A Description of the EM Algorithm

We provide a short description of the *Expectation-Maximization (EM) algorithm* for maximizing likelihood in statistical models with latent variables. Consider a probability distribution $p_{\boldsymbol{\lambda}}$ sampling $(\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{X}$ is a vector of observable random variables, $\boldsymbol{Z}$ a vector of non-observable random variables and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ a vector of parameters. Given independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of the observed random variables, the goal of maximum likelihood estimation is to select $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ maximizing the log-likelihood of the samples, namely $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Unfortunately, computing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$ involves summing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ over all possible values of $\boldsymbol{z}_i$, which commonly results in a log-likelihood function that is non-convex with respect to $\boldsymbol{\lambda}$ and therefore hard to optimize.

In this context, the EM algorithm proposes the following heuristic:

- Start with an initial guess $\boldsymbol{\lambda}^{(0)}$ of the parameters.
- For all $t \geq 0$, until convergence:
  - (E-Step) For each sample i, compute the posterior $Q_i^{(t)}(\boldsymbol{z}) := p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{Z} = \boldsymbol{z} | \boldsymbol{X} = \boldsymbol{x}_i)$.
  - (M-Step) Set $\boldsymbol{\lambda}^{(\boldsymbol{t+1})} := \arg\max_{\boldsymbol{\lambda}} \sum_i \sum_{\boldsymbol{z}} Q_i^{(t)}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z})}{Q_i^{(t)}(\boldsymbol{z})}$.

Intuitively, the E-step of the algorithm uses the current guess of the parameters, $\boldsymbol{\lambda}^{(t)}$, to form beliefs, $Q_i^{(t)}$, about the state of the (non-observable) $\boldsymbol{Z}$ variables for each sample i. Then the M-step uses the new beliefs about the state of $\boldsymbol{Z}$ for each sample to maximize with respect to $\boldsymbol{\lambda}$ a lower bound on $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Indeed, by the concavity of the log function, the objective function used in the M-step of the algorithm is a lower bound on the true log-likelihood for all values of $\boldsymbol{\lambda}$, and it equals the true log-likelihood for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. From these observations, it follows that the above alternating procedure improves the true log-likelihood until convergence.

A.1 Derivation of the EM iteration for Naive Bayes - Proof of Lemma 1

Proof of Lemma 1:

The non-observable random variables for Naive Bayes mixture is the mixture component that the sample is drawn from, i.e. $Z \in \{1, 2\}$ while the parameters are the (normalized) probabilities for each feature. Thus the posterior $Q_i^{(t)}(Z = 1) = \frac{p_{\boldsymbol{\lambda}^{(t)}}(Z=1) \cdot p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{X}=\boldsymbol{x}_i | Z=1)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{X}=\boldsymbol{x}_i)}$ can be computed as $\frac{\frac{1}{2}\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)/k}{\frac{1}{2}\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)/k + \frac{1}{2}\prod_{i=1}^n (1-\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)/k} = \frac{\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}{\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) + \prod_{i=1}^n (1-\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}$.

Computing the M-Step, we get that the updated probabilities are:

$$\begin{aligned}
\frac{1 + \boldsymbol{\lambda}^{(t+1)}}{k} &= \frac{\underset{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}}{\mathbb{E}} \left[\frac{\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}{\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) + \prod_{i=1}^n (1-\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)} \cdot \boldsymbol{x} \right]}{\underset{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}}{\mathbb{E}} \left[\frac{\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}{\prod_{i=1}^n (1+\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) + \prod_{i=1}^n (1-\boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)} \right]} \\
&= 2 \underset{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}}{\mathbb{E}} \left[\frac{\prod_{i=1}^n (1 + \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}{\prod_{i=1}^n (1 + \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) + \prod_{i=1}^n (1 - \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)} \cdot \boldsymbol{x} \right] \qquad \text{since } E[Q_i^{(t)}(Z = 1)] = E[Q_i^{(t)}(Z = 2)] = \frac{1}{2} \\
&= \underset{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}}{\mathbb{E}} \left[\frac{\prod_{i=1}^n (1 + \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) - \prod_{i=1}^n (1 - \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}{\prod_{i=1}^n (1 + \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) + \prod_{i=1}^n (1 - \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)} \cdot \boldsymbol{x} \right] \qquad \text{since } E[Q_i^{(t)}(Z = 1) \cdot \boldsymbol{x}] = -E[Q_i^{(t)}(Z = 2) \cdot \boldsymbol{x}]
\end{aligned}$$

Therefore, the iteration becomes

$$\boldsymbol{\lambda}^{(t+1)} = k \underset{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}}{\mathbb{E}} \left[\frac{\prod_{i=1}^n (1 + \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) - \prod_{i=1}^n (1 - \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)}{\prod_{i=1}^n (1 + \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i) + \prod_{i=1}^n (1 - \boldsymbol{\lambda}_i^{(t)} \cdot \boldsymbol{x}_i)} \cdot \left(\boldsymbol{x} - \boldsymbol{1}\frac{1}{k} \right) \right]. \tag{A.6}$$

Equation (A.6), can be further simplified to give the required iteration form by noting that that $\boldsymbol{x}$ is a 0/1 vector, and that

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad \text{and} \quad \tanh^{-1}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right).$$

■

A.2 Derivation of the EM iteration for Binary Features - Proof of Lemma 2

Proof of Lemma 2: The iteration at the t-th step of the algorithm from A.6 simplifies to the following form with the simplifications in notation for the binary case:

$$\boldsymbol{\lambda}^{(t+1)} = 2 \mathop{\mathbb{E}}_{\boldsymbol{x} \sim p_{\boldsymbol{\mu}}} \left[\frac{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i)}{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) + \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)} \boldsymbol{x} \right]. \tag{A.7}$$

We expand (A.7) as follows:

$$\begin{aligned}
\boldsymbol{\lambda}^{(t+1)} &= \mathbb{E}_{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}} \left[\frac{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i)}{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) + \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)} \boldsymbol{x} \right] + \\
&\quad + \mathbb{E}_{\boldsymbol{x} \sim d_{-\boldsymbol{\mu}}} \left[\frac{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i)}{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) + \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)} \boldsymbol{x} \right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}} \left[\frac{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i)}{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) + \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)} \boldsymbol{x} \right] - \\
&\quad - \mathbb{E}_{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}} \left[\frac{\prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)}{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) + \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)} \boldsymbol{x} \right] \\
&= \mathbb{E}_{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}} \left[\frac{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) - \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)}{\prod_{j=1}^n (1 + \lambda_i^{(t)} x_i) + \prod_{j=1}^n (1 - \lambda_i^{(t)} x_i)} \boldsymbol{x} \right].
\end{aligned} \tag{A.8}$$

The proof of the lemma follows using (A.8) with the following facts:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \text{ and } \tanh^{-1}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right).$$

■

B From Non-Uniform to Uniform Marginals

As we argued earlier, we can assume without loss of generality that the marginal distribution of the mixture is uniform over K for every feature.

Suppose instead that a feature i has a different marginal distribution $\boldsymbol{\pi} = \frac{\boldsymbol{\pi}^1 + \boldsymbol{\pi}^2}{2} \neq \frac{1}{k} \cdot \mathbf{1}$.

We can—agnostically with respect to the parameters $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$—process the samples we receive from the mixture to make the feature marginal uniform over K in a way that we also know a one-to-one correspondence between the parameters of the resulting Naive Bayes model and the original Naive Bayes model.

To do this for every value $j \in K$, whenever the sampled value for that feature takes value j we randomly resample the value for that feature according to a known distribution.

To explain the resampling process, we partition the set K in two sets, a set $A = \{j : \pi_j > \frac{1}{k}\}$ and $B = K \setminus A$. We will only resample feature value whenever it takes value in the set A.

Whenever, value $j \in A$ is chosen, we keep it with probability $(k\pi_j)^{-1} < 1$ and with the remaining $1 - (k\pi_j)^{-1}$ probability, we sample a different value $j' \in B$ instead with probability proportional to $\frac{1}{k} - \pi_{j'}$, i.e. value $j \in A$

is resampled to value $j' \in B$ with probability $(1-(k\pi_j)^{-1})\frac{\frac{1}{k}-\pi_{j'}}{\sum_z \frac{1}{k}-\pi_z}$. This process fixes the probability that the feature takes value $j \in A$ to be $\frac{1}{k}$. Moreover, since the remaining probability is distributed proportionally to the missing probability mass $\frac{1}{k}-\pi_{j'}$ for each value $j' \in B$, and all probabilities for coordinates $j \in A$ have been fixed, this process fixes the probabilities for all $j' \in B$ as well.

The corresponding stochastic transformation matrix is upper-diagonal and thus invertible. It's minimum eigenvalue is simply the inverse of its diagonal which is $(k \max_{j \in A} \pi_j)^{-1} \geq \frac{1}{k}$. Thus it is invertible and well-conditioned. Concluding, we can assume that the marginal distribution of the mixture is uniform over K by applying this transformation for every feature separately. We then invert back once we have computed estimates for the transformed mixture to obtain the true mixture parameters $\boldsymbol{\pi}^1$ and $\boldsymbol{\pi}^2$.

C Convergence for Two Non-Identical Binary Features - Proof of Theorem 7

We analyze here the convergence of EM iteration (2.2) when we have two binary features, i.e. $\mathrm{n}=2$, with means (μ_1, μ_2) for the first class and $(-\mu_1, -\mu_2)$ for the second class. In this case, we show that any point (μ_1', μ_2') on the curve $\mu_1'\mu_2' = \mu_1\mu_2$ is a fixed point. This is because all points on the curve have the same likelihood and it is information theoretically impossible to distinguish among them. We prove that the EM algorithm converges to this curve $\mu_1'\mu_2' = \mu_1\mu_2$ and we compute its convergence rate. Figure 2 shows experimentally how different starting points converge to different fixed points on the curve.

We now present the proof of Theorem 7.

Proof of Theorem 7. The iteration for two features with means (μ_1, μ_2) for the first class and $(-\mu_1, -\mu_2)$ for the second class can be written according to (A.8) as follows

$$\begin{aligned}
\boldsymbol{\lambda}^{(t+1)} &= \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\frac{(1+\lambda_1^{(t)}x_1)(1+\lambda_2^{(t)}x_2)-(1-\lambda_1^{(t)}x_1)(1-\lambda_2^{(t)}x_2)}{(1+\lambda_1^{(t)}x_1)(1+\lambda_2^{(t)}x_2)+(1-\lambda_1^{(t)}x_1)(1-\lambda_2^{(t)}x_2)}\begin{bmatrix}x_1\\x_2\end{bmatrix}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\frac{\lambda_1^{(t)}x_1+\lambda_2^{(t)}x_2}{1+\lambda_1^{(t)}\lambda_2^{(t)}x_1x_2}\begin{bmatrix}x_1\\x_2\end{bmatrix}\right]
\end{aligned}$$

which implies

$$\begin{aligned}
\lambda_1^{(t+1)} &= \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\frac{\lambda_1^{(t)}x_1^2+\lambda_2^{(t)}x_1x_2}{1-\lambda_1^{(t)}\lambda_2^{(t)}x_1x_2}\right] = \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\frac{\lambda_1^{(t)}+\lambda_2^{(t)}x_1x_2}{1+\lambda_1^{(t)}\lambda_2^{(t)}x_1x_2}\right] \\
&\text{and} \\
\lambda_2^{(t+1)} &= \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\frac{\lambda_1^{(t)}x_1x_2+\lambda_2^{(t)}x_2^2}{1-\lambda_1^{(t)}\lambda_2^{(t)}x_1x_2}\right] = \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\frac{\lambda_1^{(t)}x_1x_2+\lambda_2^{(t)}}{1+\lambda_1^{(t)}\lambda_2^{(t)}x_1x_2}\right].
\end{aligned}$$

We observe that the above expectations only depend on x_1x_2 and hence we can set the random variable $y = x_1x_2$. We see that the probability $y=1$ is equal to

$$\left(\frac{1+\mu_1}{2}\right)\left(\frac{1+\mu_2}{2}\right)+\left(\frac{1-\mu_1}{2}\right)\left(\frac{1-\mu_2}{2}\right)=\frac{1+\mu_1\mu_2}{2}$$

and hence the probability $y = -1$ is equal to $(1 - \mu_1\mu_2)/2$ which implies that

$$\lambda_1^{(t+1)} = \mathbb{E}_{y \sim d_{\mu_1\mu_2}}\left[\frac{\lambda_1^{(t)} + \lambda_2^{(t)} y}{1 + \lambda_1^{(t)}\lambda_2^{(t)} y}\right] \text{ and } \lambda_2^{(t+1)} = \mathbb{E}_{y \sim d_{\mu_1\mu_2}}\left[\frac{\lambda_1^{(t)} y + \lambda_2^{(t)}}{1 + \lambda_1^{(t)}\lambda_2^{(t)} y}\right] \implies$$

$$\lambda_1^{(t+1)} + \lambda_2^{(t+1)} = \mathbb{E}_{y \sim d_{\mu_1\mu_2}}\left[\frac{\left(\lambda_1^{(t)} + \lambda_2^{(t)}\right)(y+1)}{1 + \lambda_1^{(t)}\lambda_2^{(t)} y}\right] = \frac{\left(\lambda_1^{(t)} + \lambda_2^{(t)}\right)}{1 + \lambda_1^{(t)}\lambda_2^{(t)}}(1 + \mu_1\mu_2)$$

$$\lambda_1^{(t+1)} - \lambda_2^{(t+1)} = \mathbb{E}_{y \sim d_{\mu_1\mu_2}}\left[\frac{\left(\lambda_1^{(t)} - \lambda_2^{(t)}\right)(y-1)}{1 + \lambda_1^{(t)}\lambda_2^{(t)} y}\right] = \frac{\left(\lambda_1^{(t)} - \lambda_2^{(t)}\right)}{1 - \lambda_1^{(t)}\lambda_2^{(t)}}(1 - \mu_1\mu_2).$$

From these relations we observe the following

1. if $\mu_1\mu_2 < \lambda_1^{(t)}\lambda_2^{(t)}$ then $\lambda_1^{(t)} + \lambda_2^{(t)}$ decreases and $\lambda_1^{(t)} - \lambda_2^{(t)}$ increases.
2. if $\mu_1\mu_2 > \lambda_1^{(t)}\lambda_2^{(t)}$ then $\lambda_1^{(t)} + \lambda_2^{(t)}$ increases and $\lambda_1^{(t)} - \lambda_2^{(t)}$ decreases.
3. if $\mu_1\mu_2 = \lambda_1^{(t)}\lambda_2^{(t)}$ then $\lambda_1^{(t)} + \lambda_2^{(t)}$ remains the same and $\lambda_1^{(t)} - \lambda_2^{(t)}$ remains the same and hence $\lambda_1^{(t)}$ and $\lambda_2^{(t)}$ remain the same.

From these observations and using the fact that $\lambda_1^{(t)}\lambda_2^{(t)} = \frac{1}{4}\left(\lambda_1^{(t)} + \lambda_2^{(t)}\right)^2 - \left(\lambda_1^{(t)} - \lambda_2^{(t)}\right)^2$ we conclude that

1. if $\mu_1\mu_2 < \lambda_1^{(t)}\lambda_2^{(t)}$ then $\lambda_1^{(t)}\lambda_2^{(t)}$ decreases.
2. if $\mu_1\mu_2 > \lambda_1^{(t)}\lambda_2^{(t)}$ then $\lambda_1^{(t)}\lambda_2^{(t)}$ increases.
3. if $\mu_1\mu_2 = \lambda_1^{(t)}\lambda_2^{(t)}$ then $\lambda_1^{(t)}\lambda_2^{(t)}$ remains the same.

Hence the iterations $\lambda_1^{(t)}$ and $\lambda_2^{(t)}$ converge to a point with $\mu_1\mu_2 = \lambda_1^{(t)}\lambda_2^{(t)}$ and then the do not evolve, i.e. they reach a fixed point.

Finally based on the above we can easily compute the convergence rate of $\lambda_1^{(t)}\lambda_2^{(t)}$ to $\mu_1\mu_2$ to be equal to $\sqrt{1 - \mu_1\mu_2}$. □

D Convergence for many i.i.d Features - Proof of Theorem 8

To derive the convergence rate bound we follow the *sensitivity method* developed in [DTZ17] which we present here for completeness.

The main idea is to use the Mean Value Theorem with respect to the second coordinate of the function $M = M_i$ on the interval $[\lambda, \mu]$.

$$\frac{M(\lambda, \mu) - M(\lambda, \lambda)}{\mu - \lambda} = \left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi} \quad \text{with } \xi \in (\lambda, \mu)$$

But we know that $M(\lambda, \lambda) = \lambda$ and $M(\lambda, \mu) = \lambda'$ and therefore we get

$$\lambda' - \lambda \geq \left(\min_{\xi \in [\lambda, \mu]} \left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi}\right)(\mu - \lambda)$$

which is equivalent to

$$|\lambda' - \mu| \leq \left(1 - \min_{\xi \in [\lambda, \mu]} \left.\frac{\partial M(\lambda, y)}{\partial y}\right|_{y=\xi}\right)|\lambda - \mu| \tag{D.9}$$

where we have used the fact that $\lambda' < \mu$ which is comes from the fact that $M(\lambda, \mu)$ is increasing with respect to λ and that $M(\mu, \mu) = \mu$.

Therefore it suffices to lower bound $\frac{dM_i}{d\mu}(\lambda, \xi)$. Towards this direction observe that

$$\frac{dM_i}{d\mu}(\lambda, \boldsymbol{\xi}) = \sum_{p=1}^{n} \frac{\partial \lambda'_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi})$$

Then the following lemma hold.

Lemma 4. *Let μ, λ have the same sign and let $\tau = \min(|\mu|, |\lambda|)$, then*

1. *for every $i \in [n]$ we have that*
$$\frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi}) \geq 1 - \left(1 - \tau^2\right)^{\frac{n-2}{2}}.$$
2. *for every $i, p \in [n]$ with $i \neq p$ we have that*
$$\frac{\partial M_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi}) \geq 0.$$

Using the above lemma we get that

$$\begin{aligned}\frac{dM_i}{d\mu}(\lambda, \boldsymbol{\xi}) &= \sum_{p=1}^{n} \frac{\partial M_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi}) \\ &= \frac{\partial M_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi}) + \sum_{p=1, p\neq i}^{n} \frac{\partial M_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi}) \\ &\geq 1 - \left(1 - \tau^2\right)^{\frac{n-2}{2}}.\end{aligned}$$

and hence we get the following result

Remark. Our theorem does not give any information about the case $n = 1$ and $n = 2$. For $n = 1$ it is easy to see that any value of λ is a fixed point of the EM iteration. The case $n = 2$ is captured by the analysis in Section 3. We also observe that as the number of features increases the convergence rate decreases and hence the EM iteration speeds up. This is natural since the more features we have the easier it is to cluster the samples that we get to the appropriate classes.

To finish the proof of Theorem 8 we present the proof of Lemma 4.

Proof of Lemma 4. We first prove statement 2. We start observing by (2.2) that

$$\begin{aligned}M_i(\lambda, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}_{-p}\sim d_{\boldsymbol{\mu}_{-p}}} &\left[\tanh\left(\tanh^{-1}(\lambda)\left(\sum_{j=1, p\neq j}^{n} x_j + 1\right)\right)\left(\frac{1+\mu_p}{2}\right) x_i + \right. \\ &\left. + \tanh\left(\tanh^{-1}(\lambda)\left(\sum_{j=1, p\neq j}^{n} x_j - 1\right)\right)\left(\frac{1-\mu_p}{2}\right) x_i\right] \implies \\ \frac{\partial M_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}_{-p}\sim d_{\boldsymbol{\xi}_{-p}}} &\left[\tanh\left(\tanh^{-1}(\lambda)\left(\sum_{j=1, p\neq j}^{n} x_j + 1\right)\right) x_i - \right. \\ &\left. - \tanh\left(\tanh^{-1}(\lambda)\left(\sum_{j=1, p\neq j}^{n} x_j - 1\right)\right) x_i\right]\end{aligned}$$

It is easy to see that this last quantity is positive. This comes from the fact that tanh is an increasing function and hence the sign of the expression depends on x_i only. Now since $\mu > 0$ the probability mass on the positive values is greater than the probability mass on negative values and also the absolute values on the positive x_i are greater than the absolute values on negative x_i. Hence we get that

$$\frac{\partial M_i}{\partial \mu_p}(\lambda, \boldsymbol{\xi}) \geq 0.$$

We continue with the proof of statement 1. We start observing by (2.2) that

$$\begin{aligned}
M_i(\lambda, \boldsymbol{\mu}) &= \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\mu}_{-i}}} \left[\tanh\left(\tanh^{-1}(\lambda) \left(\sum_{j=1, i \neq j}^{n} x_j + 1 \right) \right) \left(\frac{1 + \mu_i}{2} \right) - \right. \\
&\qquad \left. - \tanh\left(\tanh^{-1}(\lambda) \left(\sum_{j=1, i \neq j}^{n} x_j - 1 \right) \right) \left(\frac{1 - \mu_i}{2} \right) \right] \implies \\
\frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi}) &= \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\xi}_{-i}}} \left[\tanh\left(\tanh^{-1}(\lambda) \left(\sum_{j=1, i \neq j}^{n} x_j + 1 \right) \right) + \right. \\
&\qquad \left. + \tanh\left(\tanh^{-1}(\lambda) \left(\sum_{j=1, i \neq j}^{n} x_j - 1 \right) \right) \right] \\
&= \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\xi}_{-i}}, x_i \sim d_0} \left[\tanh\left(\tanh^{-1}(\lambda) \left(\sum_{j=1}^{n} x_j \right) \right) \right]
\end{aligned}$$

From this expression and using the same argument that we used to prove part 2. of the lemma we get that $\frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi})$ is an increasing function of λ. Hence we can assume without loss of generality that $\lambda \leq \xi_i$ because otherwise the value of $\frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi})$ is even greater.

Therefore we can assume that $\min(\lambda, \min_{i \in [n]}(\xi_i) = \lambda$. Writing the EM iteration according to (A.8) and get that

$$\frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\xi}_{-i}}, x_i \sim d_0} \left[\frac{\prod_{j=1}^{n}(1 + \lambda x_i) - \prod_{j=1}^{n}(1 - \lambda x_i)}{\prod_{j=1}^{n}(1 + \lambda x_i) + \prod_{j=1}^{n}(1 - \lambda x_i)} \right].$$

Using this form of iteration we get the following sequence of bounds using $\tau = \min(\lambda, \mu)$.

$$\begin{aligned}
1 - \frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi}) &= 1 - \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\mu}_{-i}}, x_i \sim d_0} \left[\frac{\prod_{j=1}^{n}(1 + \lambda x_i) - \prod_{j=1}^{n}(1 - \lambda x_i)}{\prod_{j=1}^{n}(1 + \lambda x_i) + \prod_{j=1}^{n}(1 - \lambda x_i)} \right] \\
&= \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\xi}_{-i}}, x_i \sim d_0} \left[\frac{2 \prod_{j=1}^{n}(1 - \lambda x_i)}{\prod_{j=1}^{n}(1 + \lambda x_i) + \prod_{j=1}^{n}(1 - \lambda x_i)} \right] \\
&= \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\xi}_{-i}}, x_i \sim d_0} \left[\frac{2}{\frac{\prod_{j=1}^{n}(1 + \lambda x_i)}{\prod_{j=1}^{n}(1 - \lambda x_i)} + 1} \right] \\
&\leq \mathbb{E}_{\boldsymbol{x}_{-i} \sim d_{\boldsymbol{\xi}_{-i}}, x_i \sim d_0} \left[\sqrt{\frac{\prod_{j=1}^{n}(1 - \lambda x_i)}{\prod_{j=1}^{n}(1 + \lambda x_i)}} \right] \\
&= \frac{1}{2} \mathbb{E}_{\boldsymbol{x}_{-i,k} \sim d_{\boldsymbol{\xi}_{-i,k}}, x_i \sim d_0} \left[\sqrt{\frac{\prod_{j=1, j \neq k}^{n}(1 - \lambda x_i)}{\prod_{j=1, j \neq k}^{n}(1 + \lambda x_i)}} \left(\sqrt{\frac{1 - \lambda}{1 + \lambda}}(1 + \xi_k) + \sqrt{\frac{1 + \lambda}{1 - \lambda}}(1 - \xi_k) \right) \right] \\
&\leq \sqrt{1 - \lambda^2} \left(\mathbb{E}_{\boldsymbol{x}_{-i,k} \sim d_{\boldsymbol{\xi}_{-i,k}}, x_i \sim d_0} \left[\sqrt{\frac{\prod_{j=1, j \neq k}^{n}(1 - \lambda x_i)}{\prod_{j=1, j \neq k}^{n}(1 + \lambda x_i)}} \right] \right)
\end{aligned}$$

where for the last inequality we used some very simple algebraic computations and the fact that as we said we can use $\lambda = \min(\lambda, \min_{i\in[n]}(\xi_i))$ and hence $\lambda \le \xi_k$. So is general for greater λ we have

$$1 - \frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi}) \le \sqrt{1-\tau^2}\left(\mathbb{E}_{\boldsymbol{x}_{-i,k}\sim d_{\boldsymbol{\xi}_{-i,k}}, x_i\sim d_0}\left[\sqrt{\frac{\prod_{j=1,j\neq k}^{n}(1-\lambda x_i)}{\prod_{j=1,j\neq k}^{n}(1+\lambda x_i)}}\right]\right)$$

now inductively we can get that

$$\begin{aligned} 1 - \frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi}) &\le \left(\sqrt{1-\tau^2}\right)^{n-1}\left(\mathbb{E}_{x_i\sim d_0}\left[\sqrt{\frac{1-\lambda x_i}{1+\lambda x_i}}\right]\right) \\ &\le \frac{\left(\sqrt{1-\tau^2}\right)^{n-1}}{\sqrt{1-\lambda^2}}. \end{aligned}$$

Finally using the fact that $\frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi})$ is increasing with respect to λ we get that

$$1 - \frac{\partial M_i}{\partial \mu_i}(\lambda, \boldsymbol{\xi}) \le \left(1-\tau^2\right)^{\frac{n-2}{2}}$$

and the lemma follows. □

E Proof of Theorem 10

Proof of theorem 10: It is easy to see that:

$$|\tanh(z) - z| \le |z|^3 \tag{E.10}$$

$$\text{and } \ |\tanh^{-1}(z) - z| \le |z|^3, \text{for} |z| \le 0.9. \tag{E.11}$$

From (E.11), for sufficiently small $\|\boldsymbol{\lambda}\|_1$, we have that $\tanh^{-1}(\lambda_{ij}) = \lambda_{ij} \pm |\lambda_{ij}|^3$, for all i, j. Since $\boldsymbol{x}$ is a binary vector it follows then that

$$\tanh^{-1}(\boldsymbol{\lambda}) \cdot \boldsymbol{x} = \boldsymbol{\lambda} \cdot \boldsymbol{x} \pm \sum_{ij} |\lambda_{ij}|^3 = \boldsymbol{\lambda} \cdot \boldsymbol{x} \pm \|\boldsymbol{\lambda}\|_3^3.$$

Using this, Eq. (E.10), that $\boldsymbol{x}$ is a binary vector, and assuming $\|\boldsymbol{\lambda}\|_1$ is sufficiently small, we get that:

$$\begin{aligned} \tanh\left(\tanh^{-1}(\boldsymbol{\lambda}) \cdot \boldsymbol{x}\right) &= \boldsymbol{\lambda} \cdot \boldsymbol{x} \pm \|\boldsymbol{\lambda}\|_3^3 \pm \left(|\boldsymbol{\lambda} \cdot \boldsymbol{x}| + \|\boldsymbol{\lambda}\|_3^3\right)^3 \\ &= \boldsymbol{\lambda} \cdot \boldsymbol{x} \pm O(\|\boldsymbol{\lambda}\|_1^3). \end{aligned}$$

Plugging into the population EM update rule of Eq. (2.1) we get that:

$$\begin{aligned} \boldsymbol{\lambda}' &= k \cdot \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[(\boldsymbol{\lambda} \cdot \boldsymbol{x}) \cdot \left(\boldsymbol{x} - \frac{1}{k} \cdot \mathbf{1}\right)\right] \pm O(k \cdot \|\boldsymbol{\lambda}\|_1^3) \\ &= k \cdot \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\left(\boldsymbol{x} - \frac{1}{k} \cdot \mathbf{1}\right) \cdot \boldsymbol{x}^{\mathrm{T}}\right] \cdot \boldsymbol{\lambda} \pm O(k \cdot \|\boldsymbol{\lambda}\|_1^3). \end{aligned}$$

■

F Proof of Lemma 3

Proof of Lemma 3: Let us denote by $\Psi = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\left(\boldsymbol{x} - \frac{1}{k} \cdot \mathbf{1}\right) \cdot \boldsymbol{x}^{\mathrm{T}}\right]$. Each row of Ψ in indexed by a pair (i, j), where $i \in \{1, \dots, n\}$ corresponds to a feature and $j \in \{1, \dots, k\}$ to a possible value for that feature. Ψ is of course symmetric and positive semi-definite as it is a covariance matrix. It is easy to see that

$$\Psi_{(i,j),(i',j')} = \begin{cases} \frac{1}{k} \cdot \left(1 - \frac{1}{k}\right), & \text{if } i = i', j = j'; \\ -\frac{1}{k^2}, & \text{if } i = i', j \neq j'; \\ \frac{\mu_{ij} \cdot \mu_{i'j'}}{k^2}, & \text{if } i \neq i'. \end{cases}$$

Hence we can write Ψ as follows: $\Psi = \frac{1}{k}I - \frac{1}{k^2}J + \frac{1}{k^2}(\boldsymbol{\mu}\cdot\boldsymbol{\mu}^{\mathrm{T}} - \mathrm{diag}(\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}}))$, where I is the identity matrix, J is the block diagonal matrix that is all-1's in the diagonal blocks (corresponding to $i = i'$) and all-0's in the off-diagonal blocks (corresponding to $i \neq i'$), and $\mathrm{diag}(\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}})$ is the block diagonal matrix, whose i-th diagonal block equals $\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}}$.

Given this structure of Ψ and our assumption on the norms of the $\boldsymbol{\mu}_i$ vectors being equal, we see that:

$$\Psi\cdot\boldsymbol{\mu} = \left(\frac{1}{k} + \frac{1}{k^2}(n-1)\left\|\boldsymbol{\mu}_1\right\|_2^2\right)\cdot\boldsymbol{\mu},$$

where the $\frac{1}{k}\cdot\boldsymbol{\mu}$ came from multiplying with $\frac{1}{k}I$, the multiplication by $\frac{1}{k^2}J$ contributed 0 as $\mathbf{1}\cdot\boldsymbol{\mu}_i = 0$, for all i, and $\frac{1}{k^2}(n-1)\left\|\boldsymbol{\mu}_1\right\|_2^2\cdot\boldsymbol{\mu}$ came from multiplying by $\frac{1}{k^2}(\boldsymbol{\mu}\cdot\boldsymbol{\mu}^{\mathrm{T}} - \mathrm{diag}(\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}}))$, using that the norms of all $\boldsymbol{\mu}_i$ are equal. So $\boldsymbol{\mu}$ is an eigenvector with eigenvalue $\left(\frac{1}{k} + \frac{1}{k^2}(n-1)\left\|\boldsymbol{\mu}_1\right\|_2^2\right) > \frac{1}{k}$.

Next consider any unit vector y that is orthogonal to μ, and let us compute $y^{\mathrm{T}}\Psi y$. If y were an eigenvector, this quadratic form would equal its eigenvalue. Because y is orthogonal to μ we have that:

$$y^{\mathrm{T}}\Psi y = \frac{1}{k} - \frac{1}{k^2}y^{\mathrm{T}}(J + \mathrm{diag}(\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}}))y$$

Note that each diagonal block of matrix $J + \mathrm{diag}(\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}})$ is a positive semidefinite matrix, and because this matrix is diagonal, it follows that $y^{\mathrm{T}}(J + \mathrm{diag}(\boldsymbol{\mu}_i\cdot\boldsymbol{\mu}_i^{\mathrm{T}}))y \geq 0$. Hence, $y^{\mathrm{T}}\Psi y \leq 1/k$.

It follows that all other eigenvalues of Ψ are $\leq 1/k$. Hence $\boldsymbol{\mu}$ is the principle eigenvector. ■

G Applications of Theorem 8 to mixtures of Gaussians

Proof of Theorem 9 using Theorem 8. We have that

$$\boldsymbol{\lambda}' = \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\tanh\left(\tanh^{-1}(\lambda)\left(\sum_{j=1}^{n} x_j\right)\right)\boldsymbol{x}\right] \implies$$

$$\lambda' = \mathbb{E}_{\boldsymbol{x}\sim d_{\boldsymbol{\mu}}}\left[\tanh\left(\tanh^{-1}(\lambda)\left(\sum_{j=1}^{n} x_j\right)\right)\frac{\sum_{i=1}^{n} x_i}{n}\right] \implies$$

$$\lambda' = \mathbb{E}_{X\sim B\left(n,\frac{1+\mu}{2}\right)}\left[\tanh\left(\tanh^{-1}(\lambda)\left(2X-n\right)\right)\frac{2X-n}{n}\right]$$

Now we use the well known result that as $n\to\infty$ the distribution of the variable X converges shifted and multiplied by the appropriate factors, to a Gaussian with mean $\bar{\mu} = \frac{\mu}{\sqrt{n}}$. Therefore substituting μ with $\bar{\mu} = \frac{\mu}{\sqrt{n}}$ and $\bar{\lambda} = \frac{\lambda}{\sqrt{n}}$ and doing the calculations we get that

$$\lambda' = \mathbb{E}_{X\sim B\left(n,\frac{1+\mu}{2}\right)}\left[\tanh\left(\frac{\tanh^{-1}(\frac{\lambda}{\sqrt{n}})}{\frac{\lambda}{\sqrt{n}}}\lambda\sqrt{1-\frac{\mu^2}{n}}\left(\frac{X-\sqrt{n}(\sqrt{n}+\mu)/2}{\sqrt{n(1-\mu)/2}}\right)\right)\sqrt{1-\frac{\mu^2}{n}}\left(\frac{X-\sqrt{n}(\sqrt{n}+\mu)/2}{\sqrt{n(1-\mu)/2}}\right)\right].$$

Now the we observe that the above sum as $n\to\infty$ converges to an integral. Also using the Central Limit theorem we have that the probability density according to which we compute the expectation converges to the normal distribution as $n\to\infty$. Putting these two together we take that the limit $n\to\infty$ becomes

$$\lambda' = \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\tanh\left(\lambda\left(x+\mu\right)\right)\left(x+\mu\right)\right].$$

Which is the EM iteration for the balanced mixture of two isotropic Gaussians. We now apply Theorem 8 and we get that the convergence rate of the above iteration is equal to

$$\left(1+\frac{\min(\lambda,\mu)}{n}\right)^{\frac{n-2}{2}}$$

which if we take the limit $n \to \infty$ becomes exactly

$$\exp\left(-\frac{\min(\lambda^{(t)}, \mu)^2}{2\sigma^2}\right)$$

and the theorem follows. □

H Convergence of Power Pretrained EM for i.i.d. Features

Proof of Theorem 11. We recall the general EM update with arbitrary initialization

$$\boldsymbol{\lambda}^{(t+1)} = \mathop{\mathbb{E}}_{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}}\left[\tanh\left(\tanh^{-1}(\boldsymbol{\lambda}^{(t)}) \cdot \boldsymbol{x}\right) \boldsymbol{x}\right].$$

We first bound the difference in the update between starting with an estimation $\boldsymbol{\lambda}_1^{(t)}$ on the line spanned by $\boldsymbol{\mu}$ and starting with an estimation $\boldsymbol{\lambda}_2^{(t)}$ that is η close to $\boldsymbol{\lambda}_1^{(t)}$ in ℓ_∞ norm. For simplicity we use $\boldsymbol{\lambda}_1$ for $\boldsymbol{\lambda}_1^{(t)}$ and $\boldsymbol{\lambda}_1'$ for $\boldsymbol{\lambda}_1^{(t+1)}$ and the same for $\boldsymbol{\lambda}_2^{(t)}$, $\boldsymbol{\lambda}_2^{(t+1)}$. Let $j \in [n]$ we have that

$$\left|\frac{\partial \lambda_i'}{\partial \lambda_j}\right| = \left|\mathop{\mathbb{E}}_{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}}\left[\tanh'\left(\tanh^{-1}(\boldsymbol{\lambda}^{(t)}) \cdot \boldsymbol{x}\right) \tanh^{-1'}(\lambda_j) x_i x_j\right]\right| \leq \frac{1}{1 - \lambda_j^2}. \tag{H.12}$$

Where the last inequality follows by the computation of the derivative of $\tanh^{-1}$ and the fact that $\tanh'(\cdot) \leq 1$. Now since we have assumed that $|\mu_j|$ is bounded away from 1 we can easily get that the same is true for $|\lambda_j|$ and hence we have that $\left|\frac{\partial \lambda_i'}{\partial \lambda_j}\right| \leq c$ for some constant c. Finally using Taylor's theorem and the fact that $\|\boldsymbol{\lambda}\|_\infty \leq \|\boldsymbol{\lambda}\|_2 \leq \sqrt{n}\,\|\boldsymbol{\lambda}\|_\infty$ we have

$$\|\boldsymbol{\lambda}_1' - \boldsymbol{\lambda}_2'\|_2 \leq c n^{3/2} \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 . \tag{H.13}$$

This implies that as far as the estimations $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ are suffieciently close we will have that the updated estimations $\boldsymbol{\lambda}_1', \boldsymbol{\lambda}_2'$ will be close too. Now assume that we start two executions of the original EM algorithm from the estimations $\boldsymbol{\lambda}_1^{(0)}$ parallel to $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}_2^{(0)}$ with $\left\|\boldsymbol{\lambda}_1^{(0)} - \boldsymbol{\lambda}_2^{(0)}\right\|_2 \leq \eta$. Using (H.13) we have that

$$\left\|\boldsymbol{\lambda}_1^{(t)} - \boldsymbol{\lambda}_2^{(t)}\right\|_2 \leq \left(c n^{3/2}\right)^t \eta.$$

Then from Theorem 8 we have that there exist a $\kappa \in (0, 1)$ such that

$$\left\|\boldsymbol{\lambda}_1^{(t+1)} - \boldsymbol{\mu}\right\|_2 \leq \kappa \left\|\boldsymbol{\lambda}_1^{(t)} - \boldsymbol{\mu}\right\|_2$$

which implies

$$\begin{aligned}
\left\|\boldsymbol{\lambda}_2^{(t+1)} - \boldsymbol{\mu}\right\|_2 &\leq \left\|\boldsymbol{\lambda}_2^{(t+1)} - \boldsymbol{\lambda}_1^{(t+1)}\right\|_2 + \left\|\boldsymbol{\lambda}_1^{(t+1)} - \boldsymbol{\mu}\right\|_2 \\
&\leq c n^{3/2} \left\|\boldsymbol{\lambda}_2^{(t)} - \boldsymbol{\lambda}_1^{(t)}\right\|_2 + \kappa \left\|\boldsymbol{\lambda}_1^{(t)} - \boldsymbol{\mu}\right\|_2 \\
&\leq \kappa \left\|\boldsymbol{\lambda}_2^{(t)} - \boldsymbol{\mu}\right\|_2 + \left(c n^{3/2} + \kappa\right) \left\|\boldsymbol{\lambda}_2^{(t)} - \boldsymbol{\lambda}_1^{(t)}\right\|_2 \\
&\leq \kappa \left\|\boldsymbol{\lambda}_2^{(t)} - \boldsymbol{\mu}\right\|_2 + \left(c n^{3/2} + \kappa\right) \left(c n^{3/2}\right)^t \eta.
\end{aligned}$$

Now if we achieve $\left\|\boldsymbol{\lambda}_2^{(T)} - \boldsymbol{\mu}\right\| \leq \varepsilon$ for some T then it is easy to see that $\left\|\boldsymbol{\lambda}_2^{(t)} - \boldsymbol{\mu}\right\| \leq \varepsilon$ for every $t > T$. Let $\delta = \min\{\mu, 1 - \mu\}$. Using Theorem 8 and the fact that after Step 2. of Power Pretrained EM we will have that the projection of $\boldsymbol{\lambda}^{(0)}$ to the line will be at least δ, with high probability over the randomness of $\tilde{\boldsymbol{\lambda}}^{(0)}$, we have that $\kappa \leq (1-\delta)^{\frac{n-2}{2}}$ and hence if $\eta \leq \varepsilon^{\frac{6 \log n}{n}}$ and $T \geq \frac{2 \log(1/\varepsilon)}{n\delta}$ we have that $\left\| \boldsymbol{\lambda}_2^{(T)} - \boldsymbol{\mu} \right\| \leq \varepsilon$.

Therefore the only thing that is left, is to prove that after the first steps of Power Pretrained EM we will have $\left\| \boldsymbol{\lambda}_1^{(0)} - \boldsymbol{\lambda}_2^{(0)} \right\|_2 \leq \eta$ for some $\eta \leq \varepsilon$. To do so we oberve that using Theorem 10 and Lemma 3 the first steps of Power Pretrained EM are equivalent with power iteration with eigenvalue gap $\rho = 1 + (n-1)\delta$. Also since the Power Pretrained EM starts with a random direction we will have that initially the projection of $\tilde{\boldsymbol{\lambda}}^{(0)}$ to any eigenvector $\boldsymbol{v}$ of the covariance matrix of $d_{\boldsymbol{\mu}}$, is $\frac{c'}{\sqrt{n}}$ with probability at least $1/\text{poly}(n)$. At every step the ratio

$$\frac{\langle \frac{\tilde{\boldsymbol{\lambda}}^{(L)}}{\left\| \tilde{\boldsymbol{\lambda}}^{(L)} \right\|_2}, \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2} \rangle}{\max_{\boldsymbol{v} \nparallel \boldsymbol{\mu}} \left\{ \langle \frac{\tilde{\boldsymbol{\lambda}}^{(L)}}{\left\| \tilde{\boldsymbol{\lambda}}^{(L)} \right\|_2}, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2} \rangle \right\}}$$

increases by ρ and when it reaches $1/\varepsilon$ we have the desired estimated. Therefore it suffices to set $L \geq \frac{\log(n/\varepsilon)}{\log(1+(n-1)\delta)}$ to get to an estimation $\boldsymbol{\lambda}^{(0)}$ such that the closest point $\boldsymbol{\lambda}_1^{(0)}$ parallel to $\boldsymbol{\mu}$ satisfies $\left\| \boldsymbol{\lambda}^{(0)} - \boldsymbol{\lambda}_1^{(0)} \right\| \leq \varepsilon$ and the theorem follows. □

I Convergence with Finite Samples

Proof of Theorem 12:

Consider a point $\boldsymbol{\lambda} \in \Lambda$. The population EM iteration starting from that point is $M(\boldsymbol{\lambda}) = \underset{\boldsymbol{x} \sim d_{\boldsymbol{\mu}}}{\mathbb{E}} \left[\tanh\left(\tanh^{-1}(\boldsymbol{\lambda}) \cdot \boldsymbol{x} \right) \boldsymbol{x} \right]$.

The corresponding finite sample iteration is $\bar{M}(\boldsymbol{\lambda}) = \frac{1}{N} \sum_{i=1}^{N} \tanh\left(\tanh^{-1}(\boldsymbol{\lambda}) \cdot \boldsymbol{x}_i \right) \boldsymbol{x}_i$.

Consider the function $f(\boldsymbol{x}, \boldsymbol{\lambda}) = \tanh\left(\tanh^{-1}(\boldsymbol{\lambda}) \cdot \boldsymbol{x} \right) \boldsymbol{x}$

Since for any fixed $\boldsymbol{\lambda}$, the function $f(\cdot, \boldsymbol{\lambda})$ is in $[-1,1]^n$, the empirical expectation concentrates and it holds that $\left\| \bar{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda}) \right\| \leq c$ with probability $1-\delta$ after $N = \Omega(\frac{n \log 1/\delta}{c^2})$ samples.

Moreover, working similarly to (H.12), we can show that for any set of samples $\boldsymbol{x}_i$, the function $\frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i, \cdot)$ has a Lipschitz constant $n/(1-\lambda_{max}^2)$. This implies that for any starting point $\boldsymbol{\lambda}^{(0)}$ such that $\left\| \boldsymbol{\lambda} - \boldsymbol{\lambda}^{(0)} \right\| \leq c(1-\lambda_{max}^2)/n$, it holds that $\left\| \bar{M}(\boldsymbol{\lambda}) - \bar{M}(\boldsymbol{\lambda}^{(0)}) \right\| \leq c$.

Now consider the discrete set of points

$$\bar{\Lambda} \in \{-1, -1 + c(1-\lambda_{max}^2)/n^2, ..., 1\}^n \cap \Lambda.$$

Setting $\delta = \delta'/|\bar{\Lambda}|$, we get by a union bound that with probability $1-\delta'$ it holds that $\left\| \bar{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda}) \right\| \leq c$ for all points $\boldsymbol{\lambda} \in \bar{\Lambda}$.

Moreover, for any other point $\boldsymbol{\lambda} \in \Lambda$, there exists a point $\boldsymbol{\lambda}' \in \bar{\Lambda}$, such that $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\| \leq c(1-\lambda_{max}^2)/n$. This implies that $\left\| \bar{M}(\boldsymbol{\lambda}) - \bar{M}(\boldsymbol{\lambda}') \right\| \leq c$ and $\|M(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda}')\| \leq c$. This shows that $\left\| M(\boldsymbol{\lambda}) - \bar{M}(\boldsymbol{\lambda}) \right\| \leq 3c$.

Therefore, we get that $\left\| \boldsymbol{\mu} - \bar{M}(\boldsymbol{\lambda}) \right\| \leq \kappa \|\boldsymbol{\mu} - \boldsymbol{\lambda}\| + \varepsilon + 3c$. This shows that the finite sample iteration converges with additional error at most $3c/(1-\kappa)$. By the choice of δ, the number of samples required to achieve this error is $\Omega(\frac{n^2 \log(n^2/c(1-\lambda_{max}^2))}{c^2}) = \tilde{\Omega}(\frac{n^2}{c^2})$. Therefore, with N samples in total we get error $\tilde{O}(\frac{n}{\sqrt{N}})$. ■

J Figures

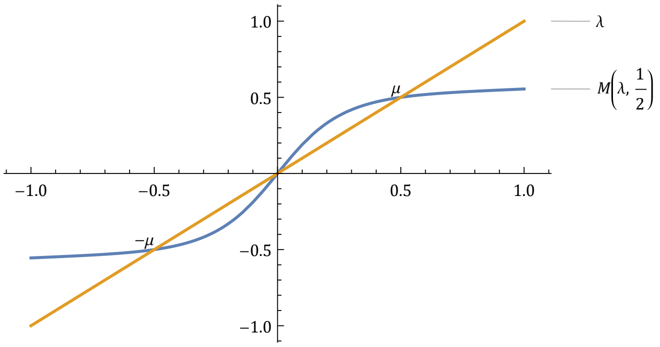


Figure 5: The EM iteration $M(\lambda, \mu)$ for $n = 5$ and $\mu = 1/2$.

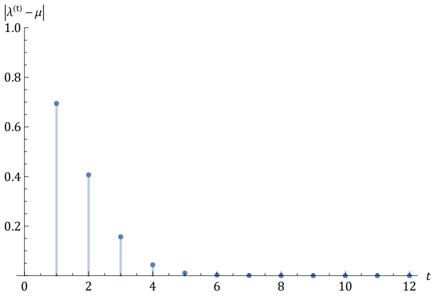


Figure 6: The evolution of $\left|\lambda^{(t)} - \mu\right|$ for $n = 5$, $\mu = 1/2$ and $\lambda^{(0)} = 1/10$.

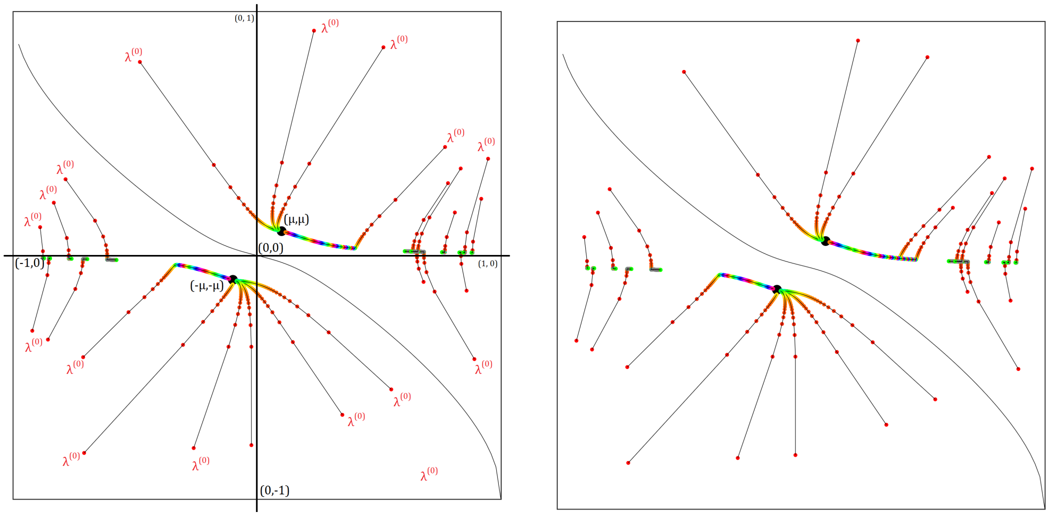
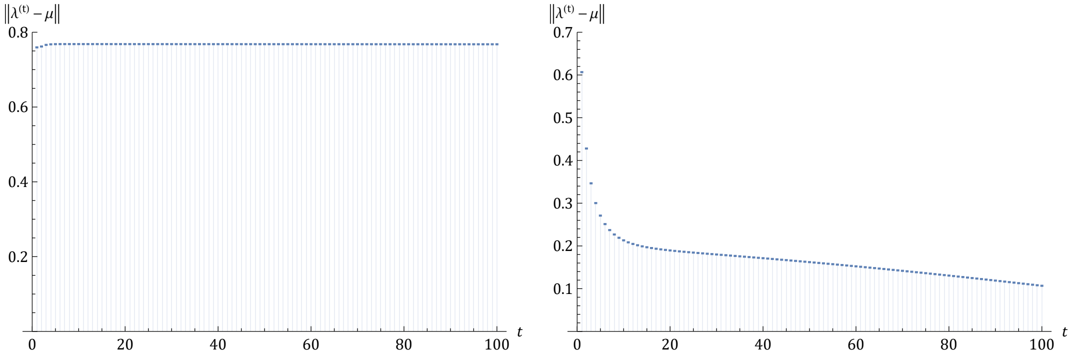
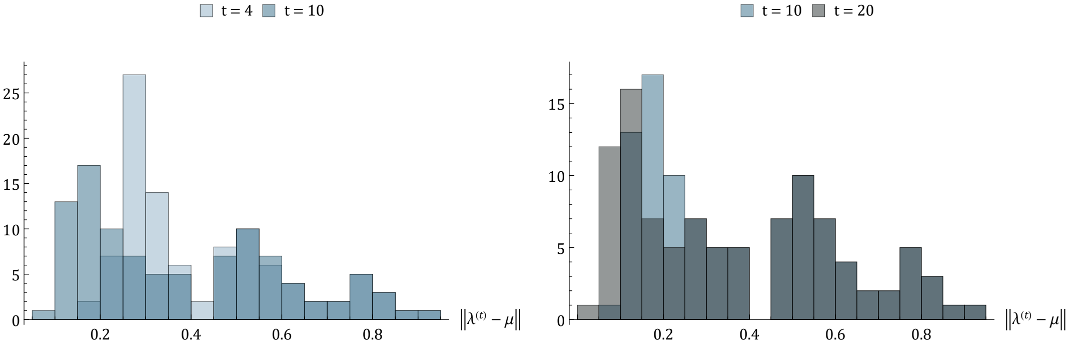


Figure 7: In this figure we see the EM execution for $n = 5$, $\boldsymbol{\mu} = (\mu, \mu, \mu, \mu, \mu)$ with $\mu = 1/10$ and initial guesses of the form $(\lambda_1, \lambda_2, \lambda_2, \lambda_2, \lambda_2)$. The plane that we present here is the (λ_1, λ_2) plane. The red dots that are endpoints of path represent different initial guesses $\boldsymbol{\lambda}^{(0)}$ and the rest of the path represents the execution of the algorithm. The length of the paths is 100. The curved continuous black line separates the region of attraction to the fixed points (μ, μ) and $(-\mu, -\mu)$. As we can see there is a region near the $(0, 1)$ corner of the plane where the convergence of EM is very slow and after 100 steps the progress of EM is very small.



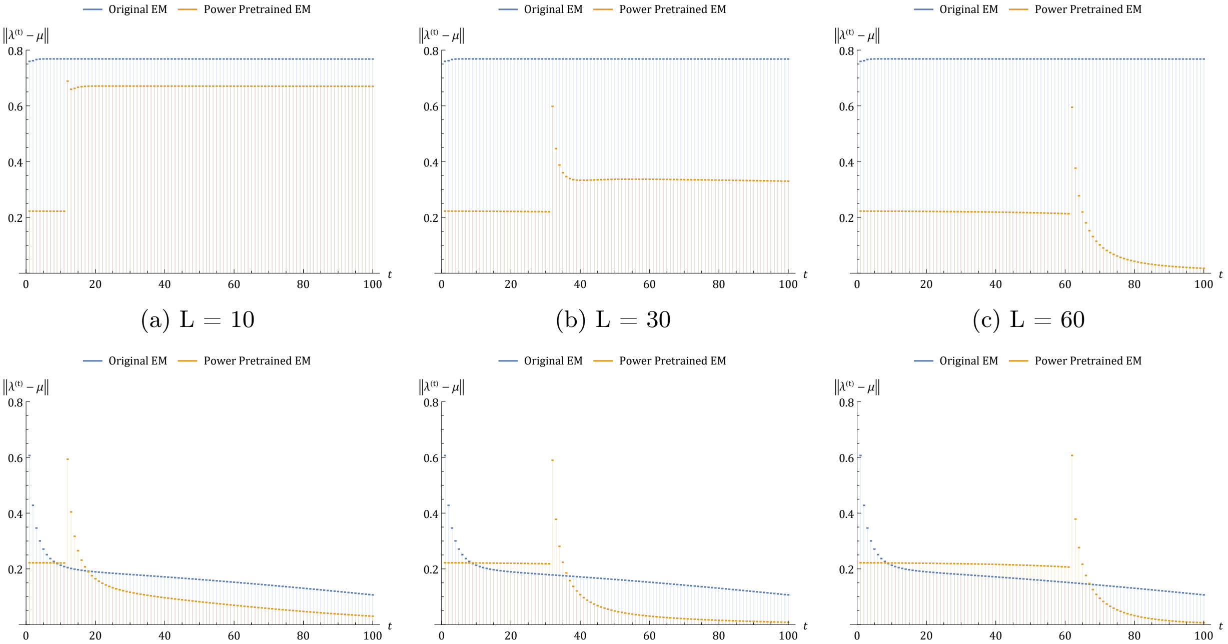
(a) (b)

Figure 8: This figure shows the two most typical scenarios that can be observed when running EM with random initialization for $n = 5$ and $\boldsymbol{\mu} = (0.1, 0.1, 0.1, 0.1, 0.1)$. When we say random initialization we mean that $\boldsymbol{\lambda}^{(0)}$ is picked uniformly at random from $[0, 1]^n$. Both behaviors appear frequently enougn under a random initialization but in general (a) is more frequent that (b).



(a) (b)

Figure 9: This figure shows the distribution of distances $\|\boldsymbol{\lambda}^t - \boldsymbol{\mu}\|$ for several steps t. In figure (a) we see steps $t = 4$, $t = 10$ and in figure (b) we see steps $t = 10$, $t = 20$. The set up is the same as before with $n = 5$ and $\boldsymbol{\mu} = (0.1, 0.1, 0.1, 0.1, 0.1)$.



(a) L = 10

(b) L = 30

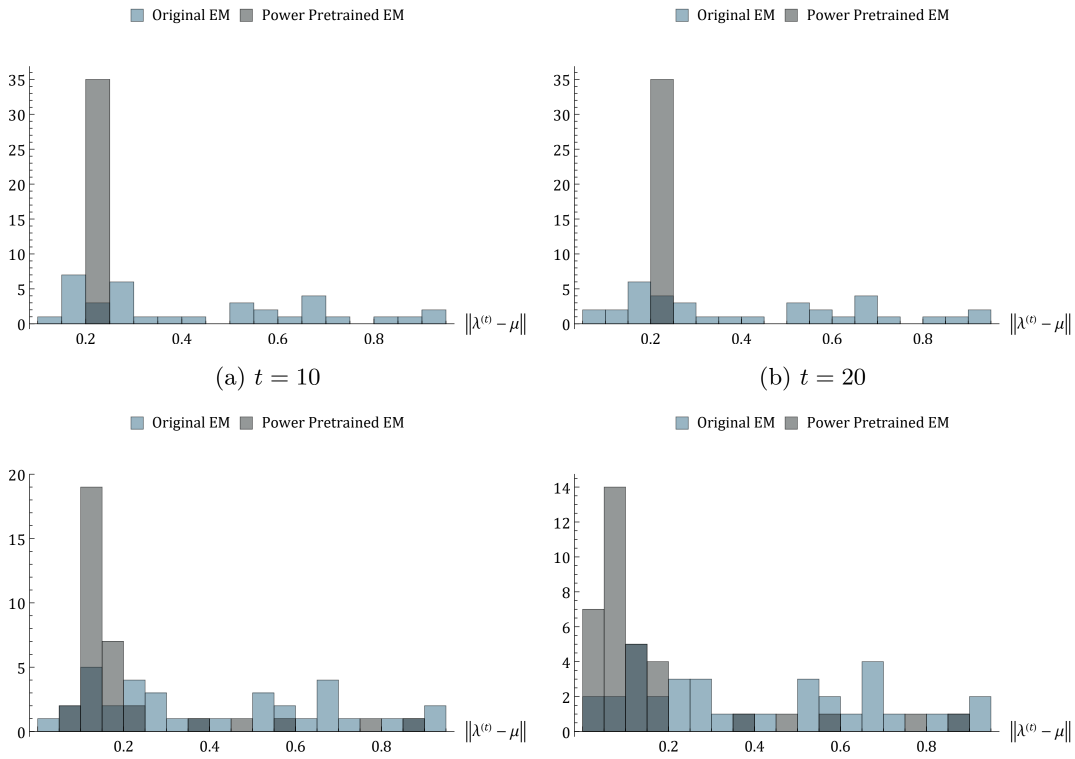
(c) L = 60

(d) L = 10

(e) L = 30

(f) L = 60

Figure 10: The figures (a)-(c) correspond to the initial vector that was used for Figure 8 (a) shown for different values of the parameter L and the figures (d)-(f) correspond to the initial vector that was used for Figure 8 (b) again shown for different values of the parameter L. The set up is again the same with $n = 5$ and $\boldsymbol{\mu} = (0.1, 0.1, 0.1, 0.1, 0.1)$.



(a) $t = 10$

(b) $t = 20$

(c) $t = 40$

(d) $t = 50$

Figure 11: This figures show the distribution of distances $\|\boldsymbol{\lambda}^t - \boldsymbol{\mu}\|$ for several steps t and for both the original and the Power Pretrained EM algorithm. The Power Pretrained EM algorithm goes from step 2. to step 3. for $t = 30$. The set up is again the same with $n = 5$ and $\boldsymbol{\mu} = (0.1, 0.1, 0.1, 0.1, 0.1)$.

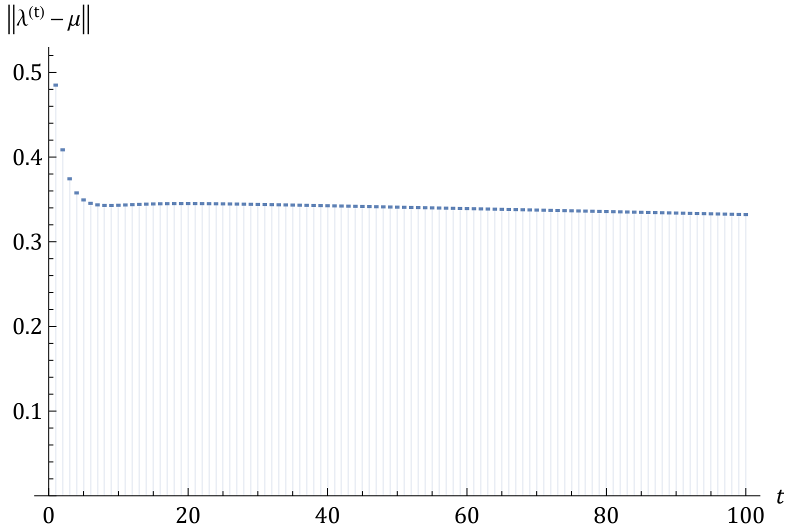
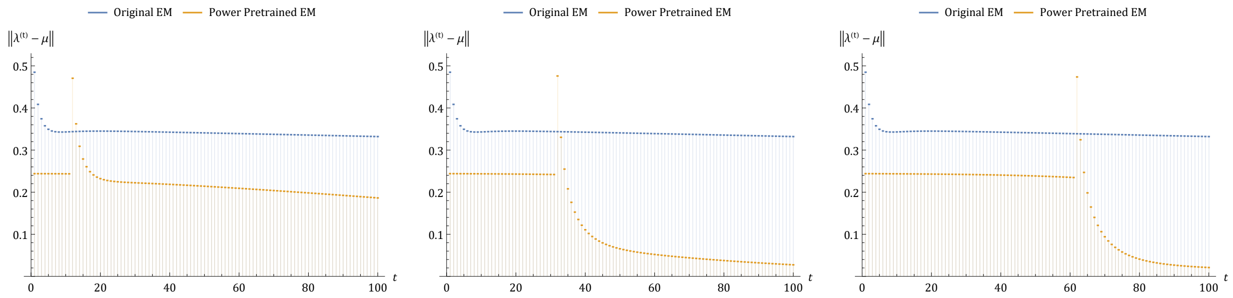
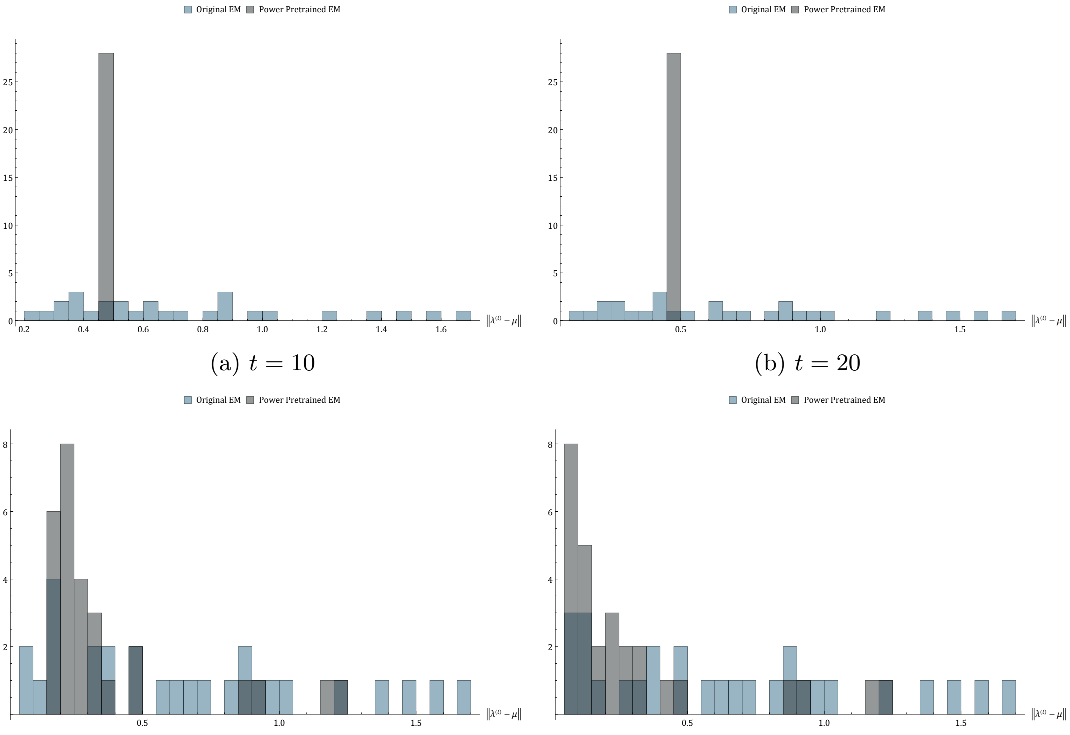


Figure 12: This figure shows the most typical scenario that can be observed when running EM with random initialization for $n = 5$ and $\boldsymbol{\mu} = (0.053, 0.16, 0.09, 0.13, 0.06)$. When we say random initialization we mean again that $\boldsymbol{\lambda}^{(0)}$ is picked uniformly at random from $[0, 1]^n$.



(a) L = 10 (b) L = 30 (c) L = 60

Figure 13: The figures (a)-(c) correspond to the initial vector that was used for Figure 12 shown for different values of the parameter L. The set up is again the same with $n = 5$ and $\boldsymbol{\mu} = (0.053, 0.16, 0.09, 0.13, 0.06)$.



(a) $t = 10$ (b) $t = 20$

(c) $t = 40$ (d) $t = 50$

Figure 14: This figures show the distribution of distances $\|\boldsymbol{\lambda}^t - \boldsymbol{\mu}\|$ for several steps t and for both the original and the Power Pretrained EM algorithm. The Power Pretrained EM algorithm goes from step 2. to step 3. for $t = 30$. The set up is again the same with $n = 5$ and $\boldsymbol{\mu} = (0.053, 0.16, 0.09, 0.13, 0.06)$.