Figure 4 shows the evolution of "errors" during 10 representative runs of belief propagation on 20–dimensional problems. The first graph plots, for each k, the relative root–mean–squared error between the mean of πf and the mean of $\alpha(q_1^{(k)} q_2^{(k)}/p_0)$ – the approximation generated by belief propagation at the k^{th} iteration. By relative root–mean–squared error, we are referring to the root–mean–squared difference between the two vectors divided by the root–mean–squared value of the first vector. As indicated by our analysis, if belief

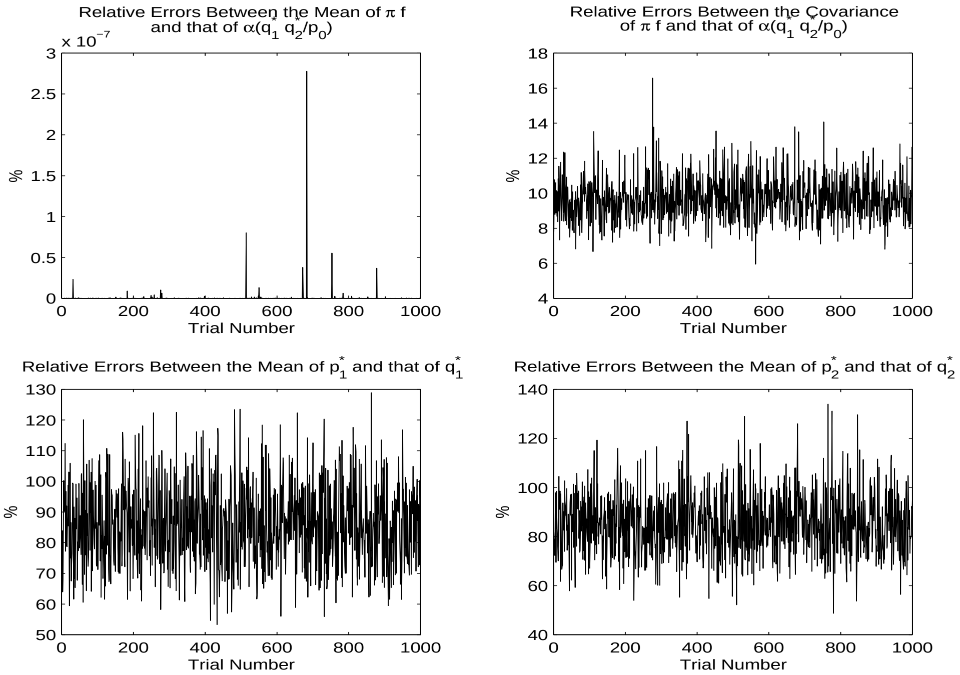

Figure 5: Errors after 50 iterations. Densities generated at the 50th iteration are used as a proxy for the fixed points of belief propagation.

propagation converges, this error converges to zero. The second chart plots relative root–mean–squared error between the covariance of πf and that of $\alpha(q_1^{(k)} q_2^{(k)}/p_0)$. (These two matrices are diagonal, and we treat the diagonal elements as components of vectors when measuring root–mean–squared error.) Although these covariances converge, in agreement with Theorem 1, the ultimate errors are far from zero. The two final graphs plot relative root–mean–squared errors between the means of p_1^* and p_2^* and those of $q_1^{(k)}$ and $q_2^{(k)}$, respectively. Again, even if these means converge, the ultimate errors can be large.

Figure 5 plots data from 1000 different experiments involving 50–dimensional problems. In each experiment, belief propagation was executed for 50 iterations. In measuring errors, densities generated after 50 iterations are assumed to be equal to the stationary points q_1^* and q_2^*. The horizontal axes are labeled with indices of the problem instances. These graphs exhibit the same phenomenon as that observed in the case of 20–dimensional problems – the errors between the mean of πf and that of $\alpha(q_1^* q_2^*/p_0)$ are very close to zero, while errors associated with other statistics vary dramatically.

It is worth noting that in all the reported experiments, the sequences of the mean vectors appeared to converge. However, in larger problems, we have observed divergent cases, though they are very rare. This suggests that there may be sufficient conditions for convergence that are almost always satisfied. Finding such conditions remains an open problem.

8 Closing Remarks

We have shown that, when densities are Gaussian, belief propagation often converges and the mean associated with the limit of convergence coincides with that of the desired posterior density. It is intriguing to note that, in the context of communications, the objective is to choose a code word $\overline{x}$ that comes close to the transmitted code x. One natural way to do this involves assigning to $\overline{x}$ the code word that maximizes the conditional density f, i.e., the one that has the highest chance of being correct. In the Gaussian case that we have studied, this corresponds to the mean of f – a quantity that is computed correctly by belief propagation!

It will be interesting to explore generalizations of the line of analysis presented in this paper. One direction might be to expand the arguments to encompass belief propagation on general network topologies with Gaussian densities. A more interesting – and probably more challenging – pursuit would be to develop theory pertaining to more general (non–Gaussian) densities.

As a parting note, let us suggest that, even in the context of Gaussian densities, belief propagation may prove to be useful. Let us reconsider, for example, a coupled hidden Markov model as that described in Section 3. Suppose now that prior distributions associated with this coupled hidden Markov model are Gaussian. Then, although computation of the conditional mean is tractable via traditional methods, belief propagation may provide a more efficient alternative, as we will now explain.

Since the densities are Gaussian, the mean of x conditioned on y_1 and y_2 is given by

$$\mu = \left(\Sigma_1^{-1} + \Sigma_2^{-1} - I\right)^{-1} \left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right).$$

Computation of this mean may be carried out via inversion of the relevant symmetric positive definite matrices, which takes on the order of $n^{2.81}$ operations. Additional computation might also be required to obtain μ_1, μ_2, Σ_1 and Σ_2.

Let us consider an alternative approach that uses belief propagation. Recall that belief propagation computes sequences $q_1^{(k)}$ and $q_2^{(k)}$. As discussed in Section 3, computation of $q_1^{(k)}$ and $q_2^{(k)}$ at each iteration can be done efficiently via belief propagation – the procedure requires $O(n)$ operations per iteration. If the algorithm converges, the mean of the resulting approximation coincides with μ. Hence, if the algorithm converges within s iterations, or at least comes very close to the limit point, we can obtain a very good approximation to μ in $O\left(sn\right)$ operations. If s is not too large, this can result in substantial computational savings. Unfortunately, we do have a bound on the proximity between μ and the mean of the density generated by belief propagation after s iterations. Nevertheless, our experimental results suggest that belief propagation converges fairly quickly. The notion that belief propagation might compute the mean of a posterior distribution more quickly than traditional approaches raises a tantalizing possibility that the algorithm and potential variants might be able to accelerate the solution of many similar tasks in numerical computation.

Acknowledgments

This research was supported in part by a Stanford University Terman Award. The authors are grateful to David Forney and Michael Jordan for pointers to literature on turbo decoding and inference at early stages in this research. We also thank Michael Saunders for some useful discussions on linear algebra. Finally, we thank anonymous reviewers for detailed thoughtful comments and suggestions.

A Lemmas on Matrix Algebra

In this section, we collect together some useful lemmas on matrix algebra. These results will be used throughout the appendices. The first lemma states an inequality due to Bellman [4].

Lemma 9 *If A is a symmetric positive definite matrix, then for all x and y*

$$(x'Ax)\left(y'A^{-1}y\right) \geq (x'y)^2 .$$

It is easy to see that matrix inversion and the δ operator do not commute. The next lemma reflects possible consequences of reordering.

Lemma 10 *If A is a symmetric positive definite matrix, then*

$$\left(\delta\left(A^{-1}\right)\right)^{-1} \leq \delta\left(A\right).$$

Proof: By letting $x = y = e_i$, where e_i is the unit vector whose i^{th} component is equal to one, we have

$$A_{ii}\left(A^{-1}\right)_{ii} = (x'Ax)\left(y'A^{-1}y\right) \geq (x'y)^2 = 1,$$

where the inequality follows from Lemma 9. It follows that

$$\frac{1}{\left(A^{-1}\right)_{ii}} \leq A_{ii},$$

for all i, which immediately leads to the desired result. ■

The next lemma states an inequality due to Bergstrom [4].

Lemma 11 *Let A and B be symmetric positive definite matrices. Let $A(i)$ and $B(i)$ denote the sub-matrices (also symmetric positive definite) obtained by deleting the i^{th} row and column. Then,*

$$\frac{|A|}{|A(i)|} + \frac{|B|}{|B(i)|} \leq \frac{|A+B|}{|A(i)+B(i)|},$$

where $|M|$ denotes the determinant of a matrix M.

Next, we have a lemma that reflects potential consequences of distributing a certain combination of matrix inversions and the δ operator among addends in a sum.

Lemma 12 *Let A and B be symmetric positive definite matrices. Then,*

$$\left(\delta\left(A^{-1}\right)\right)^{-1} + \left(\delta\left(B^{-1}\right)\right)^{-1} \leq \left(\delta\left((A+B)^{-1}\right)\right)^{-1}.$$

Proof: For any nonsingular matrix A, it is well-known [27] that

$$\left(A^{-1}\right)_{ii} = \frac{|A\left(i\right)|}{|A|}$$

for all i. It therefore follows from Lemma 11 that

$$\frac{1}{\left(A^{-1}\right)_{ii}} + \frac{1}{\left(B^{-1}\right)_{ii}} \leq \frac{1}{\left((A+B)^{-1}\right)_{ii}}$$

for all i. Equivalently,

$$\left(\delta\left(A^{-1}\right)\right)^{-1} + \left(\delta\left(B^{-1}\right)\right)^{-1} \leq \left(\delta\left((A+B)^{-1}\right)\right)^{-1}.$$

■

B Proof of Lemmas 2 and 3

This appendix contains the proof of Lemmas 2 and 3. We first prove the following result, which will be used to prove Lemma 3.

Lemma 13 *For all $D \in \mathcal{D}$, $\mathcal{F}_1(D)$ and $\mathcal{F}_2(D)$ are positive definite diagonal matrices.*

Proof: It suffices to prove this result for $\mathcal{F}_1(D)$. The proof for $\mathcal{F}_2(D)$ is similar. Since $\Sigma_1^{-1} - I$ is positive definite (Assumption 1(b)), $A_{\Sigma_1,D}$ is well-defined and positive definite for all $D \in \mathcal{D}$. In addition, it follows from Lemma 10 that

$$\left(\delta\left(A_{\Sigma_1,D}\right)\right)^{-1} \leq \delta\left(A_{\Sigma_1,D}^{-1}\right),$$

which implies that

$$\left(\delta\left(A_{\Sigma_1,D}\right)\right)^{-1} \leq \delta\left(\Sigma_1^{-1}\right) + D^{-1} - I.$$

Therefore,

$$\left(\delta\left(A_{\Sigma_1,D}\right)\right)^{-1} + I - D^{-1} \leq \delta\left(\Sigma_1^{-1}\right),$$

or equivalently,

$$\left(\left(\delta\left(A_{\Sigma_1,D}\right)\right)^{-1} + I - D^{-1}\right)^{-1} \geq \left(\delta\left(\Sigma_1^{-1}\right)\right)^{-1}.$$

Using the definition of $\mathcal{F}_1$, it follows that

$$\mathcal{F}_1(D) \geq \left(\delta\left(\Sigma_1^{-1}\right)\right)^{-1},$$

which implies that $\mathcal{F}_1(D)$ is a positive definite diagonal matrix. ■

Here is the proof of Lemma 3.

Proof: It suffices to prove this result for T_1. The proof for T_2 is similar. Recall that for any density g,

$$T_1 g = \alpha\left(\left(\pi\frac{p_1^* g}{p_0}\right)\frac{p_0}{g}\right)$$

Consider a density $g \in \mathcal{G}$ with $g \sim N(\mu_g, \Sigma_g)$. Since $\Sigma_1^{-1} - I$ is positive definite (Assumption 1(b)), A_{Σ_1,Σ_g} is positive definite. Thus, it follows from Lemma 1 that

$$\alpha\left(\frac{p_1^* g}{p_0}\right) \sim N\left(A_{\Sigma_1,\Sigma_g}\left(\Sigma_1^{-1}\mu_1 + \Sigma_g^{-1}\mu_g\right), A_{\Sigma_1,\Sigma_g}\right),$$

which implies that

$$\alpha\left(\pi\frac{p_1^* g}{p_0}\right) \sim N\left(A_{\Sigma_1,\Sigma_g}\left(\Sigma_1^{-1}\mu_1 + \Sigma_g^{-1}\mu_g\right), \delta\left(A_{\Sigma_1,\Sigma_g}\right)\right).$$

It follows from the definition of $\mathcal{F}_1(\Sigma_g)$ and Lemma 13 that the matrix

$$\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} + I - \Sigma_g^{-1}$$

is well-defined and positive definite. Application of Lemma 1 implies that $T_1 g$ is a Gaussian density whose covariance matrix is given by

$$\left(\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} + I - \Sigma_g^{-1}\right)^{-1},$$

which is simply $\mathcal{F}_1(\Sigma_g)$. Lemma 1 also tells us that the mean of the density $T_1 g$ is given by

$$\mathcal{F}_1(\Sigma_g)\left(\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} A_{\Sigma_1,\Sigma_g}\left(\Sigma_1^{-1}\mu_1 + \Sigma_g^{-1}\mu_g\right) - \Sigma_g^{-1}\mu_g\right),$$

or equivalently,

$$\mathcal{F}_1(\Sigma_g)\left(\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} A_{\Sigma_1,\Sigma_g} - I\right)\Sigma_g^{-1}\mu_g + \mathcal{F}_g(\Sigma_g)\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} A_{\Sigma_1,\Sigma_g}\Sigma_1^{-1}\mu_1.$$

Since

$$\mathcal{F}_1(\Sigma_g) = \left(\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} + I - \Sigma_g^{-1}\right)^{-1},$$

it follows that

$$\left(\delta\left(A_{\Sigma_1,\Sigma_g}\right)\right)^{-1} = \left(\mathcal{F}_1(\Sigma_g)\right)^{-1} + \Sigma_g^{-1} - I = A^{-1}_{\mathcal{F}_1(\Sigma_g),\Sigma_g}.$$

Using this fact, the mean of the density $T_1 g$ can be written as

$$\mathcal{F}_1(\Sigma_g)\left(A^{-1}_{\mathcal{F}_1(\Sigma_g),\Sigma_g} A_{\Sigma_1,\Sigma_g} - I\right)\Sigma_g^{-1}\mu_g + \mathcal{F}_g(\Sigma_g) A^{-1}_{\mathcal{F}_1(\Sigma_g),\Sigma_g} A_{\Sigma_1,\Sigma_g}\Sigma_1^{-1}\mu_1,$$

which is simply $\mathcal{H}_1(\mu_g, \Sigma_g)$. Therefore,

$$T_1 g \sim N\left(\mathcal{H}_1(\mu_g, \Sigma_g), \mathcal{F}_1(\Sigma_g)\right).$$

■

It is obvious that Lemma 2 is a direct corollary of Lemma 3 and 13.

C Proof of Lemma 4

(a) *Continuity :* Continuity of the operator $\mathcal{F}$ on its domain $\mathcal{D} \times \mathcal{D}$ follows immediately from its definition.

(b) *Monotonicity:* Let $B_2 = X_2^{-1} - Y_2^{-1}$, and note that $B_2 \geq 0$ since $X_2 \leq Y_2$. Starting with the definition of A_{Σ_1,X_2}, we have

$$A_{\Sigma_1,X_2}^{-1} = \Sigma_1^{-1} + X_2^{-1} - I = \Sigma_1^{-1} + Y_2^{-1} - I + B_2 = A_{\Sigma_1,Y_2}^{-1} + B_2,$$

which implies that

$$A_{\Sigma_1,X_2} = \left(A_{\Sigma_1,Y_2}^{-1} + B_2\right)^{-1}.$$

Since $B_2 \geq 0$, Lemma 12 (Appendix A) asserts that

$$\left(\delta\left(A_{\Sigma_1,X_2}\right)\right)^{-1} \geq \left(\delta\left(A_{\Sigma_1,Y_2}\right)\right)^{-1} + \left(\delta\left(B_2^{-1}\right)\right)^{-1},$$

and since B_2 is diagonal,

$$\left(\delta\left(A_{\Sigma_1,X_2}\right)\right)^{-1} \geq \left(\delta\left(A_{\Sigma_1,Y_2}\right)\right)^{-1} + X_2^{-1} - Y_2^{-1}.$$

It follows that

$$\left(\delta\left(A_{\Sigma_1,X_2}\right)\right)^{-1} + I - X_2^{-1} \geq \left(\delta\left(A_{\Sigma_1,Y_2}\right)\right)^{-1} + I - Y_2^{-1},$$

which implies that $\left(\mathcal{F}_1\left(X_2\right)\right)^{-1} \geq \left(\mathcal{F}_1\left(Y_2\right)\right)^{-1}$, or equivalently, $\mathcal{F}_1(X_2) \leq \mathcal{F}_1(Y_2)$. An analogous argument shows that $\mathcal{F}_2(X_1) \leq \mathcal{F}_2(Y_1)$. Hence, $\mathcal{F}\left(X_1, X_2\right) \leq \mathcal{F}\left(Y_1, Y_2\right)$.

(c) *Boundedness:* It follows from the definition of A_{Σ_1,D_2}, that

$$\left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} = \left(\delta\left(\left(\Sigma_1^{-1} + D_2^{-1} - I\right)^{-1}\right)\right)^{-1} = \left(\delta\left(\left(U+V\right)^{-1}\right)\right)^{-1},$$

where $U = \Sigma_1^{-1} - I$ and $V = D_2^{-1}$. From Assumption 1(b), we know that U is positive definite. It follows from Lemma 10 and 12 (Appendix A) that

$$\left(\delta\left(U^{-1}\right)\right)^{-1} + \left(\delta\left(V^{-1}\right)\right)^{-1} \leq \left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} \leq \delta\left(U+V\right).$$

Since $V = D_2^{-1}$ is diagonal,

$$\left(\delta\left(U^{-1}\right)\right)^{-1} + D_2^{-1} \leq \left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} \leq \delta\left(\Sigma_1^{-1}\right) - I + D_2^{-1},$$

and since U is positive definite,

$$D_2^{-1} < \left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} \leq \delta\left(\Sigma_1^{-1}\right) - I + D_2^{-1}.$$

It follows that

$$I < \left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} + I - D_2^{-1} \leq \delta\left(\Sigma_1^{-1}\right).$$

Since $\mathcal{F}_1(D_2) = \left((\delta(A_{\Sigma_1,D_2}))^{-1} + I - D_2^{-1}\right)^{-1}$, we have

$$\left(\delta\left(\Sigma_1^{-1}\right)\right)^{-1} \leq \mathcal{F}_1(D_2) < I.$$

An analogous argument shows that

$$\left(\delta\left(\Sigma_2^{-1}\right)\right)^{-1} \leq \mathcal{F}_2(D_1) < I,$$

Hence,

$$\left(\overline{D}_1, \overline{D}_2\right) \leq \mathcal{F}(D_1, D_2) < (I, I),$$

where $\overline{D}_1 = \left(\delta\left(\Sigma_1^{-1}\right)\right)^{-1}$ and $\overline{D}_2 = \left(\delta\left(\Sigma_2^{-1}\right)\right)^{-1}$.

(d) *Scaling:* We will begin by establishing that

$$\beta\delta(A_{\Sigma_1,D_2}) \leq \delta(A_{\Sigma_1,\beta D_2}).$$

By definition, we have

$$A_{\Sigma_1,D_2} = \left(\Sigma_1^{-1} - I + D_2^{-1}\right)^{-1} = \left(\beta\left(\Sigma_1^{-1} - I\right) + D_2^{-1} + (1-\beta)\left(\Sigma_1^{-1} - I\right)\right)^{-1}.$$

Application of Lemma 12 (Appendix A) implies that

$$(\delta(A_{\Sigma_1,D_2}))^{-1} \geq \left(\delta\left(\left(\beta\left(\Sigma_1^{-1} - I\right) + D_2^{-1}\right)^{-1}\right)\right)^{-1} + (1-\beta)\left(\delta\left(\left(\Sigma_1^{-1} - I\right)^{-1}\right)\right)^{-1},$$

Since $\Sigma_1^{-1} - I$ is positive definite (Assumption 1(b)), we have

$$(\delta(A_{\Sigma_1,D_2}))^{-1} \geq \left(\delta\left(\left(\beta\left(\Sigma_1^{-1} - I\right) + D_2^{-1}\right)^{-1}\right)\right)^{-1},$$

which implies that

$$\delta(A_{\Sigma_1,D_2}) \leq \delta\left(\left(\beta\left(\Sigma_1^{-1} - I\right) + D_2^{-1}\right)^{-1}\right).$$

However,

$$\delta\left(\left(\beta\left(\Sigma_1^{-1} - I\right) + D_2^{-1}\right)^{-1}\right) = \frac{1}{\beta}\delta\left(\left(\Sigma_1^{-1} - I + (\beta D_2)^{-1}\right)^{-1}\right) = \frac{1}{\beta}\delta(A_{\Sigma_1,\beta D_2}),$$

which implies that

$$\beta\delta(A_{\Sigma_1,D_2}) \leq \delta(A_{\Sigma_1,\beta D_2}).$$

The bound on $\beta\delta(A_{\Sigma_1,D_2})$ implies that

$$(\delta(A_{\Sigma_1,\beta D_2}))^{-1} \leq \frac{1}{\beta}(\delta(A_{\Sigma_1,D_2}))^{-1}.$$

It follows that

$$\begin{aligned}
\left(\mathcal{F}_1\left(\beta D_2\right)\right)^{-1} &= \left(\delta\left(A_{\Sigma_1,\beta D_2}\right)\right)^{-1} + I - \left(\beta D_2\right)^{-1} \\
&\le \frac{1}{\beta}\left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} + I - \frac{1}{\beta}D_2^{-1} \\
&< \frac{1}{\beta}\left(\delta\left(A_{\Sigma_1,D_2}\right)\right)^{-1} + \frac{1}{\beta}I - \frac{1}{\beta}D_2^{-1} \\
&= \frac{1}{\beta}\left(\mathcal{F}_1\left(D_2\right)\right)^{-1}
\end{aligned}$$

Therefore,

$$\beta\mathcal{F}_1\left(D_2\right) < \mathcal{F}_1\left(\beta D_2\right).$$

An analogous argument shows that

$$\beta\mathcal{F}_2\left(D_1\right) < \mathcal{F}_2\left(\beta D_1\right),$$

and the result follows. ■

D Proof of Lemma 7

The proof of Lemma 7 relies on the following two results. Because they are of standard flavor we state them without proof.

Lemma 14 *Let $\{y_k\}$, $\{\alpha_k\}$, and $\{\beta_k\}$ be sequences of non-negative real numbers such that*

$$y_{k+1} \le \alpha_k y_k + \beta_k,$$

for all $k \ge 0$. If

$$\lim_{k\to\infty} \alpha_k = \alpha^* \quad \text{and} \quad \lim_{k\to\infty} \beta_k = 0,$$

where $0 < \alpha^* < 1$, then $\lim_{k\to\infty} y_k = 0$.

Lemma 15 *If A is any matrix such that $\rho\left(A\right) \ne 0$ and $\rho\left(A\right) < 1$, then there exist a constant C such that*

$$\left\|A^n\right\| \le C\rho\left(A\right)^n$$

for all n.

Here is the proof of Lemma 7.

Proof: Let us first assume that $\rho\left(A\right) \ne 0$. Let the sequence $\{\overline{x}_k\}$ be defined by

$$\overline{x}_{k+1} = A\overline{x}_k + b,$$

for all $k \ge 0$ with $\overline{x}_0 = x_0$. It follows that

$$x_{k+1} - \overline{x}_{k+1} = A\left(x_k - \overline{x}_k\right) + \left(A_k - A\right)x_k + \left(b_k - b\right)$$

for all $k \geq 0$. Using the above recursion, one can show that

$$x_{k+1} - \overline{x}_{k+1} = \sum_{i=0}^{k} A^i \left(A_{k-i} - A\right) x_{k-i} + A^i \left(b_{k-i} - b\right),$$

which implies that

$$\begin{aligned} \|x_{k+1} - \overline{x}_{k+1}\| &\leq \sum_{i=0}^{k} \|A^i\| \|A_{k-i} - A\| \|x_{k-i}\| + \|A^i\| \|b_{k-i} - b\| \\ &\leq \sum_{i=0}^{k} \|A^i\| \|A_{k-i} - A\| \|x_{k-i} - \overline{x}_{k-i}\| + \\ &\quad \sum_{i=0}^{k} \|A^i\| \|A_{k-i} - A\| \|\overline{x}_{k-i}\| + \|A^i\| \|b_{k-i} - b\| \\ &\leq C\left(\sum_{i=0}^{k} \rho\left(A\right)^i \|A_{k-i} - A\| \|x_{k-i} - \overline{x}_{k-i}\|\right) + \\ &\quad C\left(\sum_{i=0}^{k} \rho\left(A\right)^i \|A_{k-i} - A\| \|\overline{x}_{k-i}\| + \rho\left(A\right)^i \|b_{k-i} - b\|\right) \end{aligned}$$

where the last inequality follows from Lemma 15. Define the sequence $\{z_k\}$ by

$$\begin{aligned} z_{k+1} &= C\left(\sum_{i=0}^{k} \rho\left(A\right)^i \|A_{k-i} - A\| \|x_{k-i} - \overline{x}_{k-i}\|\right) + \\ &\quad C\left(\sum_{i=0}^{k} \rho\left(A\right)^i \|A_{k-i} - A\| \|\overline{x}_{k-i}\| + \rho\left(A\right)^i \|b_{k-i} - b\|\right). \end{aligned}$$

From the definition of z_k, it follows that

$$\begin{aligned} z_{k+1} &= \rho\left(A\right) z_k + C\Big(\|A_k - A\| \|x_k - \overline{x}_k\| + \|A_k - A\| \|\overline{x}_{k-i}\| + \|b_k - b\|\Big) \\ &\leq \Big(\rho\left(A\right) + C\|A_k - A\|\Big) z_k + C\Big(\|A_k - A\| \|\overline{x}_{k-i}\| + \|b_k - b\|\Big) \end{aligned}$$

where the last inequality follows from the fact that

$$\|x_k - \overline{x}_k\| \leq z_k,$$

for all $k \geq 0$. Since the sequence $\{A_k\}$ converges to A, it follows that

$$\lim_{k\to\infty} \rho\left(A\right) + C\|A_k - A\| = \rho\left(A\right) < 1.$$

Moreover, since $\rho\left(A\right) < 1$, the sequence $\{\overline{x}_k\}$ converges. Thus,

$$\lim_{k\to\infty} C\Big(\|A_k - A\| \|\overline{x}_{k-i}\| + \|b_k - b\|\Big) = 0.$$

It follows from Lemma 14 that the sequence $\{z_k\}$ converges to 0. Since $\|x_k - \overline{x}_k\| \leq z_k$ for all k, and the sequence $\{\overline{x}_k\}$ converges, it follows that the sequence $\{x_k\}$ also converges.

Thus, we have established convergence of the sequence $\{x_k\}$ when $\rho(A) \neq 0$. The proof for the case when $\rho(A) = 0$ is similar. The only modification is in the result of Lemma 15. In this case, we have

$$\|A^n\| \leq C\rho(A)^n$$

for sufficiently large n. It is now easy to see that the above argument still works in this case. ■

E Proof of Proposition 2

The proof of Proposition 2 relies on the following three lemmas. The first relates stability of $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2})$ to that of its two sub–matrices.

Lemma 16 *If*

$$\rho\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*} - I\right)\rho\left(A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2} - I\right) < 1,$$

then $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$.

Proof: Recall that the matrix $\mathbf{T}_{\Sigma_1,\Sigma_2}$ is defined by

$$\mathbf{T}_{\Sigma_1,\Sigma_2} = \left(\begin{array}{cc} 0 & A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*} - I \\ A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2} - I & 0 \end{array}\right).$$

Let $\mathbf{M}$ be a diagonal matrix defined by

$$\mathbf{M} = \left(\begin{array}{cc} A^{1/2}_{C_1^*,C_2^*} & 0 \\ 0 & A^{1/2}_{C_1^*,C_2^*} \end{array}\right).$$

The definition of $\mathbf{T}_{\Sigma_1,\Sigma_2}$ and $\mathbf{M}$ implies that

$$\mathbf{M}\mathbf{T}_{\Sigma_1,\Sigma_2}\mathbf{M}^{-1} = \left(\begin{array}{cc} 0 & A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*} - I \\ A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*} - I & 0 \end{array}\right).$$

It is easy to see that

$$\begin{aligned}\rho\left(\mathbf{M}\mathbf{T}^2_{\Sigma_1,\Sigma_2}\mathbf{M}^{-1}\right) &= \rho\left(\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*} - I\right)\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*} - I\right)\right) \vee \\ &\quad \rho\left(\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*} - I\right)\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*} - I\right)\right).\end{aligned}$$

Since $A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}$ and $A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*}$ are symmetric, and $\rho(AB) = \rho(B'A')$ for all matrices A and B, we have

$$\begin{aligned}&\rho\left(\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*} - I\right)\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*} - I\right)\right) \\ &\quad = \rho\left(\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*} - I\right)\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*} - I\right)\right).\end{aligned}$$

Hence,

$$\begin{aligned}
\rho\left(\mathbf{M}\mathbf{T}^2_{\Sigma_1,\Sigma_2}\mathbf{M}^{-1}\right) &= \rho\left(\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}-I\right)\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*}-I\right)\right)\\
&\leq \left\|\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}-I\right)\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*}-I\right)\right\|_2\\
&\leq \left\|A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}-I\right\|_2\left\|A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*}-I\right\|_2\\
&= \rho\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}-I\right)\rho\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*}-I\right)
\end{aligned}$$

where the last equality follows from the symmetry of $A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}$ and $A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}$ $A^{-1/2}_{C_1^*,C_2^*}$. Note that

$$\rho\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}-I\right)=\rho\left(A^{1/2}_{C_1^*,C_2^*}\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}-I\right)A^{-1/2}_{C_1^*,C_2^*}\right).$$

Since eigenvalues are invariant under similarity transformations,

$$\rho\left(A^{-1/2}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}A^{-1/2}_{C_1^*,C_2^*}-I\right)=\rho\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}-I\right).$$

An analogous argument shows that

$$\rho\left(A^{-1/2}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}A^{-1/2}_{C_1^*,C_2^*}-I\right)=\rho\left(A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}-I\right).$$

Therefore,

$$\rho\left(\mathbf{M}\mathbf{T}^2_{\Sigma_1,\Sigma_2}\mathbf{M}^{-1}\right)\leq\rho\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}-I\right)\rho\left(A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}-I\right).$$

The result then follows from the fact that

$$\rho\left(\mathbf{M}\mathbf{T}^2_{\Sigma_1,\Sigma_2}\mathbf{M}^{-1}\right)=\rho\left(\mathbf{T}^2_{\Sigma_1,\Sigma_2}\right)=\left(\rho(\mathbf{T}_{\Sigma_1,\Sigma_2})\right)^2.$$

■

For any positive reals a and b, we have $0 < a/(a+b) < 1$. The next lemma generalizes this result to the case of positive definite matrices.

Lemma 17 *Suppose that A and B are symmetric positive definite matrices. Let $C = A\,(A+B)^{-1}$. If λ is an eigenvalue of C, then $0 < \lambda < 1$.*

Proof: Note that C is well-defined since $A+B$ is a symmetric positive definite matrix. Let λ be an eigenvalue of C with an associated eigenvector u, which may be a complex vector. By definition,

$$Cu = A\,(A+B)^{-1}\,u = \lambda u$$

Thus,

$$\begin{aligned}
u &= \lambda\,(A+B)\,A^{-1}u\\
&= \lambda\left(I+BA^{-1}\right)u
\end{aligned}$$

Pre-multiplying the above equation by B^{-1}, it follows that

$$B^{-1}u=\lambda\left(B^{-1}+A^{-1}\right)u.$$

After pre-multiply the above equation by u^H, the conjugate transpose of u, we have

$$u^H B^{-1} u = \lambda \left(u^H B^{-1} u + u^H A^{-1} u \right).$$

From linear algebra, if R is any real symmetric matrix, then for any vector x (possibly complex), $x^H R x$ is always a real number. Thus, λ is a real number. This implies that u is a real vector. Hence, the above equation can be written as

$$u' B^{-1} u = \lambda \left(u' A^{-1} u + u' B^{-1} u \right)$$

Since A and B are symmetric positive definite matrices, so are A^{-1} and B^{-1}. It follows that $0 < \lambda < 1$. ■

The final lemma shows convergence of the mean vectors when $\Sigma_1 = \Sigma_2$.

Lemma 18 *If $\Sigma_1 = \Sigma_2 = \Sigma$, then the sequence of covariance matrices converge to a unique fixed point (C, C) in $\mathcal{D} \times \mathcal{D}$ and $C \leq \delta(\Sigma)$. Moreover, $\rho(\mathbf{T}_{\Sigma_1,\Sigma_2}) < 1$.*

Proof: Recall from Theorem 1 that the sequence of covariance matrices converges to the unique fixed point (C_1^*, C_2^*) of $\mathcal{F}$ in $\mathcal{D} \times \mathcal{D}$. Let us first show that $C_1^* = C_2^*$. Since (C_1^*, C_2^*) is the fixed point of $\mathcal{F}$, it follows that (C_1^*, C_2^*) satisfy

$$C_1^* = \mathcal{F}_1(C_2^*) = \left(\left(\delta\left(A_{\Sigma_1, C_2^*} \right) \right)^{-1} + I - (C_2^*)^{-1} \right)^{-1},$$

and

$$C_2^* = \mathcal{F}_2(C_1^*) = \left(\left(\delta\left(A_{C_1^*, \Sigma_2} \right) \right)^{-1} + I - (C_1^*)^{-1} \right)^{-1}.$$

Since Σ_1 and Σ_2 are equal, it follows from symmetry that $C_1^* = C_2^* = C$ for some C in $\mathcal{D}$. Thus,

$$C^{-1} = \delta\left(\left(\Sigma^{-1} + C^{-1} - I \right)^{-1} \right)^{-1} + I - C^{-1}.$$

Since $C < I$ (Lemma 4(c)), $C^{-1} - I$ is positive definite. Hence, it follows from Lemma 12 (Appendix A) that

$$C^{-1} \geq (\delta(\Sigma))^{-1} + \left(\delta\left(\left(C^{-1} - I \right)^{-1} \right) \right)^{-1} + I - C^{-1} = (\delta(\Sigma))^{-1},$$

which implies that $C \leq \delta(\Sigma)$.

Recall that the matrix $\mathbf{T}_{\Sigma_1,\Sigma_2}$ is defined by

$$\mathbf{T}_{\Sigma_1,\Sigma_2} = \begin{pmatrix} 0 & A^{-1}_{C_1^*,C_2^*} A_{\Sigma_1,C_2^*} - I \\ A^{-1}_{C_1^*,C_2^*} A_{C_1^*,\Sigma_2} - I & 0 \end{pmatrix}.$$

Since $\Sigma_1 = \Sigma_2 = \Sigma$ and $C_1^* = C_2^* = C$, we have that

$$A^{-1}_{C_1^*,C_2^*} A_{\Sigma_1,C_2^*} - I = A^{-1}_{C_1^*,C_2^*} A_{C_1^*,\Sigma_2} - I = \left(2C^{-1} - I \right) A_{\Sigma,C} - I.$$

Moreover,

$$\begin{aligned}\left(2C^{-1}-I\right)A_{\Sigma,C}-I &= \left(2C^{-1}-I\right)\left(\Sigma^{-1}+C^{-1}-I\right)^{-1}-I\\ &= 2\left(C^{-1}-\frac{1}{2}I\right)\left(\Sigma^{-1}-\frac{1}{2}I+C^{-1}-\frac{1}{2}I\right)^{-1}-I\end{aligned}$$

Using the fact that $\Sigma^{-1}-\frac{1}{2}I$ (Assumption 1(b)) and $C^{-1}-\frac{1}{2}I$ (Lemma 4(c)) are positive definite, it follows from Lemma 17 that all eigenvalues of the matrix

$$\left(C^{-1}-\frac{1}{2}I\right)\left(\Sigma^{-1}-\frac{1}{2}I+C^{-1}-\frac{1}{2}I\right)^{-1}$$

are in $(0,1)$. This implies that

$$\rho\left(2\left(C^{-1}-\frac{1}{2}I\right)\left(\Sigma^{-1}-\frac{1}{2}I+C^{-1}-\frac{1}{2}I\right)^{-1}-I\right)<1,$$

or equivalently

$$\rho\left(\left(2C^{-1}-I\right)A_{\Sigma,C}-I\right)<1.$$

Therefore,

$$\rho\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}-I\right)=\rho\left(A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}-I\right)=\rho\left(\left(2C^{-1}-I\right)A_{\Sigma,C}-I\right)<1.$$

It follows from Lemma 16 that $\rho\left(\mathbf{T}_{\Sigma_1,\Sigma_2}\right)<1$. ■

Proof of Proposition 2

It is not hard to verify that Σ_1 and Σ_2 satisfy Assumption 1. Let (C,C) denotes the unique fixed point of the sequences of covariance matrices under belief propagation when the covariance matrices of p_1^* and p_2^* are Σ (Lemma 18). It follows from Theorem 1 that C satisfies the following equation

$$\left(C^{-1}+C^{-1}-I\right)^{-1}=\delta\left(\left(\Sigma^{-1}+C^{-1}-I\right)^{-1}\right).$$

Also, let the matrix $\mathbf{T}_{\Sigma,\Sigma}$ be defined by

$$\mathbf{T}_{\Sigma,\Sigma}=\begin{pmatrix}0 & A^{-1}_{C,C}A_{\Sigma,C}-I\\ A^{-1}_{C,C}A_{\Sigma,C}-I & 0\end{pmatrix}$$

Now, let C_1^* and C_2^* be defined by

$$C_1^*=\left(C^{-1}+\gamma I\right)^{-1},\quad \text{and}\quad C_2^*=\left(C^{-1}-\gamma I\right)^{-1}$$

It is a standard fact in linear algebra [27] that for any symmetric positive definite matrix Σ, $\Sigma_{ii}\leq\lambda_{max}(\Sigma)$ for all i. Since $C\leq\delta(\Sigma)$ (Lemma 18), it follows that C_1^* and C_2^* are positive definite. Moreover,

$$\begin{aligned}\left((C_1^*)^{-1}+(C_2^*)^{-1}-I\right)^{-1} &= \left(C^{-1}+C^{-1}-I\right)^{-1}\\ &= \delta\left(\left(\Sigma^{-1}+C^{-1}-I\right)^{-1}\right)\\ &= \delta\left(\left(\Sigma_1^{-1}+(C_2^*)^{-1}-I\right)^{-1}\right),\end{aligned}$$

where the last equality follows from the definition of Σ_1 and C_2^*. A similar argument shows that

$$\left((C_1^*)^{-1} + (C_2^*)^{-1} - I\right)^{-1} = \delta\left(\left((C_1^*)^{-1} + \Sigma_2^{-1} - I\right)^{-1}\right).$$

It follows from Theorem 1 that (C_1^*, C_2^*) is the unique fixed point of the sequences of covariance matrices generated by belief propagation when the covariance matrices of p_1^* and p_2^* are Σ_1 and Σ_2, respectively. We also know from Theorem 2 that the associated sequence of mean vectors will converge if $\rho\left(\mathbf{T}_{\Sigma_1,\Sigma_2}\right) < 1$ where

$$\mathbf{T}_{\Sigma_1,\Sigma_2} = \begin{pmatrix} 0 & A^{-1}_{C_1^*,C_2^*} A_{\Sigma_1,C_2^*} - I \\ A^{-1}_{C_1^*,C_2^*} A_{C_1^*,\Sigma_2} - I & 0 \end{pmatrix}$$

However, it is immediate from the definition of (Σ_1, Σ_2) and (C_1^*, C_2^*) that

$$A_{C_1^*,C_2^*} = A_{C,C}, \quad A_{\Sigma_1,C_2^*} = A_{\Sigma,C}, \quad \text{and} \quad A_{C_1^*,\Sigma_2} = A_{\Sigma,C}.$$

Thus, $\mathbf{T}_{\Sigma,\Sigma} = \mathbf{T}_{\Sigma_1,\Sigma_2}$. Since $\rho\left(\mathbf{T}_{\Sigma,\Sigma}\right) < 1$ by Lemma 18, the desired result follows. ■

F Proof of Proposition 3

Our proof of Proposition 3 relies on two lemmas. The first deals with eigenvalues of a product of a symmetric positive definite matrix and a positive definite diagonal matrix – quantities that appear in the definition of L_i and U_i.

Lemma 19 *Let A be a symmetric positive definite matrix, and let D be a positive definite diagonal matrix. Then,*

$$\lambda_{min}\left(A^{-1}D\right) = \min_{\|u\|_2=1} \frac{u'Du}{u'Au},$$

and

$$\lambda_{max}\left(A^{-1}D\right) = \max_{\|u\|_2=1} \frac{u'Du}{u'Au}.$$

Proof: It suffices to prove this result for $\lambda_{min}\left(A^{-1}D\right)$. The proof for $\lambda_{max}\left(A^{-1}D\right)$ is similar. Since A is a symmetric positive definite matrix,

$$A = U\Lambda U',$$

for some orthogonal matrix U and positive definite diagonal matrix Λ. For any real number p, let the matrix A^p be defined by

$$A^p = U\Lambda^p U'.$$

The reader can easily verify that the normal rules of exponentiation apply in this case. From linear algebra, if K is any symmetric positive definite matrix, then

$$\lambda_{min}(K) = \min_{\|v\|_2=1} v'Kv = \min_{v\neq 0} \frac{v'Kv}{v'v}.$$

Since eigenvalues are preserved under similarity transformation,

$$\begin{aligned}\lambda_{min}\left(A^{-1}D\right) &= \lambda_{min}\left(A^{1/2}A^{-1}DA^{-1/2}\right)\\ &= \lambda_{min}\left(A^{-1/2}DA^{-1/2}\right)\\ &= \min_{v\neq 0}\frac{v'A^{-1/2}DA^{-1/2}v}{v'v}\end{aligned}$$

where the last equality follows from the fact that $A^{-1/2}DA^{-1/2}$ is a symmetric positive definite matrix. Hence,

$$\begin{aligned}\lambda_{min}\left(A^{-1}D\right) &= \min_{v\neq 0}\frac{v'A^{-1/2}DA^{-1/2}v}{v'v}\\ &= \min_{v\neq 0}\frac{v'A^{-1/2}DA^{-1/2}v}{v'A^{-1/2}AA^{-1/2}v}\\ &= \min_{u\neq 0}\frac{u'Du}{u'Au}\end{aligned}$$

where the last equality follows from the identification $u = A^{-1/2}v$ and the fact that $A^{-1/2}$ is nonsingular. Since the ratio on the right-hand side of the above equation is scale-invariant,

$$\lambda_{min}\left(A^{-1}D\right) = \min_{\|u\|_2=1}\frac{u'Du}{u'Au}.$$

■

A second lemma provides a bound on the fixed point (C_1^*, C_2^*) of $\mathcal{F}$.

Lemma 20 *For $i = 1, 2$, we have*

$$(\delta(\Sigma_i))^{-1} \leq (C_i^*)^{-1} \leq \delta\left(\Sigma_i^{-1}\right).$$

Proof: Since (C_1^*, C_2^*) is the fixed point of $\mathcal{F}$, it follows that

$$C_1^* = \left(\left(\delta\left(A_{\Sigma_1,C_2^*}\right)\right)^{-1} + I - (C_2^*)^{-1}\right)^{-1} \text{ and } C_2^* = \left(\left(\delta\left(A_{C_1^*,\Sigma_2}\right)\right)^{-1} + I - (C_1^*)^{-1}\right)^{-1}.$$

Using the definition of A_{Σ_1,C_2^*}, we have

$$(C_1^*)^{-1} = \delta\left(\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)^{-1}\right)^{-1} + I - (C_2^*)^{-1}.$$

Since $C_2^* < I$ (Lemma 4(c)), $(C_2^*)^{-1} - I$ is a positive definite diagonal matrix. Thus, it follows from Lemma 10 (Appendix A) that

$$(C_1^*)^{-1} \leq \delta\left(\Sigma_1^{-1}\right) + \delta\left((C_2^*)^{-1} - I\right) + I - (C_2^*)^{-1} = \delta\left(\Sigma_1^{-1}\right).$$

Application of Lemma 12 (Appendix A) implies that

$$(C_1^*)^{-1} \geq (\delta(\Sigma_1))^{-1} + \left(\delta\left(\left((C_2^*)^{-1} - I\right)^{-1}\right)\right)^{-1} + I - (C_2^*)^{-1} = (\delta(\Sigma_1))^{-1}.$$

Therefore,

$$(\delta(\Sigma_1))^{-1} \leq (C_1^*)^{-1} \leq \delta\left(\Sigma_1^{-1}\right).$$

The result for C_2^* can be established via entirely analogous means. ∎

Equipped with our lemmas, we now move on to prove Proposition 3.

Proof of Proposition 3

Recall that the matrix $\mathbf{T}_{\Sigma_1,\Sigma_2}$ is defined by

$$\mathbf{T}_{\Sigma_1,\Sigma_2} = \begin{pmatrix} 0 & A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*} - I \\ A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2} - I & 0 \end{pmatrix}.$$

We will first find an upper bound for the eigenvalues of $A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}$ and $A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}$. Any matrix A and its transpose A' possess the same eigenvalues. It therefore suffices to consider the eigenvalues of $A_{\Sigma_1,C_2^*}A^{-1}_{C_1^*,C_2^*}$ and $A_{C_1^*,\Sigma_2}A^{-1}_{C_1^*,C_2^*}$.

Let λ be any eigenvalue of $A_{\Sigma_1,C_2^*}A^{-1}_{C_1^*,C_2^*}$ with an associated eigenvector v such that $\|v\|_2 = 1$. By definition,

$$A_{\Sigma_1,C_2^*}A^{-1}_{C_1^*,C_2^*}v = \lambda v,$$

which implies that

$$\lambda = \frac{v'A^{-1}_{C_1^*,C_2^*}v}{v'A^{-1}_{\Sigma_1,C_2^*}v} = \frac{v'\left((C_1^*)^{-1} + (C_2^*)^{-1} - I\right)v}{v'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)v}.$$

It follows from Lemma 20 that

$$\frac{v'\left((\delta(\Sigma_1))^{-1} + (C_2^*)^{-1} - I\right)v}{v'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)v} \leq \lambda \leq \frac{v'\left(\delta\left(\Sigma_1^{-1}\right) + (C_2^*)^{-1} - I\right)v}{v'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)v},$$

which implies that

$$\min_{\|u\|_2=1} \frac{u'\left((\delta(\Sigma_1))^{-1} + (C_2^*)^{-1} - I\right)u}{u'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)u} \leq \lambda \leq \max_{\|u\|_2=1} \frac{u'\left(\delta\left(\Sigma_1^{-1}\right) + (C_2^*)^{-1} - I\right)u}{u'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)u}.$$

Since $(C_2^*)^{-1} - I$ is positive definite (Lemma 4(c)) and

$$\min_{\|u\|_2=1} \frac{u'\left((\delta(\Sigma_1))^{-1} + (C_2^*)^{-1} - I\right)u}{u'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)u} \leq 1 \leq \max_{\|u\|_2=1} \frac{u'\left(\delta\left(\Sigma_1^{-1}\right) + (C_2^*)^{-1} - I\right)u}{u'\left(\Sigma_1^{-1} + (C_2^*)^{-1} - I\right)u},$$

we have

$$\min_{\|u\|_2=1} \frac{u'(\delta(\Sigma_1))^{-1}u}{u'\Sigma_1^{-1}u} \leq \lambda \leq \max_{\|u\|_2=1} \frac{u'\delta\left(\Sigma_1^{-1}\right)u}{u'\Sigma_1^{-1}u}.$$

From Lemma 19 and the definition of L_1 and U_1 given in Proposition 3, it follows that

$$1 - L_1 = \lambda_{min}\left(\Sigma_1(\delta(\Sigma_1))^{-1}\right) \leq \lambda \leq \lambda_{max}\left(\Sigma_1\delta\left(\Sigma_1^{-1}\right)\right) = U_1 + 1.$$

Since λ is arbitrary, we conclude that all eigenvalues of $A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}$ are between $1-L_1$ and U_1+1. An analogous argument shows that all eigenvalues of $A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}$ are bounded by $1-L_2$ and U_2+1. Using the fact that L_i and U_i are non-negative for $i=1,2$ (Lemma 19), it follows that

$$\rho\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}-I\right)\leq(L_1\vee U_1),$$

and

$$\rho\left(A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}-I\right)\leq(L_2\vee U_2).$$

Consequently, if $\prod_{i=1}^{2}(L_i\vee U_i)<1$, then

$$\rho\left(A^{-1}_{C_1^*,C_2^*}A_{\Sigma_1,C_2^*}-I\right)\rho\left(A^{-1}_{C_1^*,C_2^*}A_{C_1^*,\Sigma_2}-I\right)<1,$$

and the desired result follows from Lemma 16. ■

G Proof of Proposition 4

Let

$$U^1_{\Sigma_1,\Sigma_2}=\inf\left\{\gamma\in\Re:\max_{\|u\|_2=1}\frac{u'\left(\delta\left(\Sigma_i^{-1}\right)-\Sigma_i^{-1}\right)u}{u'\Sigma_i^{-1}u+\gamma-1}<1,\quad i=1,2\right\},$$

and let

$$U_{\Sigma_1,\Sigma_2}=1\vee U^1_{\Sigma_1,\Sigma_2}.$$

Since Σ_1 and Σ_2 are symmetric positive definite matrices, it is follows that $U^1_{\Sigma_1,\Sigma_2}$, and thus U_{Σ_1,Σ_2}, is well defined. Furthermore, it is easy to verify that Σ_1^β and Σ_2^β satisfy Assumption 1 for all $\beta>U_{\Sigma_1,\Sigma_2}$. It follows from Lemma 19 (Appendix F) that

$$\begin{aligned}\lambda_{max}\left(\Sigma_1^\beta\delta\left(\left(\Sigma_1^\beta\right)^{-1}\right)\right)-1 &= \lambda_{max}\left(\left(\Sigma_1^{-1}+(\beta-1)I\right)^{-1}\delta\left(\Sigma_1^{-1}+(\beta-1)I\right)\right)-1\\ &= \max_{\|u\|_2=1}\frac{u'\delta\left(\Sigma_1^{-1}\right)u+\beta-1}{u'\Sigma_1^{-1}u+\beta-1}-1\\ &= \max_{\|u\|_2=1}\frac{u'\left(\delta\left(\Sigma_i^{-1}\right)-\Sigma_i^{-1}\right)u}{u'\Sigma_i^{-1}u+\beta-1}\end{aligned}$$

Thus, for all $\beta>U_{\Sigma_1,\Sigma_2}$,

$$0\leq\lambda_{max}\left(\Sigma_1^\beta\delta\left(\left(\Sigma_1^\beta\right)^{-1}\right)\right)-1<1.$$

A similar argument shows that

$$0\leq\lambda_{max}\left(\Sigma_2^\beta\delta\left(\left(\Sigma_2^\beta\right)^{-1}\right)\right)-1<1.$$

Application of Lemma 19 (Appendix F) also implies that

$$\lambda_{min}\left(\Sigma_1^\beta\left(\delta\left(\Sigma_1^\beta\right)\right)^{-1}\right)=\min_{\|u\|_2=1}\frac{u'\left(\delta\left(\Sigma_1^\beta\right)\right)^{-1}u}{u'\left(\Sigma_1^\beta\right)^{-1}u}.$$

Thus, it follows from Lemma 10 (Appendix A) that

$$0 < \lambda_{min} \left(\Sigma_1^\beta \left(\delta \left(\Sigma_1^\beta \right) \right)^{-1} \right) \leq \min_{\|u\|_2=1} \frac{u' \delta \left(\left(\Sigma_1^\beta \right)^{-1} \right) u}{u' \left(\Sigma_1^\beta \right)^{-1} u} \leq 1,$$

from which it follows that

$$0 \leq 1 - \lambda_{min} \left(\Sigma_1^\beta \left(\delta \left(\Sigma_1^\beta \right) \right)^{-1} \right) < 1.$$

A similar argument shows that

$$0 \leq 1 - \lambda_{min} \left(\Sigma_2^\beta \left(\delta \left(\Sigma_2^\beta \right) \right)^{-1} \right) < 1.$$

Hence, for all $\beta > U_{\Sigma_1, \Sigma_2}$,

$$\prod_{i=1}^{2} \left(1 - \lambda_{min} \left(\Sigma_i^\beta \left(\delta \left(\Sigma_i^\beta \right) \right)^{-1} \right) \right) \vee \left(\lambda_{max} \left(\Sigma_i^\beta \delta \left(\left(\Sigma_i^\beta \right)^{-1} \right) \right) - 1 \right) < 1.$$

The desired result follows from Proposition 3. ■

H Proof of Proposition 5

For any symmetric positive definite matrix K, let $\delta_{max}(K)$ denote the largest diagonal element of K. Since

$$\lambda_{min}(K) = \min_{\|u\|_2=1} u'Ku \quad \text{and} \quad \lambda_{max}(K) = \max_{\|u\|_2=1} u'Ku,$$

it follows that

$$\lambda_{min}(K) \leq \delta_{max}(K) \leq \lambda_{max}(K).$$

Application of Lemma 19 and the above inequality implies that

$$\begin{aligned} \lambda_{max} \left(\Sigma_1 \delta \left(\Sigma_1^{-1} \right) \right) &= \max_{\|u\|_2=1} \frac{u' \delta \left(\Sigma_1^{-1} \right) u}{u' \Sigma_1^{-1} u} \\ &\leq \frac{\max_{\|u\|_2=1} u' \delta \left(\Sigma_1^{-1} \right) u}{\min_{\|u\|_2=1} u' \Sigma_1^{-1} u} \\ &= \frac{\delta_{max} \left(\Sigma_1^{-1} \right)}{\lambda_{min} \left(\Sigma_1^{-1} \right)} \\ &\leq \frac{\lambda_{max} \left(\Sigma_1^{-1} \right)}{\lambda_{min} \left(\Sigma_1^{-1} \right)} \\ &= \frac{\lambda_{max}(\Sigma_1)}{\lambda_{min}(\Sigma_1)}. \end{aligned}$$

An analogous argument shows that

$$\begin{aligned}
\lambda_{min}\left(\Sigma_1\left(\delta\left(\Sigma_1\right)\right)^{-1}\right) &= \min_{\|u\|_2=1}\frac{u'\left(\delta\left(\Sigma_1\right)\right)^{-1}u}{u'\Sigma_1^{-1}u} \\
&\geq \frac{\min_{\|u\|_2=1}u'\left(\delta\left(\Sigma_1\right)\right)^{-1}u}{\max_{\|u\|_2=1}u'\Sigma_1^{-1}u} \\
&= \frac{1/\delta_{max}\left(\Sigma_1\right)}{\lambda_{max}\left(\Sigma_1^{-1}\right)} \\
&\geq \frac{1/\lambda_{max}\left(\Sigma_1\right)}{\lambda_{max}\left(\Sigma_1^{-1}\right)} \\
&= \frac{\lambda_{min}\left(\Sigma_1\right)}{\lambda_{max}\left(\Sigma_1\right)}.
\end{aligned}$$

For $i = 1, 2$, let L_i and U_i be defined as in Proposition 3. Since

$$L_1 = 1 - \lambda_{min}\left(\Sigma_1\left(\delta\left(\Sigma_1\right)\right)^{-1}\right) \quad \text{and} \quad U_1 = \lambda_{max}\left(\Sigma_1\delta\left(\Sigma_1^{-1}\right)\right) - 1,$$

it follows that

$$L_1 \vee U_1 \leq \left(1 - \frac{\lambda_{min}\left(\Sigma_1\right)}{\lambda_{max}\left(\Sigma_1\right)}\right) \vee \left(\frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)} - 1\right) = \frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)} - 1.$$

Using exactly the same argument as above, one can show that

$$L_2 \vee U_2 \leq \frac{\lambda_{max}\left(\Sigma_2\right)}{\lambda_{min}\left(\Sigma_2\right)} - 1.$$

Suppose that

$$\frac{\lambda_{min}\left(\Sigma_1\right)}{\lambda_{max}\left(\Sigma_1\right)} + \frac{\lambda_{min}\left(\Sigma_2\right)}{\lambda_{max}\left(\Sigma_2\right)} > 1.$$

After multiplying the above inequality by $\lambda_{max}\left(\Sigma_1\right)\lambda_{max}\left(\Sigma_2\right)/\lambda_{min}\left(\Sigma_1\right)\lambda_{min}\left(\Sigma_2\right)$, we have

$$\frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)} + \frac{\lambda_{max}\left(\Sigma_2\right)}{\lambda_{min}\left(\Sigma_2\right)} > \frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)}\frac{\lambda_{max}\left(\Sigma_2\right)}{\lambda_{min}\left(\Sigma_2\right)}.$$

Therefore,

$$\frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)}\frac{\lambda_{max}\left(\Sigma_2\right)}{\lambda_{min}\left(\Sigma_2\right)} - \frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)} - \frac{\lambda_{max}\left(\Sigma_2\right)}{\lambda_{min}\left(\Sigma_2\right)} + 1 < 1,$$

or equivalently,

$$\left(\frac{\lambda_{max}\left(\Sigma_1\right)}{\lambda_{min}\left(\Sigma_1\right)} - 1\right)\left(\frac{\lambda_{max}\left(\Sigma_2\right)}{\lambda_{min}\left(\Sigma_2\right)} - 1\right) < 1.$$

This implies that

$$\prod_{i=1}^{2}\left(L_i \vee U_i\right) < 1.$$

The desired result follows from Proposition 3. ■

References

[1] D. Agrawal and A. Vardy, "The Turbo Decoding Algorithm and Its Phase Trajectories." To appear in *IEEE Trans. on Information Theory.*

[2] S. M. Aji, G. B. Horn, and R. J. McEliece, "On the convergence of iterative decoding on graphs with a single cycle," *Proc. CISS 1998*, Princeton, N.J., March 1998.

[3] S. M. Aji and R. J. McEliece, "Generalized Distributive Law." To appear in *IEEE Trans. on Information Theory.*

[4] R. Bellman, "Notes on Matrix Theory–IV: An inequality due to Bergstrom," *Am. Math. Monthly*, vol. 62, pp. 172-173, 1955.

[5] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallel concatenated coding schemes," *IEEE Trans. on Information Theory*, vol. 42, 2, pp. 409-428, Mar. 1996.

[6] G. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding: Turbo codes," *Proc. 1993 Int. Conf. Commun.*, Geneva, Switzerland, May 1993, pp. 1064-1070.

[7] H. El Gamal and A. R. Hammons, "Analyzing the Turbo Decoder Using the Gaussian Approximation." Submitted to *IEEE Trans. on Information Theory.*

[8] G. Forney, F. Kschischang, and B. Marcus, "Iterative decoding of tail-biting trelisses," presented at the 1998 Information Theory Workshop. San Diego: Feb. 9 - 11, 1998, pp. 11-12.

[9] B. Frey, "Turbo Factor Analysis," *Proc. Advances in Neural Information Processing Systems 12*, Dec. 1999.

[10] B. Frey, "Turbo Factor Analysis," Submitted to *Neural Computation*, May 1999.

[11] B. Frey and F. Kschischang, "Probability propagation and iterative decoding," *Proc. 34th Annual Allerton Conf. on Communications, Control, and Computing*, pp. 482-493, October 1996.

[12] R. G. Gallager, *Low–Density Parity–Check Codes.* Cambridge, MA: MIT Press, 1963.

[13] F. V. Jensen, *An Introduction to Bayesian Networks.* New York: Springer-Verlag, 1996.

[14] F. Kschischang and B. Frey, "Iterative Decoding of Compound Codes by Probability Propagation in Graphical Models," *IEEE Journal on Selected Areas in Commun.*, vol. 16, 1, Jan. 1998.

[15] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm." To appear in the *IEEE Trans. on Information Theory.*

[16] S. L. Lauritzen and D. J. Spiegelhalter, "Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems," *Journal of the Royal Statistical Society, Series B*, vol. 50, pp. 157-224, 1988.

[17] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman, "Analysis of Low Density Codes and Improved Designs Using Irregular Graphs." *Proc. 30th ACM STOC*, May 23-26, 1998.

[18] D. J. C. MacKay, "Good Error-Correcting Codes Based on Very Sparse Matrices," *IEEE Trans. on Information Theory*, vol. 45, 2, pp. 399-431, Mar. 1999.

[19] R. J. McEliece, D. J. C. MacKay, and J-F. Cheng, "Turbo Decoding as an Instance of Pearl's "Belief Propagation" Algorithm," *IEEE Journal on Selected Areas in Commun.*, vol. 16, 2, pp. 140-152, Feb. 1998.

[20] R. J. McEliece, "Coding Theory and Probability Propagation in Loopy Bayesian Networks," invited talk at the *13th Conf. on Uncertainty in Artificial Intelligence*, Aug. 1997.

[21] K. Murphy, Y. Weiss, and M. Jordan, "Loopy-belief propagation for approximate inference: an empirical study." *Proceedings of the 15th Conf. on Uncertainty in Artificial Intelligence*, Aug. 1999.

[22] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann, 1988.

[23] T. Richardson, "The Geometry of Turbo-Decoding Dynamics." To appear in the *IEEE Trans. on Information Theory.*

[24] T. Richardson, A. Shokrollahi, and R. Urbanke, "Design of Provably Good Low-Density Parity Check Codes." Submitted to the *IEEE Trans. on Information Theory.*

[25] T. Richardson and R. Urbanke, "The Capacity of Low-Density Parity Check Codes under Message-Passing Decoding." Submitted to the *IEEE Trans. on Information Theory.*

[26] P. Rusmevichientong, B. Van Roy, "An Analysis of Turbo Decoding with Gaussian Densities," *Proc. Advances in Neural Information Processing Systems 12*, Dec. 1999.

[27] G. Strang, *Linear Algebra and Its Applications.* Orlando, FL: Harcourt Brace Jovanovich, 1988.

[28] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Computation*, 12, pp.1-41, 2000.

[29] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *Proc. Advances in Neural Information Processing Systems 12*, Dec. 1999.

[30] Y. Weiss and W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology." Available at *http://www.cs.berkeley.edu /~ywoeiss/gaussTR.pdf.*

[31] N. Wiberg, *Codes and Decoding on General Graphs.* PhD thesis, Linköping University, Sweden, 1996.

[32] N. Wiberg, H.-A. Loeliger, and R. Kötter, "Codes and iterative decoding on general graphs," *European Trans. on Telecommun.*, vol. 6, pp. 513-525, Sep./Oct. 1995.

Biographies

Benjamin Van Roy

Benjamin Van Roy received the SB degree in computer science and engineering and the SM and PhD degrees in electrical engineering and computer science, all from the Massachusetts Institute of Technology. He is currently Assistant Professor in the Departments of Management Science and Engineering and Electrical Engineering at Stanford University, with a courtesy appointment in the Department of Computer Science. His research interests revolve around problems of information processing and decision making in complex systems.

He has received several awards, including a Digital Equipment Corporation Scholarship (1992-1994), the MIT George C. Newton Award (1993), for the best undergraduate electrical engineering laboratory project, the MIT Morris J. Levin Memorial Award (1995), for an outstanding Master's thesis, the MIT George M. Sprowls Award (1998), for the best doctoral dissertation in computer science, and an NFS CAREER Award (2000). In 1997, a software product developed by his team at the Unica Corporation was named Call Center Magazine Product of the Year. At Stanford, he has been named a Frederick E. Terman Fellow and a David Morgenthaler II Faculty Scholar.

Paat Rusmevichientong

Paat Rusmevichientong received the BA degree in mathematics from the University of California, Berkeley in 1997. He is currently a graduate student in the Department of Management Science and Engineering at Stanford University. His research interests include tractable inference in complex systems and decentralized decision making.

He has received the University of California Regents and Chancellors Scholarship (1995-1997), the UC Berkeley Dorothea Klumpke Roberts Prize (1997), for truly exceptional scholarship in mathematics, and the Stanford Graduate Fellowship (1997-present).