Statistical Machine Learning Notes 12

Expectation Maximization and Mixtures of Gaussians

Instructor: Justin Domke

Contents

1 Introduction 1

2 Preliminary: Jensen's Inequality 2

3 Expectation Maximization in General 2

3.1 M-Step 3

3.2 E-Step 4

4 Convergence 5

5 Mixtures of Gaussians 6

5.1 E-Step 7

5.2 M-Step 7

6 Examples 8

6.1 2 clusters 9

6.2 3 clusters 12

6.3 5 clusters 15

1 Introduction

Suppose we have decided to fit a maximum-likelihood model. How should we do it? There are three major approaches:

1. Closed-form solution.
2. Apply a nonlinear optimization tool. In general, there is no guarantee of convergence to the global optima, though there are such guarantees in many special cases.
3. Expectation Maximization

2 Preliminary: Jensen's Inequality

For any distribution $q(\mathbf{x})$, and for any concave function $f(\mathbf{x})$

$$f\big(\mathbf{x}\sum_{\mathbf{x}} q(\mathbf{x})\big) \geq \sum_{\mathbf{x}} q(\mathbf{x}) f(\mathbf{x}).$$

Intuitively speaking, this says if we compute f at the expected value of $\mathbf{x}$, this will be at least as high as the expected value of f.

3 Expectation Maximization in General

Consider fitting some model $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$, but we only have data for $\mathbf{x}$. (We will return to this issue of when this might come up later on.) What to do? A reasonable approach is to maximize the log-likelihood of $\mathbf{x}$ alone. Here, we assume that $\mathbf{z}$ is discrete.

$$\sum_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}}; \boldsymbol{\theta}) = \sum_{\hat{\mathbf{x}}} \log \sum_{\mathbf{z}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) \tag{3.1}$$

How to maximize this with respect to $\boldsymbol{\theta}$? One option is, well, just to maximize it. For a given $\boldsymbol{\theta}$, we can compute the right-hand side of Eq. 3.1. So, there is nothing to prevent us from just taking the gradient of this expression, and running gradient descent (or whatever). There is absolutely nothing wrong with this approach. However, in certain cases, it can get quite complicated

Now, consider any distribution $r(\mathbf{z}|\mathbf{x})$. We have

$$\log p(\mathbf{x}; \boldsymbol{\theta}) = \log\big(\sum_{\mathbf{z}} r(\mathbf{z}|\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}))}{r(\mathbf{z}|\mathbf{x})}\big)$$

Applying Jensen's inequality, we have

$$\log p(\mathbf{x};\boldsymbol{\theta}) \geq \sum_{\mathbf{z}} r(\mathbf{z}|\mathbf{x}) \log \frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}))}{r(\mathbf{z}|\mathbf{x})}.$$

Putting this all together, we can create a lower-bound on the log-likelihood

$$l(\boldsymbol{\theta}) = \sum_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}};\boldsymbol{\theta}) \geq q(\boldsymbol{\theta}, r) = \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \log \frac{p(\hat{\mathbf{x}},\mathbf{z};\boldsymbol{\theta}))}{r(\mathbf{z}|\hat{\mathbf{x}})}.$$

The basic idea of EM is very simple. Instead of directly maximizing l, maximize the lower bound q. One quick way to do this would be to just pick some random distribution r, and then maximize with respect to $\boldsymbol{\theta}$. This works, but we can do better. Why not also maximize with respect to the distribution r? Thus, EM alternates with respect to two steps:

- Maximize q with respect to r. This is called the "Expectation" (E) step.
- Maximize q with respect to $\boldsymbol{\theta}$. This is called the "Maximization" (M) step.

The names for these steps may seem somewhat confusing at the moment, but will become clear later on. Before continuing on to the

Abstract Expectation Maximization

- Iterate Until Convergence:
  - For all $\hat{\mathbf{x}}$, $r(\mathbf{z}|\hat{\mathbf{x}}) \leftarrow p(z|\hat{\mathbf{x}};\boldsymbol{\theta})$ (E-step)
  - $\boldsymbol{\theta} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}},\mathbf{z};\boldsymbol{\theta})$ (M-step)

Let's look out how these rules emerge, and try to understand the properties of EM. First, let us consider the the M-step.

3.1 M-Step

So far, we haven't really discussed the point of EM. We introduce a lower-bound on l, then alternate between two different optimization step. So what? Why should this help us? Consider the M-Step.

$$\boldsymbol{\theta} \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}},\mathbf{z};\boldsymbol{\theta}).$$

This is essentially just a regular maximum likelihood problem, just weighted by $r(\mathbf{z}|\hat{\mathbf{x}})$. This can be very convenient. As we will see below, in some circumstances where maximizing l with respect to $\boldsymbol{\theta}$ would entail a difficult numerical optimization, the above optimization can be done in closed-form. This increased convenience of this is the real gain of our odd lower-bounding strategy.

3.2 E-Step

On the other hand, we appear to pay a price for the convenience of the M-step. Namely, we aren't fitting l any more, just a lower-bound on l. How steep is this price? Well, let's consider the maximization with respect to r. What we need to do is, independently for all $\hat{\mathbf{x}}$, maximize

$$\sum_z r(\mathbf{z}|\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) - \sum_z r(\mathbf{z}|\hat{\mathbf{x}}) \log r(\mathbf{z}|\hat{\mathbf{x}})$$

Theorem 1. *This is maximized by setting*

$$r(\mathbf{z}|\hat{\mathbf{x}}) = p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta}).$$

Proof. Our problem is

$$\begin{array}{ll} \max\limits_{r(\mathbf{z}|\hat{\mathbf{x}})} & \sum\limits_z r(\mathbf{z}|\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) - \sum\limits_z r(\mathbf{z}|\hat{\mathbf{x}}) \log r(\mathbf{z}|\hat{\mathbf{x}}) \\ \text{s.t.} & r(\mathbf{z}|\hat{\mathbf{x}}) \geq 0 \\ & \sum\limits_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) = 1. \end{array}$$

Form the Lagrangian

$$\mathcal{L} = \sum_z r(\mathbf{z}|\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) - \sum_z r(\mathbf{z}|\hat{\mathbf{x}}) \log r(\mathbf{z}|\hat{\mathbf{x}}) - \sum_{\mathbf{z}} \lambda_z r(\mathbf{z}|\hat{\mathbf{x}}) + \nu\big(1 - \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}})\big)$$

For a given set of Lagrange multipliers $\{\lambda_{\mathbf{z}}\}$, ν we can find the maximizing $r(\mathbf{z}|\hat{\mathbf{x}})$ by

$$\frac{d\mathcal{L}}{dr(\mathbf{z}|\hat{\mathbf{x}})} = \log p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) - \log r(\mathbf{z}|\hat{\mathbf{x}}) - 1 - \lambda_z - \nu = 0$$

From this it follows that

$$r(\mathbf{z}|\hat{\mathbf{x}}) \propto p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}).$$

Now the only way that this can be true while also having r be a normalized distribution is that it is the conditional distribution of p.

$$r(\mathbf{z}|\hat{\mathbf{x}}) = p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta})$$

□

So this proof explains the meaning of "expectation step". All that we need to do, in order to maximize with respect to r, is set each conditional distribution $r(\mathbf{z}|\hat{\mathbf{x}})$ to be the conditional distribution $p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta})$.

4 Convergence

How well should EM work compared to a traditional maximization of the likelihood in Eq. 3.1? To answer this, consider the value of q at the completion of the E-step.

$$\begin{aligned}
q(\boldsymbol{\theta}, r) &= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \log \frac{p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})}{r(\mathbf{z}|\hat{\mathbf{x}})} \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta}) \log \frac{p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta})} \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta}) \log \frac{p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta}) p(\hat{\mathbf{x}}; \boldsymbol{\theta})\big)}{p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta})} \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta}) \log p(\hat{\mathbf{x}}; \boldsymbol{\theta}) \\
&= \sum_{\hat{\mathbf{x}}} \log p(\hat{\mathbf{x}}; \boldsymbol{\theta}) \\
&= l(\boldsymbol{\theta})
\end{aligned}$$

After the completion of the E-step our lower bound on the likelihood is tight. This is fantastic news. Suppose that EM converges to some r and $\boldsymbol{\theta}$. By the definition of convergence, $\boldsymbol{\theta}$ must be a (local) maxima of q. However, from the tightness of the bound, we know that if it was possible to locally improve l, it would also be possible[1] to locally improve q. Thus, at convergence, we have also achieved a local maxima of l. Thus, we should not expect to pay a penalty in accuracy from the convenience of EM.

On the other hand, note that Expectation-Maximization does not remove the issue of local minima. Both in theory and in practice, local minima remain a real problem.

5 Mixtures of Gaussians

The above is all very abstract. Let's remember the supposed advantages of EM. We have claimed that there are some circumstances where a direct maximization of $l(\boldsymbol{\theta})$ would be "hard" (in implementation, running time, or otherwise), while performing each of the steps in EM is "easy". This would all be more credible if we actually saw at least one such case. The example we consider here is by far the most common example of EM, namely for learning mixtures of Gaussians.

A mixture of Gaussians is defined by

$$p(\mathbf{x}) = \sum_z \pi_z \mathcal{N}(\mathbf{x}; \mu_z, \Sigma_z).$$

Where $\pi_z \geq 0$ and $\sum_z \pi_z = 1$.

The basic argument for mixtures of Gaussians is that they can approximate almost anything.

Now, once again, we could directly maximize the log likelihood of a mixture of Gaussians. It would be a little painful, but certainly doable. The following application of EM gives a simpler algorithm.

$$\begin{aligned}
\frac{dq}{d\boldsymbol{\theta}} &= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \frac{d}{d\boldsymbol{\theta}} \log \frac{p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})}{r(\mathbf{z}|\hat{\mathbf{x}})} \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \frac{d}{d\boldsymbol{\theta}} \log p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} r(\mathbf{z}|\hat{\mathbf{x}}) \frac{1}{p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})} \frac{d}{d\boldsymbol{\theta}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} \frac{p(\mathbf{z}|\hat{\mathbf{x}}; \boldsymbol{\theta})}{p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})} \frac{d}{d\boldsymbol{\theta}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} \frac{1}{p(\hat{\mathbf{x}}; \boldsymbol{\theta})} \frac{d}{d\boldsymbol{\theta}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})
\end{aligned}$$

$$\begin{aligned}
\frac{dl}{d\boldsymbol{\theta}} &= \sum_{\hat{\mathbf{x}}} \frac{d}{d\boldsymbol{\theta}} \log \sum_{\mathbf{z}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) \\
&= \sum_{\hat{\mathbf{x}}} \frac{1}{\sum_{\mathbf{z}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})} \frac{d}{d\boldsymbol{\theta}} \sum_{\mathbf{z}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta}) \\
&= \sum_{\hat{\mathbf{x}}} \sum_{\mathbf{z}} \frac{1}{p(\hat{\mathbf{x}}; \boldsymbol{\theta})} \frac{d}{d\boldsymbol{\theta}} p(\hat{\mathbf{x}}, \mathbf{z}; \boldsymbol{\theta})
\end{aligned}$$

---

[1] To be precise, it is possible to show that not only will $q(\boldsymbol{\theta}, r) = l(\boldsymbol{\theta})$ at convergence, but also that $dq(\boldsymbol{\theta}, r)/d\boldsymbol{\theta} = dl(\boldsymbol{\theta})/d\boldsymbol{\theta}$.

$$p(z) = \pi_z$$

$$p(\mathbf{x}|z) = \mathcal{N}(\mathbf{x}; \mu_z, \Sigma_z)$$

$$p(\mathbf{x}, z) = \pi_z \mathcal{N}(\mathbf{x}; \mu_z, \Sigma_z)$$

This is quite a clever idea. We do not "really" have hidden variables z, but we introduce these for the convenience of applying the EM algorithm.

5.1 E-Step

$$r(z|\hat{\mathbf{x}}) \leftarrow p(z|\hat{\mathbf{x}}) = \frac{p(\hat{\mathbf{x}}, z)}{p(\hat{\mathbf{x}})} = \frac{\pi_z \mathcal{N}(\mathbf{x}; \mu_z, \Sigma_z)}{\sum_{z'} \pi_{z'} \mathcal{N}(\mathbf{x}; \mu_{z'}, \Sigma_{z'})}$$

This is not hard at all and can simply be computed in closed-form.

5.2 M-Step

Do the maximization

$$\begin{aligned} &\arg \max_{\{\mu_z\},\{\Sigma_z\}} \sum_{\hat{\mathbf{x}}} \sum_{z} r(z|\hat{\mathbf{x}}) \log p(\hat{\mathbf{x}}, z; \boldsymbol{\theta}). \\ =&\arg \max_{\{\mu_z\},\{\Sigma_z\}} \sum_{\hat{\mathbf{x}}} \sum_{z} r(z|\hat{\mathbf{x}}) \log\big(\pi_z \mathcal{N}(\hat{\mathbf{x}}; \mu_z, \Sigma_z)\big) \\ =&\arg \max_{\{\mu_z\},\{\Sigma_z\}} \sum_{\hat{\mathbf{x}}} \sum_{z} r(z|\hat{\mathbf{x}}) \log \mathcal{N}(\hat{\mathbf{x}}; \mu_z, \Sigma_z) + \sum_{\hat{\mathbf{x}}} \sum_{z} r(z|\hat{\mathbf{x}}) \log \pi_z \end{aligned}$$

Maximizing with respect to π_z is quite easy[2]. This is done by setting

---

[2] Consider the Lagrangian

$$\mathcal{L} = \sum_{\hat{\mathbf{x}}} \sum_{z} r(z|\hat{\mathbf{x}}) \log \pi_z + \lambda(1 - \sum_{z} \pi_z)$$

Here, we aren't explicitly enforcing that $\pi_z \geq 0$. As we see below, however, we can get away with this, since the logarithm of π_z becomes infinitely large as $\pi_z \rightarrow 0$. For some fixed Lagrange multiplier λ, we can find the best weights π_z by

$$\pi_z = \frac{1}{\sum_{\hat{\mathbf{x}}} \sum_z r(z|\hat{\mathbf{x}})} \sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}}).$$

This can be accomplished by doing (separately for each z)

$$\arg\max_{\boldsymbol{\mu}_z, \Sigma_z} \sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}}) \log \mathcal{N}(\hat{\mathbf{x}}; \mu_z, \Sigma_z)$$

It turns out that this can be done by setting

$$\boldsymbol{\mu}_z \leftarrow \frac{1}{\sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}})} \sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}})\hat{\mathbf{x}}$$

$$\Sigma_z \leftarrow \frac{1}{\sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}})} \sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}})(\hat{\mathbf{x}} - \boldsymbol{\mu}_z)(\hat{\mathbf{x}} - \boldsymbol{\mu}_z)^T.$$

Note the similarity of these updates to a fitting a normal Gaussian distribution. The only difference here is that we compute the weighted mean and the weighted covariance matrix.

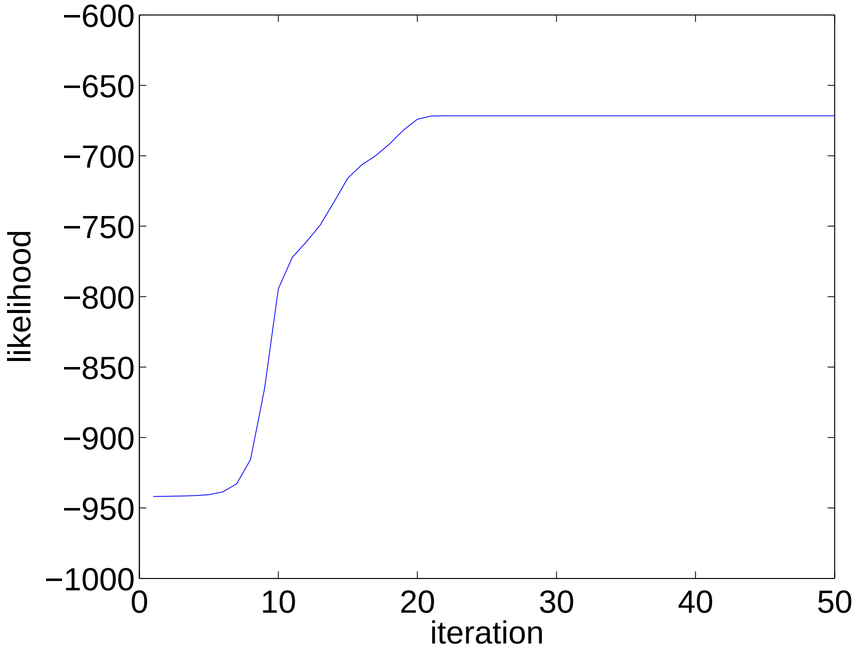
6 Examples

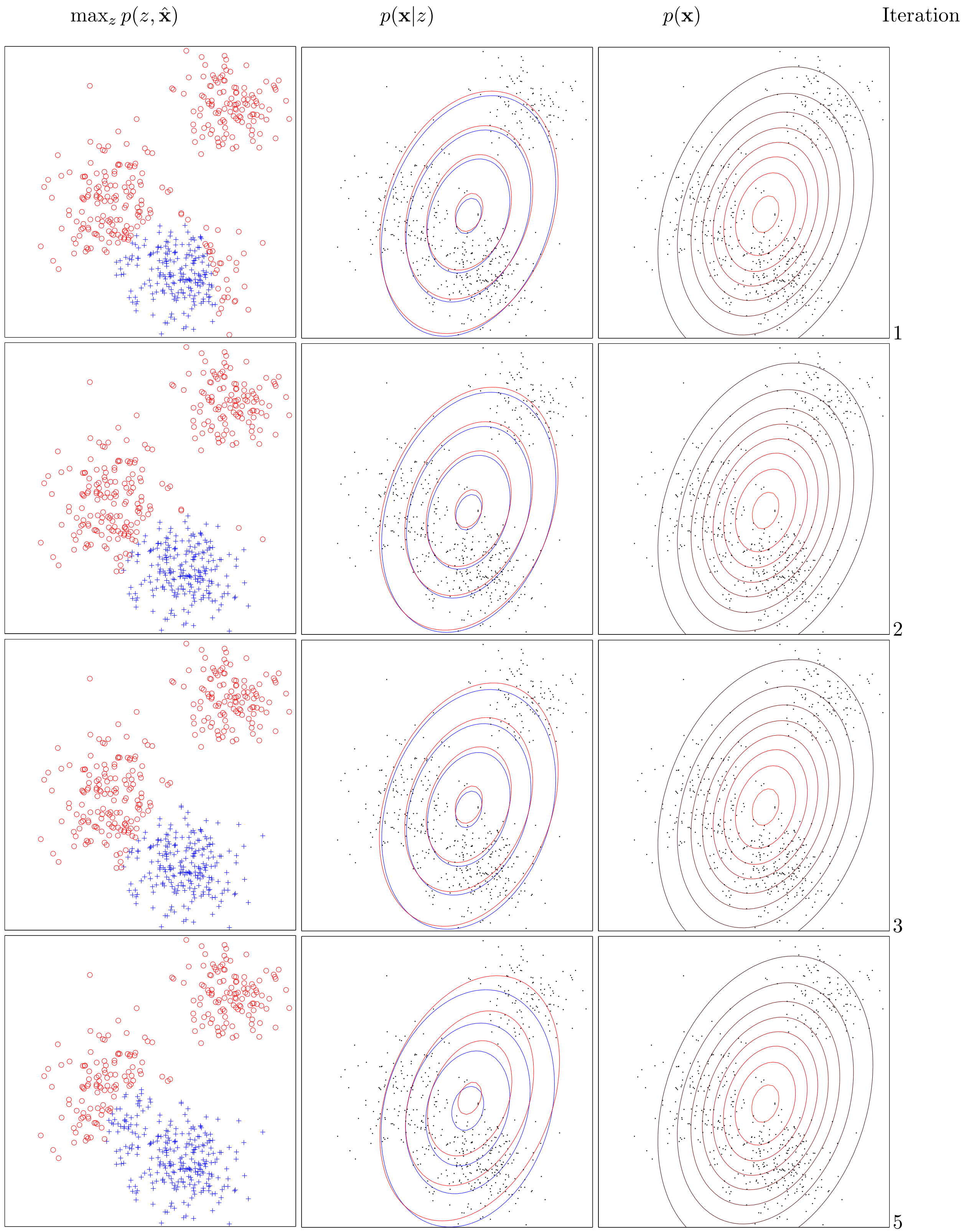
$$\frac{d\mathcal{L}}{d\pi_z} = \sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}}) \frac{1}{\pi_z} - \lambda = 0.$$

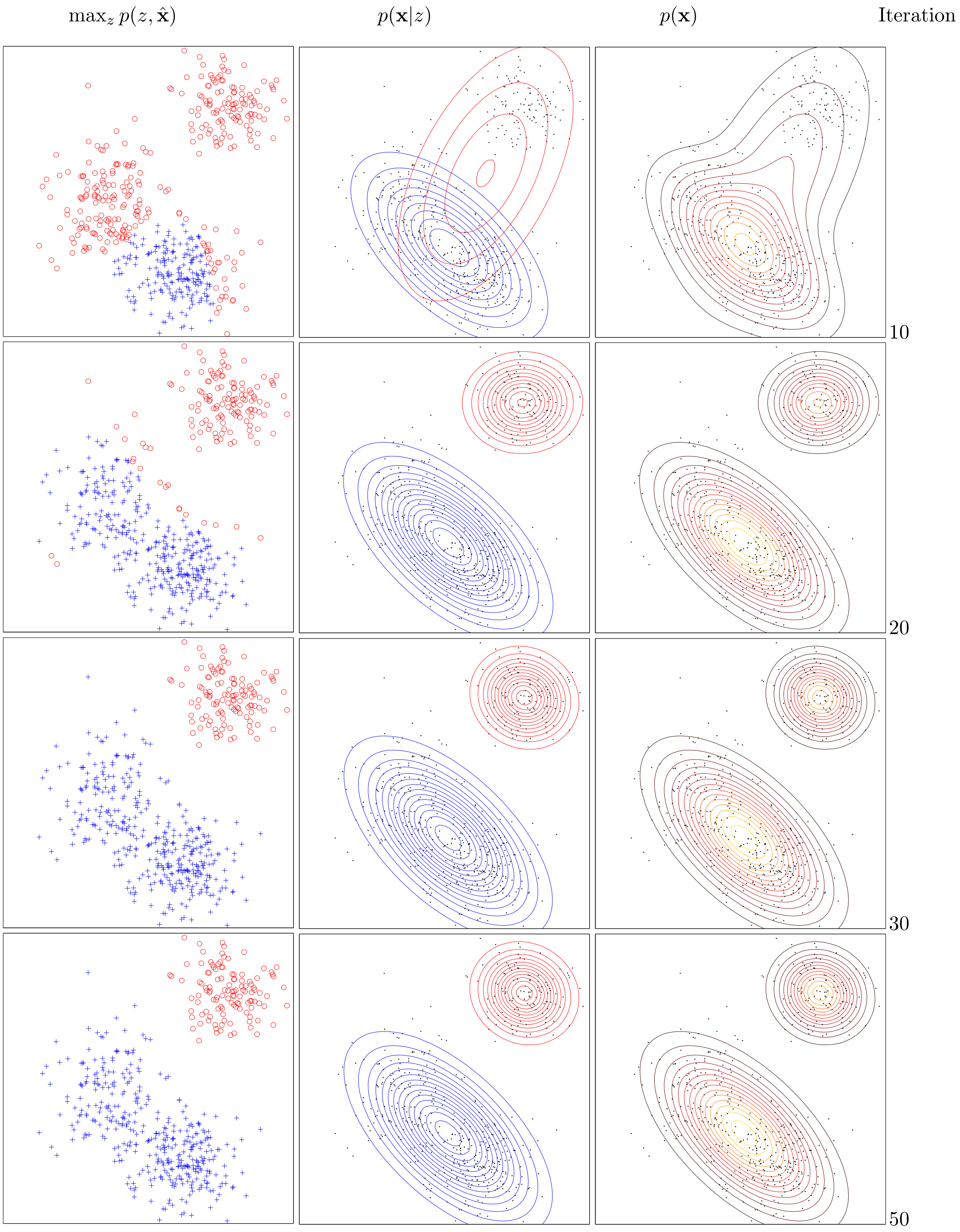
This results in the condition that $\pi_z \propto \sum_{\hat{\mathbf{x}}} r(z|\hat{\mathbf{x}})$. However, since we know that $\sum_z \pi_z = 1$, we can just normalize these scores and recover the answer.

6.1 2 clusters

In this first experiment, we use only two Gaussians. We can see that little happens in the first few iterations, as the two Gaussians greatly overlap. As they start to cover different areas around 10 iterations, the likelihood rapidly increases.

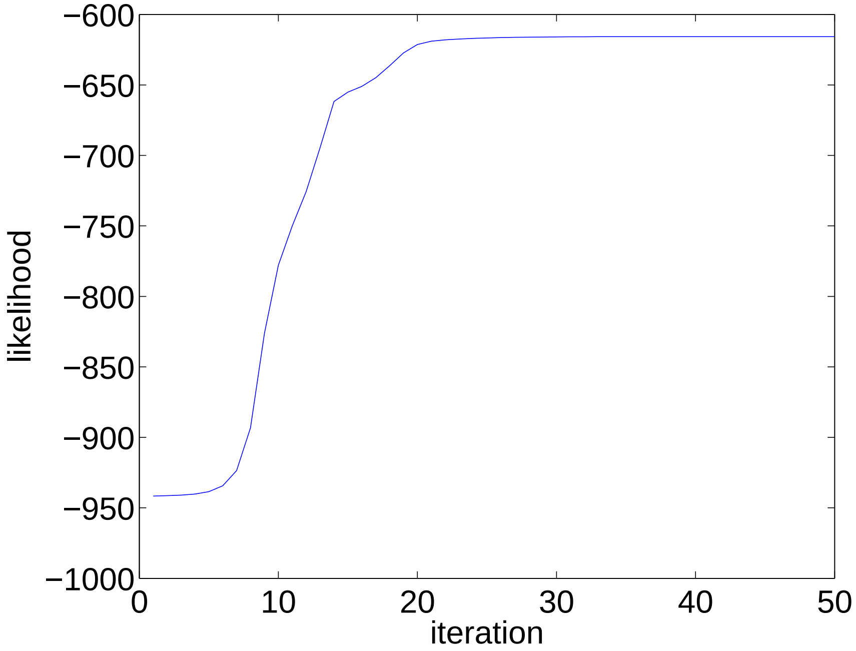


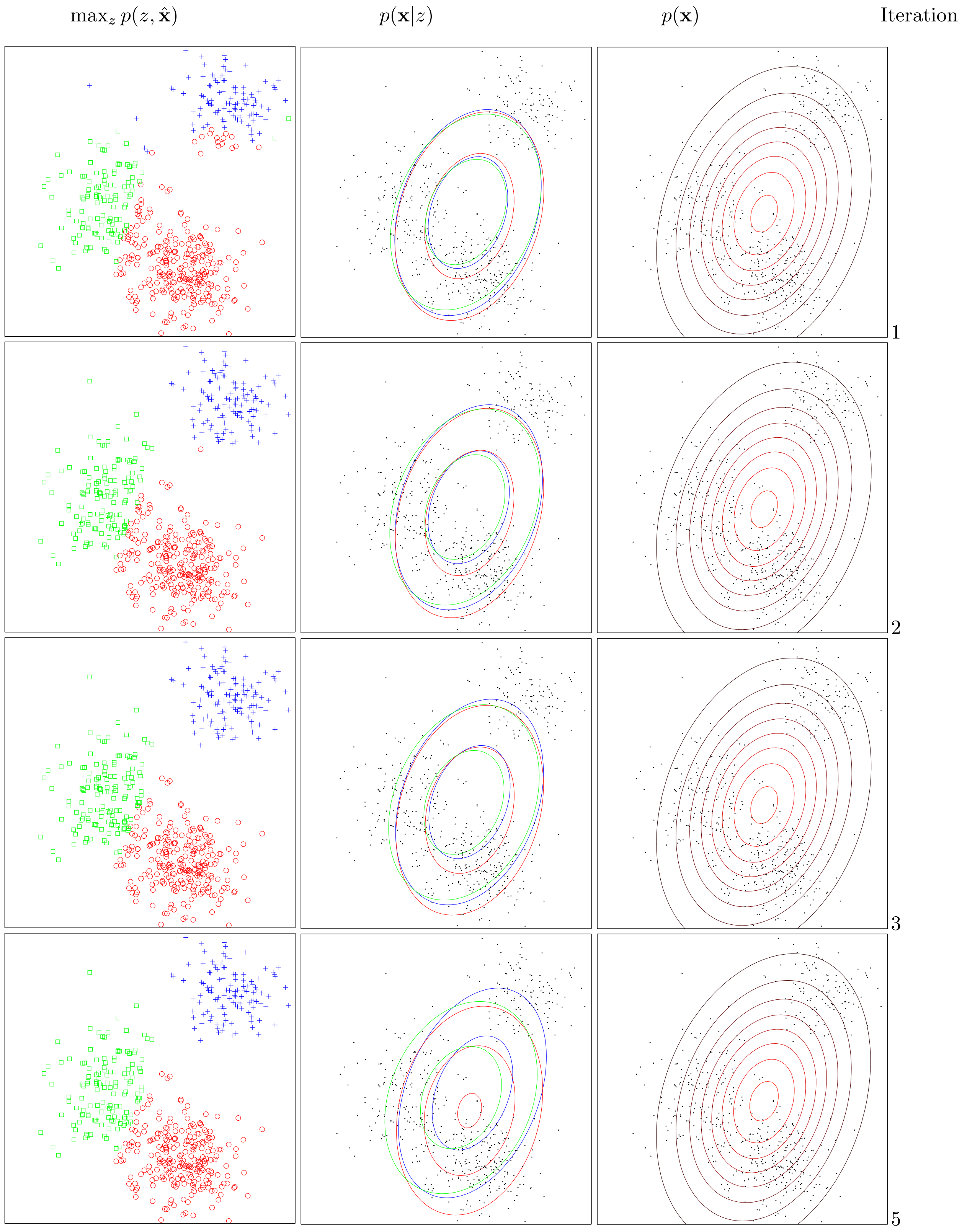

$\max_z p(z, \hat{\mathbf{x}})$
$p(\mathbf{x}|z)$
$p(\mathbf{x})$
Iteration
1
2
3
5

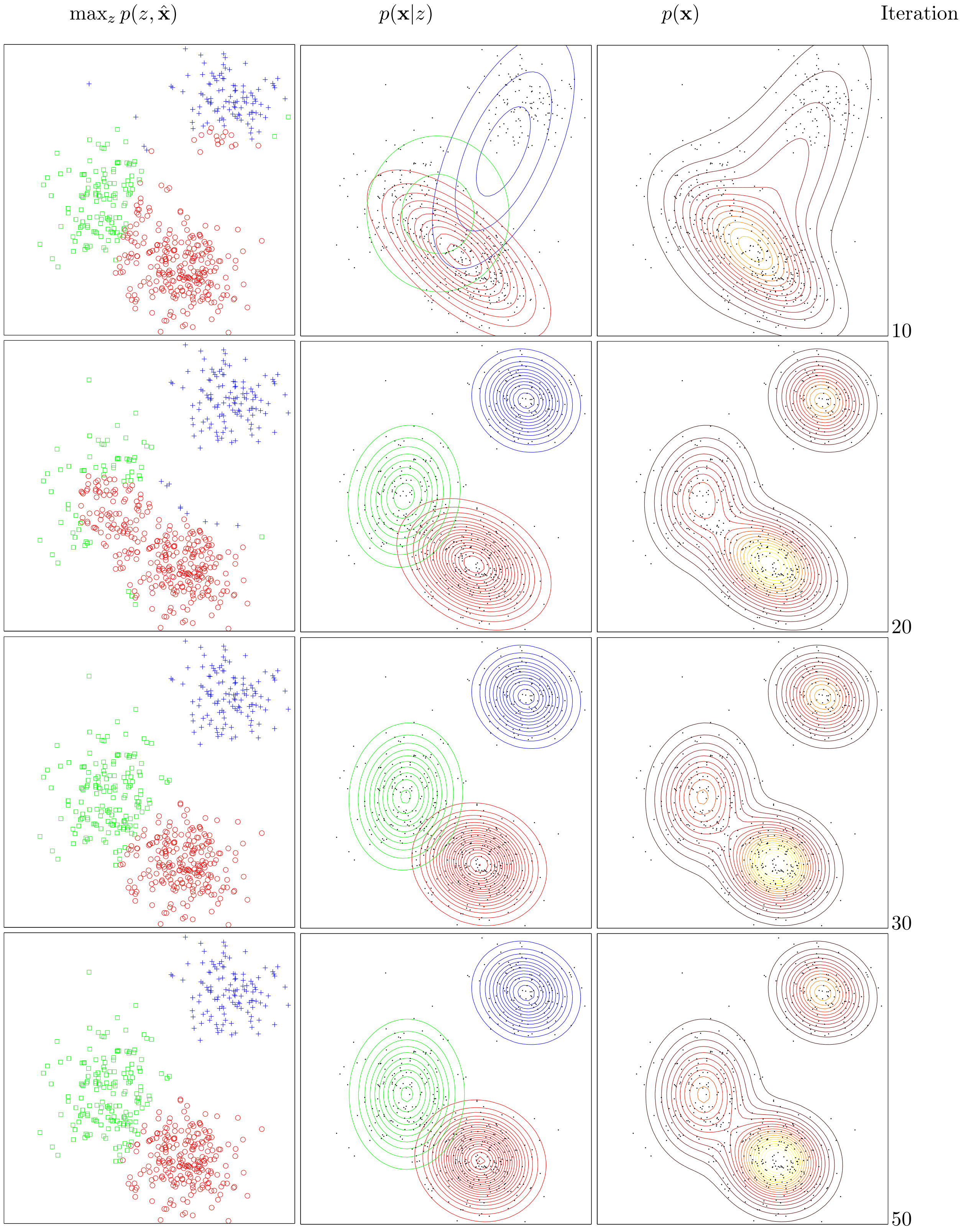

$\max_z p(z, \hat{\mathbf{x}})$
$p(\mathbf{x}|z)$
$p(\mathbf{x})$
Iteration
10
20
30
50

6.2 3 clusters

Finally, the experiment is repeated with a mixture of 3 Gaussians. Here, we see a major increase in the likelihood of the final mixture.

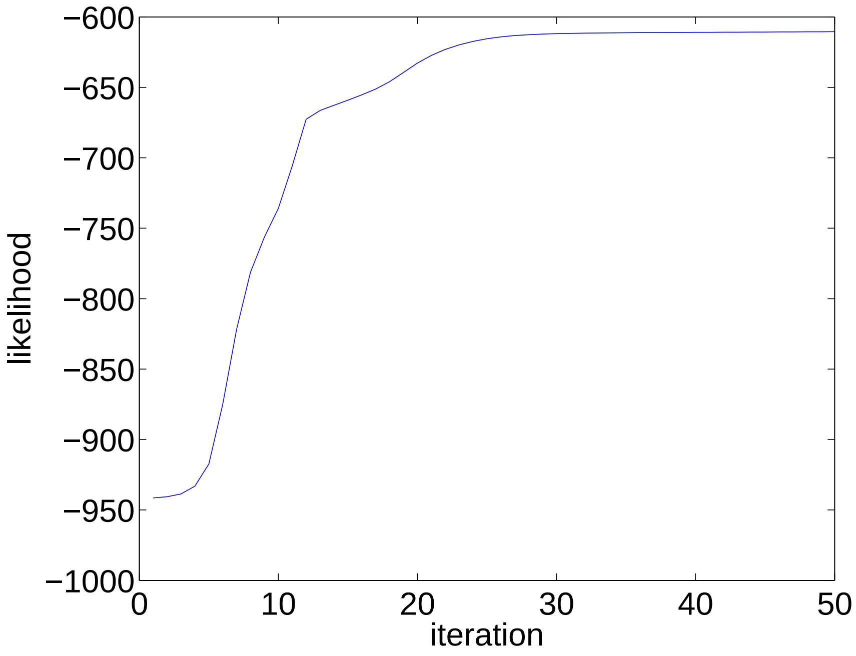


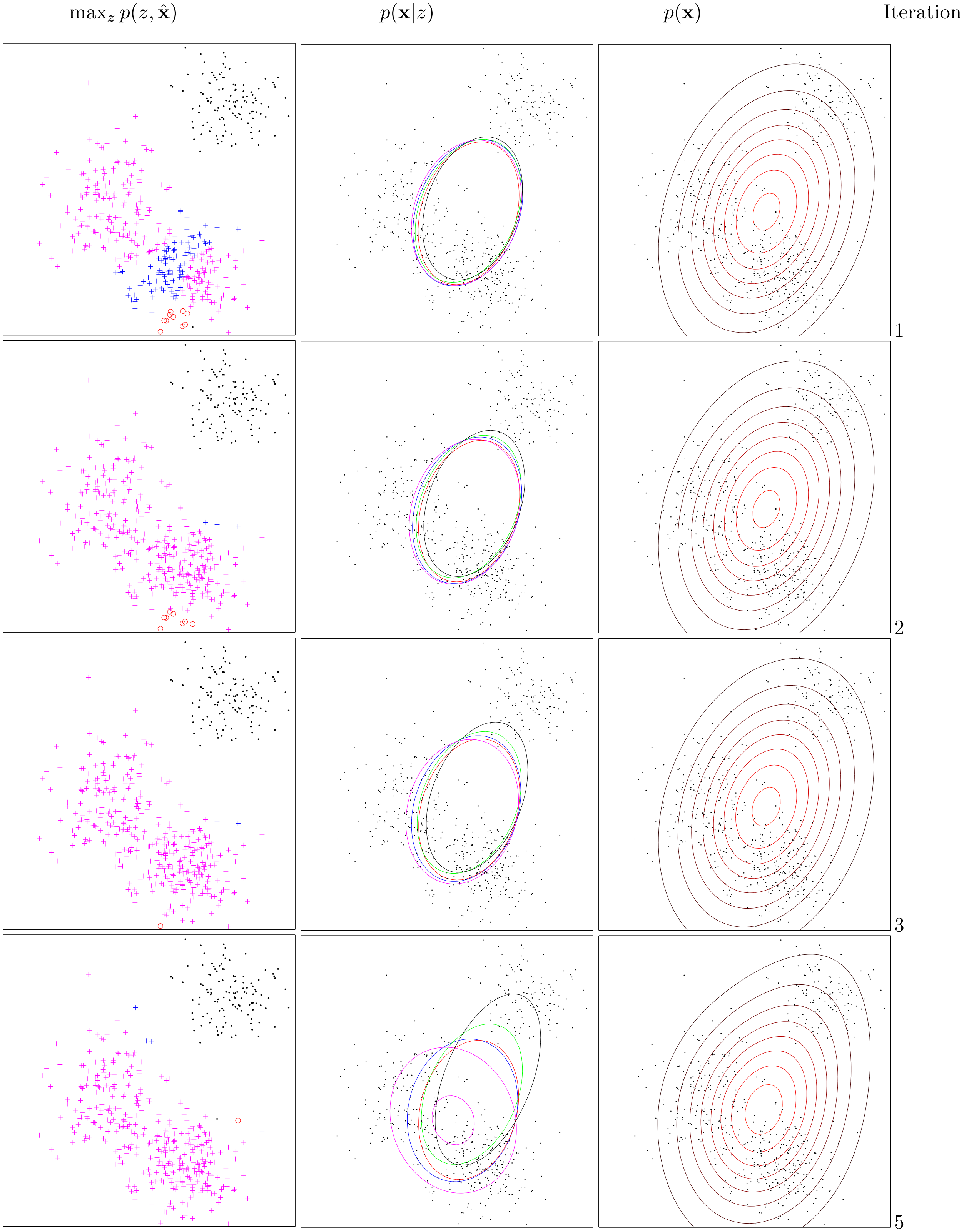

$\max_z p(z, \hat{\mathbf{x}})$
$p(\mathbf{x}|z)$
$p(\mathbf{x})$
Iteration
1
2
3
5

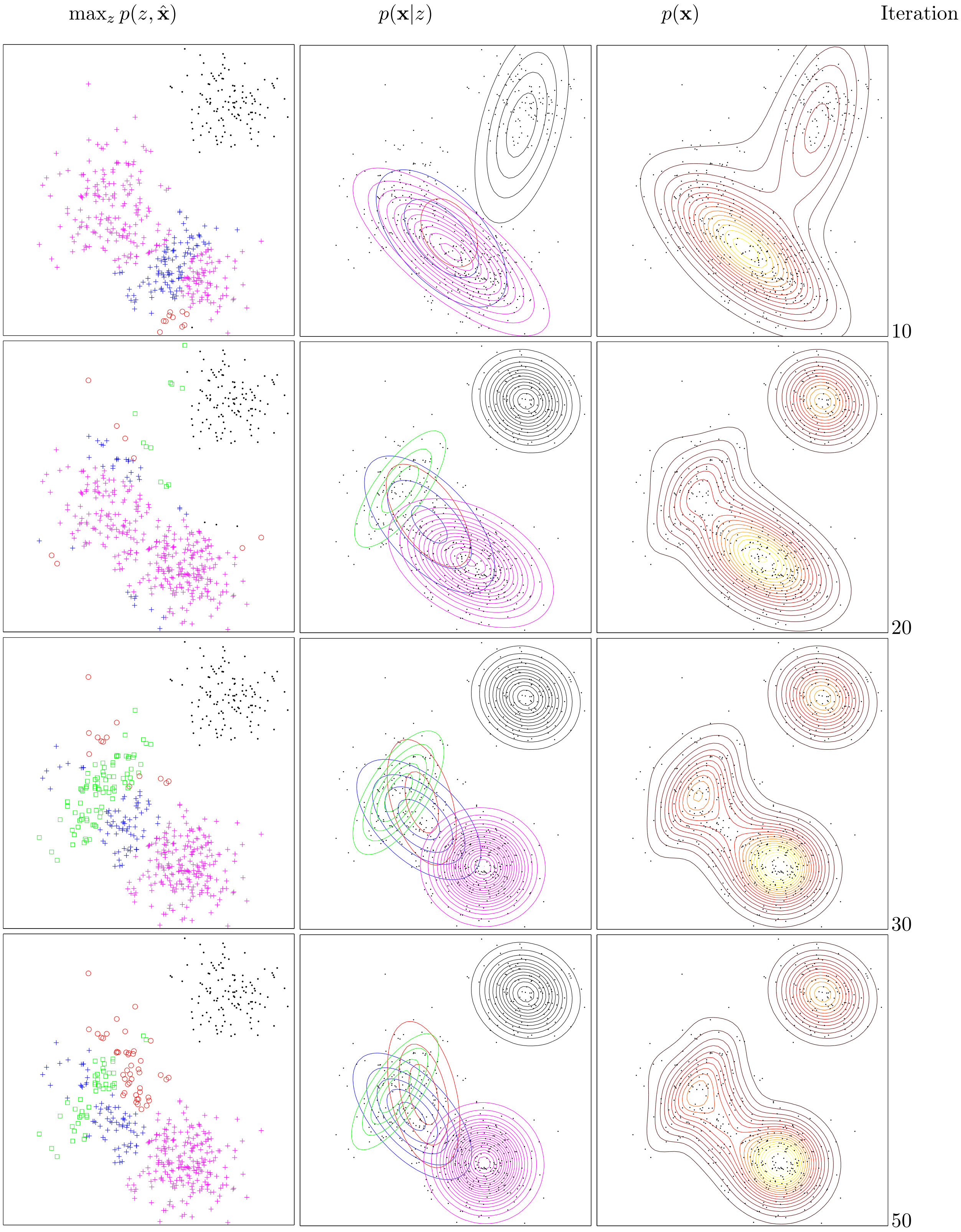

$\max_z p(z, \hat{\mathbf{x}})$
$p(\mathbf{x}|z)$
$p(\mathbf{x})$
Iteration
10
20
30
50

6.3 5 clusters

Finally, the experiment is repeated with a mixture of 5 Gaussians. In this case, only a small increase in the likelihood obtains over with mixture of 3 Gaussians.




$\max_z p(z, \hat{\mathbf{x}})$
$p(\mathbf{x}|z)$
$p(\mathbf{x})$
Iteration
1
2
3
5


$\max_z p(z, \hat{\mathbf{x}})$
$p(\mathbf{x}|z)$
$p(\mathbf{x})$
Iteration
10
20
30
50