

# EM for Spherical Gaussians

Karthekeyan Chandrasekaran

Hassan Kingravi

December 4, 2007

## 1 Introduction

In this project, we examine two aspects of the behavior of the EM algorithm for mixtures of spherical Gaussians; 1) the benefit of spectral projection for such mixtures, and 2) the general behavior of the EM algorithm under certain separability criteria. Our current results are for mixtures of two Gaussians, although these can be extended. In the case of 1), we show that the value of the  $Q$  function for EM (see below) increases in general when one performs spectral projection, and for 2), we give empirical results under different separability conditions, and make a conjecture concerning these.

## 2 The EM Algorithm

The Expectation Maximization algorithm is a means of estimating the maximum likelihood of some data given a distribution. We will sketch a general outline of the method, which follows Bilmes' tutorial [1]. Assume we are given a density function  $p(x|\Theta)$  that is governed by a set of parameters  $\Theta$ . We are also given a set of data of size  $N$ , i.e.  $X = \{x_1, \dots, x_N\}$ , which is drawn from this distribution, s.t. the individual vectors are i.i.d with distribution  $p$ . The likelihood function is then defined as

$$p(x|\Theta) = \prod_{i=1}^N p(x_i|\Theta) = L(\Theta|X)$$

We think of this function as the *likelihood* of the parameters given the data, and we wish to find the  $\Theta$  that maximizes  $L$ . (Note that one can maximize the log likelihood here if it makes analysis easier; as it does in the Gaussian case.) EM allows us to maximize this by assuming that the data that is observed is incomplete, and thus has unobserved values. Assume that  $X$  is observed, and  $Y$  is the unobserved data, and  $Z$  is the concatenation of this, i.e. the complete data. Then:

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)p(x|\Theta)$$

Note that our new likelihood function becomes  $L(\Theta|Z) = L(\Theta|X, Y) = p(X, Y|\Theta)$ , i.e. the complete data likelihood. This function is a random variable, since the missing information  $Y$  is unknown. The EM algorithm first finds the expected value of  $\log p(X, Y|\Theta)$  with respect to the unknown data  $Y$  given  $X$  and the current parameter estimates. We denote this as:

$$Q(\Theta, \Theta^{(i-1)}) = E[\log p(X, Y|\Theta)|X, \Theta^{(i-1)}] = \int_{y \in Y^*} \log p(X, Y|\Theta) f(y|X, \Theta^{(i-1)}) dy$$

where  $\Theta^{(i-1)}$  are the current parameter estimates,  $Y^*$  is the space of values that  $y$  can take on, and  $f(y|X, \Theta^{(i-1)})$  is the distribution which governs the random variable  $Y$ .

Once the expectation is evaluated, we proceed to maximize it. This step can be written as follows:

$$\Theta^{(i)} = \underset{\Theta}{\operatorname{argmax}} Q(\Theta, \Theta^{(i-1)}).$$

The EM update equations for spherical Gaussians do both expectation and maximization simultaneously. However, for our current purposes, we are interested in the  $Q$  function for spherical Gaussians. Assume you have the following probabilistic model:

$$p(x|\Theta) = \sum_{i=1}^M \alpha_i p_i(x|\theta_i)$$

where the parameters are  $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$ , the  $p_i$ 's are density functions characterized by  $\theta_i$ , and the  $\alpha_i$ 's (the mixing weights) sum up to 1. Then [1] shows that, given a guess of parameters  $\Theta'$  (for assumptions concerning the guesses, see the next section):

$$Q(\Theta, \Theta') = \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l|x_i, \Theta') + \sum_{l=1}^M \sum_{i=1}^N \log(p_l(x_i|\theta_l)) p(l|x_i, \Theta')$$

Here  $p(l|x_i, \Theta')$  means the probability of the  $l$ th distribution given the data point and the parameters, whereas  $p_l(x_i|\theta_l)$  denotes the probability of  $x_i$  coming from the  $l$ th distribution. In a spherical Gaussian with  $\theta = (\mu, \sigma)$ , this is given by:

$$p(l|x, \mu, \sigma) = p(x|\mu, \sigma) = \frac{1}{((2\pi)^{\frac{1}{2}} \sigma)^d} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

### 3 The Benefit of Spectral Projection for Mixtures of Two Spherical Gaussians

In this section, we will show the benefit of spectral projection for the first step of the EM algorithm.

#### 3.1 Notation

- $x$  A point drawn from the first Gaussian (i.e. the one we are interested in.)
- $d$  The dimension of the original space.
- $k$  The dimension of the SVD subspace (and thus the number of Gaussians.)
- $\pi_{SVD}(y)$  The projection of  $y$  onto the SVD subspace, for arbitrary point  $y$ .
- $\alpha_i$  The probability that point  $x$  comes from the  $i$ th Gaussian in  $n$ -dimensional space
- $\hat{\alpha}_i$  The probability that point  $x$  comes from the  $i$ th Gaussian in  $k$ -dimensional space
- $\mu_i$  The true mean of the  $i$ th Gaussian.
- $\sigma_i$  The true variance of the  $i$ th Gaussian.
- $\mu'_i$  The first guess for the mean of the  $i$ th Gaussian.
- $\Delta\mu_k$  The distance between the first and the  $k$ th mean, i.e.  $\|\mu_1 - \mu_k\|^2$

#### 3.2 Assumptions

As noted,  $x$  comes from the first Gaussian. We are trying to prove that if  $x$  comes from the first Gaussian, a spectral projection will make the value of the  $Q$  function used in EM greater in the first iteration. To achieve this, we use assume the following:

- Currently, assume  $\forall i, j \leq k, \sigma_i = \sigma_j = \sigma$ .
- We have an initial guess for the weights, which we note holds only in the first step of the EM algorithm. We denote this by  $w_1^0 = w_2^0 = \dots = w_k^0 = \frac{1}{k}$ , where the superscript 0 refers to the 0th iteration of the algorithm.
- We use a concentration lemma to bound the expected values of our points to our guess for the means  $\mu'_i$ . The assumption we use here is that each respective mean  $\mu'_i$  is chosen s.t. the lemma below applies; roughly speaking, this corresponds to saying that the distance between the guess for the  $i$ th mean and the real  $i$ th mean is not too high. Since  $k$ -means clustering is often performed on the data as a preprocessing step to initialize the means, this is not overly restrictive.

### 3.3 Concentration Inequalities

Let  $x \in D_1$ .

Then:

$$\sigma^2 d - \sigma^2 \sqrt{d \log(m)} \leq \mathbb{E}(\|x - \mu_1\|^2) \leq \sigma^2 d + \sigma^2 \sqrt{d \log(m)}$$

and

$$\begin{aligned} \Delta\mu^2 + \sigma^2 d - \sigma^2 \sqrt{d \log(m)} - 2\Delta\mu\sigma(d \log(m))^{\frac{1}{4}} &\leq \mathbb{E}(\|x - \mu_2\|^2) \\ &\leq \Delta\mu^2 + \sigma^2 d + \sigma^2 \sqrt{d \log(m)} + 2\Delta\mu\sigma(d \log(m))^{\frac{1}{4}} \end{aligned}$$

Note that bounds similar to the last two items above hold for their SVD projections in dimension  $k$ . Since the means for spherical Gaussians are contained in the SVD subspace, they do not move upon projection [2]. This fact helps us in the proofs in the next section.

### 3.4 Q Function Verification for the 2-Dimensional Case

Let us define the aforementioned  $Q$  function in terms of two components, as below:

$$\begin{aligned} Q(\Theta, \Theta') &= R(\Theta') + S(\Theta, \Theta') \\ \text{s.t. } R(\Theta') &= \sum_{l=1}^M \sum_{i=1}^N \log(\alpha_l) p(l|x_i, \Theta') \text{ and } S(\Theta, \Theta') = \sum_{l=1}^M \sum_{i=1}^N \log(p_l(x_i|\theta_l)) p(l|x_i, \Theta') \end{aligned}$$

We now prove two theorems.

**Theorem 1.** *Let  $\hat{p}(l|x_i, \Theta')$  be the probability of a distribution  $l \in \{1, 2\}$  after the projection of a dataset generated by a mixture of two Gaussians meeting the assumptions stated above. If  $x_i \in D_l$ , then  $\hat{p}(l|x_i, \Theta') \geq p(l|x_i, \Theta')$ , where the latter is the probability in the original space.*

*Proof.* For the purposes of this proof, let  $l = 1$ . A similar result can then be proved for  $l = 2$ .

We can rewrite the probabilities in the following way:

$$\begin{aligned} \hat{p}(l|x_i, \Theta') &\geq p(l|x_i, \Theta') \\ \Rightarrow \frac{e^{-\|\pi_{SVD}(x-\mu_1)\|^2}}{e^{-\|\pi_{SVD}(x-\mu_1)\|^2} + e^{-\|\pi_{SVD}(x-\mu_2)\|^2}} &\geq \frac{e^{-\|x-\mu_1\|^2}}{e^{-\|x-\mu_1\|^2} + e^{-\|x-\mu_1\|^2}} \\ \Rightarrow \frac{e^{\|\pi_{SVD}(x-\mu_2)\|^2}}{e^{\|\pi_{SVD}(x-\mu_1)\|^2} + e^{\|\pi_{SVD}(x-\mu_2)\|^2}} &\geq \frac{e^{\|x-\mu_2\|^2}}{e^{\|x-\mu_1\|^2} + e^{\|x-\mu_1\|^2}} \end{aligned}$$

Assuming the concentration lemmata hold, w.h.p,  $\|\pi_{SVD}(x-\mu_2)\|^2$  is dominated to a much greater extent by  $\Delta\mu$  than its counterpart in the original space, because all the other components shrink according to the dimension. Therefore, given a large enough  $\Delta\mu$ , the theorem follows.  $\square$

**Theorem 2.** *Assume a data set is generated by a mixture of two Gaussians meeting the assumptions stated above. Then, for the projection of the data onto the SVD subspace spanned by the first two principal components, guarantees a higher  $S$  value in the first step of the EM algorithm.*

*Proof.* Let  $k = 2$ . In this case, this allows us to express the probabilities in terms of each other, i.e.  $\alpha_2 = 1 - \alpha_1 = 1 - \alpha$ . The fact that we will not be able to do this in higher dimensions will pose an issue. Note further that this allows us to state that  $\hat{\alpha} = \alpha$ , where  $\hat{\alpha}$  represents the mixing weights in the projected space.

$$\hat{S}(\hat{\theta}', \hat{\theta}^0) \geq S(\theta', \theta^0)$$

Here,  $\theta^0$  refers to the respective parameters' initial setting, before the first step of the EM algorithm. We will prove the inequality for each point and since  $S$  is the summation across the whole set of sample points, we would have proved the above inequality in its entirety. The above becomes:

$$\begin{aligned}
&\Rightarrow -\hat{\alpha}(\|\pi_{SVD}(x - \mu'_1)\|^2) - (1 - \hat{\alpha})(\|\pi_{SVD}(x - \mu'_2)\|^2) \geq -\alpha(\|x - \mu'_1\|^2) - (1 - \alpha)(\|x - \mu'_2\|^2) \\
&\Rightarrow \alpha(\sigma^2 2 - \sigma^2 \sqrt{2 \log(m)}) + (1 - \alpha)(\Delta\mu^2 + \sigma^2 2 - \sigma^2 \sqrt{2 \log(m)} - 2\Delta\mu\sigma(2 \log(m))^{\frac{1}{4}}) \\
&\leq \alpha(\sigma^2 d + \sigma^2 \sqrt{d \log(m)}) + (1 - \alpha)(\Delta\mu^2 + \sigma^2 d + \sigma^2 \sqrt{d \log(m)} + 2\Delta\mu\sigma(d \log(m))^{\frac{1}{4}}) \\
&\quad \Rightarrow 2\sigma^2 - \sqrt{2 \log(m)}\sigma^2 - 2\Delta\mu\sigma(2 \log(m))^{\frac{1}{4}} + 2\Delta\mu\sigma(2 \log(m))^{\frac{1}{4}}\alpha \\
&\quad \leq d\sigma^2 + \sqrt{d \log(m)}\sigma^2 + 2\Delta\mu\sigma(d \log(m))^{\frac{1}{4}} - 2\Delta\mu(d \log(m))^{\frac{1}{4}}\alpha \\
&\Rightarrow 2\Delta\mu\sigma((\log(m))^{\frac{1}{4}})(2^{\frac{1}{4}} + d^{\frac{1}{4}})\alpha \leq (d - 2)\sigma^2 + \sqrt{\log(m)}(\sqrt{2} + \sqrt{d})\sigma^2 + 2\Delta\mu\sigma((\log(m))^{\frac{1}{4}})(2^{\frac{1}{4}} + d^{\frac{1}{4}})
\end{aligned}$$

Since  $0 < \alpha < 1$ , it's easy to see that the inequality holds.  $\square$

We would now like to prove a similar inequality for the  $R$  function. If we can do this, we will have proved that the  $Q$  value is higher after spectral projection than before it. Unfortunately, we run into a problem here. The issue is that the  $\alpha_l$ 's i.e.  $\alpha_l = \frac{1}{N} \sum_{i=1}^N p(l|x_i, \Theta')$  are enclosed within the log function, which makes it more difficult to quantify their behavior. However, intuitively, it seems that the  $R$  should increase after projection, because all the  $\log(\alpha_l)$  term does is assign a weight to the probabilities, which we have proved *do* increase after projection. If this does not work, one can try proving the whole  $Q$  directly; as can be seen in the proof for Theorem 2, there is a lot of slack in the inequality. For the moment, however, we will leave this as a conjecture.

**Conjecture 1.** *The value of  $R(\Theta')$  increases after spectral projection.*

## 4 The Behavior of the EM algorithm Under Separability Conditions

In this section, we present a few graphs that seem to outline how the convergence of the EM algorithm to the global optimum relates to the separability between two univariate component gaussians. We pose several questions related to interpreting the graph.

In the graphs attached below, the variance of either component is set to 1. This is without loss of generality, since these graphs are just scaled up based on the percentage of overlap, which is determined by the increase in variances. We refrain from letting the two components having different variances since the presented simplest model itself is yet to be completely understood.

Here are some simple observations:

1. As we expect of any clustering algorithm, the performance of the EM algorithm degrades with increase in overlap.
2. When the performance is good, i.e., when EM converges to the global optimum with high probability for random initializations, the bad choices of the initialization points seem to be the ones in which both the means are initialized to the same value.

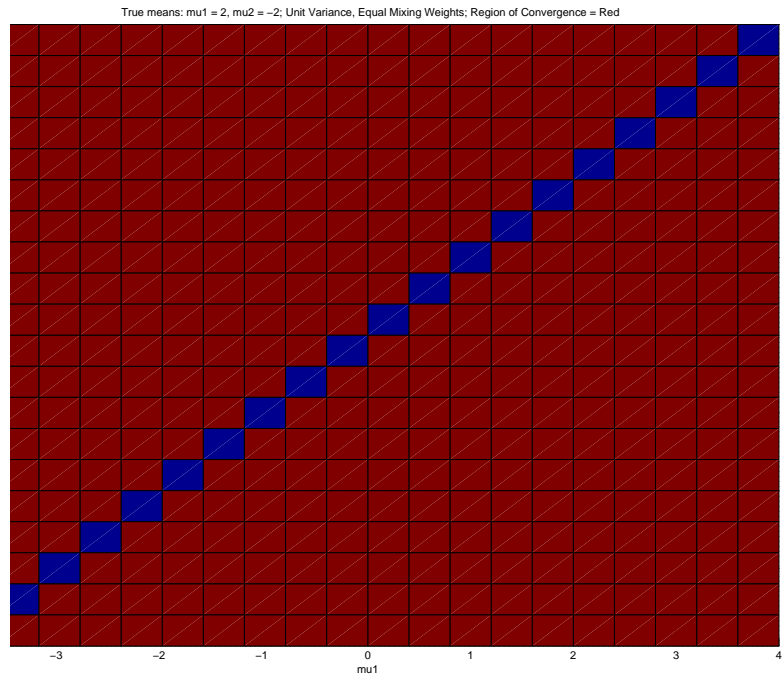
Based on these observations, one can conjecture the following.

**Conjecture 2.** *For an equal mixture of two univariate gaussians, one can identify the true means using the EM algorithm with very high probability using random initial guesses only if the overlap between the two gaussians is less than 7 percent i.e., the distance between the means is  $3\sigma$  if  $\sigma$  is the standard deviation of the two distributions.*

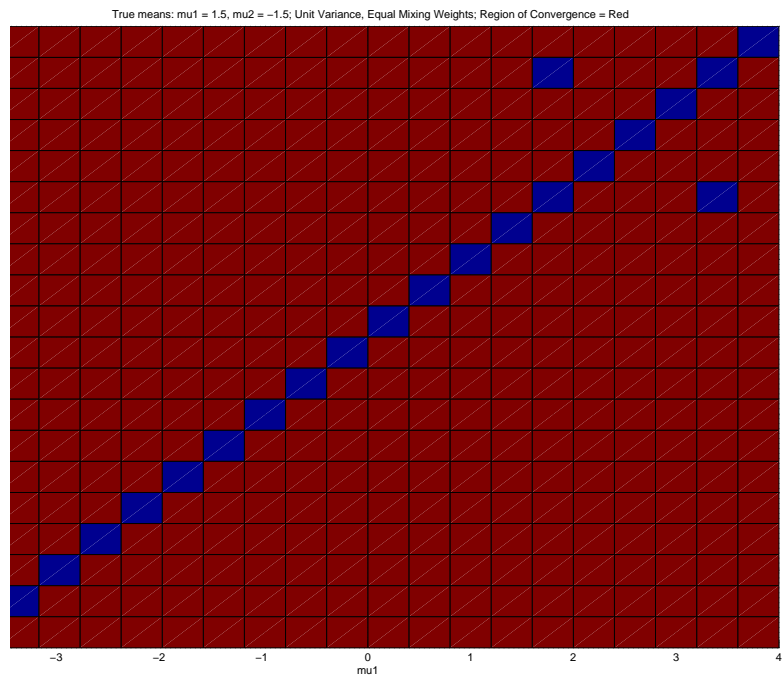
We also noted some interesting behavior that wasn't really expected. The graph with 7% overlap shows some initialization points where the performance of EM is bad. It was found by decreasing the granularity of the grid that there exist quite a few bad initialization points; however, the area of convergence still remains large. This number of bad initialization points is expected to increase in higher dimensions. However, we were unable to come up with obvious explanations for the existence of such points. It would be interesting to see some further analysis on this point.

## References

- [1] J. Bilmes. *A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*. Technical Report, University of Berkeley, ICSI-TR-97-021, 1997.
- [2] R. Kannan and S. Vempala. *Spectral Algorithms*. Unpublished notes. <http://www.cc.gatech.edu/~vempala/spectral/course.html>

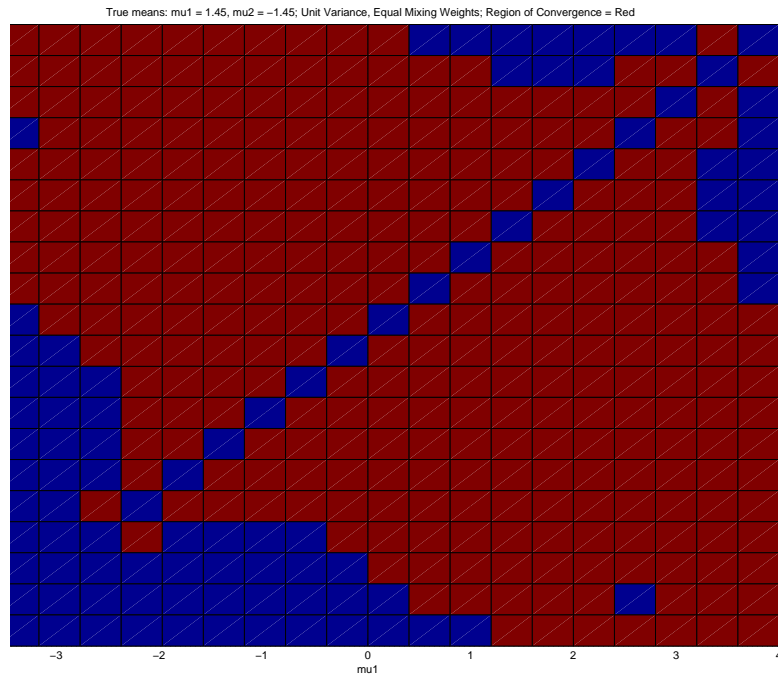


(a) 2% Overlap

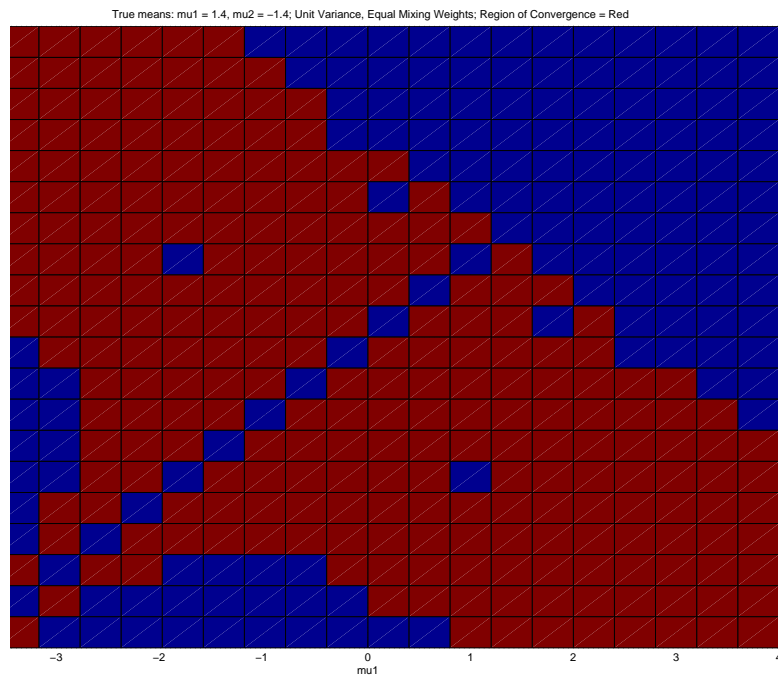


(b) 7% Overlap

Figure 1: Empirical Results for Convergence: Low Overlap

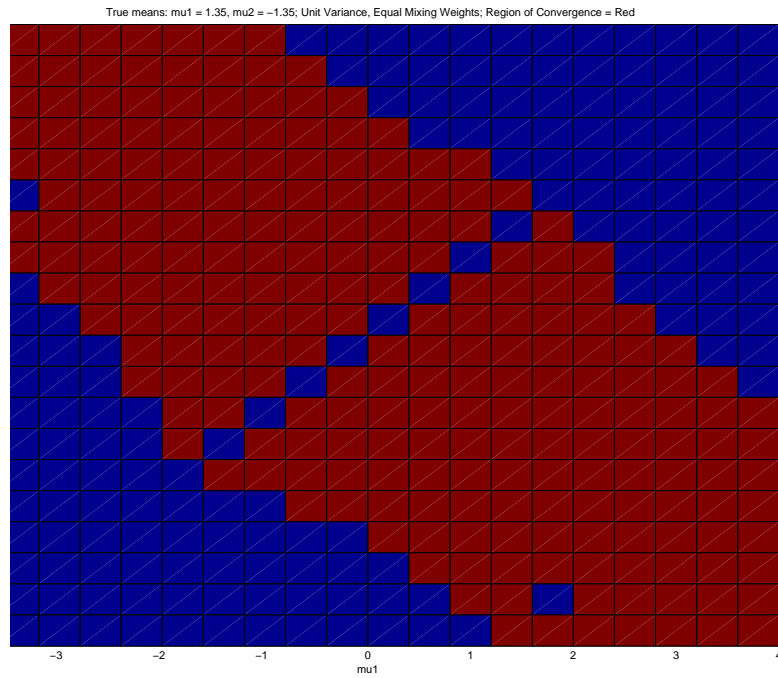


(a) 7.5% Overlap

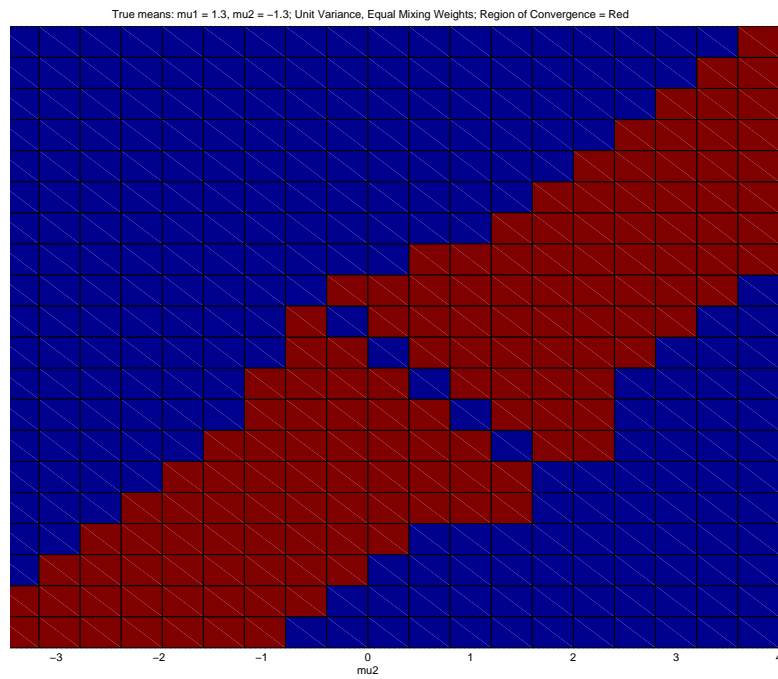


(b) 8% Overlap

Figure 2: Empirical Results for Convergence: Medium Overlap

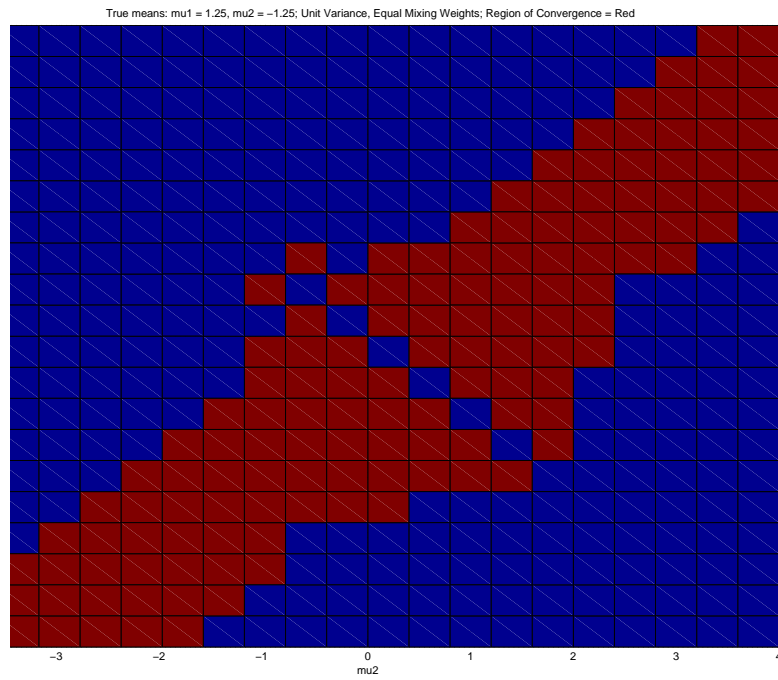


(a) 9% Overlap

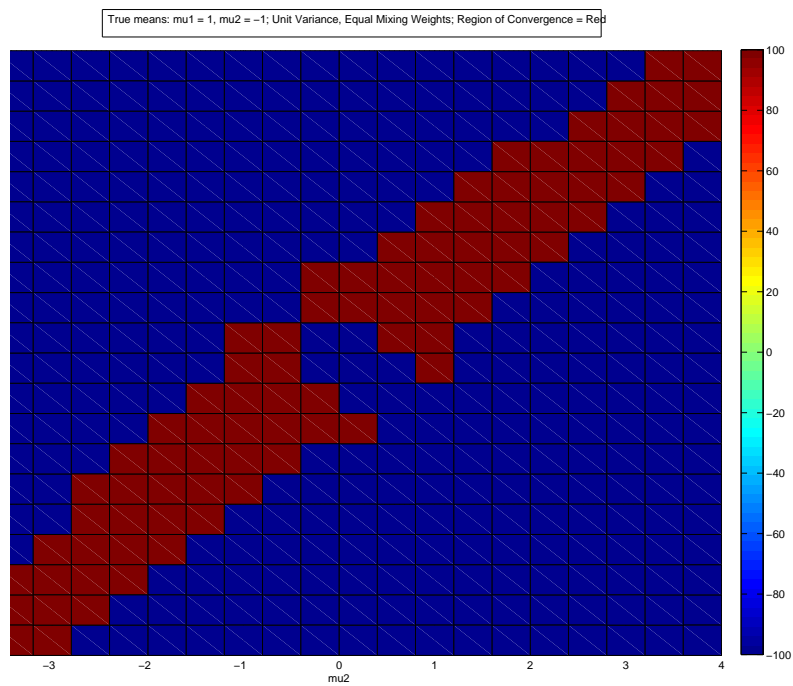


(b) 10% Overlap

Figure 3: Empirical Results for Convergence: High Overlap



(a) 11% Overlap



(b) 15% Overlap

Figure 4: Empirical Results for Convergence: Higher Overlap