Comparing EM with GD in Mixture Models of Two Components: Supplementary Material

Guojun Zhang and Pascal Poupart
Computer Science, Waterloo AI Institute, University of Waterloo
Vector Institute
{guojun.zhang,ppoupart}@uwaterloo.ca

George Trimponias
Noah's Ark Lab
Huawei
g.trimponias@huawei.com

A Proofs for mixtures of Bernoullis

A.1 Derivation of (4.14)

From (4.10), the update of $\boldsymbol{\mu}_1$ is:

$$M(\boldsymbol{\mu}_1)_i = Z_1^{-1} \int \tilde{q}_1(\mathbf{x})\mathbf{x}d\mathbf{x} = \frac{\mu_{1i}}{\overline{x}_i} Z_1^{-1} F_i, \quad \text{(A.1)}$$

where $F_i = \pi_1^* \mu_{1i}^* B_{1i} + \pi_2^* \mu_{2i}^* B_{2i}$ and B_{1i}, B_{2i} defined in (4.15). So,

$$\begin{aligned} M(\boldsymbol{\lambda})_i - \lambda_i &= 2S_i^{-1}\mu_i^*(M(\boldsymbol{\mu}_1)_i - \mu_{1i}) \\ &= 2S_i^{-1}\mu_i^*\mu_{1i}\overline{x}_i^{-1}Z_1^{-1}(F_i - Z_1\overline{x}_i). \end{aligned} \quad \text{(A.2)}$$

Bringing in the definition of Z_1 in (4.12), we have

$$\begin{aligned} F_i - Z_1\overline{x}_i &= \pi_1^* B_{1i}(\mu_{1i}^* - \overline{x}_i(1 + \pi_2^*\lambda_i)) + \\ &+ \pi_2^* B_{2i}(\mu_{2i}^* - \overline{x}_i(1 - \pi_1^*\lambda_i)). \end{aligned} \quad \text{(A.3)}$$

With the definitions of $\overline{x}_i$, μ_i^* and λ_i, we obtain:

$$F_i - Z_1\overline{x}_i = 2\pi_1^*\pi_2^*\mu_i^*(1-\overline{x}_i)^{-1}(1-\mu_{1i})(B_{1i} - B_{2i}),$$

which, combined with (A.2), yields (4.14).

A.2 Proof of Theorem 4.5

In this section, we prove the following theorem:

Theorem 4.4. *For $m = D = 2$, given $\sigma_{12} \neq 0$ and $\overline{\mathbf{x}} \in (0,1)^D$, with EM algorithm, $\pi_1 = \epsilon$, $\boldsymbol{\mu}_2 = \overline{\mathbf{x}}$ and uniform random initialization for $\boldsymbol{\mu}_1$, $\boldsymbol{\lambda}$ will converge to the positive regions at a linear rate with probability 1. Therefore, EM will almost surely escape one-cluster regions.*

We assume that $\sigma_{12} > 0$ because the $\sigma_{12} < 0$ can be similarly proved by relabeling $x_2 \to 1 - x_2$. The theorem is equivalent to showing that $\mathbf{b}$ converges to the regions where $b_1 b_2 > 0$, due to (4.13) and Corollary 4.1. We also call these regions as positive regions. It is not hard to derive from (4.10) and (2.9) that the EM update is:

$$b_1 \leftarrow b_1 + Z_1^{-1}\sigma\Lambda_1 b_2, \quad \text{(A.4)}$$
$$b_2 \leftarrow b_2 + Z_1^{-1}\sigma\Lambda_2 b_1, \quad \text{(A.5)}$$

with $\Lambda_i = \mu_{1i}(1 - \mu_{1i})$.

We first notice some properties of σ_{12}, as can be easily seen from its definition. For convenience, in the following proof we define $\sigma := \sigma_{12} S_1^{-1} S_2^{-1}$ which we call the normalized covariance.

Lemma A.1. *If $\sigma_{12} > 0$, then $\sigma_{12} < \overline{x}_1(1 - \overline{x}_2)$, $\sigma_{12} < \overline{x}_2(1 - \overline{x}_1)$.*

Proof. Trivial from the definition of σ_{12}. □

A direct consequence is:

Corollary A.1. *If $\sigma_{12} > 0$, then $\sigma^2 S_1 S_2 < 1$.*

In the following lemma, we show that in a neighborhood of the origin, $\mathbf{b}$ almost always converges to the positive regions.

Lemma A.2 (Convergence with small $||\mathbf{b}||$ initialization, two features). *Assume $\sigma_{12} > 0$. $\exists \delta > 0$ small enough, with a random $\mathbf{b}$ initialized from the L_1 ball $||\mathbf{b}||_1 < \delta$ and the update function defined by EM, $\mathbf{b}$ converges to the positive regions $\{(b_1, b_2)|b_1 b_2 > 0\}$ at a linear rate.*

Proof. In this case, Λ_i and Z are roughly constant, and

$$\mathbf{b}' = \begin{bmatrix} b_1' \\ b_2' \end{bmatrix} = \mathbf{A}(\mathbf{b})\mathbf{b} = \begin{bmatrix} 1 & \sigma_{12}\Sigma_2^{-1} \\ \sigma_{12}S_1^{-1} & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \quad \text{(A.6)}$$

The eigensystem of $\mathbf{A}(\mathbf{b})$ is:

$$\lambda_1 = 1 + \sigma_{12}\sqrt{S_1^{-1}S_2^{-1}},\ \mathbf{v}_1 = (\sqrt{\sigma_{12}S_2^{-1}}, \sqrt{\sigma_{12}S_1^{-1}}).$$
$$\lambda_2 = 1 - \sigma_{12}\sqrt{S_1^{-1}S_2^{-1}},\ \mathbf{v}_2 = (-\sqrt{\sigma_{12}S_2^{-1}}, \sqrt{\sigma_{12}S_1^{-1}}).$$

Applying $\mathbf{A}(\mathbf{b})$ for enough number of times, $\mathbf{b}$ will converge to a multiple of $\mathbf{v}_1$, where $\sigma b_1 b_2 > 0$.

Let us make the argument above more concrete. Expand $\mathbf{b}$ as:

$$\mathbf{b} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2. \tag{A.7}$$

Since $c_1 = 0$ is a measure zero set, we have $c_1 \neq 0$ almost everywhere. WLOG, we assume $c_1 > 0$. Denote

$$\mathbf{A}(\mathbf{b}) = \begin{bmatrix} 1 & \sigma Z^{-1}\Lambda_1 \\ \sigma Z^{-1}\Lambda_1 & 1 \end{bmatrix}, \tag{A.8}$$

$$\mathbf{A}(0) = \begin{bmatrix} 1 & \sigma_{12} S_2^{-1} \\ \sigma_{12} S_1^{-1} & 1 \end{bmatrix}. \tag{A.9}$$

We first prove that each element of $\mathbf{A}(\mathbf{b}) - \mathbf{A}(0)$ is bounded. This can be done by noticing:

$$\mathbf{A}(\mathbf{b}) - \mathbf{A}(0) = \begin{bmatrix} 0 & \sigma(Z^{-1}\Lambda_1 - S_1) \\ \sigma(Z^{-1}\Lambda_2 - S_2) & 0 \end{bmatrix},$$

and that

$$Z^{-1}\Lambda_i - S_i = Z^{-1}(b_i(1-2\overline{x}_i) - b_i^2 - \sigma b_1 b_2 S_i).$$

So, $\mathbf{A}(\mathbf{b}) - \mathbf{A}(0) = O(\delta)$ in $||\mathbf{b}||_1 < \delta$. From this fact, one can show in $||\mathbf{b}||_1 < \delta$,

$$\mathbf{A}(\mathbf{b})\mathbf{b} \succeq \mathbf{A}(0)\mathbf{b} - c\delta^2 \mathbf{v}_1, \tag{A.10}$$

with c some constant. WLOG, we assume $\mathbf{b}$ is still in the negative region and thus $||\mathbf{b}||_1$ decreases, so our approximation is still valid. Applying EM for k times and by use of (A.10), we know the result is at least (the generalized inequality is defined by the positive cone $\mathbb{R}^2_{++}$, see, e.g., [1])

$$\begin{aligned} &\mathbf{A}(0)^k \mathbf{b} - c\delta^2(\lambda_1^{k-1} + \cdots + \lambda_1 + 1)\mathbf{v}_1 \\ = \;& (c_1\lambda_1^k - c\delta^2 \frac{\lambda_1^k - 1}{\lambda_1 - 1})\mathbf{v}_1 + c_2 \lambda_2^k \mathbf{v}_2. \end{aligned} \tag{A.11}$$

In the above analysis we used $\mathbf{A}(0)\mathbf{u} \succeq 0$ for $\mathbf{u} \succeq 0$. For δ small enough, we know almost surely $c_1 > c\delta^2/(\lambda_1 - 1)$. Therefore, at almost everywhere $\mathbf{A}(\mathbf{b})^k \mathbf{b}$ converges to the positive regions at a linear rate.

The choice of L_1 ball is irrelevant, since in finite vector space all L_p norms are equivalent. □

Now, let us show that in the worst case $\mathbf{b}$ shrinks to a neighborhood of the origin. Hence, combined with Lemma A.2, we finish the proof. First, rewrite (A.4) and (A.5) as:

$$b_1 \leftarrow Z^{-1}(b_1 + \sigma S_1 b_2 + \sigma(1-2\overline{x}_1)b_1 b_2), \tag{A.12}$$

$$b_2 \leftarrow Z^{-1}(b_2 + \sigma S_2 b_1 + \sigma(1-2\overline{x}_2)b_1 b_2). \tag{A.13}$$

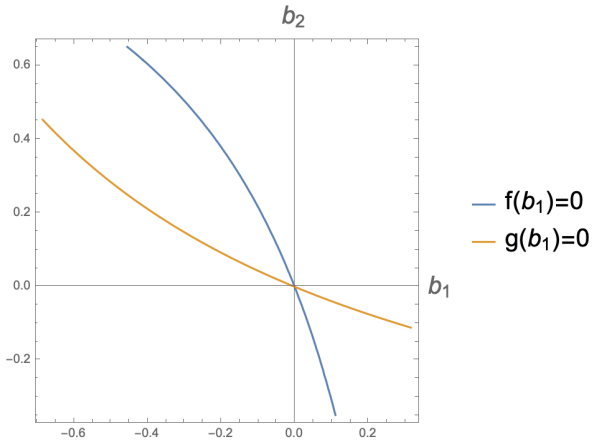

Figure 1: Contours $f(b_1) = 0$ and $g(b_1) = 0$.

The two contours $b_1' = 0, b_2' = 0$ are respectively:

$$C_{b_1'=0} : b_2 = f(b_1) = -\frac{b_1}{\sigma(1-2\overline{x}_1)b_1 + \sigma S_1}, \tag{A.14}$$

$$C_{b_2'=0} : b_2 = g(b_1) = \frac{-\sigma S_2 b_1}{1 + \sigma(1-2\overline{x}_2)b_1}. \tag{A.15}$$

f, g are both linear fractional functions of b_1, an example of which is depicted in Figure 1. The derivatives are:

$$\begin{aligned} f'(b_1) &= -\frac{S_1}{\sigma((1-2\overline{x}_1)b_1 + S_1)^2}, \\ g'(b_1) &= -\frac{\sigma S_2}{(1+\sigma(1-2\overline{x}_2)b_1)^2}, \end{aligned} \tag{A.16}$$

therefore, f, g are both decreasing if $\sigma > 0$. It follows that $f'(0) = -(\sigma S_1)^{-1}$ and $g'(0) = -\sigma S_2$. From Corollary A.1,

$$\frac{|g'(0)|}{|f'(0)|} < 1. \tag{A.17}$$

Now, let us look at the secant lines crossing the origin and $f(-\overline{x}_1)$, $g(-\overline{x}_1)$ separately. From (A.14) and (A.15),

$$f(-\overline{x}_1) = \frac{1}{\sigma \overline{x}_1}, \tag{A.18}$$

$$g(-\overline{x}_1) = \frac{\sigma S_2 \overline{x}_1}{1 - \sigma\overline{x}_1(1-2\overline{x}_2)}, \tag{A.19}$$

and one can obtain $f(1-\overline{x}_1)$ and $g(1-\overline{x}_1)$ similarly. So, $-b_1/b_2$ is bounded in the following region:

$$\{(b_1, b_2) | b_1 b_2 < 0, f(b_1)g(b_1) < 0\}, \tag{A.20}$$

and the bound is given by the slopes of the tangent lines at $(0,0)$ and the secant lines.

Another important point to notice is that $f(b_1)$ and $g(b_1)$ intersect exactly once. Which can be proved from Lemma A.1, (A.14) and (A.15):

Lemma A.3. *Assume $\sigma > 0$, $f(b_1) = g(b_1)$ has exactly one solution $b_1 = 0$ in the feasible region $b_1 \in [-\overline{x}_1, 1 - \overline{x}_1]$.*

Proof. The solution $b_1 = 0$ is obvious. For $b_1 \neq 0$, $f(b_1) = g(b_1)$ is equivalent to:

$$\sigma^2 S_2 \left((1 - 2\overline{x}_1)b_1 + S_1\right) - \sigma(1 - 2\overline{x}_2)b_1 = 1. \quad \text{(A.21)}$$

We will show that the left hand side is always less than one. This is a linear function, so we only need to show it at both end points. At $b_1 = -\overline{x}_1$, the left hand side can be simplified as:

$$\sigma\overline{x}_1(1 - \overline{x}_2) + \sigma\overline{x}_1\overline{x}_2\left(\sigma\overline{x}_1(1 - \overline{x}_2) - 1\right) < 1,$$

where we used Lemma A.1. Similarly, at $b_1 = 1 - \overline{x}_1$, the left hand side of (A.21) is:

$$\sigma\overline{x}_2(1 - \overline{x}_1) + \sigma(1 - \overline{x}_1)(1 - \overline{x}_2)\left(\sigma\overline{x}_2(1 - \overline{x}_1) - 1\right) < 1.$$

□

From Lemma A.3, the feasible region of $\mathbf{b}$ is divided into four parts by $f(b_1)$ and $g(b_1)$. If $f(b_1)g(b_1) > 0$, then EM update goes to the positive region. Otherwise, $f(b_1)g(b_1) < 0$. In this region, neither b_1 nor b_2 changes the sign. Because $-b_1/b_2$ is bounded, from (A.4) and (A.5), $||\mathbf{b}||_1^{(t+1)} \leqslant q||\mathbf{b}||_1^{(t)}$ with $0 < q < 1$ being a constant. So, in the worst case, $\mathbf{b}$ will converge to the neighborhood of the origin at a linear rate, and then shift to the positive regions at a linear rate, according to Lemma A.2. This lemma can be used because the EM update is not singular: it does not map a measure nonzero set to a measure zero set. Hence, after finitely many steps, at the neighborhood of the origin, the random distribution at the beginning is still random.

Figure 2 is an example of the trajectory.

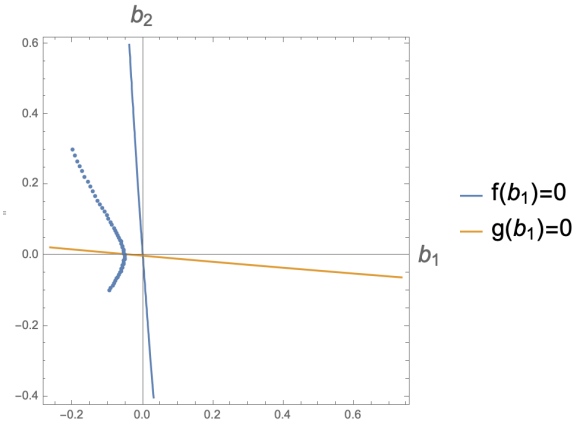

Figure 2: An example of the trajectory. $\mathbf{b}$ moves from the second orthant to a positive region, by shrinking its norm and rotating.

B A General Conjecture

In this appendix, we propose a general conjecture for mixtures of Bernoullis:

Conjecture B.1. *For any number of clusters and a general number of features, with random initialization around k-cluster regions, EM will almost always converge to an m-cluster point.*

Besides empirical evidence, we can show theoretical guarantees at $m = 2$. In this case, the problem reduces to showing the convergence to positive regions proposed in Section 4.2.2. The convergence to the positive regions is observed empirically for mixtures of two Bernoullis, which always happens with random initialization.

The first result shows that if $||\boldsymbol{\lambda}||$ is small, $\boldsymbol{\lambda}$ will converge to the positive regions:

Proposition B.1 (Convergence with small $||\boldsymbol{\lambda}||$ initialization, general). *For mixtures of two Bernoullis, $\exists\delta > 0$ small enough, with a random $\boldsymbol{\lambda}$ initialized from the L_1 ball $||\boldsymbol{\lambda}||_1 < \delta$ and the update function defined by EM, $\boldsymbol{\lambda}$ will converge to the positive regions.*

Proof. Around $||\boldsymbol{\lambda}|| \sim 0$, $Z_1 \sim 1$. Expanding (4.14) to the linear order, we find that the update of $\boldsymbol{\lambda}$ can be linearized as:

$$M(\boldsymbol{\lambda}) = \mathbf{A}\boldsymbol{\lambda}, \quad \text{(B.1)}$$

where $A_{ii} = 1$ and $A_{ij} = (2\mu_i^*)^2\pi_1^*\pi_2^* S_i^{-1} Z_1^{-1} > 0$ for $i \neq j$. After enough iterations, $\boldsymbol{\lambda}$ will converge to the linear span of the largest eigenvector of $\mathbf{A}$.

From the Perron-Frobenius theorem [2], $\mathbf{A}$ has a unique largest real eigenvalue, and $\text{eig}_{\max}(\mathbf{A}) \geqslant \min_i \sum_j A_{ij} > 1$. Also, the maximal eigenvector of $\mathbf{A}$, $\mathbf{v}_{\max}$, is a multiple of an all positive vector. Therefore, we can prove the proposition in a similar fashion as the proof of Lemma A.2. □

This proposition also tells us that EM has an effect of rotating $\boldsymbol{\lambda}$ to the positive regions. It is interesting to observe that such unstable fixed point $\boldsymbol{\lambda} = 0$ is analogous to the strict saddle points studied in [3]. It might be possible to use stable manifold theorem to prove our conjecture at $m = 2$.

Another special case is when the $\max_i \lambda_i$ increases or $\min_i \lambda_i$ decreases. For a given $\boldsymbol{\lambda}$, we can order the components. WLOG, we assume $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \lambda_i < 0 = \lambda_{i+1} = \cdots = \lambda_j < \lambda_{j+1} \leqslant \ldots \leqslant \lambda_D$. We can show the following proposition:

Proposition B.2. *For mixtures of two Bernoullis, assume $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \lambda_i < 0 = \lambda_{i+1} = \cdots = \lambda_j < \lambda_{j+1} \leqslant \ldots \leqslant \lambda_D$, if $M(\boldsymbol{\lambda})_1 < \lambda_1$ or $M(\boldsymbol{\lambda})_D > \lambda_D$, then $\boldsymbol{\lambda}$ will eventually converge to the positive regions. Otherwise, we have $M(\boldsymbol{\lambda})_1 \geqslant \lambda_1$ and $M(\boldsymbol{\lambda})_D \leqslant \lambda_D$.*

Proof. From the definitions of B_{1i} and B_{2i}, (4.15), we have

$$B_{11} - B_{21} \geqslant \ldots \geqslant B_{1D} - B_{2D}. \tag{B.2}$$

$M_1(\boldsymbol{\lambda}) < \lambda_1$, from (4.14), tells us that $B_{11} - B_{21} < 0$, and thus every λ_i decreases. As each λ_i decreases, $B_{1i} - B_{2i}$ will get smaller as well. Therefore, all λ_i's decrease at least as a linear function. Since the feasible region is bounded, $\boldsymbol{\lambda}$ will converge to $-\mathbb{R}^D_{++}$ eventually.

Similarly, if $M_D(\boldsymbol{\lambda}) > \lambda_D$, we know that $\boldsymbol{\lambda}$ will converge to $\mathbb{R}^D_{++}$ eventually.

Otherwise, we must have $B_{11} - B_{21} \geqslant 0$ and $B_{1D} - B_{2D} \leqslant 0$, yielding $M_1(\boldsymbol{\lambda}) \geqslant \lambda_1$ and $M_D(\boldsymbol{\lambda}) \leqslant \lambda_D$. □

The two patterns $M(\boldsymbol{\lambda})_1 < \lambda_1$ and $M(\boldsymbol{\lambda})_D > \lambda_D$ have been observed in experiments very frequently, while the case with $M(\boldsymbol{\lambda})_1 \geqslant \lambda_1$ and $M(\boldsymbol{\lambda})_D \leqslant \lambda_D$ needs some further understanding.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[2] Carl D Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

[3] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.