Proceedings of Machine Learning Research vol 178:1–71, 2022 35th Annual Conference on Learning Theory

Where Does EM Converge in Latent Gaussian Tree Models?

Yuval Dagan DAGAN@CSAIL.MIT.EDU
MIT, EECS

Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
MIT, EECS

Anthimos Vardis Kandiros KANDIROS@MIT.EDU
MIT, EECS

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the optimization landscape of the log-likelihood function and the convergence of the Expectation-Maximization (EM) algorithm in latent Gaussian tree models, i.e. tree-structured Gaussian graphical models whose leaf nodes are observable and non-leaf nodes are unobservable. We show that the unique non-trivial stationary point of the population log-likelihood is its global maximum, and establish that the expectation-maximization algorithm is guaranteed to converge to it in the single latent variable case. Our results for the landscape of the log-likelihood function in general latent tree models provide support for the extensive practical use of maximum likelihood based-methods in this setting. Our results for the EM algorithm extend an emerging line of work on obtaining global convergence guarantees for this celebrated algorithm. We show our results for the non-trivial stationary points of the log-likelihood by arguing that a certain system of polynomial equations obtained from the EM updates has a unique non-trivial solution. The global convergence of the EM algorithm follows by arguing that all trivial fixed points are higher-order saddle points.

1. Introduction

Estimating latent variable models is a widely-studied task in Statistics and Machine Learning. It is also a daunting one, computationally and statistically. Even if the underlying, fully observable distribution is an exponential family and therefore has a concave log-likelihood function, marginalizing out the latent variables makes the log-likelihood non-concave, in most cases. In the same exponential-family setting, under mild conditions, the population (i.e. infinite sample) log-likelihood of the complete model has a unique maximum at the true model parameters, yet even in this setting very little is understood about the landscape of the partially observable model's log-likelihood or its stationary points.

A widely-applicable method for estimating latent variable models is the Expectation-Maximization (EM) algorithm of Dempster et al. (1977). Given a parametric family of distributions $\{p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})\}_{\boldsymbol{\theta} \in \Theta}$, where variables $\boldsymbol{X}$ are observable and variables $\boldsymbol{Y}$ are unobservable, and given independent observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ from some model in this family, the EM algorithm starts with some initialization $\boldsymbol{\theta}^{(0)}$ and iteratively performs a sequence of interleaved "E-steps" and "M-steps," a consecutive pair of which are called an "EM update." Specifically, for all $t \geq 0$, the algorithm updates the current vector of parameters $\boldsymbol{\theta}^{(t)}$ by performing the following:

- **(E-step)** for each sample i, compute a posterior belief about the values of the unobservable variables by setting, for all $\boldsymbol{y}$, $Q_i^{(t)}(\boldsymbol{y}) = p_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}_i)$;
- **(M-step)** update the parameters to $\boldsymbol{\theta}^{(t+1)} \in \arg\max_{\boldsymbol{\theta}} \sum_i \int_{\boldsymbol{y}} Q_i^{(t)}(\boldsymbol{y}) \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{X}=\boldsymbol{x}_i, \boldsymbol{Y}=\boldsymbol{y})}{Q_i^{(t)}(\boldsymbol{y})} d\boldsymbol{y}$.

Notice that, by Jensen's inequality, the function maximized in the M-step of the algorithm lower bounds the log-likelihood of the samples and, by the choice made in the E-step of the algorithm, this lower bound equals the log-likelihood of the samples at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Thus, whenever the EM update results in $\boldsymbol{\theta}^{(t+1)} \neq \boldsymbol{\theta}^{(t)}$, the likelihood of the samples increases. Moreover, when the class of models $\{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta} \in \Theta}$ is an exponential family, the M-step becomes a concave maximization problem, making the algorithm quite attractive in this setting.

Despite its wide use and study, with north of 66k citations, relatively little is known about EM's behavior. Conditions have been identified under which the EM iterates converge to or cluster at stationary points of the log-likelihood — see e.g. Wu (1983); Tseng (2004); Chrétien and Hero (2008), or exhibit *local* convergence to the maximum of the likelihood — see e.g. Redner and Walker (1984); Balakrishnan et al. (2017); Zhao et al. (2020); Kwon and Caramanis (2020a). Conditions under which EM exhibits *global* convergence to the maximum of the likelihood are rare (Wu, 1983) with a surge of recent results inching towards establishing global convergence guarantees in more and more settings — from balanced mixtures of two Gaussians (Xu et al., 2016; Daskalakis et al., 2017) to balanced mixtures of two truncated Gaussians (Nagarajan and Panageas, 2020), balanced mixtures of two Laplace distributions (Barazandeh and Razaviyayn, 2018), unbalanced mixtures of two Gaussians (Xu et al., 2018), binary variable naive Bayes models (Daskalakis et al., 2018), and mixtures of linear regression models (Kwon et al., 2019; Klusowski et al., 2019; Kwon and Caramanis, 2020b; Kwon et al., 2021) — and towards understanding the role of overparametrization in EM's global convergence (Xu et al., 2018; Dwivedi et al., 2020).

To the best of our knowledge, recent works on the global convergence of EM are for single-latent-variable models. Extending this recent line of work, our paper studies the convergence of EM in Latent Gaussian Tree Models (LGTMs), i.e. tree-structured Gaussian Graphical Models whose leaf variables are observable and non-leaf variables are unobservable. Latent tree models in general, and LGTMs in particular have found wide use in scientific and applied domains due to their combined expressiveness and tractability of inference; see e.g. Mourad et al. (2013); Zwiernik (2018) for recent surveys. Some notable applications of LGTMs are in phylogenetics, where they have been used to model the evolution of continuous traits (Felsenstein, 1973; Hiscott et al., 2016; Truell et al., 2021), in network tomography, to model network delays (Castro et al., 2004; Eriksson et al., 2010; Bhamidi et al., 2010), and in linguistics, for modelling the evolution of languages using acoustic data (Ringe et al., 2002; Shiers et al., 2017).

Given observations from a latent tree model, a long line of research has studied whether the structure of the model can be recovered and, if the structure is known, whether the parameters of the model can be recovered. Most techniques with provable guarantees are based on defining and estimating tree metrics from the samples; see e.g. Erdős et al. (1999); Felsenstein (2004); Daskalakis et al. (2006); Roch (2006, 2010); Roch and Sly (2017) and their references. On the practical front, however, some of the most popular packages are based on maximum likelihood estimation; see e.g. Yang (1997); Stamatakis (2006). Even

when the latent structure is known, however, the landscape of the likelihood function is not well understood (Felsenstein, 1973; Truell et al., 2021). This is true even in the population limit, i.e. when infinitely many samples are available, even when the tree is trivial, i.e. there is a single latent node, and even when the latent tree model is a LGTM. For this paradigmatic and widely used family of models we study the following question:

Main Question: *Given a LGTM model $p_{\boldsymbol{\theta}^*}(\boldsymbol{X}, \boldsymbol{Y})$ on a tree T whose leaves $\boldsymbol{X}$ are observable and internal nodes $\boldsymbol{Y}$ are unobservable, can we characterize the stationary points of the population log likelihood $\ell_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{x} \sim p_{\boldsymbol{\theta}^*}(\boldsymbol{X})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x})]$? Does it have spurious stationary points $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ and under what conditions does EM converge to $\boldsymbol{\theta}^*$?*

We study the afore-described questions in the setting where all the nodes of T are single-dimensional Gaussian variables and assume, without loss of generality, that they have zero mean and that the setting is ferromagnetic, i.e. assume that for every pair of variables their correlation lies in $(0, 1)$. Our first main result is the following (the formal version is Theorem 15, combined with Lemma 14):

Informal Theorem 1 *In the setting of our main question, and that of the preceding paragraph, suppose that $\boldsymbol{\theta}$ is a stationary point of the population log-likelihood $\ell_{\boldsymbol{\theta}^*}(\cdot)$ that is non-trivial, i.e. in model $p_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{Y})$ there is no edge of the tree whose endpoints have correlation in $\{0, 1\}$. Then $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.*

Our result guarantees that, if gradient-descent, EM, or similar method converges to a stationary point of the population log-likelihood that is non-trivial, then this point indexes the true model. While there are other criteria that can be used, our result implies that the stationarity of the log-likelihood can be used as an alternative, post-hoc criterion to argue that the correct model has been identified. In particular, our result substantiates the extensive use of maximum-likelihood-based methods in practice (Yang, 1997; Stamatakis, 2006), and the experimental evidence that EM succeeds with high probability in this setting (Wang and Zhang, 2006). Next, we study whether we can guarantee that EM converges to a non-trivial stationary point in our setting. We show this for the case where there is a single latent node in the model (the formal statement is Theorem 3):

Informal Theorem 2 *In the setting Informal Theorem 1, suppose additionally that there is a single latent node. In this case, EM is guaranteed to converge to $\boldsymbol{\theta}^*$. If n is the number of leaves and ϵ is the desired accuracy for all the parameters (in absolute value), then the sample complexity is $O(\mathrm{poly}(n)/\epsilon^2)$ and the number of iterations is $O(\mathrm{poly}(n)\log(1/\epsilon))$.*

Proof Ideas. Starting with Informal Theorem 2, it is known that EM converges to some $\boldsymbol{\theta}$ that is a fixed point of the EM update. Hence, the proof follows by making the following arguments: (1) $\boldsymbol{\theta}^*$ is the only non-trivial fixed point of the EM update; if we parametrize the model via correlations on its edges this the same as saying that $\boldsymbol{\theta}^*$ is the only fixed point of the EM update in the interior of the parameter space Θ; and (2) While there are fixed points at the boundary of Θ, EM does not converge to any of those. To show (1), we analyze the EM update and show that its fixed points are solutions to a system of degree-2 polynomial equations. By using a simple special form of the Jacobian Conjecture we argue that these have a unique non-trivial solution. To show (2), we analyze the behavior

of the EM around the fixed points at the boundary, by computing the derivatives of the update rule around these points. This requires careful analysis since these fixed points are higher-order saddle points. For the case of general trees (Informal Theorem 1), we show that the fixed points are solutions of higher-degree polynomial equations. We establish a novel reduction of these algebraic equations to the second-degree polynomial system from the single-latent-model case, and use this reduction to certify that there is a unique fixed point at the interior of the parameter space, which is $\boldsymbol{\theta}^*$.

2. Tree models with one latent node

We start by elaborating on the simpler setting with one latent node. For simplicity, we start by anayzing the *population EM*, which amounts to running the EM algorithm on the whole population (rather than a finite sample). So simplify even more, we analyze asymptotic convergence, namely, in the limit as the number of iterations goes to infinity. Later, in Section 2.6, we describe the finite-sample and finite-iterate result.

2.1. Definitions and properties of the model

We consider the family of multivariate Gaussian distributions over zero-mean variables, with latent node y and observed nodes $x_1, \dots, x_n$, that have the property

$$\Pr[x_1, \dots, x_n, y] = \Pr[y] \prod_{i=1}^{n} \Pr[x_i \mid y] \ .$$

Each such distribution is uniquely defined by the following parameters:

$$\sigma_y^2 := \mathrm{Var}(y); \ \sigma_{x_i}^2 := \mathrm{Var}[x_i]; \ \rho_i := \frac{\mathrm{Cov}(x_i, y)}{\sqrt{\mathrm{Var}(x_i)\,\mathrm{Var}(y)}}, \text{ for } i = 1, \dots, n \ ,$$

We note that generally $\rho_i \in [-1, 1]$, yet, we analyze the ferromagnetic setting where $\rho_i \in [0, 1]$. Denote by $\mathcal{P}$ the set of all such distributions with $\rho_i \in [0, 1]$ and $\sigma_y, \sigma_{x_i} > 0$. We use the following convenient properties of distributions in $\mathcal{P}$:

Claim 1 *For any $i \neq j$, $\mathrm{Cov}(x_i, x_j) = \sigma_{x_i}\sigma_{x_j}\rho_i\rho_j$. Further, $\mathbb{E}[x_i \mid y] = \frac{\sigma_{x_i}\rho_i}{\sigma_y} y$ and*

$$\mathbb{E}[y \mid x_1 \cdots x_n] = \sigma_y \sum_{i=1}^{n} \lambda_i \frac{x_i}{\sigma_i}, \quad \textit{where } \lambda_i = \frac{\rho_i/(1-\rho_i^2)}{1 + \sum_{j=1}^{n} \rho_j^2/(1-\rho_j^2)} \ . \tag{1}$$

Given any distribution $\mu \in \mathcal{P}$, denote its marginals and conditionals as $\mu_{x_1 \cdots x_n}$, $\mu_{y|x_1 \cdots x_n}$ etc. Lastly, denote $\overline{\rho} = (\rho_1, \dots, \rho_n)$ and $\overline{x} = (x_1, \dots, x_n)$.

2.2. The expectation-maximization algorithm - over the population

We analyze the EM algorithm. We start by analyzing the *population EM*: this assumes that each iteration of the EM is executed over the *whole population*, rather than over a finite sample. This greatly simplifies the analysis (see Section 2.6 for the results on finite

sample). The population EM can be described as follows: we set $\mu^0 \in \mathcal{P}$ arbitrarily. Then, at any $t > 0$, define

$$\mu^{t+1} = \operatorname*{argmax}_{\mu \in \mathcal{P}} \mathop{\mathbb{E}}_{x_1 \cdots x_n \sim \mu^*_{x_1 \cdots x_n}} \mathop{\mathbb{E}}_{y \sim, \mu^t_{y|x_1 \cdots x_n}} [\log \Pr_{\mu}(x_1, \ldots, x_n, y)],$$

where $\Pr_\mu$ denotes the density with respect to μ. Denote by $\sigma^t_\cdot, \rho^t_i$ the parameters corresponding to μ^t and by λ^t_i the coefficients from (1). Similarly, $\sigma^*_\cdot, \rho^*_i$ and λ^*_i correspond to μ^*. We would like to understand how these parameters update in each iteration of the EM algorithm. For this purpose, we have the following lemma:

Lemma 1 *Let $\mu^{t,*}$ denote the joint distribution over $x_1 \cdots x_n, y$ such that*

$$\Pr_{\mu^{t,*}}[x_1, \cdots, x_n, y] = \Pr_{\mu^*}[x_1, \cdots, x_n] \Pr_{\mu^t}[y \mid x_1, \ldots, x_n].$$

Then, for any i, we have that

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \mathop{\mathbb{E}}_{\mu^{t,*}}[x_i y], \ \operatorname{Var}_{\mu^{t,*}}[x_i] = \operatorname{Var}_{\mu^{t+1}}[x_i], \ \operatorname{Var}_{\mu^{t,*}}[y] = \operatorname{Var}_{\mu^{t+1}}[y] \ .$$

Proof Notice that μ^{t+1} is the MLE over $\mathcal{P}$, given samples drawn from $\mu^{t,*}$. Hence, for any $\mu \in \mathcal{P}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{\overline{x}, y \sim \mu^{t,*}} \log \Pr_{\mu}(\overline{x}, y) &= \mathop{\mathbb{E}}_{\overline{x}, y \sim \mu^{t,*}} \log \Pr_{\mu}(y) \prod_i \Pr_{\mu}(x_i \mid y) \\ &= \mathop{\mathbb{E}}_{\overline{x}, y \sim \mu^{t,*}} \log \Pr_{\mu}(y) + \sum_i \mathop{\mathbb{E}}_{\overline{x}, y \sim \mu^{t,*}} \log \Pr_{\mu}(x_i \mid y). \end{aligned} \tag{2}$$

Recall that each $\mu \in \mathcal{P}$ can be decomposed as $\Pr_\mu[x, y] = \Pr_{\mu_y}[y] \prod_i \Pr_{\mu_{x_i|y}}[x_i \mid y]$, and each term in this decomposition can be chosen to maximize its corresponding summand from (2). By Gibbs inequality[1], the maximizing choice is obtained by selecting $\mu_y \sim \mu^{t,*}_y$ and $\mu_{x_i|y} \sim \mu^{t,*}_{x_i|y}$. This choice satisfies $\mu_{x_i y} \sim \mu_y \mu_{x_i|y} \sim \mu^{t,*}_y \mu^{t,*}_{x_i|y} \sim \mu^{t,*}_{x_i, y}$. Hence, the pairwise marginals between x_i and y are conserved, which concludes the proof. ■

As a corollary, we obtain the following update rules for the parameters σ^t_y, σ^t_i and ρ^t_i, using the analogous parameters of $\mu^{t,*}$, that are calculated using formulas for the conditional Gaussian measure (proof appears in Section A.3).

Lemma 2 *For any $i \neq j$, denote by $\Delta^t_{ij} = \rho^*_i \rho^*_j - \rho_i \rho_j$. For any $t > 0$ and any i, $\sigma^t_i = \sigma^*_i$,*

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \sigma^t_i \sigma^t_y \left(\lambda^t_i + \sum_{j \neq i} \rho^*_i \rho^*_j \lambda^t_j \right) = \sigma^t_i \sigma^t_y \left(\rho^t_i + \sum_{j \neq i} \Delta^t_{ij} \lambda^t_j \right) \ ,$$

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[y^2] = (\sigma^{t+1}_y)^2 = (\sigma^t_y)^2 \left(\sum_{i=1}^n (\lambda^t_i)^2 + \sum_{i \neq j \in \{1, \ldots, n\}} \lambda^t_i \lambda^t_j \rho^*_i \rho^*_j \right) = (\sigma^t_y)^2 \left(1 + \sum_{j \neq k} \Delta^t_{ij} \lambda^t_j \lambda^t_k \right)$$

$$\rho^{t+1}_i = \frac{\lambda^t_i + \sum_{j \neq i} \lambda^t_j \rho^*_i \rho^*_j}{\sqrt{\sum_{i=1}^n (\lambda^t_i)^2 + \sum_{i \neq j \in \{1, \ldots, n\}} \lambda^t_i \lambda^t_j \rho^*_i \rho^*_j}} = \frac{\rho^t_i + \sum_{j \neq i} \Delta_{ij} \lambda^t_j}{\sqrt{1 + \sum_{j \neq k} \Delta_{ij} \lambda^t_j \lambda^t_k}} \ . \tag{3}$$

1. Gibb's inequality states that for any distribution P, $\operatorname{argmax}_Q \mathbb{E}_{x \sim P} \log \Pr_Q[x] = P$, where Q is taken over all the probability measures. This inequality can be similarly applied on conditional distributions.

2.3. Asymptotic Convergence of the population EM

We would like to argue that the iterates of the EM converge to μ^*, which is characterized by the parameters $\sigma_y^*, \sigma_{x_i}^*$ and ρ_i^* for $i = 1, \dots, n$. Yet, notice that one only observes samples from the marginal $\mu_{x_1 \cdots x_n}^*$, which is a function of only $\sigma_{x_i}^*$ and ρ_i^* but not of σ_y^*. Hence, we cannot expect to learn σ_y^*. With regard to the other parameters, we prove:

Theorem 3 *Assume that the correlations in the underlying distribution μ^* satisfy $\rho_i^* \in (0,1)$ for all i and that the first iterate of the population EM satisfies $\rho_i^0 \in (0,1)$ for all i. Then, the iterates ρ_i^t and σ_i^t of the population EM converge to the parameters of the underlying distribution:*

$$\lim_{t\to\infty} \rho_i^t = \rho_i^*, \quad \text{and } \sigma_i^t = \sigma_i^* \text{ for all } i = 1, \dots, n \text{ and } t \geq 1 \ .$$

We will stay under the assumptions that $\rho_i^*, \rho_i^0 \in (0,1)$ for all i throughout the proof. Recall that Lemma 2 argues that for any $t \geq 1$, $\sigma_{x_i}^t = \sigma_{x_i}^*$. Hence, it remains to argue about the convergence of ρ_i^t to ρ_i^*. We use the following definition:

Definition 4 *A point $\overline{\rho} = (\rho_1, \dots, \rho_n)$ is a stationary point of the EM if $\overline{\rho}^t = \overline{\rho}$ implies that $\overline{\rho}^{t+1} = \overline{\rho}$.*

Denote the set of all stationary points by $\mathcal{S}$. The following is proved in Section A.3:

Lemma 5 *The iterates of the population EM converge to some stationary point $\overline{\rho} \in \mathcal{S}$. Further, for any $\overline{\rho} \in \mathcal{S}$, if $\overline{\rho}^t = \overline{\rho}$ then for any i, $\mu^{t+1} = \mu^t$*

Since the EM converges only to a stationary point, it is useful to characterize the set of stationary points, as stated below:

Lemma 6 *The set $\mathcal{S}$ of stationary points of the population EM equals*

$$\mathcal{S} = \{(\rho_1^*, \dots, \rho_n^*), (0, \dots, 0)\} \cup \{\overline{\rho}^{(i)} \colon i = 1, \dots, n\}, \quad \text{where } \overline{\rho}_j^{(i)} = \begin{cases} 1 & j = i \\ \rho_i^* \rho_j^* & j \neq i \end{cases} \ .$$

We will rule out the possibility of convergence to any point that is not the true parameter.

Lemma 7 *The correlation parameters of the population EM converge to a point in $(0,1)$:*

$$0 < \lim_{t\to\infty} \rho_i^t < 1 \quad \text{for all } i = 1, \dots, n \ .$$

Combining the two lemmas above, we obtain that ρ_i^t must converge to ρ_i^*, which concludes Theorem 3. In Section 2.4 we will prove that there are no stationary points where $\rho_i \in (0,1)$ for all i. In Section A.1 we will prove that the only stationary point with some $\rho_i = 0$ is $(0, \dots, 0)$, and that the algorithm does not converge to this point. In Section A.2 we will prove that the only stationary points where some $\rho_i = 1$ are $\overline{\rho}^{(i)}$ and that the algorithm does not converge to these points. In Section 2.5 we sketch the proofs from Section A.1 and Section A.2.

2.4. No stationary points at the interior

In this section, we will prove that there are no stationary points of the EM with $\rho_i \in (0,1)$ for all i, as summarized in the following lemma:

Lemma 8 *Let $\overline{\rho} = (\rho_1, \ldots, \rho_n) \in (0,1)^n$. If $\overline{\rho} \neq \overline{\rho}^*$ then $\overline{\rho}$ is not a stationary point.*

Below, fix some stationary point $\overline{\rho}^t = (\rho_1, \ldots, \rho_n) \in (0,1)^n$ and we will prove that $\overline{\rho} = \overline{\rho}^*$. By Lemma 5, $\mathbb{E}_{\mu^{t+1}}[x_i y] = \mathbb{E}_{\mu^t}[x_i y]$ for all i, and by Lemma 2 this translates to

$$\sigma^t_{x_t}\sigma^t_y\rho^t_i = \sigma^t_i\sigma^t_y\left(\rho^t_i + \sum_{j\neq i}\Delta^t_{ij}\lambda^t_j\right) \implies \sum_{j\neq i}\Delta^t_{ij}\lambda^t_j = 0 \implies \sum_{j\neq i}\rho^*_i\rho^*_j\lambda^t_j = \sum_{j\neq i}\rho^t_i\rho^t_j\lambda^t_j \ .$$

Multiplying by λ^t_i and substitute $\rho^t_i\lambda^t_i = u^t_i$ and $\rho^*_j\lambda^t_j = u^*_j$, we obtain

$$\forall i = 1,\ldots,n\colon \quad \sum_{j\neq i}\rho^*_i\rho^*_j\lambda^t_i\lambda^t_j = \sum_{j\neq i}\rho^t_i\rho^t_j\lambda^t_i\lambda^t_j \implies \sum_{j\neq i}u^t_i u^t_j = \sum_{j\neq i}u^*_i u^*_j \ . \tag{4}$$

By the assumption that $\rho^t_i, \rho^*_i \in (0,1)$ and by definition of λ_i in (1), we have that $\lambda^t_i > 0$, hence $u^t_i, u^*_i > 0$. We will prove that this set of equations imply that $\overline{u}^t = \overline{u}^*$, as summarized below:

Lemma 9 *Let $\overline{u}, \overline{v} \in (0,\infty)^n$, and assume that for all $i = 1,\ldots,n$, $\sum_{j\neq i} u_i u_j = \sum_{j\neq i} v_i v_j$. Then $\overline{u} = \overline{v}$.*

Lemma 9 will imply that $\lambda^t_i\rho^t_i = \lambda^t_i\rho^*_i$ which implies that $\rho^t_i = \rho^*_i$ as required to conclude the proof of Lemma 8.

In the remainder of this section, we prove Lemma 9. For this purpose, we have the following lemma, which is a special case of the Jacobian Conjecture on the uniqueness of solutions of polynomial systems. Its special case for degree-2 polynomials was proven by Wang (1980). We state and prove a corollary of this statement:

Lemma 10 *Let $p_1, \ldots, p_n\colon \mathbb{R}^n \to \mathbb{R}$ a collection of quadratic polynomials. If there are two distinct vectors, $\overline{u}$ and $\overline{u}'$ such that $p_j(\overline{u}) = p_j(\overline{u}')$ for all j, then, the Jacobian matrix of $\overline{p}$ computed at $(\overline{u}+\overline{u}')/2$, $J^{\overline{p}}((\overline{u}+\overline{u}')/2)$, is singular, where*

$$J^{\overline{p}}_{i,j}(\overline{v}) = \frac{dp_i(\overline{v})}{dv_j}.$$

Proof Look at the path $\gamma\colon [0,1] \to \mathbb{R}^n$ defined by $\gamma(s) = s\overline{u} + (1-s)\overline{u}'$. Then, $p_i(\gamma(s))$ is a quadratic polynomial in s, that satisfies $p_i(\gamma(0)) = p_i(\gamma(1))$. Therefore, $1/2$ is a stationary point, hence

$$0 = \frac{d}{ds}p_i(\gamma(s))\bigg|_{s=1/2} = \sum_j \frac{dp_i}{dv_j}\bigg|_{\overline{v}=\gamma(1/2)}\frac{d\gamma_j}{ds}\bigg|_{s=1/2} = \sum_j \frac{dp_i}{dv_j}\bigg|_{\overline{v}=\gamma(1/2)}(u'_j - u_j).$$

Combining these equalities in a matrix form for all $i = 1,\ldots,n$, we derive that $0 = J^{\overline{p}}(\overline{\gamma}(1/2))(\overline{u}' - \overline{u}) = J^{\overline{p}}((\overline{u}+\overline{u}')/2)(\overline{u}' - \overline{u})$ hence $J^{\overline{p}}((\overline{u}+\overline{u}')/2)$ is singular. ■

To complete the proof of Lemma 9, let us substitute $p_i(\overline{u}) = \sum_{j\neq i} u_i u_j$. The following lemma proves that the Jacobian matrix of this system is non-singular at any point $\overline{u}$ with positive entries, which suffices to conclude the proof:

Lemma 11 *Let $u_1, \dots, u_n > 0$. Then the following matrix J is non singular:*

$$J_{ij} = \begin{cases} \sum_{k \neq i} u_k & i = j \\ u_i & i \neq j \end{cases}. \tag{5}$$

Notice that this matrix can be written as $D + \overline{v}\overline{w}^\top$, where D is diagonal and $\overline{v}, \overline{w} \in \mathbb{R}^n$. Such a matrix is singular if and only if $1 + \overline{v}^\top A^{-1} \overline{w} = 0$ Sherman and Morrison (1950). Proving that this matrix is non-singular from this formula is not immediate. For completeness, we present in Section A.3 a proof that does not rely on this formula. This concludes Lemma 9 and Lemma 8 follows.

2.5. The limiting point is bounded away from the boundary

In this section, we sketch the proof that $0 < \lim_{t\to\infty} \rho_i^t < 1$, whose full version appears in Section A.1 and Section A.2. First, by inspecting the update rule from Lemma 2, it is easy to verify that the only stationary points with some $\rho_i \in \{0, 1\}$ are $\overline{0} = (0, \dots, 0)$ and $\overline{\rho}^{(i)}$ (defined formally in Lemma 6). It remains to prove that the EM does not converge to these stationary points. To argue about $\overline{0}$, we write the Taylor series of ρ_i^{t+1} around 0 and obtain

$$\rho_i^{t+1} \geq \rho_i^t + \sum_{j \neq i} \rho_j^t \rho_i^* \rho_j^* - O\left(\sum_{j,k} \rho_j^t \rho_k^t\right) \geq \rho_i^t + \Omega\left(\max_{j \neq i} \rho_j^t\right) - O\left(\max_{j \in \{1,\dots,n\}} (\rho_j^t)^2\right).$$

This can be used to show that whenever $\overline{\rho}^t$ approaches $\overline{0}$, in the subsequent iteration it will repel from $\overline{0}$ and particularly, $\sum_i \rho_i^{t+1} \geq \sum_i \rho_i^t$. Hence, the algorithm cannot converge to $\overline{0}$.

Next, we explain why EM does not converge to $\overline{\rho}^{(i)}$. Assume w.l.o.g that $i = 1$ and recall that $\overline{\rho}^{(1)} = (1, \rho_1^* \rho_2^*, \dots, \rho_1^* \rho_n^*)$. The proof is more involved than the proof at $\overline{0}$, due to (1) one is required to derive a second-order Taylor series (compared to a first-order computed around $\overline{0}$), and (2) one would like to argue that if $\overline{\rho}^t \approx \overline{\rho}^{(1)}$ then $\|\overline{\rho}^{t+1} - \overline{\rho}^{(1)}\| \geq \|\overline{\rho}^t - \overline{\rho}^{(1)}\|$ (in some norm). Yet, this is true only for some values of $\overline{\rho}^t$: in particular, this holds whenever $\max_i |\rho_i^t - \rho_i^{(1)}| \leq O(|\rho_1^t - \rho_1^{(1)}|)$. By analyzing the update step of the EM, it can be shown that even if ρ^t does not satisfy this condition, which might happen with a bad initialization ($t = 0$), ρ^{t+1} will always satisfy it and the proof follows.

2.6. Finite sample and finite iteration

In this section we assume that m draws from the marginal distribution over the leaves are given, where sample i is denoted by $(x_1^{(i)}, \dots, x_n^{(i)})$, for $i \in \{1, \dots, m\}$. Then, the EM iterate is defined by

$$\mu^{t+1} = \operatorname*{argmax}_{\mu \in \mathcal{P}} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\substack{y \sim \mu^t \\ y | x_1^{(i)} \cdots x_n^{(i)}}} \left[\log \Pr_{\mu}\left(x_1^{(i)}, \dots, x_n^{(i)}, y\right)\right].$$

We have the following theorem, which bounds the sample complexity and the convergence time of the population EM:

Theorem 12 *Let $\alpha, \beta > 0$ and assume that $\theta_i^0, \theta_i^* \geq \alpha$ for all i and $\theta_i^0, \theta_i^* \leq 1 - \beta$ for all i. Then, there exist constants $C_1(\alpha, \beta), C_3(\alpha, \beta)$ that depend only on α and β and universal constants C_2, C_4 such that the following holds: let $\epsilon, \delta > 0$ and assume that the number of samples m is at least $m \geq (C_1(\alpha, \beta) n^{C_2} + \log(1/\delta))/\epsilon^2$ and let $T \geq C_3(\alpha, \beta) n^{C_4} \log(1/\epsilon)$. Then, with probability $1 - \delta$ over the sample, for any $t \geq T$, $|\rho_i^t - \rho_i^*| \leq \epsilon$ and $|\sigma_i^t - \sigma_i^*| \leq \epsilon$ for all $i \in \{1, \ldots, n\}$.*

3. General Tree models

We now extend the analysis for more complicated tree topologies.

3.1. Definitions and properties of the model

We consider a multivariate Gaussian latent-tree distribution, that is characterized by a tree $G = (V, E)$ with n leaves $x_1, \ldots, x_n$ (of degree 1) and m internal nodes $y_1, \ldots, y_n$ (of degree at least 2). The vertices $u \in V$ corresponds to random variables $\{z_u \colon u \in V\}$ with the joint distribution

$$\Pr[z_1, \ldots, z_{n+m}] = \prod_{(i,j) \in E} \Pr[z_i, z_j] \tag{6}$$

where z_{x_i} are observed by the algorithm and z_{y_i} are latent variables. The variables are assumed zero-mean and the distribution is uniquely characterized by the variance parameters $\sigma_u^2 := \mathrm{Var}[Z_u]$ for $u \in V$ and the correlation parameters $\rho_{uv} := \mathrm{Cov}[Z_u Z_v] / \sqrt{\mathrm{Var}(Z_u)\,\mathrm{Var}(Z_v)}$ for each edge $(u, v) \in E$. As with one latent, we assume $\rho_{uv} \in [0, 1]$. The correlation between any two nodes $u, v \in E$ can be expressed as the product of correlations along the path connecting them,

$$\rho_{z_i z_j} := \frac{\mathbb{E}[z_i z_j]}{\sigma_{z_i} \sigma_{z_j}} = \prod_{(z_u, z_v) \in P(z_i, z_j)} \rho_{z_u z_v}, \quad \text{where } P(z_i, z_j) \text{ is the path connecting } i \text{ and } j.$$

Similarly to the argument regarding topologies of one latent node, the variances of the latent nodes $\sigma_{y_j}^2$ cannot be estimated (see Remark 66).

An equivalent way to characterize the distribution is through the exponential family parametrization $(J = -\Sigma^{-1}, h)$, where $\Pr[z_1 \cdots z_{n+m}] \propto \exp\left(\sum_{ij} J_{ij} z_i z_j / 2 + \sum_i h_i z_i\right)$. Since the distribution factorizes as (6), the only non-zero entries J_{ij} correspond to edges $(i, j) \in E$ and for diagonal elements $i = j$. While in the original model $h_i = 0$ since the random variables are zero mean, in the conditional distribution of y given x this is no longer true and $h_y = -J_{yx} x$.

Lastly, we remark that the parameters are information-theoretically identifiable only when all the latent nodes have degree at least 3, yet, any graph can be modified to to satisfy this property, while retaining the distribution over observables (see Remark 67).

3.2. EM and likelihood for general trees

In this section, we analyze the EM iteration on latent tree models (see Section 2.2 for an elaborate exposition). Given an initial point $\mu^0 \in \mathcal{P}_G$, its iteration μ^t is defined as:

$$\mu^{t+1} = \max_{\mu \in \mathcal{P}_G} \mathop{\mathbb{E}}_{x_1 \cdots x_n \sim \mu^*_{x_1 \cdots x_n}} \mathop{\mathbb{E}}_{y_1, \cdots, y_m \sim \mu^t_{y|x_1 \cdots x_n}} [\log \Pr_{\mu}(x_1, \ldots, x_n, y_1, \ldots, y_m)], \tag{7}$$

where $\Pr_\mu$ denotes the density with respect to μ. Denote by $\sigma_\cdot^t, \rho_i^t, J^t$ the parameters corresponding to μ^t. Similarly, $\sigma_\cdot^*, \rho_i^*$ and λ^* correspond to μ^*. Using the same arguments as in Lemma 1, the following can be shown:

Lemma 13 *Let $\mu^{t,*}$ denote the joint distribution over $x_1 \cdots x_n, y_1, \dots, y_m$ such that*

$$\Pr_{\mu^{t,*}}[x_1, \cdots, x_n, y_1, \dots, y_m] = \Pr_{\mu^*}[x_1, \cdots, x_n] \Pr_{\mu^t}[y_1, \dots, y_m \mid x_1, \dots, x_n].$$

Then, for any $u \in V$ and any $(v, w) \in E$,

$$\mathrm{Cov}_{\mu^{t+1}}(z_v, z_w) = \mathrm{Cov}_{\mu^{t,*}}(z_v, z_w), \ \ \mathrm{Var}_{\mu^{t,*}}[z_u] = \mathrm{Var}_{\mu^{t+1}}[z_u] \ .$$

Denote by $\mathcal{S}$ the set of fixpoints of the EM, namely, the points ρ such that $\overline{\rho}^t = \overline{\rho}$ implies $\overline{\rho}^{t+1} = \overline{\rho}$. Analogously to Lemma 5, these correspond to the set of fixpoints μ of (7). We can equivalently consider the EM iteration under the natural parametrization J, h, as discussed in Section 3.1. Hence, a fixed point $\tilde{\rho}$ corresponds to some $\tilde{J}$, meaning that $\mathcal{S}$ remains unchanged if we change parametrization. The reason we choose the parameter J is that the likelihood has a more convenient form as an exponential family. The importance of $\mathcal{S}$ is further exemplified by the following lemma, which states that the notions of EM fixpoints and stationary points of the log-likelihood are equivalent. The proof is folklore.

Lemma 14 *Let $\mu^* \in P_G$ be such that $\rho_{ij}^* \in (0,1)$ for all $(i,j) \in E$ and define $L(J) := \mathbb{E}_{x \sim \mu^*} \log \Pr_{\mu(J)}(x)$. Then, for any $\tilde{J} \in \mathbb{R}_+^{(n+m)\times(n+m)}$ we have that $\nabla L(\tilde{J}) = 0$ if and only if $\tilde{J}$ is a stationary point of the update rule (7).*

3.3. Uniqueness of EM fixpoints for general trees

In this section, we prove that the only fixpoint of EM with non-degenerate edge weights is the true model μ^*. A detailed proof of all the claims in this Section is given in Section F.

Theorem 15 *Let $G = (V, E)$ be a tree and $\mu^* \in P_G$ be a distribution with $\rho_{ij}^* \in (0,1)$ for all $(i,j) \in E$. Suppose $\tilde{\rho}$ is a stationary point of the EM update rule (49) with $\tilde{\rho}_{ij} \in (0,1)$ for all $(i,j) \in E$. Then $\tilde{\rho}_{ij} = \rho_{ij}^*$ for all $(i,j) \in E$.*

We will denote by $\mu^{*,t}$ the distribution defined as $\mu^{*,t}(x,y) := \mu^*(x)\tilde{\mu}(y|x)$. We begin by exploring a simple implication of the fixpoint conditions, according to the rules (7). The proof follows by noticing that the conditional distribution of y given x is the same in $\tilde{\mu}$ and $\mu^{*,t}$.

Lemma 16 *Let $\mu^*, \tilde{\mu}$ be the distributions defined in Theorem 15. Then, for any internal nodes y_1, y_2 that are connected by and edge in G*

$$\mathop{\mathbb{E}}_{\mu^*}\left[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]\right] = \mathop{\mathbb{E}}_{\tilde{\mu}}\left[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]\right]$$
$$\mathop{\mathbb{E}}_{\mu^*}\left[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]^2\right] = \mathop{\mathbb{E}}_{\tilde{\mu}}\left[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]^2\right], \ \ \mathop{\mathbb{E}}_{\mu^*}\left[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]^2\right] = \mathop{\mathbb{E}}_{\tilde{\mu}}\left[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]^2\right]$$

Hence, all we have to do is compute these conditional expectations in a suitable way, so as to reveal some structure. Let N be the set of neighbors of y_1 in G and assume w.l.o.g. that they are all non-leaf nodes (the case of leaf neighbors is even easier). The first step will be to marginalize all the other non-leaf nodes except y_1 and y_2. Denote by Y^c the set of non-leaf nodes that are not neighbors of y_1. Then, let h'_{y_1}, h'_{y_2} be the information parameters of y_1, y_2 after the marginalization of Y^c. Since no neighbor of y_1 was marginalized, we will have $h'_{y_1} = 0$. Suppose we remove the edge (y_1, y_2) from the tree. Then the set of leaves is partitioned into two subsets. Call S_{y_2} the subset that is connected to y_2 after the removal. We show that h'_{y_2} is a linear combination of the values of the leaves in S_{y_2}. To do that, we utilize the marginalization formulas for Gaussian distributions, as well as the fact that $\tilde{J}^{-1}_{Y^cY^c}$ can be thought of as a covariance matrix of some Gaussian tree distribution and hence satisfies the multiplication over paths property.

Lemma 17 *Let y_2 be a non-leaf neighbor of y_1 and S_{y_2} be the corresponding set of leaves for the partition that y_2 belongs to. Then, the quantity $h'_{y_2} = h_{y_2} - \tilde{J}_{y_2Y^c}\tilde{J}^{-1}_{Y^cY^c}h_{Y^c}$ is a linear combination of the leaves in S_{y_2}.*

A similar result holds for any neighbor y_i of y_1. Our strategy is to build a system of equations similar to the one in Lemma 9. But in the proof of one latent, the variables of the system corresponded to the covariances of the latent node with individual leaves. Here, for neighbor y_2 we will define some variable that depends on a linear combination of the paths leading to leaves in S_{y_2}. In this direction, let's define the vector $H \in \mathbb{R}^s$, which has one entry for each node in N. We define $H_i = h'_{y_i}$.

So far we have marginalized on everyone except nodes in N. To gain specific information about the interaction of y_1, y_2, we now marginalize also over the neighbors of y_1, except y_2. Denote by N_2 that set of neighbors. Let $h''_{y_1}, h''_{y_2}, \tilde{J}''$ be the resulting parametrizations. From standard properties of the Gaussian, we can compute these as follows

$$h''_{y_1} = h_{y_1} - \tilde{J}_{y_1N_2}(\tilde{J}'_{N_2N_2})^{-1}h'_{N_2} = \sum_{y_j \in N, y_j \neq y_2} \frac{\tilde{J}_{y_1y_j}}{\tilde{J}'_{y_jy_j}} h'_{y_j},\ h''_{y_2} = h'_{y_2} - \tilde{J}_{y_2N_2}(\tilde{J}'_{N_2N_2})^{-1}h'_{N_2} = h'_{y_2}$$

This follows by the fact that the matrix $\tilde{J}'_{N_2N_2}$ is diagonal, since the neighbors of y_1 are not connected to each other, and $J_{y_2N_2} = 0$. To write this more compactly, we introduce the vector $r \in \mathbb{R}^s$, where $r_i = \tilde{J}_{y_1y_i}/\tilde{J}'_{y_iy_i}$ if $i \in N$. Notice that $r_i \neq 0$ always. Hence, this relation becomes

$$h''_{y_1} = \sum_{i \neq y_2} r_i H_i \tag{8}$$

We are now close to arriving at the desired form of the polynomial equations. Since we have marginalized over all nodes except y_1, y_2, we can now determine the conditional expectations by the formula

$$\begin{pmatrix} \mathbb{E}_{\tilde{\mu}}[y_1|x] \\ \mathbb{E}_{\tilde{\mu}}[y_2|x] \end{pmatrix} = \tilde{\Sigma}_{y_1y_2|x} \begin{pmatrix} h''_{y_1} \\ h''_{y_2} \end{pmatrix}$$

that connects the two parametrizations in a Gaussian. Here, $\tilde{\Sigma}_{y_1y_2|x}$ is the 2×2 covariance matrix of the conditional distribution of y_1, y_2 given x. The reason we have used $\tilde{\Sigma}_{y_1y_2|x}$ is that the covariance matrix does not change when we marginalize some nodes. Suppose

$$\tilde{\Sigma}_{y_1y_2|x} = \begin{pmatrix} c_1 & c_2 \\ c_3 & c_4 \end{pmatrix}$$

with $c_1c_4 - c_2c_3 \neq 0$. The reason the variances are not necessarily 1 is that we are now in the conditional model. Then, the first fixpoing condition of Lemma 16 translates to the following:

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[(c_1h''_{y_1} + c_2h''_{y_2})(c_3h''_{y_1} + c_4h_{y_2})] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(c_1h''_{y_1} + c_2h''_{y_2})(c_3h''_{y_1} + c_4h_{y_2})]$$

We observe that this relation can be viewed as a linear equation in terms of the variables $\mathbb{E}_{x\sim\mu^*}[h_1^2] - \mathbb{E}_{x\sim\tilde{\mu}}[h_1^2]$, $\mathbb{E}_{x\sim\mu^*}[h_2^2] - \mathbb{E}_{x\sim\tilde{\mu}}[h_2^2]$, $\mathbb{E}_{x\sim\mu^*}[h_1h_2] - \mathbb{E}_{x\sim\tilde{\mu}}[h_1h_2]$. The coefficients of these variables are functions of c_1, c_2, c_3, c_4. We can obtain two other such equations from the other two conditions of Lemma 16. It turns out that this 3×3 system has a unique solution if and only if $\tilde{\Sigma}_{y_1y_2|x}$ is invertible. This is the content of the following Lemma.

Lemma 18 *Let $\tilde{\Sigma}_{y_1y_2|x}$ be invertible. Then the conditions of Lemma 16 imply that*

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[h''_{y_1}h''_{y_2}] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h''_{y_1}h''_{y_2}],\ \mathop{\mathbb{E}}_{x\sim\mu^*}[(h''_{y_1})^2] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(h''_{y_1})^2],\ \mathop{\mathbb{E}}_{x\sim\mu^*}[(h''_{y_2})^2] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(h''_{y_2})^2]$$

Using the previous calculations for h''_{y_1}, h''_{y_2}, the first equality of this Lemma can be written as

$$\mathop{\mathbb{E}}_{x\sim\mu^*}\left[H_{y_2}\left(\sum_{i\neq y_2} r_iH_i\right)\right] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}\left[H_{y_2}\left(\sum_{i\neq y_2} r_iH_i\right)\right] \tag{9}$$

Now we are almost in the algebraic form of Lemma 9. All we need to do is get rid of the expectations. So, it is time to compute them. Let $i, j \in N$ be two neighbors. Then, we know that H_i, H_j are linear combinations of leaves in the partitions of i, j. Thus, by the multiplication property $\mathbb{E}[H_iH_j]$ involves the covariances of all pairs of leaves from S_{y_i} to S_{y_j}. Suppose that $H_i = (a^i)^\top X_{y_i}$, where X_{y_i} is the vector of leaves in S_{y_i}. By the path multiplication property

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[H_iH_j] = \mathop{\mathbb{E}}_{x\sim\mu^*}[(a^i)^\top X_{y_i}(a^j)^\top X_{y_j}] = \sum_{x_k\in S_{y_i}, x_l\in S_{y_j}} a^i_k a^j_l \prod_{e\in P_(x_k,x_l)} \rho^*_e$$

where $P(x_k, x_l)$ denotes the path between leaves x_k, x_l. Now, notice that all the paths from X_{y_i} to X_{y_j} will have to go through the edges connecting y_1 to y_i and y_1 to y_j. Let ρ^*_i, ρ^*_j be the correlations in these two edges. Then, using the multiplication over paths, Equation (9) can be written as

$$\rho_i^* \left(\sum_{x_r \in S_{y_i}} a_r^i \prod_{e \in P_{x_r y_i}} \rho_e^* \right) \sum_{y_j \in N, j \neq i} \rho_j^* \left(\sum_{x_s \in S_{y_j}} a_s^j \prod_{e \in P_{x_s y_j}} \rho_e^* \right)$$
$$= \tilde{\rho}_i \left(\sum_{x_r \in S_{y_i}} a_r^i \prod_{e \in P_{x_r y_i}} \tilde{\rho}_e \right) \sum_{y_j \in N, j \neq i} \tilde{\rho}_j \left(\sum_{x_s \in S_{y_j}} a_s^j \prod_{e \in P_{x_s y_j}} \tilde{\rho}_e \right) \tag{10}$$

In this form, we can define

$$w_i^* = \rho_i^* \left(\sum_{x_r \in S_{y_i}} a_r^i \prod_{e \in P(x_r, y_i)} \rho_e^* \right), \; \tilde{w}_i = \tilde{\rho}_i \left(\sum_{x_r \in S_{y_i}} a_r^i \prod_{e \in P(x_r, y_i)} \tilde{\rho}_e \right).$$

Now the condition is

$$r_j w_j^* \left(\sum_{i \neq j} r_i w_i^* \right) = r_j \tilde{w}_j \left(\sum_{i \neq j} r_i \tilde{w}_i \right) \tag{11}$$

which is exactly in the form of the system of Lemma 9. By applying this Lemma, we immediately get the following corollary.

Lemma 19 *Let $\mu^*, \tilde{\mu}$ be defined as in Theorem 15. Then for every node i, if we define S_{y_i} in reference to some neighbor y_1, it holds $w_i^* = \tilde{w}_i$.*

If a node y_1 is connected to a leaf x_j with correlation ρ_j^*, then we can extend all the previous statements and define $w_j^* = \rho_j^*, \tilde{w}_j = \tilde{\rho}_j$. This immediately implies that $\rho_j^* = \tilde{\rho}_j$. The proof of Theorem 15 relies exactly on using the equalities implied by Lemma 19 in the correct order, in order to guarantee that all correlations are the same in the two models.

Proof [Proof sketch of Theorem 15] We use the following procedure: in each iteration, we select an internal node y that only has one non-leaf neighbor in the remaining tree (there is always one such node). If e is some edge connecting y with some leaf in the tree, we declare that $\rho_e^* = \tilde{\rho}_e$ and remove this edge along with the leaf from the tree. After we do this for all such edges, the current iteration ends.

First of all, it is clear that if a node y is selected for some iteration, then for the remaining iterations it will be a leaf and not be selected. Hence, the process terminates after m steps, at which point all edges have been examined. We prove inductively that at each step the algorithm correctly infers the equality of the edges. For the base case, we already argued that edges that are adjacent to leaves will agree in the two models. For some arbitrary iteration, if y is selected and has a leaf neighbor x, then we use Lemma 19 on x with $y_1 = y$ to infer that $w_x^* = \tilde{w}_x$. From the definition of w and the inductive hypothesis, the parentheses multiplying ρ_i^* and $\tilde{\rho}_i$ are equal in the two models, which implies that $\rho_i^* = \tilde{\rho}_i$ and the proof is complete. ■

References

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

Babak Barazandeh and Meisam Razaviyayn. On the behavior of the expectation-maximization algorithm for mixture models. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 61–65. IEEE, 2018.

Shankar Bhamidi, Ram Rajagopal, and Sébastien Roch. Network delay inference from additive metrics. *Random Structures & Algorithms*, 37(2):176–203, 2010.

Rui Castro, Mark Coates, Gang Liang, Robert Nowak, and Bin Yu. Network tomography: Recent developments. *Statistical science*, 19(3):499–517, 2004.

Stéphane Chrétien and Alfred O Hero. On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Optimal phylogenetic reconstruction. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 159–168, 2006.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of em suffice for mixtures of two gaussians. In *Conference on Learning Theory*, pages 704–710. PMLR, 2017.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Bootstrapping em via em and convergence analysis in the naive bayes model. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, 2018.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J Wainwright, Michael I Jordan, and Bin Yu. Singularity, misspecification and the convergence rate of em. *The Annals of Statistics*, 48(6):3161–3182, 2020.

Péter L Erdős, Michael A Steel, László A Székely, and Tandy J Warnow. A few logs suffice to build (almost) all trees (i). *Random Structures & Algorithms*, 14(2):153–184, 1999.

Brian Eriksson, Gautam Dasarathy, Paul Barford, and Robert Nowak. Toward the practical use of network tomography for internet topology discovery. In *2010 Proceedings IEEE INFOCOM*, pages 1–9. IEEE, 2010.

Joseph Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics*, 25(5):471, 1973.

Joseph Felsenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

Gordon Hiscott, Colin Fox, Matthew Parry, and David Bryant. Efficient recycled algorithms for quantitative trait models on phylogenies. *Genome biology and evolution*, 8(5):1338–1350, 2016.

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65(6):3515–3524, 2019.

Jeongyeol Kwon and Constantine Caramanis. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pages 2425–2487. PMLR, 2020a.

Jeongyeol Kwon and Constantine Caramanis. Em converges for a mixture of many linear regressions. In *International Conference on Artificial Intelligence and Statistics*, pages 1727–1736. PMLR, 2020b.

Jeongyeol Kwon, Wei Qian, Constantine Caramanis, Yudong Chen, and Damek Davis. Global convergence of the em algorithm for mixtures of two component linear regression. In *Conference on Learning Theory*, pages 2055–2110. PMLR, 2019.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the em algorithm for learning two-component mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1405–1413. PMLR, 2021.

Raphaël Mourad, Christine Sinoquet, Nevin Lianwen Zhang, Tengfei Liu, and Philippe Leray. A survey on latent tree models and applications. *Journal of Artificial Intelligence Research*, 47:157–203, 2013.

Sai Ganesh Nagarajan and Ioannis Panageas. On the analysis of em for truncated mixtures of two gaussians. In *Algorithmic Learning Theory*, pages 634–659. PMLR, 2020.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

Don Ringe, Tandy Warnow, and Ann Taylor. Indo-european and computational cladistics. *Transactions of the philological society*, 100(1):59–129, 2002.

Sebastien Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(1): 92–94, 2006.

Sebastien Roch. Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science*, 327(5971):1376–1379, 2010.

Sebastien Roch and Allan Sly. Phase transition in the sample complexity of likelihood-based phylogeny inference. *Probability Theory and Related Fields*, 169(1):3–62, 2017.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1): 124–127, 1950.

Nathaniel Shiers, John AD Aston, Jim Q Smith, and John S Coleman. Gaussian tree constraints applied to acoustic linguistic functional data. *Journal of Multivariate Analysis*, 154:199–215, 2017.

Alexandros Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

Michael Truell, Jan-Christian Hütter, Chandler Squires, Piotr Zwiernik, and Caroline Uhler. Maximum likelihood estimation for brownian motion tree models based on one sample. *arXiv preprint arXiv:2112.00816*, 2021.

Paul Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

Stuart Sui-Sheng Wang. A jacobian criterion for separability. *Journal of Algebra*, 65(2): 453–494, 1980.

Yi Wang and Nevin Lianwen Zhang. Severity of local maxima for the em algorithm: Experiences with hierarchical latent class models. In *Probabilistic Graphical Models*, pages 301–308. Citeseer, 2006.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

Ji Xu, Daniel Hsu, and Arian Maleki. Global analysis of Expectation Maximization for mixtures of two Gaussians. In *the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

Ji Xu, Daniel J Hsu, and Arian Maleki. Benefits of over-parameterization with em. *Advances in Neural Information Processing Systems*, 31, 2018.

Ziheng Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences*, 13(5):555–556, 1997.

Ruofei Zhao, Yuanzhi Li, and Yuekai Sun. Statistical convergence of the em algorithm on gaussian mixture models. *Electronic Journal of Statistics*, 14(1):632–660, 2020.

Piotr Zwiernik. Latent tree models. In *Handbook of Graphical Models*, pages 283–306. CRC Press, 2018.

Appendix A. Deferred proofs for one latent node

A.1. The EM does not converge to 0

In this section, we will prove the following two lemmas:

Lemma 20 *The point $\overline{0} = (0, \dots, 0)$ is a stationary point and there is no other stationary point $\overline{\rho}$ with some $\rho_i = 0$.*

Lemma 21 *The iterate of the EM algorithm satisfies $\lim_{t\to\infty} \overline{\rho}^t \neq \overline{0}$.*

First, we notice that the point $\overline{0} = (0, \ldots, 0)$ is a stationary point. Indeed, by (1), $\overline{\rho}^t = (0, \ldots, 0)$ then $\lambda_i^t = 0$. Together with, Lemma 2, this implies that $\overline{\rho}^{t+1} = (0, \ldots, 0)$ which concludes that $\overline{0}$ is a stationary point. The following concludes the proof of Lemma 20:

Lemma 22 *If $\rho_{i_1}^t > 0$ for some $i_1 > 0$ then, $\rho_i^{t+1} > 0$ for all i. Consequently, if $\overline{\rho}$ satisfies $\rho_{i_0} = 0$ and $\rho_{i_1} > 0$ for some i_0 and i_1, then ρ is not a stationary point.*

Proof We start with the first part of the proof. Recall the update rule for the covariances, from Lemma 2:

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \sigma_{x_i}^t \sigma_y^t \left(\lambda_i^t + \sum_{j \neq i} \rho_i^* \rho_j^* \lambda_j^t \right).$$

Further, from (1), we know that $\lambda_i^t > 0$ whenever $\rho_i^t > 0$. In particular, $\lambda_{i_1}^t > 0$, which implies that for all i, $\mathbb{E}_{\mu^{t+1}}[x_i y] \geq \sigma_{x_i}^t \sigma_y^t \rho_i^* \rho_{i_1}^* \lambda_{i_1}^t > 0$. This concludes the first part of the proof. For the second part, notice that if $\mu^t = \mu$ then, by the first part, $\rho_{i_0}^{t+1} > 0 = \rho_{i_0}^t$, hence μ cannot be a stationary point. ■

The first part of the lemma follows from the update rule analyzed in Lemma 2, while the second part follows directly from the first part and the definition of a stationary point. This concludes Lemma 20.

For the remainder, we prove Lemma 21. The first part of Lemma 22 implies that if $\overline{\rho}^0 \neq \overline{0}$ then $\overline{\rho}^t \neq \overline{0}$ for all $t > 0$. It remains to prove that $\overline{\rho}^t$ does not converge to $\overline{0}$ and we will analyze the updates of the EM rule in the neighborhood of $\overline{\rho}^t = \overline{0}$, viewing $\overline{\rho}^{t+1}$ as a function of $\overline{\rho}^t$.

Lemma 23 *If $\overline{\rho}^t = \overline{0}$ then $\overline{\rho}^{t+1} = \overline{0}$. Further, there exists $C > 0$ such that for all $i, j, k \in \{1, \ldots, n\}$ and all $\overline{\rho} \in [0, 1/2]^n$,*

$$\left.\frac{d\rho_i^{t+1}}{d\rho_j^t}\right|_{\overline{0}} = \begin{cases} 1 & i = j \\ \rho_i^* \rho_j^* & i \neq j \end{cases}; \qquad -C \leq \left.\frac{d^2 \rho_i^{t+1}}{d\rho_j^t d\rho_k^t}\right|_{\overline{\rho}} \leq C.$$

Proof We use the formula for $\overline{\rho}^{t+1}$ from Lemma 2:

$$\rho_i^{t+1} = \frac{\rho_i^t + \sum_{j \neq i} \Delta_{ij} \lambda_j^t}{\sqrt{1 + \sum_{j \neq k} \Delta_{ij} \lambda_j^t \lambda_k^t}} := \frac{f_i(\overline{\rho}^t)}{\sqrt{g_i(\overline{\rho}^t)}}.$$

Before computing the derivatives of ρ_i^{t+1}, let us compute the derivatives of λ_i^t as a function of ρ_j^t where the formula of λ_i^t appears in (1). For any i, j, k and $\overline{\rho} \in [0, 1/2]^n$,

$$\left.\frac{d\lambda_i^t}{d\rho_j^t}\right|_{\overline{0}} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}; \quad \text{and} \quad -C \leq \left.\frac{d^2 \lambda_i^t}{d\rho_j^t d\rho_k^t}\right|_{\overline{\rho}} \leq C$$

for some constant $C > 0$. Using the fact that $\lambda_i^t = 0$ if $\rho_i^t = 0$, we derive that

$$\left.\frac{df_i(\overline{\rho}^t)}{d\rho_j^t}\right|_{\overline{0}} = \begin{cases} 1 & j = i \\ \Delta_{ij} = \rho_i^* \rho_j^* & j \neq i \end{cases}, \quad f_i(\overline{0}) = \overline{0}, \quad \left.\frac{dg_i(\overline{\rho}^t)}{d\rho_j^t}\right|_{\overline{0}} = 0, \quad g_i(\overline{0}) = 1.$$

Hence,

$$\left.\frac{d\rho_i^{t+1}}{d\rho_j^t}\right|_{\overline{0}} = \frac{df_i(\overline{\rho}^t)}{d\rho_j^t} = \begin{cases} 1 & j = i \\ \rho_i^*\rho_j^* & j \neq i \end{cases}.$$

Similarly, the second derivatives of ρ_i^{t+1} are bounded, using the bounds on the second derivatives of λ_j^t. ■

This implies that if $\overline{\rho}^t$ approaches $\overline{0}$, then $\overline{\rho}^{t+1}$ repels from $\overline{0}$:

Lemma 24 *There exists some $c > 0$ such that if $\max_i \rho_i^t \leq c$ then $\sum_i \rho_{i+1}^t \geq \sum_i \rho_i^t$.*

Proof Using Lemma 23, we can write ρ_i^{t+1} as a Taylor series around $\overline{0}$

$$\rho_i^{t+1} \geq \rho_i^t + \sum_{j\neq i} \rho_j\rho_i^*\rho_j^* - \sum_{j,k} \frac{C}{2}\rho_j^t\rho_k^t.$$

Summing over i, we derive that

$$\sum_i \rho_i^{t+1} \geq \sum_i \rho_i^t + \sum_{i\neq j} \rho_j\rho_i^*\rho_j^* - \sum_{i,j,k} \frac{C}{2}\rho_j^t\rho_k^t.$$

While the second term in the right hand side is $\Omega(\max_j \rho_j^t)$, the third term is $O(\max_j(\rho_j^t)^2)$. In particular, if the constant $c > 0$ from the definition of this lemma is sufficiently small, then the second term dominates the third and the proof follows. ■

To conclude the proof of Lemma 21, assume toward contradiction that $\overline{\rho}^t \to \overline{0}$. Let $c > 0$ be the parameter from Lemma 24, and let $T > 0$ be the iteration such that for any $t \geq T$, $\max_i \rho_i^t < c$. From Lemma 24, this implies that for any $t \geq T$, $\sum_j \rho_j^t \geq \sum_j \rho_j^T$. From Lemma 22, $\sum_j \rho_j^T > 0$. In particular, $\liminf_{t\to\infty} \sum_j \rho_j^t \geq \sum_j \rho_j^T > 0$ which implies that $\lim_{t\to\infty} \overline{\rho}^t \neq \overline{0}$, as required.

A.2. Stationary points with some $\rho_i = 1$

In this Section, we analyze the case of stationary points such that there exists at least one i with $\rho_i 1$. We show that EM will never converge to any of those points. Let us start with the case where there are at least two i, j such that $\rho_i = \rho_j = 1$. In that case, let μ be the distribution at this point. Also, let

$$L(\rho) = \mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_\rho(\overline{x})$$

be the log-likelihood function. The following lemma shows that the EM algorithm will never converge to this stationary point.

Lemma 25 *We have that*

$$\lim_{\rho_i\to 1, \rho_j\to 1} L(\rho) = -\infty$$

Proof We can write the log-likelihood function as follows:

$$\begin{aligned} L(\rho) &= \mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_{\rho}(\overline{x}) \\ &= \mathop{\mathbb{E}}_{x\sim\mu^*} \log \left(\Pr_{\rho}(x_i, x_j) \Pr_{\rho} \left(\overline{x}_{[n]\setminus\{i,j\}} \mid x_i, x_j \right) \right) \\ &= \mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_{\rho}(x_i, x_j) + \mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_{\rho} \left(\overline{x}_{[n]\setminus\{i,j\}} \mid x_i, x_j \right) \end{aligned}$$

The second term is clearly upper bounded over the whole region. For the first term, it is clear that

$$\mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_{\rho}(x_i, x_j) = -KL(\mu^*(x_i, x_j), \mu(x_i, x_j; \rho)) - \mathop{\mathbb{E}}_{x\sim\mu^*} \log \mu^*(x_i, x_j)$$

and it is clear that

$$\lim_{\rho_i\to 1, \rho_j \to 1} KL(\mu^*(x_i, x_j), \mu(x_i, x_j; \rho)) = \infty$$

while the second term is constant. It follows that

$$\lim_{\rho_i\to 1, \rho_j \to 1} \mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_{\rho}(x_i, x_j) = -\infty$$

which completes the proof. ■

Since the EM is guaranteed to improve the value of the likelihood at every step, it is impossible to converge to this point.

Now, let us consider the case where there is exaclty one i with $\rho_i = 1$. Without loss of generality, suppose $\rho_1 = 1$. The reason this is a fixpoint is that if $\rho_1^t = 1$, then $y = x_1$ in the conditional model of y given x, which means that $\rho^{t+1} = 1$. Then, by the fixpoint equations we immediately get that in the fixpoint we should have

$$\tilde{\rho}_i = \rho_i^* \rho_1^*$$

for $i \neq 1$. From now on, when we refer to $\tilde{\rho}$, these will be the values that are implied. We would like to show that when we start running EM from a point in the interior of the region, we will not converge to this stationary point. However, we already know that if we start from a point with $\rho_1 = 1$, then inevitably in the next iteration we will converge to the stationary point values for the other variables as well. Hence, this stationary point will be a saddle point of the log-likelihood.

In particular, we will prove the following Theorem.

Theorem 26 *Let ρ be a stationary point with $\rho_1 = 1$. Then, there exists an $\epsilon_0 > 0$, such that the following condition holds: if we start running EM from any initial point ρ^0 that satisfies $\|\rho^0 - \rho\|_2 < \epsilon_0$ and such that $\rho_i^0 < 1$ for all i, then after a finite number of iterations $t = t(\rho_0)$ we have*

$$\|\rho^t - \rho\|_2 > \epsilon_0$$

We stress that the number of iterations t depends on the initial value ρ_0. Let us argue why Theorem 26 implies that EM does not converge to ρ from any initial point inside the region. Indeed, if EM converged for some initial value ρ^0, then for any $\epsilon > 0$, and in particular for $\epsilon = \epsilon_0$, there exists some iteration T such that for $t \geq T$ we get

$$\|\rho^t - \rho\| \leq \epsilon$$

However, if we apply Theorem 26 for $\rho^0 = \rho^T$, we get that there should exists some $t > T$ with $\|\rho^t - \rho\| > \epsilon_0$. This is a contradiction to the convergence claim. Hence, it suffices to prove Theorem 26. The first step is to establish the following Lemma. It's purpose is to show that if the distances of all the coordinates are roughly equal, then ρ_1 will move away from the fixpoint value of 1.

Lemma 27 *Let ρ^t be the current iteration of EM and define*

$$\epsilon := \max_i |\tilde{\rho}_i - \rho_i^t|$$

Then, there are constants $C, K > 0$ such that

$$\rho_1^{t+1} \leq \rho_1^t - C|\rho_1^t - 1|^2 + K\epsilon^3$$

Proof The proof consists of viewing the update rule as a function of the previous step and writing the Taylor expansion of the this function around the fixpoint ρ. In particular, we define the function

$$f(\rho) = \frac{\rho_1 + \sum_{j\neq 1}(\rho_1\rho_j - \rho_1^*\rho_j^*)\lambda_j}{\sqrt{1 + \sum_{j,k:\ j\neq k}(\rho_j\rho_k - \rho_j^*\rho_k^*)\lambda_j\lambda_k}}$$

where

$$\lambda_j = \frac{\frac{\rho_j}{1-\rho_j^2}}{1 + \sum_k \frac{\rho_k^2}{1-\rho_k^2}}$$

Clearly, we have that

$$f(\rho^t) = \rho^{t+1}$$

We will compute the Taylor expansion of f around the fixpoint ρ. We start by computing the first derivatives of f at the fixpoint $\tilde{\rho}$. To do that, we need the derivatives of λ_j with ρ_i. We start with the case $i = j$

$$\begin{aligned}
\frac{\partial\lambda_1}{\partial\rho_i} &= \frac{\frac{\partial}{\partial\rho_i}\left(\frac{\rho_i}{1-\rho_i^2}\right)}{1 + \sum_j \frac{\rho_j^2}{1-\rho_j^2}} - \frac{\partial}{\partial\rho_i}\left(1 + \sum_j \frac{\rho_j^2}{1-\rho_j^2}\right)\frac{\left(\frac{\rho_i}{1-\rho_i^2}\right)}{\left(1 + \sum_j \frac{\rho_j^2}{1-\rho_j^2}\right)^2} \\
&= \frac{\frac{1}{1-\rho_i^2} + \frac{2\rho_i^2}{(1-\rho_i^2)^2}}{1 + \sum_j \frac{\rho_j^2}{1-\rho_j^2}} - \left(\frac{2\rho_i}{(1-\rho_i^2)^2}\right)\frac{\left(\frac{\rho_i}{1-\rho_i^2}\right)}{\left(1 + \sum_j \frac{\rho_j^2}{1-\rho_j^2}\right)^2} \\
&= \frac{\lambda_1}{\rho_i} + 2\frac{\rho_i\lambda_1}{1-\rho_i^2} - 2\frac{\lambda_1^2}{1-\rho_i^2}
\end{aligned}$$

Let's evaluate the limit as $\rho \to \tilde{\rho}$. We know that as $\rho_1 \to 1$, we will have $\lambda_1 \to 1$ and $\lambda_j \to 0$ for $j \neq 1$. Based on these observations, we have that

$$\begin{aligned}
\lim_{\rho_1 \to 1} \frac{\rho_1 - \lambda_1}{1-\rho_1^2} &= \frac{\rho_1 - \frac{\frac{\rho_1}{1-\rho_1^2}}{1+\sum_j \frac{\rho_j^2}{1-\rho_j^2}}}{1-\rho_1^2} \\
&= \lim_{\rho_1 \to 1} \frac{\rho_1 + \rho_1 \sum_j \frac{\rho_j^2}{1-\rho_j^2} - \frac{\rho_1}{1-\rho_1^2}}{\left(1+\sum_j \frac{\rho_j^2}{1-\rho_j^2}\right)(1-\rho_1^2)} \\
&= \lim_{\rho_1 \to 1} \left(\rho_1 + \frac{\rho_1^3 - \rho_1}{1-\rho_1^2} + \rho_1 \sum_{j \neq 1} \frac{\rho_j^2}{1-\rho_j^2} \right) \\
&= \sum_{j \neq 1} \frac{\rho_j^2}{1-\rho_j^2}
\end{aligned}$$

Hence,

$$\lim_{\rho \to \tilde{\rho}} \frac{\partial \lambda_1}{\partial \rho_1} = 2 \sum_{j \neq 1} \frac{\tilde{\rho}_j^2}{1-\tilde{\rho}_j^2} + 1 = 2 \sum_{j \neq 1} \frac{(\rho_j^*)^2 (\rho_1^*)^2}{1-(\rho_j^*)^2(\rho_1^*)^2} + 1$$

For $j \neq 1$, we have

$$\begin{aligned}
\frac{\partial \lambda_j}{\partial \rho_i} &= \frac{\partial}{\partial \rho_i} \frac{\frac{\rho_j}{1-\rho_j^2}}{1+\sum_k \frac{\rho_k^2}{1-\rho_k^2}} = -\frac{\frac{\rho_j}{1-\rho_j^2}}{\left(1+\sum_k \frac{\rho_k^2}{1-\rho_k^2}\right)^2} \frac{\partial}{\partial \rho_i} \left(\frac{\rho_i^2}{1-\rho_i^2} \right) \\
&= -\frac{\frac{\rho_j}{1-\rho_j^2}}{\left(1+\sum_k \frac{\rho_k^2}{1-\rho_k^2}\right)^2} \frac{2\rho_i}{(1-\rho_i^2)^2} \\
&= -2 \frac{\lambda_1 \lambda_j}{1-\rho_i^2}
\end{aligned}$$

We know that as $\rho_i \to 1$, we have $\lambda_1 \to 1$. Also,

$$\lim_{\rho_i \to 1} \frac{\lambda_j}{1-\rho_i^2} = \rho_j / (1-\rho_j^2)$$

Hence,

$$\lim_{\rho \to \tilde{\rho}} \frac{\partial \lambda_j}{\partial \rho_i} = -2 \frac{\rho_j^* \rho_1^*}{1-(\rho_j^* \rho_1^*)^2}$$

What we mostly care about is that these derivatives are bounded. Now, we are ready to compute the various derivatives of f at the fixpoint. We will use the notation $\Delta_{jk} :=$

$\rho_j^*\rho_k^* - \rho_j\rho_k$.

$$\frac{\partial f}{\partial \rho_1} = \frac{\frac{\partial}{\partial \rho_1}\left(\rho_1 + \sum_{j\neq 1}\Delta_{1j}\lambda_j\right)}{\sqrt{1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k}} - \frac{1}{2}\frac{\rho_1+\sum_{j\neq 1}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}\frac{\partial}{\partial\rho_1}\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)$$

$$= \frac{1+\sum_{j\neq 1}\Delta_{1j}\frac{\partial\lambda_j}{\partial\rho_1} - \sum_{j\neq 1}\lambda_j\rho_j}{\sqrt{1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k}}$$

$$- \frac{1}{2}\frac{\rho_1+\sum_{j\neq 1}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}\left(2\sum_{j\neq 1}\frac{\partial\lambda_1\lambda_j}{\partial\rho_1}\Delta_{1j} - 2\sum_{j\neq 1}\lambda_1\lambda_j\rho_j - \sum_{j\neq k\neq 1}\frac{\partial\lambda_j\lambda_k}{\partial\rho_1}\Delta_{jk}\right)$$

Now, let us examine what happends when we plug in the fixpoint. We know that $\Delta_{1j}\to 0$ and $\lambda_j\to 0$ for $j\neq 1$, hence the entire first term tends to 1. As for the second term, we have

$$\lim_{\rho_1\to 1}\frac{\rho_1+\sum_{j\neq 1}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}} = 1$$

Also, notice that

$$\frac{\partial\lambda_1\lambda_j}{\partial\rho_1} = \frac{\partial\lambda_1}{\partial\rho_1}\lambda_j + \frac{\partial\lambda_j}{\partial\rho_1}\lambda_1$$

is bounded, hence

$$\lim_{\rho\to\tilde{\rho}}\sum_{j\neq 1}\frac{d\lambda_1\lambda_j}{d\rho_i}\Delta_{1j} = 0$$

It is also clear that

$$\lim_{\rho_i\to 1}\sum_{j\neq 1}\lambda_1\lambda_j\rho_j = 0$$

Lastly, we have

$$\frac{\partial\lambda_j\lambda_k}{d\rho_i} = \lambda_j\frac{d\lambda_k}{d\rho_i} + \lambda_k\frac{\partial\lambda_j}{d\rho_i}\to 0$$

Hence, the second term tends to 0. Hence,

$$\lim_{\rho\to\tilde{\rho}}\frac{\partial f}{\partial\rho_1} = 1$$

Let's now calculate $\partial f/\partial\rho_u$ for some $u\neq 1$. Using the exact same calculation, we arrive at the formula

$$\frac{\partial f}{\partial\rho_u} = \frac{\sum_{j\neq 1}\Delta_{1j}\frac{\partial\lambda_j}{\partial\rho_u} - \lambda_u\rho_1}{\sqrt{1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k}}$$

$$- \frac{1}{2}\frac{\rho_1+\sum_{j\neq 1}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}\left(2\sum_{j\neq u}\frac{d\lambda_u\lambda_j}{\partial\rho_u}\Delta_{uj} - 2\sum_{j\neq u}\lambda_j\lambda_u\rho_j - \sum_{j\neq k\neq u}\frac{\partial\lambda_j\lambda_k}{\partial\rho_u}\Delta_{jk}\right)$$

This time, both the first and the second terms converge to 0, hence

$$\lim_{\rho \to \tilde{\rho}} \frac{\partial f}{\partial \rho_u} = 0$$

Now, we move on to the second derivatives. First of all, let's compute the ones for λ_1. We just want to ensure that they will be bounded.

$$\begin{aligned}
\frac{\partial^2 \lambda_j}{\partial^2 \rho_1} &= \frac{\partial}{\partial \rho_1}\left(-2\frac{\lambda_1 \lambda_j}{1-\rho_1^2}\right) \\
&= -2\frac{(\partial \lambda_1/\partial \rho_1)\lambda_j}{1-\rho_1^2} - 2\frac{(\partial \lambda_j/\partial \rho_1)\lambda_1}{1-\rho_1^2} - 4\frac{\lambda_1 \lambda_j \rho_1}{(1-\rho_1^2)^2} \\
&= -2\frac{(\partial \lambda_1/\partial \rho_1)\lambda_j}{1-\rho_1^2} + 4\frac{\lambda_j \lambda_1^2}{(1-\rho_1^2)^2} - 4\frac{\lambda_1 \lambda_j \rho_1}{(1-\rho_1^2)^2} \\
&= -2\frac{(\partial \lambda_1/\partial \rho_1)\lambda_j}{1-\rho_1^2} + 4\frac{\lambda_1 \lambda_j}{1-\rho_1^2}\left(\frac{\lambda_1 - \rho_1}{1-\rho_1^2}\right)
\end{aligned}$$

The first term is constant in the limit and the second term is also constant, since the limit $\lim_{\rho_1 \to 1}(\lambda_1 - \rho_1)/(1-\rho_1^2)$ has been shown to be constant. Similarly, we have

$$\begin{aligned}
\frac{\partial^2 \lambda_j}{\partial \rho_1 \partial \rho_u} &= \frac{\partial}{\partial \rho_u}\left(-2\frac{\lambda_1 \lambda_j}{1-\rho_1^2}\right) \\
&= -2\frac{(\partial \lambda_1/\partial \rho_u)\lambda_j + (\partial \lambda_j/\partial \rho_u)\lambda_1}{1-\rho_1^2} \\
&= 4\frac{\lambda_1 \lambda_j \lambda_u}{(1-\rho_1^2)(1-\rho_u^2)} + 4\frac{\lambda_j \lambda_u \lambda_1}{(1-\rho_1^2)(1-\rho_u^2)}
\end{aligned}$$

It's clear that both terms tend to 0. Finally, let's calculate

$$\begin{aligned}
\frac{\partial^2 \lambda_1}{\partial^2 \rho_1} &= \frac{\partial}{\partial \rho_1}\left(\frac{\lambda_1}{\rho_1} + 2\frac{\rho_1 \lambda_1}{1-\rho_1^2} - 2\frac{\lambda_1^2}{1-\rho_1^2}\right) \\
&= \frac{\partial}{\partial \rho_1}\left(\frac{\lambda_1}{\rho_1}\right) + 2\frac{\partial \lambda_1}{\partial \rho_1}\frac{\rho_1 - \lambda_1}{1-\rho_1^2} + 2\lambda_1 \frac{\partial}{\partial \rho_1}\left(\frac{\rho_1 - \lambda_1}{1-\rho_1^2}\right)
\end{aligned}$$

By taking the limit $\rho_1 \to 1$ we verify easily that the second derivative is bounded.

Now, let's compute the second derivative $\partial^2 f/\partial^2 \rho_1$. To make the presentation easier, we derive each of the two terms separately.

$$\begin{aligned}
&\frac{\partial}{\partial \rho_1}\left(\frac{1 + \sum_{j \neq 1} \Delta_{1j}\frac{\partial \lambda_j}{\partial \rho_1} - \sum_{j \neq 1}\lambda_j \rho_j}{\sqrt{1 + \sum_{j,k:\, j \neq k} \Delta_{jk}\lambda_j \lambda_k}}\right) = \frac{\sum_{j \neq 1} \Delta_{1j}\frac{\partial^2 \lambda_j}{\partial^2 \rho_1} - \sum_{j \neq 1}\frac{\partial \lambda_j}{\partial \rho_1}\rho_j - \sum_{j \neq 1}\frac{\partial \lambda_j}{\partial \rho_1}\rho_j}{\sqrt{1 + \sum_{j,k:\, j \neq k} \Delta_{jk}\lambda_j \lambda_k}} \\
&- \frac{1}{2}\frac{1 + \sum_{j \neq 1} \Delta_{1j}\frac{\partial \lambda_j}{\partial \rho_1} - \sum_{j \neq 1}\lambda_j \rho_j}{\left(1 + \sum_{j,k:\, j \neq k} \Delta_{jk}\lambda_j \lambda_k\right)^{3/2}}\left(2\sum_{j \neq 1}\frac{d\lambda_1 \lambda_j}{\partial \rho_1}\Delta_{1j} - 2\sum_{j \neq 1}\lambda_1 \lambda_j \rho_j - \sum_{j \neq k \neq i}\frac{\partial \lambda_j \lambda_k}{\partial \rho_1}\Delta_{jk}\right)
\end{aligned}$$

In the first term, the final result once we take the limit is

$$-2\sum_{j\neq 1}\rho_j\frac{\partial\lambda_j}{\partial\rho_1}\mid_{\rho=\tilde{\rho}}$$

As for the second term, the rightmost parenthesis is 0, as shown in the calculation of the first derivative. Now, let's move to the second term. We want

$$\frac{\partial}{\partial\rho_1}\frac{1}{2}\frac{\rho_1+\sum_{j\neq i}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}\left(2\sum_{j\neq 1}\frac{d\lambda_1\lambda_j}{\partial\rho_1}\Delta_{1j}-2\sum_{j\neq 1}\lambda_1\lambda_j\rho_j-\sum_{j\neq k\neq i}\frac{\partial\lambda_j\lambda_k}{\partial\rho_1}\Delta_{jk}\right)$$

We can view this as a product of three terms and use the product rule. Notice that the third term is 0 in the fixpoint, so when deriving the first two terms we will get 0 in the final expression. Hence, the only term that matters is

$$\frac{1}{2}\frac{\rho_1+\sum_{j\neq i}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}\frac{\partial}{\partial\rho_1}\left(2\sum_{j\neq 1}\frac{d\lambda_1\lambda_j}{\partial\rho_1}\Delta_{1j}-2\sum_{j\neq 1}\lambda_1\lambda_j\rho_j-\sum_{j\neq k\neq i}\frac{\partial\lambda_j\lambda_k}{\partial\rho_1}\Delta_{jk}\right)$$

$$=\frac{1}{2}\frac{\rho_1+\sum_{j\neq i}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}\left(2\sum_{j\neq 1}\frac{\partial^2\lambda_1\lambda_j}{\partial^2\rho_1}\Delta_{1j}-2\sum_{j\neq 1}\frac{d\lambda_1\lambda_j}{\partial\rho_1}\rho_j-2\sum_{j\neq 1}\frac{d\lambda_1\lambda_j}{\partial\rho_1}\rho_j-\sum_{j\neq k\neq i}\frac{\partial^2\lambda_j\lambda_k}{\partial^2\rho_1}\Delta_{jk}\right)$$

First of all, as usual we have

$$\lim_{\rho\to\tilde{\rho}}\frac{\rho_1+\sum_{j\neq i}\Delta_{1j}\lambda_j}{\left(1+\sum_{j,k:\ j\neq k}\Delta_{jk}\lambda_j\lambda_k\right)^{3/2}}=1$$

Also, we have

$$\lim_{\rho\to\tilde{\rho}}\sum_{j\neq 1}\frac{\partial^2\lambda_1\lambda_j}{\partial^2\rho_1}\Delta_{1j}=0$$

$$\lim_{\rho\to\tilde{\rho}}\sum_{j\neq 1}\frac{\partial\lambda_1\lambda_j}{\partial\rho_1}\rho_j=\lim_{\rho\to\tilde{\rho}}\sum_{j\neq 1}\left(\lambda_1\frac{\partial\lambda_j}{\partial\rho_1}+\lambda_j\frac{d\lambda_1}{\partial\rho_1}\right)\rho_j=\sum_{j\neq 1}\rho_j\frac{\partial\lambda_j}{\partial\rho_1}\mid_{\rho=\tilde{\rho}}$$

Hence, we get the same expression as in the first term and these two cancel each other. We are only left with computing the term

$$\lim\sum_{j\neq k\neq 1}\frac{\partial^2\lambda_j\lambda_k}{\partial^2\rho_1}\Delta_{jk}$$

If we set $g(x)=x/(1-x^2)$, then we get

$$\begin{aligned}\lim_{\rho\to\tilde{\rho}}\sum_{j\neq k\neq 1}\frac{\partial^2\lambda_j\lambda_k}{\partial^2\rho_1}\Delta_{jk}&=\lim_{\rho\to\tilde{\rho}}\sum_{j\neq k\neq 1}\left(\frac{\partial^2\lambda_j}{\partial^2\rho_1}\lambda_k+\frac{\partial^2\lambda_k}{\partial^2\rho_1}\lambda_j+2\frac{\partial\lambda_j}{\partial\rho_1}\frac{d\lambda_k}{d\rho_i}\right)\Delta_{jk}\\&=8\sum_{j\neq k\neq 1}g(\rho_1^*\rho_j^*)g(\rho_1^*\rho_k^*)\rho_j^*\rho_k^*(1-(\rho_1^*)^2)\end{aligned}$$

Hence,

$$\frac{\partial^2 f}{\partial^2 \rho_1} = -4 \sum_{j \neq k \neq 1} g(\rho_1^* \rho_j^*) g(\rho_1^* \rho_k^*) \rho_j^* \rho_k^* (1 - (\rho_1^*)^2) < 0$$

Let's define

$$C := 2 \sum_{j \neq k \neq 1} g(\rho_1^* \rho_j^*) g(\rho_1^* \rho_k^*) \rho_j^* \rho_k^* (1 - (\rho_1^*)^2)$$

Clearly C is a constant depending only on the true model. Now, let's compute the other mixed derivatives. Let $u \neq 1$. We will compute $\partial^2 f / \partial \rho_1 \partial \rho_u$. Again, for convenience we will derive each of the two terms of $\partial \rho_1' / \partial \rho_1$ separately. We have

$$\frac{\partial}{\partial \rho_u} \left(\frac{1 + \sum_{j \neq 1} \Delta_{1j} \frac{\partial \lambda_j}{\partial \rho_1} - \sum_{j \neq 1} \lambda_j \rho_j}{\sqrt{1 + \sum_{j,k:\, j \neq k} \Delta_{jk} \lambda_j \lambda_k}} \right) = \frac{\sum_{j \neq 1} \Delta_{1j} \frac{\partial^2 \lambda_j}{\partial \rho_1 \partial \rho_u} - \frac{d \lambda_u}{\partial \rho_1} \rho_1 - \sum_{j \neq 1} \frac{\partial \lambda_j}{\partial \rho_u} \rho_j - \lambda_u}{\sqrt{1 + \sum_{j,k:\, j \neq k} \Delta_{jk} \lambda_j \lambda_k}}$$
$$- \frac{1}{2} \frac{1 + \sum_{j \neq 1} \Delta_{1j} \frac{\partial \lambda_j}{\partial \rho_1} - \sum_{j \neq 1} \lambda_j \rho_j}{\left(1 + \sum_{j,k:\, j \neq k} \Delta_{jk} \lambda_j \lambda_k\right)^{3/2}} \left(2 \sum_{j \neq u} \frac{d \lambda_u \lambda_j}{\partial \rho_u} \Delta_{uj} - 2 \sum_{j \neq u} \lambda_u \lambda_j \rho_j - \sum_{j \neq k \neq u} \frac{\partial \lambda_j \lambda_k}{\partial \rho_u} \Delta_{jk} \right)$$

From the first term, the only term that survives in the limit is

$$- \frac{d \lambda_u}{\partial \rho_1} \rho_1 \mid_{\rho = \tilde{\rho}} = - \frac{d \lambda_u}{\partial \rho_1} \mid_{\rho = \tilde{\rho}}$$

We turn to the second term that needs differentiation. This is similar to what we had previously, where the only term that matters is

$$\frac{1}{2} \frac{\rho_1 + \sum_{j \neq i} \Delta_{1j} \lambda_j}{\left(1 + \sum_{j,k:\, j \neq k} \Delta_{jk} \lambda_j \lambda_k\right)^{3/2}} \frac{\partial}{\partial \rho_u} \left(2 \sum_{j \neq 1} \frac{d \lambda_1 \lambda_j}{\partial \rho_1} \Delta_{1j} - 2 \sum_{j \neq 1} \lambda_1 \lambda_j \rho_j - \sum_{j \neq k \neq i} \frac{\partial \lambda_j \lambda_k}{\partial \rho_1} \Delta_{jk} \right)$$
$$= \frac{1}{2} \frac{\rho_1 + \sum_{j \neq i} \Delta_{1j} \lambda_j}{\left(1 + \sum_{j,k:\, j \neq k} \Delta_{jk} \lambda_j \lambda_k\right)^{3/2}} \times$$
$$\times \left(2 \sum_{j \neq 1} \frac{\partial^2 \lambda_1 \lambda_j}{\partial \rho_1 \partial \rho_u} \Delta_{1j} - 2 \frac{d \lambda_1 \lambda_u}{\partial \rho_1} \rho_1 - 2 \sum_{j \neq 1} \frac{d \lambda_1 \lambda_j}{\partial \rho_u} \rho_j - 2 \lambda_1 \lambda_u - \sum_{j \neq k \neq 1} \frac{\partial^2 \lambda_j \lambda_k}{\partial^2 \rho_1 \rho_u} \Delta_{jk} - 2 \sum_{j \neq u, 1} \frac{\partial \lambda_j \lambda_u}{\partial \rho_1} \rho_j \right)$$

The only terms that survive in the second parenthesis are

$$-2 \frac{\partial \lambda_1 \lambda_u}{\partial \rho_1} \rho_1$$

In the limit, this is equal to

$$-2 \frac{\partial \lambda_u}{\partial \rho_1} \mid_{\rho = \tilde{\rho}}$$

We see that this exactly cancels with what we got in the first term. Hence,

$$\frac{\partial^2 f}{\partial \rho_1 \partial \rho_u} \mid_{\rho = \tilde{\rho}} = 0$$

Similarly, for $u \neq v \neq 1$ we get

$$\frac{\partial^2 f}{\partial \rho_u d\rho_v} = 0$$

By similar calculations we can easily show that the third derivatives will also be bounded in a neighborhood of $\tilde{\rho}$. By using the second order Taylor Theorem, we get that

$$\begin{aligned}\rho_1^{t+1} &= 1 + (\rho_1^t - 1)\frac{\partial f}{\partial \rho_1} \mid_{\rho=\tilde{\rho}} + \frac{1}{2}\frac{\partial^2 f}{\partial^2 \rho_1} \mid_{\rho=\tilde{\rho}} (1-\rho_1^t)^2 + O(\epsilon^3) \\ &= \rho_1^t - C|\rho_1^t - 1|^2 + O(\epsilon^3)\end{aligned}$$

■

Using Lemma 27, it is clear that there is a second order term that is pushing ρ_1^t away from 1. Notice that this implies that we do not have a strict saddle point. However, the third order term depends on the distance of all the other coordinates from their fixpoint values. Hence, in order for the second order term to dominate the third order term, we need to prove that $|\rho_i^t - \rho_i^*\rho_1^*|$ is roughly the same size as $1 - \rho_1^t$ for all $i \neq 1$. Obviously, with a bad initialization, this might not happen. However, we can prove that after one iteration of the EM, this will always happen, no matter the starting point. This is the content of the following lemma.

Lemma 28 *Let ρ^t be the current iteration of EM. Then, there exists an $\epsilon > 0$ such that if $\|\rho^t - \tilde{\rho}\|_\infty < \epsilon$, then there exists an absolute constant $M > 0$ such that for all $i \neq 1$*

$$|\rho_i^{t+1} - \rho_i^*\rho_1^*| \leq M(1 - \rho_1^{t+1})$$

Proof Let us fix an $i \neq 1$ and define the functions

$$\begin{aligned} f_i(\rho) &:= \frac{\rho_i + \sum_{j\neq i}(\rho_i\rho_j - \rho_i^*\rho_j^*)\lambda_j}{\sqrt{1 + \sum_{j,k:\, j\neq k}(\rho_j\rho_k - \rho_j^*\rho_k^*)\lambda_j\lambda_k}} \\ f_1(\rho) &:= \frac{\rho_1 + \sum_{j\neq 1}(\rho_1\rho_j - \rho_1^*\rho_j^*)\lambda_j}{\sqrt{1 + \sum_{j,k:\, j\neq k}(\rho_j\rho_k - \rho_j^*\rho_k^*)\lambda_j\lambda_k}} \\ r_i(\rho) &:= \frac{f_i(\rho) - \rho_i^*\rho_1^*}{f_1(\rho) - 1} \end{aligned}$$

We will show that the limit

$$\lim_{\rho\to\tilde{\rho}} r_i(\rho)$$

exists and is bounded. First, let's fix the coordinates ρ_i for $i \neq 1$ and have $\rho_1 \to 1$. This means that $\lambda_1 \to 1$ and $\lambda_j \to 0$ for $j \neq 1$. Notice however that we do not have $\Delta_{ij} \to 0$, as the values of the other coordinates are not in the fixpoint yet. However, we will see that

this will not matter for the computation of the limit. We have by L'Hospital's rule

$$\lim_{\rho_1 \to 1} r_i(\rho) = \lim_{\rho_1 \to 1} \frac{\frac{\partial f_i(\rho)}{\partial \rho_1}}{\frac{\partial f_1(\rho)}{\partial \rho_1}}$$
$$= \sum_{j \neq i} \frac{\partial \lambda_j}{\partial \rho_1} \Delta_{ij} - \lambda_1 \rho_i - \sum_{j \neq 1} \frac{\partial \lambda_j}{\partial \rho_1} \Delta_{1j}$$

Now, we take the limit with respect to the remaining variables, which gives

$$\lim_{\rho \to \tilde{\rho}} r_i(\rho) = \sum_{j \neq i} \frac{\partial \lambda_j}{\partial \rho_1} \mid_{\rho = \tilde{\rho}} \rho_i^* \rho_j^* (1 - (\rho_1^*)^2) - \rho_i^* \rho_1^*$$

This is a finite quantity that depends only on the true parameters, hence it is a constant. Since the function $r_i(\rho)$ is continuous, we get that there exists an $\epsilon > 0$ such that if $\|\rho - \tilde{\rho}\|_\infty < \epsilon$ then

$$|r_i(\rho)| \leq M$$

for some $M > 0$. Substituting $\rho = \rho^t$ gives

$$\left| \frac{\rho_i^{t+1} - \rho_i^* \rho_1^*}{\rho_1^{t+1} - 1} \right| \leq M$$

which is what we wanted to prove.

■

With Lemmas 27 and 28 in our hands, we can proceed to prove Theorem 26. The idea is the following: if in the starting point ρ^0 the distance of ρ_1^0 from 1 is much less that the distances of the other coordinates from the fixpoint values, then in the following iteration they will all come closer to the fixpoint, so that all the distances are comparable. This means that even if the algorithm starts from a bad angle, it will quickly correct itself to a good angle in the next iteration. After that, in the following iterations, the second order term computed in Lemma 27 will dominate the third order term and repel ρ_1 away from 1, until it gets out of the ball.

Proof [Proof of Theorem 26] Let ϵ_1 be the value of ϵ from Lemma 27 and ϵ_2 be the value from Lemma 28. Also, let ϵ_3 be small enough so that

$$C\epsilon^2/2 > KM\epsilon^3$$

for all $\epsilon < \epsilon_3$. We choose $\epsilon_0 = \min(\epsilon_1, \epsilon_2, \epsilon_3)$. Let ρ^0 be a starting point of EM with $\|\rho^0 - \tilde{\rho}\|_2 \leq \epsilon_0$ and $\rho_i^0 < 1$ for all i.

Let $t \geq 1$. If $\|\rho^{t'} - \tilde{\rho}\|_2 > \epsilon_0$ for some $t' \leq t$, then we have nothing to prove. Suppose $\|\rho^{t'} - \tilde{\rho}\|_2 < \epsilon_0$ for all $t' \leq t$. Then, by applying Lemma 28 we get that

$$|\rho_i^t - \rho_i^* \rho_1^*| \leq M(1 - \rho_1^t)$$

for all $i \neq 1$. This means that by applying Lemma 27 and by the particular choice of ϵ_0, we have that

$$\rho_1^{t+1} \leq \rho_1^t - C(1 - \rho_1^t)^2 + KM(1 - \rho_1^t)^3 \leq \rho_1^t - \frac{C}{2}(1 - \rho_1^t)^2$$

This implies that

$$1-\rho_1^{t+1} \geq 1-\rho_1^t + \frac{C}{2}(1-\rho_1^t)^2$$

This holds for all times up to t. If we set $\alpha_t = 1-\rho_1^t$, this implies the recursive relation

$$\alpha^{t+1} \geq \alpha^t + \frac{C}{2}(\alpha^t)^2$$

It is obvious that α^t is an increasing function of t, meaning that $\alpha^t \geq \alpha^0$. This means that $\alpha^{t+1} \geq \alpha^t + \frac{C}{2}(\alpha^0)^2$. Since $\alpha^0 = 1-\rho_1^0 > 0$, we have that $\lim_{t\to\infty} \alpha^t = \infty$. This implies that there exists a finite time T such that

$$1-\rho_1^t = \alpha^t > \epsilon_0$$

This implies that $\|\rho^t - \tilde{\rho}\|_2 > \epsilon_0$ and the proof is complete.

■

A.3. Deffered proofs from Section 2

Proof [Proof of Lemma 2] Using Lemma 1 and Claim 1, we derive that

$$\begin{aligned}\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \mathop{\mathbb{E}}_{\mu^{t,*}}[x_i y] = \mathop{\mathbb{E}}_{\mu_x^*}\left[\mathop{\mathbb{E}}_{\mu_{y|x}^t}[x_i y]\right] = \mathop{\mathbb{E}}_{\mu_x^*}\left[x_i \mathop{\mathbb{E}}_{\mu_{y|x}^t}[y]\right] = \mathop{\mathbb{E}}_{\mu_x^*}\left[x_i \sum_{j=1}^n \frac{\sigma_y^t}{\sigma_j^t}\lambda_j^t x_j\right] \\ = \sum_{j=1}^n \frac{\sigma_y^t}{\sigma_j^t}\lambda_j^t \mathop{\mathbb{E}}_{\mu_x^*}[x_i x_j] = \sigma_y^t\sigma_i^t\lambda_i^t + \sum_{j\neq i}\sigma_y^t\sigma_i^t\lambda_j^t\rho_i^*\rho_j^* = \sigma_y^t\sigma_i^t\left(\lambda_i^t + \sum_{j\neq i}\lambda_j^t\rho_i^*\rho_j^*\right) \quad .\end{aligned} \tag{12}$$

This concludes the first expression for $\mathbb{E}_{\mu^{t+1}}[x_i y]$. Next, we would like to show that the second expression equals the first expression. First, recall that in (12) we have shown that

$$\mathop{\mathbb{E}}_{\mu_x^*}\left[\mathop{\mathbb{E}}_{\mu_{y|x}^t}[x_i y]\right] = \sigma_y^t\sigma_i^t\left(\lambda_i^t + \sum_{j\neq i}\lambda_j^t\rho_i^*\rho_j^*\right) \quad .$$

If we substitute μ^t instead of μ^*, we derive that

$$\mathop{\mathbb{E}}_{\mu_x^t}\left[\mathop{\mathbb{E}}_{\mu_{y|x}^t}[x_i y]\right] = \sigma_y^t\sigma_i^t\left(\lambda_i^t + \sum_{j\neq i}\lambda_j^t\rho_i^t\rho_j^t\right) \quad .$$

Since the right hand side equals

$$\mathop{\mathbb{E}}_{\mu^t}[x_i y] = \sigma_i^t\sigma_y^t\rho_i^t,$$

we derive that

$$\sigma_i^t\sigma_y^t\rho_i^t - \sigma_i^t\sigma_y^t\left(\lambda_i^t + \sum_{j\neq i}\lambda_j^t\rho_i^t\rho_j^t\right) = 0 \quad . \tag{13}$$

Adding the left hand side of (13) to (12), we derive that

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \sigma_i^t \sigma_y^t \left(\rho_i^t + \sum_{j \neq i} (\rho_i^* \rho_j^* - \rho_i^t \rho_j^t) \lambda_j^t \right),$$

which is exactly the second expression for $\mathbb{E}_{\mu^{t+1}}[x_i y]$. Next, we compute the variance for y, again, using Lemma 1 and Claim 1:

$$\begin{aligned} \mathop{\mathbb{E}}_{\mu^{t+1}}[y^2] = \mathop{\mathbb{E}}_{\mu^{t,*}}[y^2] = \mathop{\mathbb{E}}_{\mu_x^*}\left[\mathop{\mathbb{E}}_{\mu_{y|x}^t}[y^2] \right] = \mathop{\mathbb{E}}_{\mu_x^*}\left[\left(\sum_{j=1}^n \frac{\sigma_y^t}{\sigma_j^t} \lambda_j^t x_j \right)^2 \right] \\ = \sum_{i=1}^n \sum_{j=1}^n \frac{\sigma_y^t}{\sigma_i^t} \lambda_i^t \frac{\sigma_y^t}{\sigma_j^t} \lambda_j^t \mathop{\mathbb{E}}_{\mu_x^*}[x_i x_j] = (\sigma_y^t)^2 \left(\sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \rho_i^* \rho_j^* \right). \end{aligned} \tag{14}$$

This derives the first expression for the variance of y. We derive the second expression using a similar logic as in the calculation in the second expression for $\mathbb{E}_{\mu^{t+1}}[x_i y]$. First, substituting ρ^t instead of ρ^* in (14), we derive that

$$(\sigma_y^t)^2 = \mathop{\mathbb{E}}_{\mu^t}[y^2] = \mathop{\mathbb{E}}_{\mu_x^t}\left[\mathop{\mathbb{E}}_{\mu_{y|x}^t}[y^2] \right] = (\sigma_y^t)^2 \left(\sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \rho_i^t \rho_j^t \right).$$

This implies that

$$(\sigma_y^t)^2 \left(1 - \sum_{i=1}^n (\lambda_i^t)^2 - \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \rho_i^t \rho_j^t \right) = 0.$$

Adding this to the expression in (14), this yields the second expression for the covariance of y, as required.

Lastly, the expression of ρ_i^{t+1} is derived, by definition of ρ_i,

$$\rho_i^{t+1} = \frac{\mathbb{E}_{\mu^{t+1}}[x_i y]}{\sqrt{\mathrm{Var}_{\mu^{t+1}}[x_i] \, \mathrm{Var}_{\mu^{t+1}}[y]}},$$

and by substituting the expressions for the covariance of x_i and y, the variance of y, and using Lemma 1 which argues that $\mathrm{Var}_{\mu^{t+1}}[x_i] = \sigma_i^2$. ■

Proof [Proof of Lemma 5] Yuval: Elaborate if there's time The convergence of $\overline{\rho}^t$ to some $\overline{\rho} \in \mathcal{S}$ follows directly from Wu (1983). For the second part of the lemma, assume that $\overline{\rho} \in \mathcal{S}$ and by definition of a stationary point, $\overline{\rho}^{t+1} = \overline{\rho}^t$. Further, by Lemma 2, we know that $\sigma_{x_i}^{t+1} = \sigma_{x_i}^* = \sigma_{x_i}^t$ for all t. This implies that $\mu_{x_1 \cdots x_n}^t = \mu_{x_1 \cdots x_n}^{t+1}$. In particular, $\mathrm{KL}(\mu^* \| \mu^t) = \mathrm{KL}(\mu^* \| \mu^{t+1})$. Yet, Wu (1983) implies that if $\mu^{t+1} \neq \mu^t$ then $\mathrm{KL}(\mu^* \| \mu^{t+1}) < \mathrm{KL}(\mu^* \| \mu^t)$. Hence, $\mu^{t+1} = \mu^t$ as required. ■

Proof [Proof of Lemma 11] We will analyze the transpose of the matrix in (5), assuming that $J_{ij} = u_j$ for any $i \neq j$. Indeed, the original matrix is singular if and only if its transpose is.

We would like to show that there is no nontrivial solution a for $Ja = 0$. We have

$$(Ja)_i = \sum_{j\neq i} a_j u_j + a_i(\sum_{j\neq i} u_j) = \sum_j a_j u_j + a_i(\sum_j u_j - 2u_i).$$

Substitute $s = \sum_j u_j$ and $u = \sum_j a_j u_j$, we have

$$(Ja)_i = u + a_i(s - 2u_i).$$

To solve the system of equations $Ja = 0$, we assume without loss of generality that $u_1 = \max_j u_j$ and divide into cases according to u_1. First, assume that $u_1 < s/2$. Then,

$$(Ja)_i = u + a_i(s - 2u_i) = 0 \tag{15}$$

implies

$$a_i = \frac{-u}{s - 2u_i}.$$

The equation $\sum_i u_i a_i = u$ implies

$$\sum_i u_i a_i = \sum_i \frac{-uu_i}{s - 2u_i} = u. \tag{16}$$

Notice that $u \neq 0$. Assume towards contradiction that $u = 0$. Then, we have

$$0 = a_i(s - 2u_i).$$

Dividing by $s-2u_i$ we get $a_i = 0$ for all i. Since we look for nontrivial solutions a to $Ja = 0$, we assume that $a \neq 0$ which implies that $u \neq 0$. Dividing (16) by u we get

$$\sum_i \frac{-u_i}{s - 2u_i} = 1. \tag{17}$$

By the assumption $u_1 < s/2$ we have $s-2u_i \geq s-2u_1 > 0$. In particular, the left hand side of (17) is negative while the right hand side is positive, hence there is no solution! Next, assume that $u_1 > s/2$. In this case, we have $u_i < s/2$ for all $i \geq 2$, since $\sum_i u_i = s$. Here, (17) is still valid, and is equivalent to

$$\frac{u_1}{2u_1 - s} - 1 = \sum_{i>1} \frac{u_i}{s - 2u_i},$$

which is equivalent to

$$\frac{s - u_1}{2u_1 - s} = \sum_{i>1} \frac{u_i}{s - 2u_i} \quad .$$

Here, we will show that the left hand side is strictly greater than the right hand side, which implies that there is no equality. First, notice that for any $i \geq 2$, $2u_1 - s < s - 2u_i$. Indeed, this follows from $s = \sum_j u_j > u_1 + u_i$, since all u_j are positive and $n \geq 3$. We derive that

$$\frac{s - u_1}{2u_1 - s} = \sum_{i\geq 2} \frac{u_i}{2u_1 - s} > \sum_{i\geq 2} \frac{u_i}{s - 2u_i} \quad ,$$

which arrives at a contradiction. Lastly, assume that $u_1 = s/2$. Since $s = \sum_{i=1}^{n} u_i$ and $n \geq 3$, then for all $i \geq 2$ we have $u_i < s/2$. Recall (15) which states that

$$0 = (Ja)_i = u + a_i(s - 2u_i).$$

Applying with $i = 1$ we obtain

$$u = 0.$$

For all $i \geq 2$ we have

$$a_i(s - 2u_i) = 0$$

which implies that $a_i = 0$ for $i \geq 2$. By definition of u and by the computations above,

$$0 = u = \sum_i a_i u_i = a_1 u_1.$$

Since $u_1 > 0$ we have

$$a_1 = 0.$$

We conclude that $a_i = 0$ for all i which implies that there is no nontrivial solution for $Ja = 0$, as required. ■

Appendix B. Upper and Lower bounds for EM iteration

We will denote by ρ^t, θ^t the vector of correlations, covariances at time t. We also denote by μ_ρ the density of the distribution on leaves with correlations ρ. Our first task will be to show that if we initialize EM at a constant distance away from the optimum, it will always remain in a bounded distance within that optimum.

Theorem 29 *Suppose that* $\min_i \rho_i^0 \geq c_1, 1 - \max_i \rho_i^0 \geq c_2$, *where* $c_1, c_2 \in (0,1)$ *are constants. Suppose we have access to* m *i.i.d. samples* $x^{(1)}, \ldots, x^{(k)}$ *from the distribution* μ_x^*. *Let* ρ_i^t *be the correlations produced by the EM iteration run using these samples. Then, with probability at least* $1 - \delta$, *there are absolute constants* $C, C' > 0$ *such that for all* $t > 0$, i *and* $k = 5$*:*

$$\frac{C}{n^{k+2}} \leq \rho_i^t \leq 1 - \frac{C'}{n^{4k+9}}$$

We will break the proof of Theorem 29 into multiple lemmas over the next few Sections.

B.1. Two correlations cannot be close to 1

We will first be concerned with establishing that all ρ_i remain bounded away from 1 in all iterations. Our first Lemma contributes in this direction. It says that no 2 coordinates i, j can have ρ_i, ρ_j simultaneously close to 1.

Lemma 30 *Suppose that* $\min_i \rho_i^0 \geq c_1, 1 - \max_i \rho_i^0 \geq c_2$, *where* $c_1, c_2 \in (0,1)$ *are constants. Suppose we have access to* m *i.i.d. samples* $x^{(1)}, \ldots, x^{(k)}$ *from the distribution* μ_x^*. *Let* ρ_i^t *be the correlations produced by the EM iteration run using these samples. Then, with probability*

at least $1-\delta$*, there exists a constant* $c = c(c_1, c_2, \rho^*)$ *such that for all* t*, there exists at most one* $i \in [n]$ *such that* $\rho_i^t > 1 - c/(n^2\eta)$*, where*

$$\eta := \sqrt{\frac{\log(n/\delta)}{m}}$$

Proof We use the fact that EM always improves the value of the likelihood function. We consider the likelihood of a given observation on the leaves as a function of the standard deviations at the leaves σ_{x_i} and the correlations ρ_i between x_i, y. Let σ_x be the vector of standard deviations and ρ the vector of correlations. The empirical likelihood function can be written as

$$L(\rho, \sigma_x; x^{(1)}, \dots, x^{(k)}) = \frac{1}{m}\sum_{k=1}^{m} \log \mu_{\rho,\sigma_x}(x^{(k)})$$

From now on, we will omit the dependence on the samples $x^{(1)}, \dots, x^{(k)}$ whenever it is implied. We have that

$$\begin{aligned} L(\rho, \sigma_x) &= \frac{1}{m}\sum_{k=1}^{m} \log \mu_{\rho,\sigma_x}(x^{(k)}) \\ &= -\frac{1}{2}\log\left(2\pi\left|\Sigma_{\rho,\sigma_x}\right|\right) - \frac{1}{2m}\sum_{k=1}^{m}(x^{(k)})^\top \Sigma_{\rho,\sigma_x}^{-1} x^{(k)} \\ &= -\frac{1}{2}\log\left(2\pi\left|\Sigma_{\rho,\sigma_x}\right|\right) - \frac{1}{2}tr\left(\Sigma_{\rho,\sigma_x}^{-1}\hat{\Sigma}\right) \end{aligned} \tag{18}$$

where $\hat{\Sigma}$ is the empirical covariance matrix of the samples. Recall the closed form expression for KL of Gaussians.

$$KL(\mathcal{N}(0,\Sigma_1)||\mathcal{N}(0,\Sigma_2)) = \frac{1}{2}\left(\log\frac{|\Sigma_2|}{|\Sigma_1|} - n + tr(\Sigma_2^{-1}\Sigma_1)\right)$$

From this, we conclude that

$$L(\rho, \sigma_x) = -\frac{n}{2} - \frac{1}{2}\log\left(2\pi\left|\hat{\Sigma}\right|\right) - KL(\mathcal{N}(0,\hat{\Sigma})||\mathcal{N}(0,\Sigma_{\rho,\sigma_x}))$$

Let us start by lower bounding $L(\rho^0, \sigma_x^0)$. We are going to upper bound $KL(\mathcal{N}(0,\hat{\Sigma})||\mathcal{N}(0,\Sigma_{\rho^0,\sigma_x^0})))$. The matrix Σ_{ρ,σ_x} can be written as a sum of a rank 1 matrix and a diagonal matrix. We will use this to evaluate determinants and inverses of these matrices. Specifically, for any ρ, if σ_x is the all 1's vector, we have that

$$\Sigma_{\rho,\sigma_x} = \begin{pmatrix} 1-\rho_1^2 & 0 & \dots & 0 \\ 0 & 1-\rho_2^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1-\rho_n^2 \end{pmatrix} + \rho\rho^\top \tag{19}$$

We will denote this diagonal matrix as $diag(1-\rho^2)$ for convenience. From now on, we will omit the dependence on σ_x if we have unit variances (which happens for ρ^0 and ρ^* for

example). Using the Sherman-Morrison formula, we get that

$$\Sigma_\rho^{-1} = diag(1-\rho^2)^{-1} - \frac{diag(1-\rho^2)^{-1}\rho\rho^\top diag(1-\rho^2)^{-1}}{1+\rho^\top diag(1-\rho^2)^{-1}\rho} \tag{20}$$

We will first try to bound $tr(\Sigma_{\rho^0}^{-1}\Sigma_{\rho^*})$ and then show that it is close to $tr(\Sigma_{\rho^0}^{-1}\hat{\Sigma})$ Thus, after some algebraic manipulations, we have that

$$\begin{aligned} tr(\Sigma_{\rho^0}^{-1}\Sigma_{\rho^*}) &= n + \frac{\left(\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}\right)^2 - \left(\sum_i \frac{\rho_i^0\rho_i^*}{1-(\rho_i^0)^2}\right)^2 + \sum_i \frac{(\rho_i^0)^2(\rho_i^*)^2}{(1-(\rho_i^0)^2)^2} - \sum_i \frac{(\rho_i^0)^4}{(1-(\rho_i^0)^2)^2}}{1+\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}} \\ &= n + \frac{\left(\sum_i \frac{(\rho_i^0)(\rho_i^0-\rho_i^*)}{1-(\rho_i^0)^2}\right)\left(\sum_i \frac{(\rho_i^0)(\rho_i^0+\rho_i^*)}{1-(\rho_i^0)^2}\right) + \sum_i \frac{(\rho_i^0)^2((\rho_i^*)^2-(\rho_i^0)^2)}{(1-(\rho_i^0)^2)^2}}{1+\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}} \end{aligned}$$

Now, notice that the function $x \mapsto x^2/(1-x^2)$ is increasing. This implies, using the assumption of the Lemma, that

$$\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2} \geq n\frac{c_1^2}{1-c_1^2}$$

Also, since the function $x \mapsto x/(1-x^2)$ is increasing, this implies that

$$\sum_i \frac{(\rho_i^0)(\rho_i^0+\rho_i^*)}{1-(\rho_i^0)^2} \leq 2n\frac{1-c_2}{1-(1-c_2)^2}$$

and

$$\left|\sum_i \frac{(\rho_i^0)(\rho_i^0-\rho_i^*)}{1-(\rho_i^0)^2}\right| \leq 2n\frac{1-c_2}{1-(1-c_2)^2}$$

Also, for all i we have

$$|(\rho_i^*)^2 - (\rho_i^0)^2| = |\rho_i^* - \rho_i^0||\rho_i^* + \rho_i^0| \leq 2$$

By monotonicity again, this implies

$$\left|\sum_i \frac{(\rho_i^0)^2((\rho_i^*)^2-(\rho_i^0)^2)}{(1-(\rho_i^0)^2)^2}\right| \leq 2\sum_i \frac{(\rho_i^0)^2}{(1-(\rho_i^0)^2)^2} \leq 2n\frac{(1-c_2)^2}{(1-(1-c_2)^2)^2}$$

Putting all these facts together, we conclude that there exists a constant $C = C(c_1, c_2)$, such that

$$\left|\frac{\left(\sum_i \frac{(\rho_i^0)(\rho_i^0-\rho_i^*)}{1-(\rho_i^0)^2}\right)\left(\sum_i \frac{(\rho_i^0)(\rho_i^0+\rho_i^*)}{1-(\rho_i^0)^2}\right) + \sum_i \frac{(\rho_i^0)^2((\rho_i^*)^2-(\rho_i^0)^2)}{(1-(\rho_i^0)^2)^2}}{1+\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}}\right| \leq Cn \tag{21}$$

Now, let's focus on $tr(\Sigma_{\rho^0}^{-1}\hat{\Sigma})$. By standard Chernoff bounds, we have that with probability at least $1-\delta$, for all i,j

$$\left|(\Sigma_{\rho^*})_{ij} - \hat{\Sigma}_{ij}\right| = O\left(\sqrt{\frac{\log(n/\delta)}{m}}\right) := \eta$$

We have

$$tr\left(\Sigma_{\rho^0}^{-1}(\hat{\Sigma} - \Sigma_{\rho^*})\right) = \sum_{i,j}(\Sigma_{\rho^0}^{-1})_{ij}\left((\Sigma_{\rho^*})_{ij} - \hat{\Sigma}_{ij}\right)$$

Hence, all we have to do is bound the entries of $\Sigma_{\rho^0}^{-1}$. For this, we can use the Sherman-Morrison formula, which we also used earlier. Let's start with a non-diagonal element $i \neq j$. Then, the formula gives

$$\left|(\Sigma_{\rho^0})_{ij}\right| = \frac{\frac{\rho_i^0}{1-(\rho_i^0)^2}\frac{\rho_j^0}{1-(\rho_j^0)^2}}{1+\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}} \leq \frac{1}{c_2^2}$$

For $i=j$, we have

$$\left|(\Sigma_{\rho^0})_{ii}\right| = \left|\frac{1}{1-(\rho_i^0)^2} - \frac{\left(\frac{\rho_i^0}{1-(\rho_i^0)^2}\right)^2}{1+\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}} \leq \frac{1}{c_2^2}\right| \leq \frac{1}{c_2} + \frac{1}{c_2^2}$$

Hence, with probability at least $1-\delta$

$$\left|tr\left(\Sigma_{\rho^0}^{-1}(\hat{\Sigma} - \Sigma_{\rho^*})\right)\right| \leq Cn^2\eta \tag{22}$$

where C is some constant. Now, let's calculate the determinant. We use the matrix determinant lemma to compute $|\Sigma_\rho|$.

$$|\Sigma_\rho| = (1+\rho^\top diag(1-\rho^2)^{-1}\rho)|diag(1-\rho^2)| = (1+\sum_i \frac{\rho_i^2}{1-\rho_i^2})\prod_i(1-\rho_i^2)$$

This gives us

$$\log\frac{|\Sigma_{\rho^0}|}{|\Sigma_{\rho^*}|} = \log\frac{1+\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2}}{1+\sum_i \frac{(\rho_i^*)^2}{1-(\rho_i^*)^2}} + \sum_i \log\frac{1-(\rho_i^0)^2}{1-(\rho_i^*)^2}$$

The function $x \mapsto \log(1-x^2)$ is decreasing, hence we have

$$\sum_i \log\frac{1-(\rho_i^0)^2}{1-(\rho_i^*)^2} \leq n\log\frac{1-c_1^2}{c_2}$$

We also have

$$\sum_i \frac{(\rho_i^0)^2}{1-(\rho_i^0)^2} \leq n\frac{1}{c_2}$$

Finally

$$\sum_i \frac{(\rho_i^*)^2}{1-(\rho_i^*)^2} \geq Cn$$

It follows that there exists a constant $C' = C'(M, \rho^*)$, such that

$$\left|\log \frac{|\Sigma_{\rho^0}|}{|\Sigma_{\rho^*}|}\right| \leq C'n \tag{23}$$

However, in the expression we have $\hat{\Sigma}$ instead of Σ^*. We use the property that

$$\frac{\partial \log |A|}{\partial A} = A^{-1}$$

We already showed that the entries of Σ_ρ^{-1} are bounded if ρ is upper and lower bounded by constants. Using this and Taylor's Theorem, we get

$$\left|\log \frac{|\Sigma_{\rho^*}|}{|\hat{\Sigma}|}\right| \leq Cn^2\eta \tag{24}$$

for some constant C.

Using inequalities (21) and (23) together with the expression of KL, we conlcude that

$$KL(\mathcal{N}(0, \Sigma_{\rho^*})||\mathcal{N}(0, \Sigma_{\rho^0})) \leq Kn$$

where $K = K(\rho^*)$. Also, using this result and inequalities (24) and (22) we get

$$KL(\mathcal{N}(0, \hat{\Sigma})||\mathcal{N}(0, \Sigma_{\rho^0})) \leq K'n^2\eta$$

Overall, by equation(18) we get that

$$L(\rho^0) \geq -\frac{n}{2} - \frac{1}{2}\log\left(2\pi\left|\hat{\Sigma}\right|\right) - K'n^2\eta$$

for some $K' = K'(c_1, c_2, \rho^*)$.

Now, suppose for some iteration t there exist two indices i, j with $\rho_i^t > 1-\epsilon, \rho_j^t > 1-\epsilon$, where ϵ will be determined in the sequel. Now, let us upper bound $L(\rho^t)$. By KL subadditivity, we have

$$\begin{aligned} L(\rho^t, \sigma_x^t) = &-\frac{n}{2} - \frac{1}{2}\log\left(2\pi\left|\hat{\Sigma}\right|\right) - KL(\mathcal{N}(0, \hat{\Sigma})||\mathcal{N}(0, \Sigma_{\rho^t, \sigma_x^t})) \\ &-\frac{n}{2} - \frac{1}{2}\log\left(2\pi\left|\hat{\Sigma}\right|\right) - KL(\mathcal{N}(0, \hat{\Sigma}_{x_i,x_j})||\mathcal{N}(0, (\Sigma_{\rho^t, \sigma_x^t})_{x_i,x_j})) \end{aligned}$$

where in the last equation we are comparing the marginals of the distributions on x_i, x_j. The marginal distribution of x_i, x_j is a gaussian with zero mean. First, we will analyze the

ideal situation where we have Σ^* instead of $\hat{\Sigma}$. Using the closed form of the KL divergence between two Gaussians, we obtain:

$$\begin{aligned}
&KL(\mathcal{N}(0,\Sigma^*_{x_i,x_j})||\mathcal{N}(0,(\Sigma_{\rho^t,\sigma_x^t})_{x_i,x_j}))\\
&=\frac{1}{2}\left(\log\frac{1-(\rho_i^t\rho_j^t)^2}{1-(\rho_i^*\rho_j^*)^2}-2+\frac{1}{(1-(\rho_i^t\rho_j^t)^2}tr\left(\begin{pmatrix}1 & -\rho_i^t\rho_j^t\\ -\rho_i^t\rho_j^t & 1\end{pmatrix}\begin{pmatrix}1 & \rho_i^*\rho_j^*\\ \rho_i^*\rho_j^* & 1\end{pmatrix}\right)\right)\\
&=\frac{1}{2}\left(\log\frac{1-(\rho_i^t\rho_j^t)^2}{1-(\rho_i^*\rho_j^*)^2}-2+2\frac{1-\rho_i^t\rho_j^t\rho_i^*\rho_j^*}{1-(\rho_i^t\rho_j^t)^2}\right)
\end{aligned}$$

Then,

$$KL(\mathcal{N}(0,\Sigma^*_{x_i,x_j})||\mathcal{N}(0,(\Sigma_{\rho^t,\sigma_x^t})_{x_i,x_j}))\geq\frac{1}{2}\left(\log(1-(\rho_i^t\rho_j^t)^2)-2+2\frac{1-\rho_i^*\rho_j^*/4}{1-(\rho_i^t\rho_j^t)^2}\right)$$

Now, for any constant $c>0$, consider the function $f(x)=x(\log x+c/x)$. This is a continuous function on $(0,1)$ and it is easy to see that it is decreasing for $x<1/e$. Let x_0 be a small enough constant such that $x_0\log x_0+c>c/2$. Then, for $x<x_0$

$$f(x)\geq f(x_0)\implies\log x+c/x\geq\frac{c/2}{x}$$

Now let $c=2(1-\rho_i^*\rho_j^*/4)$. If we set $\epsilon=c'/(\eta n^2)$ for some suitable constant c', then we have that

$$1-(\rho_i^t\rho_j^t)^2<x_0$$

and

$$KL(\mathcal{N}(0,\Sigma^*_{x_i,x_j})||\mathcal{N}(0,(\Sigma_{\rho^t,\sigma_x^t})_{x_i,x_j}))>(K'+K'')n^2\eta$$

where K'' is a constant that will be determined now. Using the exact same arguments as in the case of ρ^0, we can show that

$$\left|KL(\mathcal{N}(0,\Sigma^*_{x_i,x_j})||\mathcal{N}(0,(\Sigma_{\rho^t,\sigma_x^t})_{x_i,x_j}))-KL(\mathcal{N}(0,\Sigma^*_{x_i,x_j})||\mathcal{N}(0,(\Sigma_{\rho^t,\sigma_x^t})_{x_i,x_j}))\right|\leq K''\eta$$

for some constant C. It follows that

$$L(\rho^t,\sigma_x^t)<-\frac{n}{2}-\frac{1}{2}\log\left(2\pi\left|\hat{\Sigma}\right|\right)-K'n^2\eta<L(\rho^0)$$

This is a contradiction, since the likelihood value increases at each step of EM. This gives us the desired result.

■

B.2. No correlation can be too close to 1

In the previous Section, we showed that two corrrelations cannot be close to 1 at the same time. In this Section, building on this result, we show that in fact no correlation can be too close to 1. Hence, each ρ_i is upper bounded at all iterations of the algorithm. This is the topic of the following Lemma:

Lemma 31 *Suppose $\|\rho^0 - \rho^*\|_\infty \leq M$. Then, for all i and for all t, $\rho_i^t \leq 1 - C''/n^{4k+9}$, where C'' is an absolute constant and $k \geq 5$.*

Remark 32 *The proof is assumes that the executed dynamics are according to the population EM. The extension to the sample EM is straightforward.*

To prove it, we will essentially prove that no ρ_i will ever get really close to 1. The proof will be similar to the one that established divergence from the saddle points at the boundary. We start by stating a direct Corollary of Lemma 30.

Corollary 33 *Suppose $\|\rho^0 - \rho^*\|_\infty \leq M$. If for some t we have $\rho_1^t > 1 - c/n$, then for all $i \neq 1$ we have $\rho_i^t \leq 1 - c/n$, where c is the constant of Lemma 30.*

Our strategy now will be the following: we want to show that ρ_1 will not be very close to 1, without loss of generality. First of all, we want to argue that if it comes close to 1, it will immediately start moving away. We do this by showing that after one iteration of EM, the errors of the others will be comparable to the error of ρ_1. We showed a similar claim in the proof of divergence from stationary points. The difference here is that the errors of the other ρ_i can be close to 1, while in the original proof they were assumed to lie in some ball around the fixpoint. Hence, we prove the following Lemma.

Lemma 34 *Suppose that for some t we have $\rho_i^t \leq \alpha$ for all $i \neq 1$ and $\rho_1^t > \beta$, for some $\alpha, \beta \in (0,1)$. Define*

$$R = R(\alpha, \beta, n) := n^2 \frac{1-\beta}{\beta^4(1-\alpha)^2} + \frac{n}{\beta^2(1-\alpha)} \tag{25}$$

and assume that $(1-\beta)R \leq 1/2$. Then, for all $i \neq 1$

$$|\rho_i^{t+1} - \rho_i^* \rho_1^*| \leq C \frac{1 + \frac{2n}{1-\alpha} + R}{1 - \left(1 + \frac{2n}{1-\alpha}\right) \frac{n(1-\beta)}{\beta^2(1-\alpha)} - Cn^2 \frac{(1-\beta)}{\beta^4(1-\alpha)^2} - C\frac{n(1-\beta)R}{\beta^2(1-\alpha)}} (1 - \rho_1^{t+1})$$

where C is some absolute constant.

The first step to proving this Lemma will be to show that if all ρ_i are bounded away from 1 except ρ_1, then λ_i will be small for $i \neq 1$. This is the content of the following Lemma.

Lemma 35 *Suppose for some ρ we have $\rho_i \leq \alpha$ for all $i \neq 1$ and $\rho_1 > \beta$, for some $\alpha, \beta \in (0,1)$. Then,*

$$\lambda_1 > \frac{\rho_1}{1 + 2n\frac{1-\rho_1}{1-\alpha}}$$

and

$$\lambda_i \leq \frac{2(1-\rho_1)}{\beta^2(1-\alpha)}$$

for all $i \neq 1$.

Proof We have

$$\lambda_i = \frac{\frac{\rho_i}{1-\rho_i^2}}{1+\sum_j \frac{\rho_j^2}{1-\rho_j^2}} = \frac{\rho_i}{1+\sum_{j\neq i}\frac{\rho_j^2(1-\rho_i^2)}{1-\rho_j^2}}$$

For $i \neq 1$, this implies

$$\lambda_i \leq \frac{1}{\sum_{j\neq i}\frac{\rho_j^2(1-\rho_i^2)}{1-\rho_j^2}} \leq \frac{1-\rho_1^2}{\rho_1^2(1-\rho_i^2)}$$

We have $1-\rho_1^2 = (1-\rho_1)(1+\rho_1) < 2(1-\rho_1)$ and $1-\rho_i^2 = (1-\rho_i)(1+\rho_i) \geq 1-\alpha$. This gives

$$\lambda_i \leq \frac{2(1-\rho_1)}{\beta^2(1-\alpha)}$$

We also have

$$\lambda_1 = \frac{\rho_1}{1+\sum_{j\neq 1}\frac{\rho_j^2(1-\rho_1^2)}{1-\rho_j^2}}$$

We have that

$$\sum_{j\neq 1}\frac{\rho_j^2(1-\rho_1^2)}{1-\rho_j^2} \leq 2n\frac{1-\rho_1}{1-\alpha} \implies \lambda_1 \geq \frac{\rho_1}{1+2n\frac{1-\rho_1}{1-\alpha}}$$

■

We are now ready to prove Lemma 34.

Proof [Proof of Lemma 34] We will first prove that

$$|\rho_i^{t+1} - \rho_i^*\rho_1^*| \leq C(\alpha,\beta)|\rho_1^t - 1| \tag{26}$$

for some constant $C(\alpha,\beta)$ that will be specified in the sequel. We have that

$$|\rho_i^{t+1} - \rho_i^*\rho_1^*| = \left|\frac{\rho_i^t + \sum_{j\neq i}(\rho_i^*\rho_j^* - \rho_i^t\rho_j^t)\lambda_j^t}{\sqrt{1+\sum_{j\neq k}\lambda_j^t\lambda_k^t(\rho_j^*\rho_k^* - \rho_j^t\rho_k^t)}} - \rho_i^*\rho_1^*\right|$$

As usual, denote $\Delta_{ij}^t = \rho_i^*\rho_j^* - \rho_i^t\rho_j^t$. Also, let

$$U^t := \sum_{j\neq k}\lambda_j^t\lambda_k^t(\rho_j^*\rho_k^* - \rho_j^t\rho_k^t)$$

Hence, we can write

$$\begin{aligned}
|\rho_i^{t+1} - \rho_i^*\rho_1^*| &= \left|\frac{\rho_i^t + \sum_{j\neq i}\Delta_{ij}^t\lambda_j^t - \rho_i^*\rho_1^*\sqrt{1+U^t}}{\sqrt{1+U^t}}\right| \\
&= \left|\frac{\rho_i^t + \lambda_1(\rho_i^*\rho_1^* - \rho_i^t\rho_1^t) - \rho_i^*\rho_1^*\sqrt{1+U^t} + \sum_{j\neq i,1}\Delta_{ij}^t\lambda_j^t}{\sqrt{1+U^t}}\right| \\
&= \left|\frac{(\lambda_1 - \sqrt{1+U^t})(\rho_i^*\rho_1^* - \rho_i^t\rho_1^t) + \rho_i^t(1-\rho_1^t\sqrt{1+U^t}) + \sum_{j\neq i,1}\Delta_{ij}^t\lambda_j^t}{\sqrt{1+U^t}}\right|
\end{aligned}$$

Let's start by bounding U^t. Using lemma 35 have that for $j,k \neq 1$:

$$|\lambda_j^t \lambda_k^t \Delta_{jk}^t| \leq 2\left(\frac{2(1-\rho_1^t)}{\beta^2(1-\alpha)}\right)^2$$

If $j=1$ then

$$|\lambda_1^t \lambda_k^t(\rho_1^*\rho_k^* - \rho_1^t\rho_k^t)| \leq 2\frac{2(1-\rho_1^t)}{\beta^2(1-\alpha)}$$

Hence, we conclude that

$$|U^t| \leq n^2\frac{8(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} + 4n\frac{1-\rho_1^t}{\beta^2(1-\alpha)}$$

By Assumption(25) we have that the right hand side of the last inequality is $< 1/2$. This gives, by a simple Taylor approximation, that

$$|\sqrt{1+U^t} - 1| \leq C|U^t|$$

for some constant C. Hence,

$$\begin{aligned}|\lambda_1^t - \sqrt{1+U^t}| &\leq |\lambda_1^t - 1| + |\sqrt{1+U^t} - 1| \leq 1 - \frac{\rho_1^t}{1+2n\frac{1-\rho_1^t}{1-\alpha}} + Cn^2\frac{8(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} + 4Cn\frac{1-\rho_1^t}{\beta^2(1-\alpha)} \\ &\leq \left(1+\frac{2n}{1-\alpha}\right)(1-\rho_1^t) + Cn^2\frac{8(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} + 4Cn\frac{1-\rho_1^t}{\beta^2(1-\alpha)}\end{aligned}$$

We also have

$$|1-\rho_1^t\sqrt{1+U^t}| \leq 1-\rho_1^t + \rho_1^t|1-\sqrt{1+U^t}| \leq 1-\rho_1^t + Cn^2\frac{8(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} + 4Cn\frac{1-\rho_1^t}{\beta^2(1-\alpha)}$$

Lastly, we have

$$\sum_{j\neq i,1}\Delta_{ij}^t\lambda_j^t \leq 4n\frac{1-\rho_1^t}{\beta^2(1-\alpha)}$$

Hence,

$$\begin{aligned}|\rho_i^{t+1} - \rho_i^*\rho_1^*| &\leq \frac{\left(2+\frac{2n}{1-\alpha}\right)(1-\rho_1^t) + Cn^2\frac{8(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} + 4Cn\frac{1-\rho_1^t}{\beta^2(1-\alpha)} + 4n\frac{1-\rho_1^t}{\beta^2(1-\alpha)}}{\sqrt{1-n^2\frac{8(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} - 4n\frac{1-\rho_1^t}{\beta^2(1-\alpha)}}} \\ &\leq C(1-\rho_1^t)\left(1+\frac{2n}{1-\alpha} + n^2\frac{1-\beta}{\beta^4(1-\alpha)^2} + \frac{n}{\beta^2(1-\alpha)}\right)\end{aligned}$$

for some constant C, since we have assumed (25). If we set $C(\alpha,\beta)$ equal to the multiplier of $1-\rho_1^t$, we have obtained (26). The second step is to show that

$$1-\rho_1^{t+1} \geq K(\alpha,\beta)(1-\rho_1^t) \tag{27}$$

Inequalities (26) and (27) together give us the desired result. We have that

$$\rho_1^{t+1} = \frac{\rho_1^t + \sum_{j\neq 1}(\rho_1^*\rho_j^* - \rho_1^t\rho_j^t)\lambda_j^t}{\sqrt{1+\sum_{j\neq k}\lambda_j^t\lambda_k^t(\rho_j^*\rho_k^* - \rho_j^t\rho_k^t)}}$$

First of all, by Taylor we have that if $|x| \leq 1/2$

$$(1+x)^{-1/2} \leq 1 - \frac{x}{2} + cx^2$$

We have already shown that $|U^t|$ is smaller than a constant (using also assumption (25)). This gives

$$\begin{aligned}
\rho_1^{t+1} &\leq \left(\rho_1^t + \sum_{j\neq 1}(\rho_1^*\rho_j^* - \rho_1^t\rho_j^t)\lambda_j^t\right)\left(1 - \frac{U^t}{2} + c|U^t|^2\right) \\
&= \rho_1^t + \sum_{j\neq 1}(\rho_1^*\rho_j^* - \rho_1^t\rho_j^t)\lambda_j^t - \rho_1^t\lambda_1^t\sum_{j\neq 1}(\rho_1^*\rho_j^* - \rho_1^t\rho_j^t)\lambda_j^t - \frac{\rho_1^t}{2}\sum_{j\neq k\neq 1}\Delta_{ij}^t\lambda_j^t\lambda_k^t - \frac{U^t}{2}\sum_{j\neq 1}\Delta_{1j}^t\lambda_j^t + c|U^t|^2(\rho_1^t + \sum_{j\neq} \\
&\leq \rho_1^t + (1-\rho_1^t\lambda_1^t)\sum_{j\neq 1}\Delta_{1j}^t\lambda_j^t + Cn^2\frac{(1-\rho_1^t)^2}{\beta^4(1-\alpha)^2} + C\frac{n|U^t|(1-\rho_1^t)}{\beta^2(1-\alpha)} + C\frac{n|U^t|^2(1-\rho_1^t)}{\beta^2(1-\alpha)}
\end{aligned}$$

We know that

$$\sum_{j\neq 1}\Delta_{1j}^t\lambda_j^t \leq 4n\frac{1-\rho_1^t}{\beta^2(1-\alpha)}$$

and

$$1 - \rho_1^t\lambda_1^t = 1 - \rho_1^t + \rho_1^t(1-\lambda_1^t) \leq (1-\rho_1^t)\left(1 + \frac{2n}{1-\alpha}\right)$$

Hence, we have that

$$1 - \rho_1^{t+1} \geq (1-\rho_1^t)\left(1 - \left(1+\frac{2n}{1-\alpha}\right)\frac{n(1-\rho_1^t)}{\beta^2(1-\alpha)} - Cn^2\frac{(1-\rho_1^t)}{\beta^4(1-\alpha)^2} - C\frac{n|U^t|}{\beta^2(1-\alpha)}\right)$$

■

We now have almost all the necessary ingredients required for the proof of Lemma 31. The final Lemma we will prove says that if ρ_1 happens to come very close to 1, then after one iteration it will start moving away from 1. This is qualitatively similar to the proof of Theorem 25 in the manuscript, but now we have to make precise quantitative predictions for how close it should be to 1. This is the content of the following Lemma.

Lemma 36 *Suppose $\rho_1^t > 1 - c'/n^k > 1 - c/n$, where $c' > 0$ is a sufficiently small constant and $k \geq 5$. Then, for all $t' \geq t+1$ we have that if $\rho_1^{t'} > c'/n^k$, then for some absolute constant $C > 0$ we have*

$$\rho_1^{t'+1} \leq \rho_1^{t'} - \frac{C}{2}(1-\rho_1^{t'})^2$$

Proof We first restate, without proof, Lemma 26 from the manuscript.

Lemma 37 *Let ρ^t be the current iteration of EM and define*

$$\epsilon^t := \max(1 - \rho_1^t, \max_i |\rho_i^* \rho_1^* - \rho_i^t|)$$

Then, there exist absolute constants $C, K > 0$ such that

$$\rho_1^{t+1} \leq \rho_1^t - C|\rho_1^t - 1|^2 + Kn^3(\epsilon^t)^3$$

By Corollary 33 we have that $\rho_i^t < 1 - c/n$. Thus, if we set $\alpha = 1 - c/n$ and $\beta = 1 - c'/n^k$ we get that

$$(1-\beta)R(\alpha, \beta, n) \leq C\frac{1}{n^k}\left(\frac{1}{n^{k-4}} + n^2\right)$$

which can be made less than $1/2$ with $k \geq 2$. Thus, all the assumptions of Lemma 34 are satisfied. By applying this result, we get that for some constant C

$$|\rho_i^{t+1} - \rho_i^* \rho_1^*| \leq C\frac{n^2}{1 - \frac{C'}{n^{k-2}}}(1 - \rho_1^{t+1}) \leq C'' n^2 (1 - \rho_1^{t+1})$$

Similarly, for any $t' \geq t+1$ such that $\rho_1^{t'} > 1 - c'/n^k$ the previous calculations apply, hence we get for all $t' \geq t+1$

$$|\rho_i^{t'} - \rho_i^* \rho_1^*| \leq C\frac{n^2}{1 - \frac{C'}{n^{k-2}}}(1 - \rho_1^{t+1}) \leq C'' n^2 (1 - \rho_1^{t'})$$

Thus, for every $t' \geq t+1$ we have

$$\epsilon^{t'} \leq C'' n^2 (1 - \rho_1^{t'})$$

Hence, by applying Lemma 37, we get that

$$\rho_1^{t'+1} = \rho_1^{t'} - C(1 - \rho_1^{t'})^2 + K' n^9 (1 - \rho_1^{t'+1})^3$$

Now, if $k \geq 5$ and c' is a small enough constant, we have

$$K' n^9 (1 - \rho_1^{t'+1})^2 < C/2$$

which implies that

$$\rho_1^{t'+1} \leq \rho_1^{t'} - \frac{C}{2}(1 - \rho_1^{t'})^2$$

■

We are now ready to present the proof of the main Lemma of this section.

Proof [Proof of Lemma 31] Initially, we have $\rho_i^0 < 1 - c'/n^k$ for all i. Let t by the first time where $\rho_i^t > 1 - c'/n$ for some i. Let T be the first time after t such that $\rho_i^t \leq 1 - c'/n^k$. Then, by Lemma 36 we have that $\rho_i^{t'+1} \leq \rho_i^{t'}$ for all $T > t' \geq t+1$. Hence, we have $\rho_i^{t'} \leq max(\rho_i^t, \rho_i^{t+1})$ for all $t' < T$. Let us now try to upper bound ρ_i^{t+1} and ρ_i^t. First of all,

we can write $y^t = Ax_i^t + r^t$, where r^t is independent of x_i^t. Let v^2 be the variance of r^t. Then, by definition

$$\rho_i^t = \frac{A}{\sqrt{A^2+v^2}} = \frac{1}{\sqrt{1+(v/A)^2}}$$

We have that $A = Cov(y^t, x_i^t) \leq \sum_i \lambda_i^t \leq n$. As for v^2, we can show Vardis: add proof here that

$$v^2 \geq \frac{1}{1+\sum \frac{(\rho_i^{t-1})^2}{1-(\rho_i^{t-1})^2}}$$

Now, we know that $1-(\rho_i^{t-1})^2 > 1-c'/n^k$ for all i by definition. Hence,

$$v^2 \geq \frac{C}{n^{2k+1}}$$

Hence,

$$\rho_i^t \leq \frac{1}{\sqrt{1+C/n^{2k+3}}} \leq 1 - \frac{C'}{n^{2k+3}}$$

Using this bound, in a completely similar fashion we get that

$$\rho_i^{t+1} \leq 1 - \frac{C''}{n^{4k+9}}$$

Hence, we have that $\rho_i^{t'} \leq 1-C/n^{4k+9}$ for all $t' < T$. Now, let's examine what happens after time T. We proved in Lemma 37 that $|\rho_j^T - \rho_i^*\rho_j^*| \leq Cn^2(1-\rho_i^{T-1}) \leq C/n^{k-2}$ for all $j \neq i$. Hence, for $k \geq 2$ and small enough constant $q \in (0,1)$ we have that $\rho_j^T \leq \rho_i^*\rho_j^* + q \leq 1-c'/n^k$ for $j \neq i$. We already know that $\rho_i^T \leq 1-c'/n^k$ by definition of T. Hence, at the T-th iteration all correlations will remain bounded away from 0. We can then consider the first time after T such that some correlation becomes bigger than $1-c'/n^k$ and the same bounds that we already established apply again and so on. Hence, we conclude that for all i, t, $\rho_i^t \leq 1 - C''/n^{4k+9}$.

■

B.3. No correlation is close to 0

In this Section, we prove a lower bound for all the correlations at any iteration of the algorithm. We will prove the following Lemma.

Lemma 38 *Suppose $\|\rho^0 - \rho^*\|_\infty \leq M$. Then, there exists a constant $c = c(M, \rho^*)$ such that for all t and for all $i \in [n]$ $\rho_i > c/n^{k+2}$ for $k \geq 4$.*

Remark 39 *The proof is assumes that the executed dynamics are according to the population EM. The extension to the sample EM is straightforward.*

Proof We start by restating Lemma 22 from the manuscript.

Lemma 40 *If $\rho^t = 0$ then $\rho^{t+1} = 0$. Further, there exists $C > 0$ such that for all $i, j, k \in \{1, \ldots, n\}$ and all $\rho \in [0, 1/2]^n$,*

$$\left.\frac{d\rho_i^{t+1}}{d\rho_j^t}\right|_0 = \begin{cases} 1 & i = j \\ \rho_i^* \rho_j^* & i \neq j \end{cases}; \qquad -C \leq \left.\frac{d^2\rho_i^{t+1}}{d\rho_j^t d\rho_k^t}\right|_\rho \leq C.$$

Using Lemma 40 and Taylor's Theorem, we get that for all i

$$\rho_i^{t+1} \geq \rho_i^t + \sum_{j \neq i} \rho_j \rho_i^* \rho_j^* - \sum_{j,k} \frac{C}{2} \rho_j^t \rho_k^t. \tag{28}$$

We now prove that during all iterations t and for all $i \neq j$, we have $\rho_i^t \geq c(n)\rho_j^t$, for some constant $c(n)$ that is independent of the iteration t. For $t = 0$, we can find some constant c independent of n such that this is satisfied (because we initially start at a constant distance from the optimum). Now, suppose $t > 1$. Suppose without loss of generality that $\rho_1^{t-1} \geq \rho_i^{t-1}$ for all $i \neq 1$. Let us fix $i \neq j$. By the update rule, we have that

$$\rho_i^t \geq \frac{\lambda_1^{t-1} \rho_i^* \rho_1^*}{\sqrt{1 + 2n^2}} \geq \frac{\rho_i^* \rho_1^*}{n\sqrt{3}} \lambda_1^{t-1}$$

On the other hand, we have that

$$\rho_j^t \leq \sum_k \lambda_k^{t-1} \leq n\lambda_1^{t-1}$$

since $\lambda_1^{t-1} \geq \lambda_k^{t-1}$ for all $k \neq 1$. Thus, we conclude that

$$\rho_i^t \geq \frac{\rho_i^* \rho_1^*}{n^2\sqrt{3}} \rho_j^t \geq \frac{K}{n^2} \rho_j^t$$

so we set $c(n) := K/n^2$, where K is some constant that depends on the lower bound on ρ^* (it is lower bounded by assumption).

Now that we have proven this claim, we can use inequality(28). Suppose w.l.o.g. that $1 = \arg\max_i \rho_i^t$. Then

$$\rho_i^{t+1} \geq \rho_i^t + \sum_{j \neq i} \rho_j \rho_i^* \rho_j^* - \sum_{j,k} \frac{C}{2} \rho_j^t \rho_k^t \geq \rho_i^t + n\frac{K}{n^2}\rho_1^t - Cn^2(\rho_1^t)^2 = \rho_i^t + \frac{K}{n}\rho_1^t - Cn^2(\rho_1^t)^2$$

Suppose $\rho_1^t < c/n^k$, where k will be determined later. Then,

$$\rho_i^{t+1} \geq \rho_i^t + \rho_1^t(K/n - Cn^2/n^k) > \rho_i^t$$

if $k - 2 \geq 2$ and c is small enough. Now suppose $\rho_1^t \geq c/n^k$. Then, we have already proved that

$$\rho_i^{t+1} \geq \frac{K}{n}\lambda_1^t$$

Now, we have

$$\lambda_1^t = \frac{\rho_1^t}{1 + \sum_{j \neq 1} \frac{(\rho_j^t)^2(1-(\rho_1^t)^2)}{1-(\rho_j^t)^2}} \geq \frac{\rho_1^t}{n+1}$$

since $1-(\rho_1^t)^2 \leq 1-(\rho_j^t)^2$ for all $j \neq 1$. It follows that

$$\rho_i^{t+1} \geq \frac{cK}{n^{k+2}}$$

Hence, if the maximum exceeds the threshold, all correlations are lower bounded, otherwise they do not decrease. We conclude that for all iterations t and all i the required bound holds. ■

Appendix C. Proof of Theorem 12: finite sample and finite iterate

C.1. The finite iterate

To describe the EM update, we have the following analogue of Lemma 1 for the sample EM:

Lemma 41 *Denote by $\hat{\mu}_x$ the uniform distribution over the m samples. Let $\mu^{t,\prime}$ denote the joint distribution over $x_1 \cdots x_n, y$ such that*

$$\Pr_{\mu^{t,\prime}}[x_1, \cdots, x_n, y] = \Pr_{\hat{\mu}_x}[x_1, \cdots, x_n] \Pr_{\mu^t}[y \mid x_1, \ldots, x_n].$$

Then, for any i, we have that

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \mathop{\mathbb{E}}_{\mu^{t,\prime}}[x_i y],\ \mathrm{Var}_{\mu^{t,\prime}}[x_i] = \mathrm{Var}_{\mu^{t+1}}[x_i],\ \mathrm{Var}_{\mu^{t,\prime}}[y] = \mathrm{Var}_{\mu^{t+1}}[y] \ .$$

The proof follows the same lines as the proof of Lemma 1, where the only difference is that here we are considering the sample-EM. As a consequence, we have the following analogue of Lemma 2:

Lemma 42 *For any i, denote*

$$\hat{\sigma}_i = \sqrt{\frac{1}{m}\sum_{k=1}^m \left(x_i^{(k)}\right)^2}$$

For any $i \neq j$, denote

$$\hat{\alpha}_{ij} = \frac{1}{\hat{\sigma}_i \hat{\sigma}_j} \frac{1}{m} \sum_{k=1}^m x_i^{(k)} x_j^{(k)} \ .$$

Denote by $\Delta_{ij}^t = \hat{\alpha}_{ij} - \rho_i \rho_j$. Then, for any $t > 0$ and any i:

$$\sigma_i^t = \hat{\sigma}_i,$$

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \sigma_i^t \sigma_y^t \left(\lambda_i^t + \sum_{j \neq i} \hat{\alpha}_{ij} \lambda_j^t\right) = \sigma_i^t \sigma_y^t \left(\rho_i^t + \sum_{j \neq i} \Delta_{ij}^t \lambda_j^t\right) \ ,$$

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[y^2] = (\sigma_y^{t+1})^2 = (\sigma_y^t)^2 \left(\sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \hat{\alpha}_{ij}\right) = (\sigma_y^t)^2 \left(1 + \sum_{j \neq k} \Delta_{ij}^t \lambda_j^t \lambda_k^t\right)$$

$$\rho_i^{t+1} = \frac{\lambda_i^t + \sum_{j \neq i} \lambda_j^t \hat{\alpha}_{ij}}{\sqrt{\sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \hat{\alpha}_{ij}}} = \frac{\rho_i^t + \sum_{j \neq i} \Delta_{ij} \lambda_j^t}{\sqrt{1 + \sum_{j \neq k} \Delta_{ij} \lambda_j^t \lambda_k^t}} \ . \tag{29}$$

Proof The proof is similar to the proof of Lemma 2. First, $\hat{\sigma}_i = \sigma_i^t$ follows directly from Lemma 1. Next, Using Lemma 41 and Claim 1, we derive that

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \mathop{\mathbb{E}}_{\mu^{t,'}}[x_i y] = \mathop{\mathbb{E}}_{\hat{\mu}_{xo}}\left[\mathop{\mathbb{E}}_{\mu^t_{y|x}}[x_i y]\right] = \mathop{\mathbb{E}}_{\hat{\mu}_x}\left[x_i \mathop{\mathbb{E}}_{\mu^t_{y|x}}[y]\right] = \mathop{\mathbb{E}}_{\hat{\mu}_x}\left[x_i \sum_{j=1}^n \frac{\sigma_y^t}{\sigma_j^t}\lambda_j^t x_j\right]$$
$$= \sum_{j=1}^n \frac{\sigma_y^t}{\sigma_j^t}\lambda_j^t \mathop{\mathbb{E}}_{\hat{\mu}_x}[x_i x_j] = \sigma_y^t \sigma_i^t \lambda_i^t + \sum_{j\neq i} \sigma_y^t \sigma_i^t \lambda_j^t \hat{\alpha}_{ij} = \sigma_y^t \sigma_i^t \left(\lambda_i^t + \sum_{j\neq i} \lambda_j^t \hat{\alpha}_{ij}\right) .$$

This concludes the first expression for $\mathbb{E}_{\mu^{t+1}}[x_i y]$. The second equality is derived similarly to the proof of Lemma 2. Next, we compute the variance for y, again, using Lemma 41 and Claim 1:

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[y^2] = \mathop{\mathbb{E}}_{\mu^{t,'}}[y^2] = \mathop{\mathbb{E}}_{\hat{\mu}_x}\left[\mathop{\mathbb{E}}_{\mu^t_{y|x}}[y^2]\right] = \mathop{\mathbb{E}}_{\hat{\mu}_x}\left[\left(\sum_{j=1}^n \frac{\sigma_y^t}{\sigma_j^t}\lambda_j^t x_j\right)^2\right]$$
$$= \sum_{i=1}^n \sum_{j=1}^n \frac{\sigma_y^t}{\sigma_i^t}\lambda_i^t \frac{\sigma_y^t}{\sigma_j^t}\lambda_j^t \mathop{\mathbb{E}}_{\hat{\mu}_x}[x_i x_j] = (\sigma_y^t)^2 \left(\sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i\neq j\in\{1,\ldots,n\}} \lambda_i^t \lambda_j^t \hat{\alpha}_{ij}\right) .$$

This derives the first expression for the variance of y. The second expression is derived similarly to its analogue in Lemma 2.

Lastly, the expression of ρ_i^{t+1} is derived, by definition of ρ_i,

$$\rho_i^{t+1} = \frac{\mathbb{E}_{\mu^{t+1}}[x_i y]}{\sqrt{\mathrm{Var}_{\mu^{t+1}}[x_i]\,\mathrm{Var}_{\mu^{t+1}}[y]}},$$

and by substituting the expressions for the covariance of x_i and y, the variance of y, and using that $\mathrm{Var}_{\mu^{t+1}}[x_i] = \hat{\sigma}_i^2$. ■

From Lemma 42, it follows that the convergence rate of the EM is not affected by the variance of the nodes. In particular, the correlation parameters ρ_i^t are independent of these. Therefore, we will assume for simplicity that $\sigma_i^* = 1$ for all i.

C.2. A deterministic assumption

Note that our theorem holds with high probability. To remove the probabilistic part, we assume a deterministic assumption on the sample, that will hold with high probability. First, we present a definition:

Definition 43 *Let $\eta > 0$. We say that a sample $x^{(1)}, \ldots, x^{(m)}$ is η-representative of μ_x^* if the following hold:*

- *For all $i = 1, \ldots, n$,*

$$1 - \eta \leq \sqrt{\frac{1}{m}\sum_{k=1}^m \left(x_i^{(k)}\right)^2} \leq 1 + \eta.$$

(recall that we assumed that the variance of each coordinate of μ_x^* is 1).

- *For all $i \neq j \in \{1, \dots, n\}$:*

$$\rho_i^* \rho_j^* - \eta \leq \frac{1}{m} \sum_{k=1}^{m} x_i^{(k)} x_j^{(k)} \leq \rho_i^* \rho_j^* + \eta \ .$$

- *For all $i \neq j$,*

$$\rho_i^* \rho_j^* - \eta \leq \hat{\alpha}_{ij} = \frac{\frac{1}{m}\sum_{k=1}^{m} x_i^{(k)} x_j^{(k)}}{\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(x_i^{(k)}\right)^2}\sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(x_j^{(k)}\right)^2}} \leq \rho_i^* \rho_j^* + \eta \ .$$

We will assume that the sample is η-representative, where η is sufficiently small. We note that from Chernoff-Hoeffding bound, the sample is η-representative with probability $1 - \delta$, if $m \geq \Omega(\log(n/\delta)/\eta^2)$.

C.3. Iterates are bounded away from 0 and 1

We start by arguing that in all iterates, the correlations ρ_i^t are always bounded away from 0 and 1, assuming finite sample: (proof is in Section B)

Lemma 44 *Let $\alpha, \beta > 0$ and assume that $\rho_i^0, \rho_i^* \geq \alpha$ for all i while $\rho_i^0, \rho_i^* \leq 1 - \beta$ for all β. Then, there exist constants, $C(\alpha, \beta)$ $C_1(\alpha, \beta)$ and $C_2(\alpha, \beta)$, that depend only on α and β and a universal constant C' such that for all $t \geq 0$ and all i,*

$$\frac{C_1(\alpha, \beta)}{n^{12}} \leq \rho_i^t \leq 1 - \frac{C_2(\alpha, \beta)}{n^{29}} \ ,$$

assuming that $\eta \leq C(\alpha, \beta)/n^{C'}$.

Next, we would use the fact that the iterates are always bounded away from 0 and 1 for the remainder of the proof. For that purpose, we have the following definition:

Definition 45 *We say that the iterates of the EM are (A, B)-bounded if $A \leq \rho_i^* \leq 1 - B$ and $A \leq \rho_i^t \leq 1 - B$ for all i and $t \geq 0$.*

C.4. Sample-EM is close to the population EM

Next, we will argue that one iterate of the sample EM close to one iterate of the population EM, if the sample is η-representative.

Lemma 46 *Assume that the sample is η-representative, for some $\eta > 0$. Fix values of $\rho_1^t, \dots, \rho_n^t$ and let $\rho_1^{t+1}, \dots, \rho_n^{t+1}$ denote the value of the next iterate according to the finite-sample update of Lemma 42. Similarly, denote by $\tilde{\rho}_1^{t+1}, \dots, \tilde{\rho}_n^{t+1}$ the result of applying the* population EM *update, as described in Lemma 2. Assume that the iterates of the EM are (A, B)-bounded, that $A \leq 1/n$ for some $A > 0$ and assume that $\eta \leq A^2/2$. Then,*

$$\left|\rho_i^{t+1} - \tilde{\rho}_i^{t+1}\right| \leq \eta \left(\frac{4\sqrt{8}n^3}{A^3} + \frac{\sqrt{8}n}{A}\right)$$

Proof Fix $i \in \{1, \ldots, n\}$. Let us denote the numerator and the denominator, for both the expressions for ρ_i^{t+1} and $\tilde{\rho}_i^{t+1}$, as follows:

$$X = \lambda_i^t + \sum_{j \neq i} \lambda_j^t \hat{\alpha}_{ij}, \quad X' = \lambda_i^t + \sum_{j \neq i} \lambda_j^t \rho_i^* \rho_j^*$$

and

$$Y = \sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \hat{\alpha}_{ij}, \quad Y' = \sum_{i=1}^n (\lambda_i^t)^2 + \sum_{i \neq j \in \{1,\ldots,n\}} \lambda_i^t \lambda_j^t \rho_i^* \rho_j^* \ .$$

Then, the quantity that we wish to bound is:

$$\begin{aligned} \left| \frac{X}{\sqrt{Y}} - \frac{X'}{\sqrt{Y'}} \right| &= \left| \frac{X}{\sqrt{Y}} - \frac{X}{\sqrt{Y'}} + \frac{X}{\sqrt{Y'}} - \frac{X'}{\sqrt{Y'}} \right| \leq |X| \left| \frac{1}{\sqrt{Y}} - \frac{1}{\sqrt{Y'}} \right| + \frac{1}{\sqrt{Y'}} \left| X - X' \right| \\ &= \frac{X \left| \sqrt{Y} - \sqrt{Y'} \right|}{\sqrt{YY'}} + \frac{1}{\sqrt{Y'}} \left| X - X' \right| = \frac{X \left| Y - Y' \right|}{\sqrt{YY'} \left| \sqrt{Y} + \sqrt{Y'} \right|} + \frac{1}{\sqrt{Y'}} \left| X - X' \right| \ . \end{aligned} \tag{30}$$

Our goal is to upper bound X, lower bound Y and Y', and upper bound $|X - X'|$ and $|Y - Y'|$, as computed below. First, notice that

$$|X - X'| = \left| \left(\lambda_i^t + \sum_{j \neq i} \lambda_j^t \hat{\alpha}_{ij} \right) - \left(\lambda_i^t + \sum_{j \neq i} \lambda_j^t \rho_i^* \rho_j^* \right) \right| \leq \sum_{j \neq i} \lambda_j^t |\hat{\alpha}_{ij} - \rho_i^* \rho_j^*| \leq \eta \sum_{j=1}^n \lambda_j^t, \tag{31}$$

where the last inequality follows by assumption of this lemma that $|\hat{\alpha}_{ij} - \rho_i^* \rho_j^*| \leq \eta$. Recall the expression for λ_i^t in (1), and notice that we can bound it as follows:

$$\lambda_i^t = \frac{\rho_i^t / (1 - (\rho_i^t)^2)}{1 + \sum_{j=1}^n (\rho_j^t)^2 / (1 - (\rho_j^t)^2)} \leq \frac{\rho_i^t / (1 - (\rho_i^t)^2)}{1 + (\rho_i^t)^2 / (1 - (\rho_i^t)^2)} = \rho_i^t \leq 1.$$

Then, $\sum_i \lambda_i \leq n$, which implies that

$$|X - X'| \leq \eta n.$$

Next:

$$|Y - Y'| = \left| \sum_{j \neq i} \lambda_i^t \lambda_j^t \hat{\alpha}_{ij} - \sum_{j \neq i} \lambda_i^t \lambda_j^t \rho_i^* \rho_j^* \right| \leq \sum_{j \neq i} \lambda_i^t \lambda_j^t |\hat{\alpha}_{ij} - \rho_i^* \rho_j^*| \leq \eta n^2,$$

using the fact that $\lambda_i^t \leq 1$ and the assumption of this lemma that $|\hat{\alpha}_{ij} - \rho_i^* \rho_j^*| \leq \eta$. Further,

$$X = \lambda_i^t + \sum_{j \neq i} \lambda_j^t \hat{\alpha}_{ij} \leq \sum_{j=1}^n \lambda_j^t \leq n.$$

Further,

$$Y' \geq \sum_{i=1}^n \sum_{j=1}^n \lambda_i^t \lambda_j^t \rho_i^* \rho_j^* \geq \sum_{i=1}^n \sum_{j=1}^n A^2 = \left(\sum_{i=1}^n \lambda_i^t \right)^2 A^2,$$

using the assumption that $\rho_i^* \geq A$ for all i. Similarly, we derive that

$$Y \geq \sum_{i=1}^n \sum_{j=1}^n \lambda_i^t \lambda_j^t \hat{\alpha}_{ij} \geq \sum_{i=1}^n \sum_{j=1}^n \lambda_i^t \lambda_j^t A^2/2 = \left(\sum_{i=1}^n \lambda_i^t\right)^2 A^2/2,$$

using the assumptions that $\rho_i^* \geq A$ for all i, and that $|\hat{\alpha}_{ij} - \rho_i^* \rho_j^*| \leq A^2/2$. It remains to lower bound the value of $\sum_i \lambda_i^t$:

$$\sum_i \lambda_i^t = \frac{\sum_{i=1}^n \rho_i^t/(1-(\rho_i^t)^2)}{1+\sum_{i=1}^n (\rho_i^t)^2/(1-(\rho_i^t)^2)} \geq \frac{\sum_{i=1}^n \rho_i^t/(1-(\rho_i^t)^2)}{1+\sum_{i=1}^n \rho_i^t/(1-(\rho_i^t)^2)} \geq \frac{\sum_{i=1}^n \rho_i^t}{1+\sum_{i=1}^n \rho_i^t} \geq \frac{nA}{1+nA} \ ,$$

where, for the last two inequalities, we used the fact that if $a \geq a' > 0$ and $b \geq 0$ then $a/(a+b) \geq a'/(a'+b)$. We will further use the assumption that $A \geq 1/n$ to derive that the right hand side is lower bounded by $1/2$. We derive that

$$Y, Y' \geq A^2/8.$$

Substituting our estimates from (30), we derive that, for some universal constant $C > 0$,

$$\left|\frac{X}{\sqrt{Y}} - \frac{X'}{\sqrt{Y'}}\right| \leq \frac{n \cdot \eta n^2}{A^2/8 \cdot 2(A/\sqrt{8})} + \frac{1}{A/\sqrt{8}}\eta n = \eta\left(\frac{4\sqrt{8}n^3}{A^3} + \frac{\sqrt{8}n}{A}\right) \ .$$

■

C.5. KL is comparable to the parameter ℓ_2 distance

In our argument, we will prove that in each iteration, the KL divergence between the iterate and the true distribution shrinks by a constant factor. In order to argue about that, we would like to claim that the KL divergence between two models is comparable to the ℓ_2 distance between their correlation parameters ρ_i, provided that those are bounded away from 0 and 1:

Lemma 47 *Assume that the iterates of the EM are (A, B)-bounded. Then, for some universal constant $C > 0$ and for all $t \geq 0$,*

$$(n/AB)^{-C}\|\rho^t - \rho^*\|^2 \leq \mathrm{KL}(\mu_x^*\|\mu_x^t) \leq (n/(AB))^C\|\rho^t - \rho^*\|^2.$$

In the sections below we will prove statements that are slightly more general (which will also be used further in the proof).

C.5.1. Upper bounding the KL

We start by proving the following lemma:

Lemma 48 *Assume that $\rho_i^*, \rho_i^t \leq 1 - B$ for all i and some $B > 0$. Then,*

$$\left| \frac{d\,\mathrm{KL}(\mu_x^* \| \mu_x^t)}{d\rho_i^t} \right| \leq \mathrm{poly}(n, 1/B) \cdot \|\rho^t - \rho^*\|_2.$$

Consequently, for ρ and ρ',

$$\left| \mathrm{KL}(\mu_x^* \| \mu_x^\rho) - \mathrm{KL}(\mu_x^* \| \mu_x^{\rho'}) \right| \leq \mathrm{poly}(n, 1/B) \max(\|\rho' - \rho^*\|, \|\rho - \rho^*\|) \|\rho' - \rho\|.$$

Lastly,

$$\mathrm{KL}(\mu_x^* \| \mu_x^t) \leq \mathrm{poly}(n, 1/B) \cdot \|\rho^t - \rho^*\|_2^2.$$

Proof We start by proving the first inequality. We use a few folklore equations: first, we define for any real-valued function f of a matrix A by $df(A)/dA$ the matrix whose ij-entry is the derivative of $f(A)$ as a function of A_{ij}. The first folklore equation is:

$$\frac{d \log\det(A)}{dA} = (A^{-1})^\top. \tag{32}$$

For the second equation, assume that A is a function of some parameter λ. Then,

$$\frac{dA^{-1}}{d\lambda} = -A^{-1} \frac{dA}{d\lambda} A^{-1},$$

where $dA/d\lambda$ is a matrix whose ij entry is the derivative of A_{ij} as a function λ. Further:

$$\frac{df(A)}{d\lambda} = \mathrm{trace}\left(\left(\frac{df(A)}{dA} \right)^\top \frac{dA}{d\lambda} \right). \tag{33}$$

Lastly, if $A \in \mathbb{R}^{n \times n}$ is symmetric and $\|A\|$ is the operator norm of A then

$$\mathrm{trace}(A) \leq n \|A\|. \tag{34}$$

Indeed, this is true because $\mathrm{trace}(A)$ is the sum of its singular values of A while $\|A\|$ is the largest singular value in absolute value.

Let us continue by analyzing the KL divergence. We note that the KL divergence between two mean-zero Gaussian vectors and covariances Σ_1 and Σ_2, is

$$\mathrm{KL}(\Sigma_1 \| \Sigma_2) = \frac{1}{2} \left(-\log \frac{\det((\Sigma_2)^{-1})}{\det((\Sigma_1)^{-1})} - n + \mathrm{trace}((\Sigma_2)^{-1} \Sigma_1) \right).$$

Applying this on the covariances Σ^* and Σ^t that correspond to ρ^* and ρ^t, we get that

$$\mathrm{KL}(\mu_x^* \| \mu_x^t) = \frac{1}{2} \left(-\log \frac{\det((\Sigma^t)^{-1})}{\det((\Sigma^*)^{-1})} - n + \mathrm{trace}((\Sigma^t)^{-1} \Sigma^*) \right).$$

Let us first compute the first derivative as a function of ρ_i^t. By (33), we have that

$$\frac{d\,\mathrm{KL}(\mu_x^* \| \mu_x^t)}{d\rho_i^t} = \mathrm{trace}\left(\left(\frac{d\,\mathrm{KL}(\mu_x^* \| \mu_x^t)}{d(\Sigma^t)^{-1}} \right)^\top \frac{d(\Sigma^t)^{-1}}{d\rho_i^t} \right). \tag{35}$$

Let us expand on the first term. By (32),

$$\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^t)}{d(\Sigma^t)^{-1}} = \frac{1}{2}\frac{d}{d(\Sigma^t)^{-1}}\left(-\log\det((\Sigma^t)^{-1}) + \mathrm{trace}((\Sigma^t)^{-1}\Sigma^*)\right) = \frac{1}{2}\left(-(\Sigma^t)^\top + \Sigma^*\right) = \frac{1}{2}\left(\Sigma^* - \Sigma^t\right) \ .$$

Note that the derivative at $\rho^* = \rho^t$ equals 0, which implies that

$$\left.\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^t)}{d\rho_i^t}\right|_{\rho^t=\rho^*} = 0. \tag{36}$$

By (35), for a general ρ^t, we have that

$$\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^t)}{d\rho_i^t} = \frac{1}{2}\,\mathrm{trace}\left((\Sigma^* - \Sigma^t)^\top \frac{d(\Sigma^t)^{-1}}{d\rho_i^t}\right) = \frac{1}{2}\,\mathrm{trace}\left((\Sigma^* - \Sigma^t)\frac{d(\Sigma^t)^{-1}}{d\rho_i^t}\right) \ .$$

Let us differentiate this again. We have that

$$\frac{d^2\,\mathrm{KL}(\mu_x^*\|\mu_x^t)}{d\rho_i^t d\rho_j^t} = \frac{d}{d\rho_j^t}\frac{1}{2}\,\mathrm{trace}\left((\Sigma^* - \Sigma^t)\frac{d(\Sigma^t)^{-1}}{d\rho_i^t}\right) = \frac{1}{2}\,\mathrm{trace}\left(-\frac{d\Sigma^t}{d\rho_j^t}\frac{d(\Sigma^t)^{-1}}{d\rho_i^t} + (\Sigma^* - \Sigma^t)\frac{d^2(\Sigma^t)^{-1}}{d\rho_i^t d\rho_j^t}\right) \ .$$

In order to bound the left hand side above, we use (34), to obtain that

$$\begin{aligned}
&\mathrm{trace}\left(-\frac{d\Sigma^t}{d\rho_j^t}\frac{d(\Sigma^t)^{-1}}{d\rho_i^t} + (\Sigma^* - \Sigma^t)\frac{d^2(\Sigma^t)^{-1}}{d\rho_i^t d\rho_j^t}\right) \le n\left\|-\frac{d\Sigma^t}{d\rho_j^t}\frac{d(\Sigma^t)^{-1}}{d\rho_i^t} + (\Sigma^* - \Sigma^t)\frac{d^2(\Sigma^t)^{-1}}{d\rho_i^t d\rho_j^t}\right\| \\
&\le \left\|\frac{d\Sigma^t}{d\rho_j^t}\right\|\left\|\frac{d(\Sigma^t)^{-1}}{d\rho_i^t}\right\| + \left\|(\Sigma^* - \Sigma^t)\right\|\left\|\frac{d^2(\Sigma^t)^{-1}}{d\rho_i^t d\rho_j^t}\right\|
\end{aligned}$$

We note that each of the components above is bounded by $\mathrm{poly}(n, 1/B)$, using the formulas for Σ^t and $(\Sigma^t)^{-1}$ that appear in (19) and (20). To conclude the first part of this lemma, we use the fact that the derivative equals 0 at $\rho^t = \rho^*$ as shown in (36), and we integrate the second derivative along the path $\rho(\tau) = (1-\tau)\rho^* + \tau\rho^t$. Define μ_x^τ the distribution obtained with correlations $\rho(\tau)$, then

$$\begin{aligned}
\left|\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^t)}{d\rho_i^t}\right| &= \left|\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^t)}{d\rho_i^t} - \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^*)}{d\rho_i^t}\right| = \left|\int_0^1 \frac{d}{d\tau}\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^\tau)}{d\rho_i^t}d\tau\right| \\
&= \left|\int_0^1 \sum_{j=1}^n \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^\tau)}{d\rho_i^t d\rho_j^t}\frac{d\rho(\tau)_j}{d\tau}d\tau\right| = \left|\int_0^1 \sum_{j=1}^n \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^\tau)}{d\rho_i^t d\rho_j^t}(\rho^t - \rho^*)d\tau\right| \\
&\le \int_0^1 \sum_{j=1}^n \left|\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^\tau)}{d\rho_i^t d\rho_j^t}\right|\left|\rho^t - \rho^*\right|d\tau \le \int_0^1 \sum_{j=1}^n \mathrm{poly}(n, 1/B)\left|\rho^t - \rho^*\right|d\tau \\
&\le \mathrm{poly}(n, 1/B)\|\rho^t - \rho^*\|_1 \le \sqrt{n}\,\mathrm{poly}(n, 1/B)\|\rho^t - \rho^*\|_2.
\end{aligned}$$

This concludes the first part of the lemma. The second part of this lemma is bounded similarly, by integrating, now over the first derivative: define by $\rho(\tau) = (1-\tau)\rho + \tau\rho'$.

Then,

$$\left|\mathrm{KL}(\mu_x^*\|\mu_x^\rho) - \mathrm{KL}(\mu_x^*\|\mu_x^{\rho'})\right| = \left|\int_0^1 \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^{\rho(\tau)})}{d\tau}d\tau\right| = \left|\int_0^1 \sum_{i=1}^n \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^{\rho(\tau)})}{d\rho(\tau)_i}\frac{d\rho(\tau)_i}{d\tau}d\tau\right|$$
$$= \left|\int_0^1 \sum_{i=1}^n \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^{\rho(\tau)})}{d\rho(\tau)_i}(\rho_i' - \rho_i)d\tau\right| = \left|\int_0^1 \sum_{i=1}^n \frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^{\rho(\tau)})}{d\rho(\tau)_i}(\rho_i' - \rho_i)d\tau\right|$$
$$\le \int_0^1 \sum_{i=1}^n \left|\frac{d\,\mathrm{KL}(\mu_x^*\|\mu_x^{\rho(\tau)})}{d\rho(\tau)_i}\right| |\rho_i' - \rho_i|d\tau \le \int_0^1 \sum_{i=1}^n \mathrm{poly}(n, 1/B)\|\rho(\tau) - \rho^*\|_2|\rho_i' - \rho_i|d\tau$$
$$\le \mathrm{poly}(n, 1/B)\max_\tau \|\rho(\tau) - \rho^*\|_2\|\rho_i' - \rho_i\|_1$$
$$\le \sqrt{n}\,\mathrm{poly}(n, 1/B)\max(\|\rho - \rho^*\|_2, \|\rho' - \rho^*\|)\|\rho_i' - \rho_i\|_2 \ .$$

The last part of this lemma is obtained from the second part by substituting $\rho' = \rho^*$. ■

C.5.2. KL upper bounds bounds the parameter difference

Lemma 49 *Let μ and μ' be two distribution. Assume that $\mathrm{Var}_\mu[y] = 1$ and further that $\mathrm{Var}_\mu[x_i] = \mathrm{Var}_{\mu'}[x_i] = 1$. Then,*

$$\mathrm{KL}(\mu'\|\mu) \ge \max_i \frac{(\mathbb{E}_\mu[x_i y] - \mathbb{E}_{\mu'}[x_i y])^2}{2}.$$

We would like to lower bound $\mathrm{KL}(\mu'\|\mu)$ by $(\mathbb{E}_\mu[x_i y] - \mathbb{E}_{\mu'}[x_i y])^2$. Notice that by the data-processing inequality and by the chain rule for KL divergence,

$$\mathrm{KL}(\mu'\|\mu) \ge \mathrm{KL}(\mu'_{x_i y}\|\mu_{x_i y}) = \mathrm{KL}(\mu'_{x_i}\|\mu_{x_i}) + \mathop{\mathbb{E}}_{x_i\sim\mu_{x_i}^{t+1}}[\mathrm{KL}(\mu'_{y|x_i}\|\mu_{y|x_i})] = \mathop{\mathbb{E}}_{x_i\sim\mu_{x_i}^{t+1}}[\mathrm{KL}(\mu'_{y|x_i}\|\mu_{y|x_i})]. \tag{37}$$

We would like to compute the conditional distribution of y given x_i, for both μ and μ'. We use the following formula for conditional Gaussians:

Lemma 50 *Let Z and W be jointly distributed Gaussian variables. Then,*

$$\mathbb{E}[W \mid Z] = \mathbb{E}\,W + \mathbb{E}[ZW]\,\mathrm{Var}[Z]^{-1}Z,$$

and

$$\mathrm{Var}[W \mid Z] = \mathrm{Var}[W] - \mathbb{E}[WZ]^2\,\mathrm{Var}[Z]^{-1}.$$

Using the formula of Lemma 50,

$$\mathbb{E}[y \mid x_i] = \mathbb{E}[y] + \mathbb{E}[x_i y]\,\mathrm{Var}[x_i]^{-1}x_i,$$

while

$$\mathrm{Var}[y \mid x_i] = \mathrm{Var}[y] - \mathbb{E}[x_i y]^2/\,\mathrm{Var}[x_i].$$

Computing some values for μ and μ', we get that

$$\mathop{\mathbb{E}}_{\mu}[y \mid x_i] = \mathop{\mathbb{E}}_{\mu}[x_i y]x_i \ ; \quad \mathop{\mathbb{E}}_{\mu'}[y \mid x_i] = \mathop{\mathbb{E}}_{\mu'}[x_i y]x_i; \mathrm{Var}_\mu[y \mid x_i] = 1 - \mathop{\mathbb{E}}_{\mu}[x_i y]^2 \le 1.$$

We use the formula for the KL of two univariate Gaussians:

Lemma 51 *For two univariate Gaussians X_1, X_2 with means μ_1, μ_2 and covariances σ_1^2, σ_2^2, respectively,*

$$\mathrm{KL}(X_1 \| X_2) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2} + \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2} \geq \frac{(\mu_1 - \mu_2)^2}{2\sigma_2^2}$$

Proof The first equality is folklore and follows from a direct calculation. To get the inequality, notice that $\log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2}{2\sigma_2^2} - \frac{1}{2}$ is the KL of two Gaussians with variances σ_1 and σ_2 and the same means, hence it is nonnegative. ■

we have that

$$\mathrm{KL}(\mu'_{y|x_i} \| \mu_{y|x_i}) \geq \frac{(\mathbb{E}_{\mu'}[y \mid x_i] - \mathbb{E}_{\mu^t}[y \mid x_i])^2}{2 \operatorname{Var}_{\mu^t}[y \mid x_i]} \geq \frac{x_i^2 (\mathbb{E}_{\mu'}[x_i y] - \mathbb{E}_{\mu}[x_i y])^2}{2} \quad .$$

Substituting this in the right hand side of (37) and using the fact that $\mathbb{E}_{\mu'}[x_i^2] = 1$, we get the desired result.

C.5.3. Another upper bound on the KL

Lemma 52 *Let μ_x and μ'_x be two leaf distributions with parameters ρ and ρ', respectively, which are (A, B)-bounded, namely, for all i, $A \leq \rho_i, \rho'_i \leq 1 - B$. Assume further that $\operatorname{Var}_{\mu}[x_i] = \operatorname{Var}_{\mu'}[x_i] = 1$ for all i. Then,*

$$\mathrm{KL}(\mu'_x \| \mu_x) \geq c A^4 \max_i |\rho_i - \rho'_i|,$$

where $c > 0$ is a universal constant.

Below, we prove Lemma 52. We note that there exist i, j such that $|\rho_i \rho_j - \rho'_i \rho'_j|$ is large. Indeed, let i be the maximizer of $|\rho_i - \rho'_i|$. For the purpose of lower bounding $|\rho_i \rho_j - \rho'_i \rho'_j|$, we can assume that $\rho_i > \rho'_i$. Denote, let $\lambda = \rho_i / \rho'_i$. Denote $M = \max_j |\rho_j - \rho'_j|$. Notice that

$$\lambda - 1 = (\rho_i - \rho'_i)/\rho'_i \geq \rho_i - \rho'_i = \max_l |\rho_l - \rho'_l| = M.$$

Then, divide into cases:

- If there exist two different values, j, k, such that $\rho_j / \rho'_j, \rho_k / \rho'_k \leq \sqrt{1/\lambda}$. Then,

  $$\frac{\rho'_j \rho'_k}{\rho_j \rho_k} \geq \lambda,$$

  hence

  $$\rho'_j \rho'_k - \rho_j \rho_k = \frac{\rho'_j \rho'_k - \rho_j \rho_k}{\rho_j \rho_k} \rho_j \rho_k \geq (\lambda - 1) \rho_j \rho_k \geq M \rho_j \rho_k \geq M A^2.$$

- Otherwise, there is some j such that $\rho_j / \rho'_j \geq \sqrt{1/\lambda}$. Then,

  $$\frac{\rho_i \rho_j}{\rho'_i \rho'_j} \geq \sqrt{\lambda} \geq \sqrt{M + 1} \geq 1 + cM,$$

  for some universal constant $c > 0$. This implies that

  $$\rho_i \rho_j - \rho'_i \rho'_j = \frac{\rho_i \rho_j - \rho'_i \rho'_j}{\rho_i \rho_j} \rho_i \rho_j \geq cM \rho_i \rho_j \geq cM A^2.$$

This derives that there exist some j, k such that $|\rho_j \rho_k - \rho'_j \rho'_k| \geq cA^2 M$ for some universal constant $c > 0$. This implies that

$$|\mathop{\mathbb{E}}_{\mu}[x_j x_k] - \mathop{\mathbb{E}}_{\mu'}[x_j x_k]| \geq cA^2 M.$$

From this point onwards, the proof is analogous to the proof of Lemma 49. We start by arguing that by the data processing inequality,

$$\mathrm{KL}(\mu'_x \| \mu_x) \geq \mathrm{KL}(\mu'_{x_j x_k} \| \mu_{x_j x_k}).$$

Then, we lower bound $\mathrm{KL}(\mu'_{x_j x_k} \| \mu_{x_j x_k})$ using the fact that $|\mathbb{E}_{\mu}[x_j x_k] - \mathbb{E}_{\mu'}[x_j x_k]|$, exactly the same way as in Lemma 49 we lower bounded $\mathrm{KL}(\mu'_{x_i y} \| \mu_{x_i y})$ using the fact that $|\mathbb{E}_{\mu}[x_i y] - \mathbb{E}_{\mu'}[x_i y]|$ is large. The proof follows.

C.5.4. Proof of Lemma 47

The proof follows directly from Lemma 48 and Lemma 52 that were proven in the previous subsections.

C.6. Contraction in KL - population EM

In this section, we prove that the KL between the true and the current model contracts by a constant factor in each iteration, for the populatoin EM. Later, we will show how to derive the same results for the sample-EM as well.

Proposition 53 *Let ρ^t denote iterate t of the population-EM, and assume that the iterates are (A, B)-bounded. Then,*

$$\mathrm{KL}(\mu^*_x \| \mu^{t+1}_x) \leq (1 - \kappa)\, \mathrm{KL}(\mu^*_x \| \mu^t_x),$$

where $\kappa = A^{C_1} B^{C_2} / n^{C_3}$ for some universal constants $C_1, C_2, C_3 > 0$.

We describe the proof of Proposition 53, in the subsections below.

C.6.1. First bound on the KL difference

We will prove the following lemma:

Lemma 54

$$\mathrm{KL}(\mu^*_x \| \mu^t_x) - \mathrm{KL}(\mu^*_x \| \mu^{t+1}_x) \geq \mathrm{KL}(\mu^{t+1} \| \mu^t)$$

First of all, we use the following well-known behavior of the EM:

Lemma 55 (Folklore.)

$$\mathrm{KL}(\mu^*_x \| \mu^t_x) - \mathrm{KL}(\mu^*_x \| \mu^{t+1}_x) \geq \mathop{\mathbb{E}}_{x,y \sim \mu^{t,*}}[\log \Pr_{\mu^{t+1}}[(x, y)]] - \mathop{\mathbb{E}}_{x,y \sim \mu^{t,*}}[\log \Pr_{\mu^t}[(x, y)]] \ . \tag{38}$$

Yuval: Prove if we have time

Proof [Proof of Lemma 54] We conclude Lemma 54 by analyzing the right hand side of (38). Notice that by Lemma 1,

$$\mathop{\mathbb{E}}_{x,y\sim\mu^{t,*}}[\log \Pr_{\mu^{t+1}}[(x,y)]] = \mathop{\mathbb{E}}_{x,y\sim\mu^{t,*}}\left[\log \Pr_{\mu^{t+1}}[y] + \sum_i \log \Pr_{\mu^{t+1}}[x_i \mid y]\right]$$

$$= \mathop{\mathbb{E}}_{x,y\sim\mu^{t+1}}\left[\log \Pr_{\mu^{t+1}}[y] + \sum_i \log \Pr_{\mu^{t+1}}[x_i \mid y]\right] = \mathop{\mathbb{E}}_{x,y\sim\mu^{t+1}}[\log \Pr_{\mu^{t+1}}[(x,y)]] \ .$$

Similarly,

$$\mathop{\mathbb{E}}_{x,y\sim\mu^{t,*}}[\log \Pr_{\mu^{t}}[(x,y)]] = \mathop{\mathbb{E}}_{x,y\sim\mu^{t+1}}[\log \Pr_{\mu^{t}}[(x,y)]] \ .$$

Hence, the right hand side of (38) equals

$$\mathop{\mathbb{E}}_{x,y\sim\mu^{t+1}}[\log \Pr_{\mu^{t+1}}[(x,y)] - \log \Pr_{\mu^{t}}[(x,y)]] = \mathrm{KL}(\mu^{t+1}\|\mu^t),$$

as required. ■

C.6.2. Bounding the KL by the covariance difference

We apply Lemma 49 with $\mu = \mu^t$ amd $\mu' = \mu^{t+1}$. Note that this lemma requires that $\mathrm{Var}_{\mu^t}[y] = 1$. Indeed, for the purpose of proving Proposition 53, we can assume that, since this proposition only argues about the x-marginal, and since the correlation parameters ρ^t and ρ^{t+1} are not affected by $\mathrm{Var}_{\mu^t}[y]$. Then, Lemma 49, in combination with Lemma 54, imply that

$$\mathrm{KL}(\mu_x^*\|\mu_x^t) - \mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) \geq \mathrm{KL}(\mu^{t+1}\|\mu^t) \geq \max_i \frac{\left(\mathbb{E}_{\mu^t}[x_i y] - \mathbb{E}_{\mu^{t+1}}[x_i y]\right)^2}{2} \ . \tag{39}$$

C.6.3. Minimal eigenvalue bound

We would like to lower bound the right hand side of (39). To do so, we start with an auxiliary lemma, bounding the minimal eigenvalue of some matrix.

Lemma 56 *Let U be a matrix with $U_{ij} = u_j$ if $i \neq j$ and $\sum_{k\neq i} u_k$ otherwise, where $u_i > 0$ for all i. Then,*

$$\sigma_{\min}(U) := \max_{a:\ \|a\|_2=1} \|Ua\|_2 \geq \frac{(\min_i u_i)^3}{\|u\|_2\|u\|_1} \cdot \frac{(n-2)^3}{128n^3} \ .$$

Our goal is to show that the minimal singular value of this matrix is bounded away from 0. In particular, denote $Ua = K$ for some unit vector $a = (a_1, \ldots, a_n)$ and $K = (K_1, \ldots, K_n)$ and our goal is to lower bound $\|K\|$. The proof has multiple ingredients. First, we denote by $s = \sum_i u_i$ and $t = \sum_i a_i u_i$. We have that

$$K_i = a_i \sum_{j\neq i} u_j + \sum_{j\neq i} a_j u_j = t + a_i(s - 2u_i),$$

hence

$$t + a_i(s - 2u_i) - K_i = 0, \tag{40}$$

In particular, we have for all $i \neq j$

$$a_i(s - 2u_i) - K_i = a_j(s - 2u_j) - K_j.$$

Let us assume that $u_1 \geq u_2 \geq \cdots \geq u_n \geq$.

If $|s - 2u_1|$ is small.

Lemma 57 *Assume that $|s - 2u_1| \leq (n-2)u_n/(8n)$. Then, $\max_i |K_i| \geq (n-2)u_n/(16n)$.*

Proof Denote $\epsilon = |s - 2u_1|$ and let us lower bound $\|K\|$. Then, we have for all i,

$$a_i(s - 2u_i) - K_i = a_1\epsilon - K_1.$$

Hence, for all $i > 1$,

$$a_i = \frac{a_1\epsilon - K_1 + K_i}{s - 2u_i}.$$

Denote $K = \max_i K_i$. Then,

$$|a_i| = \left|\frac{a_1\epsilon - K_1 + K_i}{s - 2u_i}\right| \leq \frac{|\epsilon| + 2K}{s - 2u_i}.$$

Notice that $s - 2u_i \geq s - u_1 - u_2 \geq (n-2)u_n$. Then,

$$|a_i| \leq \frac{|\epsilon| + 2K}{(n-2)u_n}.$$

Denote the right hand side by m. Then,

$$|t| = |\sum_i a_i u_i| \geq |a_1 u_1| - \sum_{i>1} |a_i| u_i \geq |a_1| u_1 - m \sum_{i>1} u_i.$$

Further, recall that $\sum_i a_i^2 = 1$ hence $a_1^2 = 1 - \sum_{i>1} a_i^2 \geq 1 - nm$ hence $|a_i| \geq \sqrt{1-nm} \geq 1/2$ assuming that $m < 1/(2n)$. Hence, $t \geq u_1/2 - m\sum_{i>1} u_i \geq u_1/2 - mnu_1 \geq u_1/4$ assuming that $m < 1/(4n)$. We derive that

$$|K_1| = |t + a_1(s - 2u_1)| \geq |t| - |a_1||s - 2u_1| \geq |t| - |\epsilon| \geq u_1/4 - |\epsilon|.$$

In particular, if $|\epsilon| \leq u_1/8$ then $K_1 \geq u_1/8$. Recall that we further assumed that $m \leq 1/(4n)$ which is satisfied whenever

$$|\epsilon| \leq \frac{(n-2)u_n}{8n}$$

and

$$K \leq \frac{(n-2)u_n}{16n}.$$

We derive that if

$$|\epsilon| \leq \frac{(n-2)u_n}{8n}$$

then

$$K \geq \frac{(n-2)u_n}{16n}.$$

■

t is large or K_i is large.

Lemma 58 *Assume that $|s-2u_1|\geq\epsilon$. Then, either $\sum_i|K_i|\geq\epsilon/2$ or $t\geq\epsilon/2n$.*

Proof We assume that $|s-2u_1|\geq\epsilon$. Then, $|s-2u_i|\geq\epsilon$ for all i. Further, recall from (40) that $|a_i(s-2u_i)+t|=|K_i|$, which implies that $|K_i|\geq|a_i(s-2u_i)|-|t|\geq|a_i|\epsilon-t$. Then, $\sum_i|K_i|\geq\epsilon\sum_i|a_i|-nt=\epsilon\|a\|_1-nt\geq\epsilon\|a\|_2-nt=\epsilon-nt$. In particular, either $t\geq\epsilon/2n$ or $\sum_i|K_i|\geq\epsilon/2$. ■

Assuming that $u_1<s/2$.

Lemma 59 *Assume that $t\neq0$ and that $s-2u_1\geq\epsilon$. Then, $\|K\|_2\geq\epsilon t/\|u\|_2$.*

Proof From (40), for all i,

$$a_i=\frac{K_i-t}{s-2u_i}.$$

From definition of t,

$$t=\sum_i a_iu_i=\sum_i\frac{u_iK_i-u_it}{s-2u_i}\leq\sum_i\frac{u_iK_i-u_it}{s-2u_i}\leq\sum_i\frac{|a_iK_i|}{s-2u_i}\leq\sum_i\frac{|u_iK_i|}{\epsilon}\leq\frac{\|u\|_2\|K\|_2}{\epsilon} \tag{41}$$

Hence

$$\|K\|_2\geq\epsilon t/\|u\|_2.$$

■

The case that $u_1>s/2$.

Lemma 60 *Assume that $2u_1-s\geq\epsilon$ and that $t\neq0$. Then,*

$$\|K\|_2\geq\frac{2(s-u_1-u_2)\epsilon|t|}{\|u\|_2s}\ .$$

Proof Using the first few equalities of (41) we get that

$$t=\sum_i\frac{u_iK_i-u_it}{s-2u_i}\ .$$

or equivalently,

$$1=\sum_i\frac{u_iK_i/t-u_i}{s-2u_i}=-\sum_i\frac{u_i}{s-2u_i}+R,\quad\text{where }R=\sum_i\frac{u_iK_i}{t(s-2u_i)}.$$

Then,

$$\sum_{i>1}\frac{u_i}{s-2u_i}-R=\frac{u_1}{2u_1-s}-1=\frac{s-u_1}{2u_1-s}=\sum_{i>1}\frac{u_i}{2u_1-s}.$$

hence

$$-R=\sum_i u_i\left(\frac{1}{2u_1-s}-\frac{1}{s-2u_i}\right)=2\sum_i u_i\frac{s-u_1-u_i}{(s-2u_i)(2u_1-s)}\geq2\sum_i u_i\frac{s-u_1-u_2}{s^2}=2\frac{s-u_1-u_2}{s}.$$

We derive that

$$\frac{2(s-u_1-u_2)}{s} \leq |R| \leq \frac{1}{|t|}\left|\sum_i \frac{u_i K_i}{|s-2u_i|}\right| \leq \frac{1}{|\epsilon t|}\left|\sum_i u_i K_i\right| \leq \frac{\|u\|_2\|K\|_2}{|\epsilon t|}.$$

In particular,

$$\|K\|_2 \geq \frac{2(s-u_1-u_2)\epsilon|t|}{\|u\|_2 s} \quad .$$

■

We would like to culminate the proof of the main lemma. From Lemma 57 we can assume that $|s-2u_1| \geq (n-2)u_n/8n := \epsilon$, otherwise

$$\|K\|_2 \geq \max_i K_i \geq (n-2)u_n/(16n)$$

and the proof follows. From Lemma 58 we can assume that $|t| \geq \epsilon/2n$, otherwise

$$\|K\|_2 \geq \|K\|_1/\sqrt{n} \geq \epsilon/(2\sqrt{n})$$

and the proof follows. Lastly, from Lemma 59 and Lemma 60 we can derive that

$$\|K\|_2 \geq \frac{\epsilon|t|}{\|u\|_2}\min\left(1, \frac{2(s-u_1-u_2)}{\|u\|_2 s}\right) \geq \frac{\epsilon|t|(n-2)u_n}{\|u\|_2 s} \geq \frac{\epsilon^2(n-2)u_n}{2n\|u\|_2 s} = \frac{(n-2)^3 u_n^3}{128n^3\|u\|_2 s} \quad .$$

C.6.4. Bounding the difference in covariance

We are ready to bound the right hand side of (39). This is stated in the following lemma:

Lemma 61 *Assume that the iterates of the population EM are (A,B) bounded, let $t \geq 0$. Then,*

$$\left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y]\right)^2 \geq c\frac{A^{12}(1-B)^8}{n^{11}}\sum_i (\rho_i^t - \rho_i^*)^2 \quad ,$$

where $c > 0$ is a universal constant.

We note that Lemma 61 together with (39) imply that

$$\mathrm{KL}(\mu_x^*\|\mu_x^t) - \mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) \geq \frac{CA^{C_1}B^{C_2}}{n^{C_3}}\sum_{i=1}^n \left(\rho_i^t - \rho_i^*\right)^2 \quad . \tag{42}$$

Below, we prove Lemma 61. In the proof below, we use a, b to denote bounds on the iterates: $a \leq \rho_i^t \leq 1-b$ for all t and i. Further, we use a^* and b^* to denote bounds on ρ_i^*: $a^* \leq \rho_i^* \leq b^*$. We recall that we assume $\sigma_{x_i}^t = \sigma_y^t = 1$ for all i. By Lemma 2, we derive that

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] = \lambda_i^t + \sum_{j\neq i}\rho_i^*\rho_j^*\lambda_j^t.$$

We can use the same lemma to derive $\mathbb{E}_{\mu^t}[x_i y]$: indeed, if we apply this lemma in a different scenario where the underlying distribution is μ^t (i.e. substituting $\mu^* = \mu^t$), then $\mu^{t+1} = \mu^t$ as well. Hence, if we substitute $\mu^* = \mu^t$ and $\mu^{t+1} = \mu^t$ in this lemma we get that

$$\mathop{\mathbb{E}}_{\mu^t}[x_i y] = \lambda_i^t + \sum_{j\neq i} \rho_i^t \rho_j^t \lambda_j^t.$$

We get that

$$(\lambda_i^t)^2 \left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y] \right)^2 = \left(\sum_{j\neq i} \rho_i^* \rho_j^* \lambda_i^t \lambda_j^t - \rho_i^t \rho_j^t \lambda_i^t \lambda_j^t \right)^2 .$$

Substituting $u_i = \lambda_i^t \rho_i^*$ and $v_i = \lambda_i^t \rho_i^t$, we derive that

$$(\lambda_i^t)^2 \left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y] \right)^2 = \left(\sum_{j\neq i} u_i u_j - v_i v_j \right)^2 .$$

We will prove the following lemma:

Lemma 62 *If $u = (u_1, \ldots, u_n)$, $v = (v_1, \ldots, v_n)$ where $u_i, v_i > 0$ for all i, then*

$$\sum_i \left(\sum_{j\neq i} u_i u_j - v_i v_j \right)^2 \geq \sigma_{\min}(A)^2 \cdot \|u - v\|_2^2$$

where $A_{ij} = (u_i + v_i)/2$ if $i \neq j$ and $A_{ii} = \sum_{j\neq i}(u_i + v_i)/2$.

We will first use Lemma 62 to complete the proof of Lemma 61, and then we'll prove Lemma 62.

Proof [Proof of Lemma 61] First,

$$\sum_i (\rho_i^t - \rho_i^*)^2 \leq \left(\min_i \lambda_i^t \right)^{-2} \sum_i (\lambda_i^t)^2 (\rho_i^t - \rho_i^*)^2 = \left(\min_i \lambda_i^t \right)^{-2} \sum_i (u_i - v_i)^2.$$

Then, by Lemma 62,

$$\begin{aligned} \sum_i (u_i - v_i)^2 \leq \frac{1}{\sigma_{\min}(A)^2} \sum_i \left(\sum_{j\neq i} u_i u_j - v_i v_j \right)^2 &= \frac{1}{\sigma_{\min}(A)^2} \sum_i (\lambda_i^t)^2 \left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y] \right)^2 \\ &\leq \frac{1}{\sigma_{\min}(A)^2} \sum_i \left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y] \right)^2, \end{aligned}$$

as $\lambda_i^t \leq 1$. We conclude that

$$\sum_i (\rho_i^t - \rho_i^*)^2 \leq \frac{1}{(\min_i \lambda_i^t)^2 \sigma_{\min}(A)^2} \sum_i \left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y] \right)^2 . \tag{43}$$

We would like to expand on this bound. Using the expression of λ_i^t from (1), we derive that

$$\lambda_i^t = \frac{\rho_i^t/(1-(\rho_i^t)^2)}{1+\sum_{i=1}^n (\rho_j^t)^2/(1-(\rho_j^t)^2)} \ .$$

From the assumption that $0 < a \le \rho_i < b \le 1$, we derive that

$$\rho_i^t/(1-(\rho_i^t)^2) \ge \rho_i^t \ge a,$$

while

$$(\rho_j^t)^2/(1-(\rho_j^t)^2) \le 1/(1-(\rho_j^t)^2) \le 1/(1-b^2).$$

Therefore,

$$\lambda_i^t \ge \frac{a}{1+n/(1-b^2)} = \frac{a(1-b^2)}{1-b^2+n} \ge \frac{a(1-b^2)}{n+1} \ . \tag{44}$$

Further, we want to bound $\sigma_{\min}(A)$. Using Lemma 56 applied on $u \leftarrow (u+v)/2$, it suffices to estimate properties of $(u+v)/2$. First of all, using the assumption $a^* \le \rho_i^* \le b^*$,

$$\frac{u_i+v_i}{2} \ge \frac{u_i}{2} = \frac{\lambda_i^t \rho_i^*}{2} \ge \frac{a(1-b^2)a^*}{2n+2} \ .$$

Further, we have that

$$\frac{\lambda_i^t(\rho_i^t+\rho_i^*)}{2} \le 1,$$

which implies that

$$\|(\overline{u}+\overline{v})/2\|_1 \le n, \ \|(\overline{u}+\overline{v})/2\|_2 \le \sqrt{n} \ .$$

Using Lemma 56, the minimal eigenvalue of A is at least

$$\frac{(\min_i(u_i+v_i)/2)^3}{\|(u+v)/2\|_2\|(u+v)/2\|_1} \cdot \frac{(n-2)^3}{128n^3} \ge \frac{a^3(1-b^2)^3(a^*)^2}{(2n+2)^3 n^{3/2}} \cdot \frac{(n-2)^3}{128n^3} \ge c\frac{a^3(1-b)^3(a^*)^2}{n^{9/2}} \ ,$$

where $c > 0$ is some universal constant. Substituting this and the lower bound on λ_i^t (Eq. (44)), in (43), we derive that

$$\left(\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y] - \mathop{\mathbb{E}}_{\mu^t}[x_i y]\right)^2 \ge c' \frac{a^8(1-b)^8(a^*)^4}{n^{11}} \sum_i (\rho_i^t - \rho_i^*)^2$$

where $c' > 0$ is a universal constant. ■

Now we will prove Lemma 62. Define for $i = 1, \ldots, n$ functions $p_i(w) = \sum_{j \ne i} w_i w_j$ where $w = (w_1, \ldots, w_n)$. Let $p(w) = (p_1(w), \ldots, p_n(w))$, then we want to show that $\|p(u) - p(v)\|^2 \ge C \cdot \|u - v\|^2$. We use the following simple observation:

Lemma 63 *Let $p(x)$ be a second-degree polynomial of one variable x and let $p'(x)$ denote its derivative. Then, for any $s, t \in \mathbb{R}$,*

$$p(s) - p(t) = p'((s+t)/2)(s-t).$$

Proof Let $p(x) = ax^2 + bx + c$. Then, $p'(x) = 2ax + b$. Then,

$$p(s) - p(t) = a(s^2 - t^2) + b(s - t)$$

while

$$p'((s+t)/2)(s-t) = (a(s+t) + b)(s-t) = as^2 - at^2 + b(s-t) = p(s) - p(t).$$

■

We use the following corollary:

Lemma 64 *Let $p\colon \mathbb{R}^n \to \mathbb{R}^n$ be a system of second-degree polynomials, namely, $p(x) = (p_1(x), \dots, p_n(x))$ where each p_i is a second degree polynomial. Then,*

$$\overline{p}(\overline{s}) - \overline{p}(\overline{t}) = J^{p}|_{(\overline{s}+\overline{t})/2}(\overline{s} - \overline{t}), \tag{45}$$

where $J^{\overline{p}}|_{\overline{x}}$ is the Jacobian matrix of $\overline{p}$ evaluated at $\overline{x}$, namely,

$$\left(J^{\overline{p}}|_{\overline{x}}\right)_{ij} = \frac{dp_i(\overline{x})}{dx_j} \ .$$

Consequently,

$$\|\overline{p}(s) - \overline{p}(t)\| \geq \sigma_{\min}(J^{\overline{p}}|_{\overline{x}})\|s - t\|, \tag{46}$$

where $\sigma_{\min}(A)$ is the minimal singular value of a matrix A.

Proof To show the first part, look at the path $\gamma(\lambda) = \lambda\overline{t} + (1-\lambda)\overline{s}$. Then, $\gamma(1) = \overline{t}$ while $\gamma(0) = \overline{s}$. Applying Lemma 63 on the polynomials $p_i(\gamma(\lambda))$ while substituting $s = 0$ and $t = 1$, one obtains (45). Then, (46) follows from taking the norm in both sides of the above equality, and using the fact that for a matrix A and a vector $\overline{v}$, $\|A\overline{v}\| \geq \sigma_{\min}(A)\|\overline{v}\|$. ■

To complete the proof of Lemma 62, notice that the matrix A in this lemma is the Jacobian of $\overline{p}$ evaluated at $(u+v)/2$.

C.6.5. Bounding the parameter difference by the KL: conclusion of Proposition 53

To conclude Proposition 53, we would like to bound the right hand side of (42). Recall that this right hand side contains the term $\|\rho^t - \rho^*\|_2^2$. By Lemma 48, this is lower bounded by

$$\frac{B^{C_1}}{n^{C_2}}\,\mathrm{KL}(\mu_x^* \| \mu_x^t).$$

In combination with (42), this concludes the proof of Proposition 53.

C.7. Contraction for the sample EM

While we have proven in Proposition 53 that the sample-EM contracts, we now prove that the population EM contracts as well:

Lemma 65 *Let μ^t denote the t'th iterate of the sample EM, and assume that the sample is η-representative. Assume that $\eta \le \min\left(1, \sqrt{\mathrm{KL}(\mu_x^*\|\mu_x^t)}\right)(AB/n)^C$ for some sufficiently large universal constant $C > 0$. Then,*

$$\mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) \le (1-\kappa')\,\mathrm{KL}(\mu_x^*\|\mu_x^t),$$

where $\kappa \ge (AB/n)^{C'}$ for some universal constant $C' > 0$.

Proof Below, we prove Lemma 65. To argue about that, we would use the Lemma 46 which argues that the sample step is close to the population step, in parameter distance, and further, we will use Lemma 48 to argue that this implies that the KL distance after one population step is close to that after one sample step. To be more concrete, let ρ^t denote the t'th iterate of the sample EM, let $\tilde{\rho}^{t+1}$ denote the result of applying one step of the population EM on ρ^t and let ρ^{t+1} denote iterate $t+1$ of the sample EM, namely, ρ^{t+1} is obtained from ρ^t via one iterate of the sample EM. Let $\tilde{\mu}_x^{t+1}$ and μ_x^{t+1} denote the corresponding distributions. Then, from Lemma 46, we obtain that for all i,

$$|\tilde{\rho}^{t+1} - \rho^{t+1}| \le \eta\,\mathrm{poly}(1/A, 1/B, n).$$

By Lemma 48 this implies that

$$\begin{aligned}
&\left|\mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) - \mathrm{KL}(\mu_x^*\|\tilde{\mu}_x^{t+1})\right| \\
&\le \mathrm{poly}(n, 1/A, 1/B)\|\rho^t - \rho^{t+1}\| \max\left(\|\rho^{t+1} - \rho^*\|, \|\tilde{\rho}^{t+1} - \rho^*\|\right) \\
&\le \mathrm{poly}(n, 1/A, 1/B)\eta \max\left(\|\rho^{t+1} - \rho^*\|, \|\tilde{\rho}^{t+1} - \rho^*\|\right) \\
&\le \mathrm{poly}(n, 1/A, 1/B)\eta \max\left(\|\rho^t - \rho^*\| + \|\rho^t - \rho^{t+1}\|, \|\tilde{\rho}^{t+1} - \rho^*\|\right) \\
&= \mathrm{poly}(n, 1/A, 1/B)\eta\left(\|\rho^t - \rho^*\| + \|\rho^t - \rho^{t+1}\|\right) \\
&\le \mathrm{poly}(n, 1/A, 1/B)\eta\left(\|\rho^t - \rho^*\| + \mathrm{poly}(n, 1/A, 1/B)\eta\right).
\end{aligned}$$

By Lemma 47, the right hand side is at least

$$\mathrm{poly}(n, 1/A, 1/B)\eta\left(\sqrt{\mathrm{KL}(\mu_x^*\|\mu_x^t)} + \eta\right)$$

By the assumption of Lemma 65 on η, this is at least

$$2\,\mathrm{poly}(n, 1/A, 1/B)\eta\sqrt{KL(\mu_x^*\|\mu_x^t)}.$$

Again, by the assumption on this lemma on η, if the constant C in this assumption is sufficiently large, then the last term is bounded by

$$KL(\mu_x^*\|\mu_x^t)(AB/n)^{C_1},$$

where C_1 can be chosen arbitrarily large (if C is). Combining with Proposition 53, this concludes the proof: indeed, let κ be the constant from the proposition. Then,

$$\mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) \le \mathrm{KL}(\mu_x^*\|\tilde{\mu}_x^{t+1}) + \left|\mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) - \mathrm{KL}(\mu_x^*\|\tilde{\mu}_x^{t+1})\right| \le (1-\kappa)\,\mathrm{KL}(\mu_x^*\|\mu_x^t) + KL(\mu_x^*\|\mu_x^t)(AB/n)^{C_1}.$$

Recall that $\kappa = (AB/n)^{C_2}$ for some universal constant C_2. Since we can select C_1 to be arbitrarily large, we can ensure that $(AB/n)^{C_1} \leq \kappa/2$. This implies that

$$\mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) \leq (1-\kappa/2)\,\mathrm{KL}(\mu_x^*\|\mu_x^t)$$

which suffices to conclude the proof. ■

C.8. Concluding the proof of Theorem 12

Proof [Proof of Theorem 12] We will assume that the sample is η-representative, for $\eta \leq (AB/n)^C\epsilon$ for a sufficiently large $C > 0$. This can be guaranteed if the sample size is $m = O(\log(1/\delta)/\eta^2) = \mathrm{poly}(n, 1/A, 1/B)\log(1/\delta)/\epsilon^2$. With this value of η, Lemma 65 guarantees that as long as $\mathrm{KL}(\mu_x^*\|\mu_x^t) \geq \epsilon(AB/n)^{C_1}$, then,

$$\mathrm{KL}(\mu_x^*\|\mu_x^{t+1}) \leq (1-\kappa')\,\mathrm{KL}(\mu_x^*\|\mu_x^t),$$

where C_1 can be chosen arbitrarily large and κ' is polynomial in $A, B, 1/n$. The initial value of $\mathrm{KL}(\mu_x^*\|\mu_x^0)$ is bounded by $\mathrm{poly}(1/A, 1/B, n)\|\rho^0 - \rho^*\|^2 \leq n\,\mathrm{poly}(1/A, 1/B, n)$, from Lemma 47. Hence, the number of iterations that it takes $\mathrm{KL}(\mu_x^*\|\mu_x^t)$ to drop below $\epsilon(AB/n)^{C_1}$ is bounded by $\mathrm{poly}(n, 1/A, 1/B)\log(1/\epsilon)$. Using Lemma 47, if C_1 is sufficiently large then once the KL divergence drops below that value, $|\rho_i^t - \rho_i^*| \leq \epsilon$ for all i.

Lastly, notice that for all $t \geq 1$ and all i, $\sigma_i^t = \hat{\sigma}_i$, from Lemma 42. Since the sample is $\eta \leq \epsilon$ representative, this implies that $|\sigma_i^t - \sigma_i^*| \leq \eta \leq \epsilon$. This concludes the proof. ■

Appendix D. General tree model

We consider a multivariate Gaussian latent-tree distribution, that is characterized by a tree $G = (V, E)$. Each vertex $u \in V$ corresponds to a random variable z_u. Suppose the total number of nodes in the tree is $n + m$. From now on, we might refer to the node itself as the random variable, when it is clear from the context what we mean. We define a joint probability distribution over the nodes as follows:

$$\Pr[z_1, \ldots, z_{n+m}] = \prod_{(i,j)\in E} \Pr[z_i, z_j] \tag{47}$$

The vertices V are divided into two groups: nodes of degree 1(leaves), denoted by $x_1, \ldots, x_n$ and nodes of degree at least 2 (internal nodes), denoted by $y_1, \ldots, y_m$. When we want to refer to some node in the tree without wanting to specify whether it is a leaf or an internal node, we will use the symbol z. The edges $e \in E$ are divided into two groups: the ones that are between y nodes, called *internal edges* and denoted by E_{yy}, and the ones between one y node and one x node, called *boundary edges* and denoted by E_{xy}. Each leaf x_i has a variance $\sigma_{x_i}^2$ and so it corresponds to a random variable $x_i \sim N(0, \sigma_{x_i}^2)$. Likewise, each internal node y_i has a variance $\sigma_{y_i}^2$ and so it corresponds to a random variable $y_i \sim N(0, \sigma_{y_i}^2)$. For each edge $(y_i, y_j) \in E_{yy}$ we define the correlation $\rho_{y_iy_j}$ of variables y_i, y_j and similarly, for each edge $(x_i, y_j) \in E_{xy}$ we define the correlation $\rho_{x_iy_j}$ between x_i, y_j. It can be shown that in

a Gaussian Graphical Model that satisfies the product decomposition 47, it can be shown Vardis: proof should go to appendix that these parameters are enough to specify the joint distribution over all x and y.

$$\rho_{z_i z_j} := \frac{\mathbb{E}[z_i z_j]}{\sigma_{z_i}\sigma_{z_j}} = \prod_{(z_u,z_v)\in P(z_i,z_j)} \rho_{z_u z_v}$$

In other words, the correlation between a pair of nodes is the product of the correlations along the edges of the path that connects them.

Define such distribution by $\mu_{G,\overline{\rho},\overline{\sigma}}$ where G is ommitted if it is clear from context. A sample $\overline{z} \sim \mu_{\overline{\rho},\overline{\sigma}}$ can be drawn as follows: first, $z_r \sim N(0,\sigma_r^2)$ is drawn for the root r (the root can be assigned as any node of the tree). For any choice of the root, there is a unique way to direct the edges going away from the root. This defines parent-child relationships between the nodes. One assigns random values from the remaining of the nodes, from top to bottom. Assuming that the value z_u on the parent u of a node v was already set, we draw z_v as follows:

$$z_v = \sigma_v \left(\rho_{uv} \frac{z_u}{\sigma_u} + \sqrt{1-\rho_{uv}^2}\epsilon_v, \right)$$

where $\epsilon_v \sim N(0,1)$ independently of the other variables and $\rho_{uv} = \rho_e$ for the edge e that connects u and v. Up to scaling the individual variables, one can assume that $\sigma_v = 1$ for all v.

An equivalent way to characterize the distribution is through the information matrix $J = \Sigma^{-1}$. Because of the factorization of the distribution, the only non-zero entries J_{ij} will be when (i,j) is an edge or if $i = j$. In general, the distribution could have an external field h. This will not happen in the joint distribution, since we assume the means to be 0.

In the following, we will also need the conditional distribution of the internal nodes y given the leaves x. Since the model is Gaussian, we know that there exists a matrix $\lambda \in \mathbb{R}^{m\times n}$ such that

$$\mathbb{E}[y_i|x] = \sum_j \lambda_{y_i x_j} x_j \tag{48}$$

Lastly, the external field of y in the conditional distribution is given by the relation

$$h_y = -J_{yx}x$$

All these properties will prove useful in the sequel.

In an estimation setting, we observe independent samples from the latent distribution over the leaves of the tree. In particular, each sample contains the information of $(x_l)_{l=1}^n$, obtained using one draw from the marginal (joint) distribution over the leaves. The goal is to learn the parameters ρ_e for all the edges of the tree and the variances σ_l^2 for all leaves l. Vardis: we have assumed variance one I guess.

Remark 66 *The variances $\sigma_{y_v}^2$ on the internal nodes y_v of the tree cannot be estimated. This is due to the fact that samples from two distributions $\mu_{\overline{\rho},\overline{\sigma}}, \mu_{\overline{\rho},\overline{\sigma}'}$ that have the same correlations $\overline{\rho}$ and the same variances on the leaves, differ only by a scaling of the unobseved nodes. In particular, we can transfer a sample from one distribution to a sample from the other by just multiplying the hidden nodes by constants. This change does not affect the marginal distribution over the observed nodes.*

Remark 67 *One has to assume that each internal node has degree at least* 3*. Indeed, if* v *is an internal node with neighbors* u_1, u_2*, then, as long as* $\rho_{u_1 v}\rho_{v u_2}$ *remains constant, the distribution over the leaves is not affected. In case that such a node* v *exists, one can remove it while keeping the same distribution over the leaves. In particular,* u_1 *is connected with* v *and one sets* $\rho_{u_1 u_2}$ *as equal to the value* $\rho_{u_1 v}\rho_{v u_2}$ *in the old graph.*

Appendix E. EM and the likelihood function for general trees

We are interested in analyzing the landscape of the likelihood function over the space of unknown parameters J. In particular, we show that we can characterize the stationary points of the likelihood function for a general tree. To help us identify the stationary points, it is convenient to view them as fixpoints of the EM algorithm.

Let $\mathcal{P}_G$ be the set of all distributions of the form $\mu_{G,\rho,\sigma}$ for any possible values of ρ and σ. In each iteration $t = 0, 1, \ldots,$ the algorithm will hold some distribution μ^t, such that μ^0 is arbitrary, and μ^t is obtained using μ^{t-1}. The goal is to find some unknown distribution $\mu^* \in \mathcal{P}_{\mathcal{G}}$, given only samples from $\mu^*_{x_1 \cdots x_n}$, namely, $y_1, \ldots, y_m$ is not observed. The algorithm can be described as follows: we set $\mu^0 \in \mathcal{P}$ arbitrarily. Then, at any $t > 0$, define

$$\mu^{t+1} = \max_{\mu \in \mathcal{P}_G} \mathop{\mathbb{E}}_{x_1 \cdots x_n \sim \mu^*_{x_1 \cdots x_n}} \mathop{\mathbb{E}}_{y \sim \mu^t_{y|x_1 \cdots x_n}} [\log \Pr_{\mu}(x_1, \ldots, x_n, y)], \tag{49}$$

where $\Pr_\mu$ denotes the density with respect to μ. Denote by $\sigma^t_\cdot, \rho^t_i, J^t$ the parameters corresponding to μ^t and by λ^t the coefficients from (48). Similarly, $\sigma^*_\cdot, \rho^*_i$ and λ^* correspond to μ^*. We would like to understand what are the fixpoints of this iteration rule and how they relate to the stationary points of the likelihood function. For this purpose, we first describe more explicitly the update rule in each iteration. The proof follows along the same lines as Lemma 1.

Lemma 68 *Let* $\mu^{t,*}$ *denote the joint distribution over* $x_1 \cdots x_n, y_1, \ldots, y_m$ *such that*

$$\mu^{t,*}(x_1, \cdots, x_n, y_1, \ldots, y_m) = \mu^*(x_1, \cdots, x_n)\mu^t_{y_1, \ldots, y_m | x_1, \cdots, x_n}(y).$$

Then, for any x_i, y_j, y_k*, we have that*

$$\mathop{\mathbb{E}}_{\mu^{t+1}}[x_i y_j] = \mathop{\mathbb{E}}_{\mu^{t,*}}[x_i y_j], \ \mathop{\mathbb{E}}_{\mu^{t+1}}[y_j y_k] = \mathop{\mathbb{E}}_{\mu^{t,*}}[y_j y_k], \ \mathrm{Var}_{\mu^{t,*}}[x_i] = \mathrm{Var}_{\mu^{t+1}}[x_i], \ \mathrm{Var}_{\mu^{t,*}}[y_j] = \mathrm{Var}_{\mu^{t+1}}[y_j] \ .$$

We notice that the variance of the leaves remains the same at each iteration. This means that the determining quantity for the distribution in each iteration are the correlations ρ^t. In particular, during the execution of the algorithm, the variances of the internal nodes might be different than 1, however the correlations always dictate the next iteration. Hence, for any fixpoint $\tilde{\mu}$ of the procedure that is given by some parameters $\tilde{\sigma}, \tilde{J}$ can be converted into a fixpoint with the same likelihood value but with all internal nodes y having variance 1. Therefore, in the sequel when we analyze the fixpoint of this rule, we assume w.l.o.g. that all variances are equal to 1. By scaling the variances of the internal nodes, we can obtain all equivalent fixpoints.

To see how the fixpoint of this rule relates to the stationary points of the log-likelihood, let's first choose a parametrization in terms of the inverse covariance matrix J. This exponential family parametrization will enable us to compute the stationary points easily by setting the derivatives to 0.

We first define the function $L : \mathbb{R}_+^{(n+m)\times(n+m)} \mapsto \mathbb{R}$ as

$$L(J) := \mathop{\mathbb{E}}_{x\sim\mu^*} \log \Pr_{\mu(J)}(x)$$

We have the following Lemma, which connects the stationary points of L with the fixpoints of EM. The proof is standard and is omitted. Vardis: will add if we have time

Lemma 69 *Let $\mu^* \in P_G$ be such that $\rho^*_{ij} \in (0,1)$ for all $(i,j) \in E$. Then, for any $\tilde{J} \in \mathbb{R}_+^{(n+m)\times(n+m)}$ we have that $\nabla L(\tilde{J}) = 0$ if and only if $\tilde{J}$ is a stationary point of the update rule (49).*

Hence, the two notions of stationarity are equivalent and we can focus on understanding when the update rule (49) has a fixpoint.

Appendix F. Uniqueness of stationary points of EM for general trees

We would like to prove that the only stationary point of the log-likelihood if $\tilde{\rho}_{ij} \in (0,1)$ is when $\tilde{\rho} = \rho^*$. This is the content of the following Theorem.

Theorem 15 *Let $G = (V, E)$ be a tree and $\mu^* \in P_G$ be a distribution with $\rho^*_{ij} \in (0,1)$ for all $(i,j) \in E$. Suppose $\tilde{\rho}$ is a stationary point of the EM update rule (49) with $\tilde{\rho}_{ij} \in (0,1)$ for all $(i,j) \in E$. Then $\tilde{\rho}_{ij} = \rho^*_{ij}$ for all $(i,j) \in E$.*

Proof Let $\tilde{\mu}$ denote the distribution induced by the fixpoint $\tilde{\rho}$. Let $\tilde{J}$ be the information matrix and $\tilde{\Sigma}$ the covariance matrix corresponding to $\tilde{\rho}$. Also, let $\mu^{*,f}$ denote the distribution with density

$$\mu^{*,f}(x,y) = \mu^*(x)\tilde{\mu}(y|x)$$

Using Lemma 68, we get that in the fixpoint we should have for all $z_i, z_j \in V$

$$\mathrm{Var}_{\mu^{*,f}}(z_i) = \mathrm{Var}_{\tilde{\mu}}(z_i) \quad , \quad \mathrm{Cov}_{\mu^{*,f}}(z_i, z_j) = \mathrm{Cov}_{\tilde{\mu}}(z_i, z_j) \tag{50}$$

where z_i, z_j are either leaf or non-leaf nodes that are connected in the topology of G.

We will show that the only possible solution to this system of equations is $\tilde{\rho} = \rho^*$. First, let us analyze these equations for two non-leaf nodes y_1, y_2 that are connected in G. Since all the variables are zero mean, we have by the definition of $\mu^{*,f}$ that

$$\mathrm{Cov}_{\mu^{*,f}}(y_1,y_2) = \mathop{\mathbb{E}}_{\mu^{*,f}}[y_1y_2] = \mathop{\mathbb{E}}_{x\sim\mu^*}[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1y_2|x]] = \mathop{\mathbb{E}}_{x\sim\mu^*}[\mathrm{Cov}_{\tilde{\mu}}(y_1,y_2|x) + \mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]]$$

We get a similar equation for $\tilde{\mu}$, namely that

$$\mathrm{Cov}_{\tilde{\mu}}(y_1,y_2) = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[\mathrm{Cov}_{\tilde{\mu}}(y_1,y_2|x) + \mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]]$$

At this point, we notice that the first of the two terms contains the quantity $\mathrm{Cov}_{\tilde{\mu}}(y_1, y_2|x)$, which does not depend on the value x that we condition upon (this can be seen by just applying the conditioning formula for gaussians). Hence, this term will appear in both sides of the equation and can be cancelled. So, we only need to compute the second term, which will depend on the values x. A similar argument for the variance shows that

$$\begin{aligned}\mathrm{Var}_{\mu^*,f}(y_1) &= \mathop{\mathbb{E}}_{\mu^*}\left[\mathrm{Var}_{\tilde{\mu}}(y_1|x) + \mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]^2\right] \\ \mathrm{Var}_{\tilde{\mu}}(y_1) &= \mathop{\mathbb{E}}_{\tilde{\mu}}\left[\mathrm{Var}_{\tilde{\mu}}(y_1|x) + \mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]^2\right] \end{aligned} \tag{51}$$

Again, the conditional variance does not depend on the value of x, hence it will be the same for the two distributions.

After this preliminary observations, let's see how the we can reduce this problem to the one latent case. Suppose y_1 is a latent node and denote its set of neighbors in G by N. Notice that some of the neighbors will be leaves and some will be non-leaves, which prompts us to partition N into corresponding subsets N_x, N_y. Let's denote $s = |N|, s_x = |N_x|, s_y = |N_y|$. There is a natural partition of the leaves $R_1, \dots, R_s$, which is induced by removing y_1 from the graph and taking the leaves in each connected component. We will focus on the fixed point equations that we get for the covariances of y_1 with it's neighbors. To do that, we should calculate the conditional expectations of y_1 and it's neighbors, given x. We will need to be careful when calculating them, since we would like certain quantities to appear. Therefore, we will start by marginalizing out all the non-leaf nodes except y_1 and N_y. Let Y^c denote these nodes. We first compute what is the external field of y_1 and the nodes in N_y when we do this marginalization. Indeed, if h_{y_1} was the original external field of y_1 and the new external field is h'_{y_1}, then we have that

$$h'_{y_1} = h_{y_1} - \tilde{J}_{y_1 Y^c}\tilde{J}^{-1}_{Y^c Y^c} h_{Y^c}$$

However, notice that the vector $\tilde{J}_{y_1 Y^c}$ is the 0 vector, since Y^c does not contain any neighbor of y_1. Hence, the external field of y_1 does not change. Now let y_2 be a neighbor in N_y. Then,

$$h'_{y_2} = h_{y_2} - \tilde{J}_{y_2 Y^c}\tilde{J}^{-1}_{Y^c Y^c} h_{Y^c}$$

Notice that h_{y_2} is a linear combination of the leaves that are neighbors of y_2. Also, we can show that the second term of the right hand side is a linear combination of the leaves that belong to the connected component corresponding to y_2 and are not neighbors of y_2. This is established in the following Lemma.

Lemma 70 *Let y_2 be a non-leaf neighbor of y_1 and S_{y_2} be the corresponding set of leaves for the partition that y_2 belongs to. Then, the quantity $h'_{y_2} = h_{y_2} - \tilde{J}_{y_2 Y^c}\tilde{J}^{-1}_{Y^c Y^c} h_{Y^c}$ is a linear combination of the leaves in S_{y_2}.*

Proof The leaves in S_{y_2} can be partitioned into subsets A, B, where A are the leaves that are neighbors of y_2 and B the remaining leaves. Let T_2 be the topology of the connected component that y_2 belongs to, when we remove y_1. Also, let N_2 be the neighborhood of y_2 in T_2, with corresponding subsets N_{2x}, N_{2y}. Clearly, h_{y_2} is a linear combination of the leaves

in A. We will show that the second term is a linear combination of the leaves in B, thus concluding the claim. First of all, notice that $\tilde{J}_{Y^cY^c}$ is the information matrix of a gaussian model, whose graphical representation is the tree G when y_1 and all nodes in N_y have been removed. Let's call T_3 this new topology. In this topology, T_2 has been partitioned into $|N_2|$ subtrees, one for each neighbor of y_2 (because y_2 is removed). Hence, the leaves B of T_2 have been partitioned into $|N_{2y}|$ subsets $Q_1, \ldots, Q_{|N_{2y}|}$.

Since $\tilde{J}_{Y^cY^c}$ is an information matrix, it's inverse $\tilde{J}^{-1}_{Y^cY^c}$ is a covariance matrix, where the nodes have standard deviations w_i and normalized covariances (correlations) w_{ij}. Since T_3 is a forest, there is at most one path connecting each one of the nodes. Hence, it is well known that the covariances multiply across paths in this structure, namely:

$$(\tilde{J}_{Y^cY^c})^{-1}_{ij} = \begin{cases} 0 & \text{, if } i,j \text{ are not connected in } T_3 \\ w_i w_j \prod_{e \in P_{ij}} w_{ij} & \text{, if } P_{ij} \text{ is the unique path connecting } i,j \end{cases}$$

Given this description, it is easy to see that for each $i \in Y^c$, we have that $(\tilde{J}^{-1}_{Y^cY^c} h_{Y^c})_i$ is a linear combination of leaves that belong to the same component as i in T_3. Hence, for each $i \in N_{2y}$, $(\tilde{J}^{-1}_{Y^cY^c} h_{Y^c})_i$ is a linear combination of the leaves of the Q_i that i is connected to. Hence, $\tilde{J}_{y_2Y^c}\tilde{J}^{-1}_{Y^cY^c} h_{Y^c}$ is simply a linear combination of the leaves in all the Q_i's, which means it is a linear combination of the leaves in B. ■

Hence, overall the external field h'_{y_2} will be a linear combination of the leaves in S_{y_2}. The same is true for all nodes in N_y. It will be convenient to define the vector $H \in \mathbb{R}^s$, which has one entry for each node in N. If the node is a $y_i \in N_y$, then define $H_i = h'_{y_i}$. If the node is an $x_i \in N_x$, then define $H_i = x_i$. Let's focus on some $y_2 \in N_y$ and see what relations we get in the fixed point. As we say in the earlier computation, the relation for the covariance becomes

$$\mathop{\mathbb{E}}_{\tilde{\mu}}[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]] = \mathop{\mathbb{E}}_{\mu^*}[\mathop{\mathbb{E}}_{\tilde{\mu}}[y_1|x]\mathop{\mathbb{E}}_{\tilde{\mu}}[y_2|x]] \tag{52}$$

The inner expectation is common on both sides, so let's start by calculating that. Since we have already marginalized out all the nodes in Y^c, we only need to marginalize out the nodes in N_y other than y_2. For convenience, denote $N_2 = N_y \setminus \{y_2\}$. Then, if we marginalize out nodes in N_2, the resulting extenal field h''_{y_1} will be

$$h''_{y_1} = h_{y_1} - \tilde{J}_{y_1N_2}(\tilde{J}'_{N_2N_2})^{-1}h'_{N_2}$$

The reason why we write J' is that the information matrix has been altered when marginalizing Y^c. Now, notice that since the neighbors of y_1 are not connected with each other, the matrix $\tilde{J}'_{N_2N_2}$ is diagonal. This means that we have

$$h''_{y_1} = \sum_{x_i \in N_x} \tilde{J}_{y_1x_i}x_i + \sum_{y_j \in N_y, y_j \neq y_2} \frac{\tilde{J}_{y_1y_j}}{\tilde{J}'_{y_jy_j}} h'_{y_j}$$

To write this more compactly, we introduce the vector $r \in \mathbb{R}^s$, where $r_i = \tilde{J}_{y_1x_i}$ if $i \in N_x$ and $r_i = \tilde{J}_{y_1y_i}/\tilde{J}'_{y_iy_i}$ if $i \in N_y$. Notice that $r_i \neq 0$ always. With this notation, the previous equation becomes

$$h''_{y_1} = \sum_{i \neq y_2} r_i H_i \tag{53}$$

We also need to compute h''_{y_2}, which is a much easier task, since

$$h''_{y_2} = h'_{y_2} - \tilde{J}_{y_2N_2}(\tilde{J}'_{N_2N_2})^{-1}h'_{N_2} = h'_{y_2}$$

since $J_{y_2N_2}$ is the zero vector (no connections between neighbors). Hence, we have calculated the external field of y_1, y_2 in the marginal model that contains only these two nodes. Now, to calculate the conditional expectations of these two nodes, we just have to use the relation

$$\begin{pmatrix}\mathbb{E}_{\tilde{\mu}}[y_1|x]\\ \mathbb{E}_{\tilde{\mu}}[y_2|x]\end{pmatrix} = \tilde{\Sigma}_{y_1y_2|x}\begin{pmatrix}h''_{y_1}\\ h''_{y_2}\end{pmatrix}$$

that connects the external field to the mean of a gaussian. Here, $\tilde{\Sigma}_{y_1y_2|x}$ is the 2×2 covariance matrix of the conditional distribution of y_1, y_2 given x. The reason we have used $\tilde{\Sigma}_{y_1y_2|x}$ is that the covariance matrix does not change when we marginalize some nodes. Suppose

$$\tilde{\Sigma}_{y_1y_2|x} = \begin{pmatrix}c_1 & c_2\\ c_3 & c_4\end{pmatrix}$$

with $c_1c_4 - c_2c_3 \neq 0$. The reason the variances are not necessarily 1 is that we are now in the conditional model. Then, condition(52) translates to the following:

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[(c_1h''_{y_1} + c_2h''_{y_2})(c_3h''_{y_1} + c_4h_{y_2})] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(c_1h''_{y_1} + c_2h''_{y_2})(c_3h''_{y_1} + c_4h_{y_2})]$$

This implies that

$$\begin{aligned}c_1c_3(\mathop{\mathbb{E}}_{x\sim\mu^*}[(h''_{y_1})^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(h''_{y_1})^2]) + c_2c_4(\mathop{\mathbb{E}}_{x\sim\mu^*}[(h''_{y_2})^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(h''_{y_2})^2])\\ +(c_1c_4 + c_2c_3)(\mathop{\mathbb{E}}_{x\sim\mu^*}[h''_{y_1}h''_{y_2}] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h''_{y_1}h''_{y_2}]) = 0\end{aligned} \tag{54}$$

Similarly, from the variance condition on y_1 (Equations 51) we obtain

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[(c_1h''_{y_1} + c_2h''_{y_2})^2] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(c_1h''_{y_1} + c_2h''_{y_2})^2]$$

This implies

$$\begin{aligned}c_1^2(\mathop{\mathbb{E}}_{x\sim\mu^*}[(h''_{y_1})^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(h''_{y_1})^2]) + c^2(\mathop{\mathbb{E}}_{x\sim\mu^*}[(h''_{y_2})^2]\\ - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[(h''_{y_2})^2]) + 2c_1c_2(\mathop{\mathbb{E}}_{x\sim\mu^*}[h''_{y_1}h''_{y_2}] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h''_{y_1}h''_{y_2}]) = 0\end{aligned} \tag{55}$$

Similarly, for y_2 we get

$$\begin{aligned}c_3^2\left(\mathop{\mathbb{E}}_{x\sim\mu^*}[h_1^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h_1^2]\right) + c_4^2\left(\mathop{\mathbb{E}}_{x\sim\mu^*}[h_2^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h_2^2]\right)\\ +2c_3c_4\left(\mathop{\mathbb{E}}_{x\sim\mu^*}[h_1h_2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h_1h_2]\right) = 0\end{aligned} \tag{56}$$

We can think of equations (54), (55) and (56) as a 3×3 system with matrix

$$\begin{pmatrix} c_1c_3 & c_2c_4 & c_1c_4 + c_2c_3 \\ c_1^2 & c^2 & 2c_1c_2 \\ c_3^2 & c_4^2 & 2c_3c_4 \end{pmatrix}$$

where the three unknown variables are

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[h_1^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h_1^2],\ \mathop{\mathbb{E}}_{x\sim\mu^*}[h_2^2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h_2^2],\ \mathop{\mathbb{E}}_{x\sim\mu^*}[h_1h_2] - \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h_1h_2]$$

The determinant of this matrix is $-(c_1c_4 - c_2c_3)^3$, which is non-zero, since the matrix $\Sigma_{y_1y_2|x}$ is invertible. The reason is that we have assumed that in the fixpoint $\tilde{\rho}$ all correlations are strictly less than 1. Hence, we conclude that

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[h''_{y_1}h''_{y_2}] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[h''_{y_1}h''_{y_2}]$$

Based on the calculations that were done earlier (equation (53)), this equation can be written as

$$\mathop{\mathbb{E}}_{x\sim\mu^*}\left[H_{y_2}\left(\sum_{i\neq y_2} r_iH_i\right)\right] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}\left[H_{y_2}\left(\sum_{i\neq y_2} r_iH_i\right)\right]$$

If x_j is a non-leaf neighbor, the fixed point equations give us

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[x_j(\sum_{i\neq j} r_iH_i)] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[x_j(\sum_{i\neq j} r_iH_i)] \implies$$
$$\mathop{\mathbb{E}}_{x\sim\mu^*}[H_j(\sum_{i\neq j} r_iH_i)] = \mathop{\mathbb{E}}_{x\sim\tilde{\mu}}[H_j(\sum_{i\neq j} r_iH_i)]$$

The key observation here is that the coefficients r_i that appear will be the same in all equations involving y_1. The last step of the argument involves actually computing the expectation and seeing what it implies for $\mu^{*,f}$. First, let's try to compute $\mathbb{E}[H_iH_j]$ for $y_i, y_j \in N_y$. Remember that we have established already that for each $i \in N$, H_i is a linear combination of the leaves in the partition corresponding to i. Hence, computing $\mathbb{E}[H_iH_j]$ amounts to computing the covariance between all pairs of leaves from the two different subsets. Suppose $H_i = (a^i)^\top X_i$, where X_i is the vector of leaves in S_i. As we said, correlations multiply across paths, so in particular we have that

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[H_iH_j] = \mathop{\mathbb{E}}_{x\sim\mu^*}[(a^i)^\top X_i(a^j)^\top X_j] = \sum_{x_k\in S_i, x_l\in S_j} a_k^i a_l^j \mathop{\mathbb{E}}_{x\sim\mu^*}[x_k^i x_l^j]$$
$$= \sum_{x_k\in S_i, x_l\in S_j} a_k^i a_l^j \prod_{e\in P_{x_k,x_l}} \theta_e^*$$

where $P(x_k, x_l)$ denotes the path between leaves x_k, x_l. Now, notice that all the paths from X_i to X_j will have to go through the edges connecting y_1 to i and y_1 to j. Let ρ_i^*, ρ_j^* be

the correlations in these two edges. Then, there is a convenient factorization that can be written as follows

$$\mathop{\mathbb{E}}_{x\sim\mu^*}[H_iH_j] = \rho_i^*\rho_j^*\left(\sum_{x_k\in S_i} a_k^i \prod_{e\in P_{x_k,y_i}} \rho_e^*\right)\left(\sum_{x_l\in S_j} a_l^j \prod_{e\in P_{x_l,y_j}} \rho_e^*\right)$$

The exact same relations hold for $\tilde{\rho}$, since the topology is the same. Hence, by writing out the condition

$$\begin{aligned}&\rho_i^*\left(\sum_{x_r\in S_i} a_r^i \prod_{e\in P_{x_r y_i}} \rho_e^*\right)\sum_{z_j\in N, j\neq i}\rho_j^*\left(\sum_{x_s\in S_j} a_s^j \prod_{e\in P_{x_s z_j}} \rho_e^*\right)\\ &= \tilde{\rho}_i\left(\sum_{x_r\in S_i} a_r^i \prod_{e\in P_{x_r y_i}} \tilde{\rho}_e\right)\sum_{z_j\in N, j\neq i}\tilde{\rho}_j\left(\sum_{x_s\in S_j} a_s^j \prod_{e\in P_{x_s z_j}} \tilde{\rho}_e\right)\end{aligned} \tag{57}$$

The reason we used the notation z_j is that this neighbor could be either a leaf or an internal node. Hence, for each $i \in N_y$ we set

$$w_i^* = \rho_i^*\left(\sum_{r\in X^i} a_r^i \prod_{e\in P_{ri}} \rho_e^*\right)$$

and similarly

$$\tilde{w}_i = \tilde{\rho}_i\left(\sum_{r\in X^i} a_r^i \prod_{e\in P_{ri}} \tilde{\rho}_e\right).$$

Notice that if i or j is a leaf neighbor, then we get the same expression, except that the parenthesis will be 1 and we will simply have ρ_i^* or ρ_j^* as the variable. Hence, this definition can be extended to all $i \in N$. Now, given these parametrizations, the fixed point conditions are

$$r_j w_j^*\left(\sum_{i\neq j} r_i w_i^*\right) = r_j\tilde{w}_j\left(\sum_{i\neq j} r_i\tilde{w}_i\right) \tag{58}$$

for each $i \in N$. But this is exactly the system that we got for one latent.. As we proved in Lemma 9, the only solution for this system is $r_i\tilde{w}_i = r_i w_i^*$ for each i, which implies that $\tilde{w}_i = w_i^*$ for each i. Notice that if N_x is nonempty, then this implies that $\rho_e^* = \tilde{\rho}_e$ for all edges e between y_1 and it's neighboring leaves.

We can use this result to show that $\tilde{\rho} = \rho^*$. Our argument is inductive. At each step, we pick an internal node y that only has one non-leaf neighbor (it's parent). Then, for all edges of the form $e = (y, x_i)$ for some x_i, we have that $\rho_e^* = \tilde{\rho}_e$ by the previous argument. Once we establish that, we remove all the leaf nodes that are connected to y from the tree along with their corresponding edges. This means that in the next iteration, y will be a leaf, so it will no longer be selected. This means that this process terminates after m steps.

The correctness of this procedure can be proven inductively as follows: for the base case, we already saw that edges that are connected to leaves will be equal in the two models. In

each step, for the node y that is selected at that step, we know from the induction hypothesis that all of it's leaves in the remaining tree are either true leaves, or internal nodes who have already been selected. This means that are descendants of y have already been proven to be equal in the two models. Then, for each leaf neighbor z of the node y (it might not be a true leaf in the original tree) we have parameters $w^*, \tilde{w}$. These are proven to be equal by the previous arguments. Let e be the edge connecting z, y. Then w^* is a multiple of ρ_e^* and $\tilde{w}$ is a multiple of $\tilde{\rho}_e$. The multiplier for both of these is the same in both quantities, since it only depends on descendant edges, which are proven to be equal for the two models. It follows that $\rho_e^* = \tilde{\rho}_e$ and the induction stpe is complete.

■