Preliminaries

Given: a set of observations $\mathcal{D} = \{x_i\}_{i=1}^n, x_i \in \mathcal{X}$

Objective: find a model $\hat{f} \in \mathcal{F}$ that models the phenomenon well

Requirements:

(i) the ability to generalize well

(ii) the ability to incorporate prior knowledge and assumptions

(iii) scalability

Terminology through an example: $\mathcal{D} = \{3.1, 2.4, -1.1, 0.1\}$

What is the data generator?

$\mathcal{F} = \text{Gaussian}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$
STATISTICAL FRAMEWORK

Model inference: Observations + Knowledge and Assumptions + Optimization
**MAXIMUM A POSTERIORI (MAP) INFEERENCE**

**Idea:**

$$f_{\text{MAP}} = \arg \max_{f \in \mathcal{F}} \{p(f | \mathcal{D})\},$$

where $p(f | \mathcal{D})$ is called the posterior distribution.

**How do we calculate it?**

$$p(f | \mathcal{D}) = \frac{p(\mathcal{D} | f) \cdot p(f)}{p(\mathcal{D})}$$

where $p(\mathcal{D} | f) = \text{likelihood}$, $p(f) = \text{prior}$, and $p(\mathcal{D}) = \text{data distribution.}$
MAXIMUM A POSTERIORI (MAP) INFERENCE

Finding the data distribution:

\[ p(\mathcal{D}) = \begin{cases} 
\sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\
\int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous}
\end{cases} \]

We can now simplify the process if we observe that

\[ p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \propto p(\mathcal{D}|f) \cdot p(f) \]
MAXIMUM LIKELIHOOD (ML) INFERENCE

Express the posterior distribution as

\[ p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f) \cdot p(f)}{p(\mathcal{D})} \]

\[ \propto p(\mathcal{D}|f) \cdot p(f) \]

Now, ignore \( p(f) \) to get

\[ f_{ML} = \underset{f \in \mathcal{F}}{\text{arg max}} \{ p(\mathcal{D}|f) \} \]

There are technical problems with this approach, but also reasons to use it.
Example: \( \mathcal{D} = \{2, 5, 9, 5, 4, 8\} \) is an i.i.d. sample from \( \text{Poisson}(\lambda), \lambda \in \mathbb{R}^+ \)

Find \( \lambda \)

Solution: Poisson probability mass function is \( p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \)

\[ \lambda_{ML} = \arg \max_{\lambda \in (0, \infty)} \{ p(\mathcal{D}|\lambda) \}. \]

Likelihood: \( p(\mathcal{D}|\lambda) = p(\{x_i\}_{i=1}^n | \lambda) \)

\[ = \prod_{i=1}^n p(x_i | \lambda) \]

\[ = \frac{\lambda \sum_{i=1}^n x_i \cdot e^{-n\lambda}}{\prod_{i=1}^n x_i!}. \]
EXAMPLE: ML INFERENCE

Likelihood: $p(D|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$

Log-likelihood: $ll(D, \lambda) = \ln \lambda \sum_{i=1}^{n} x_i - n \lambda - \sum_{i=1}^{n} \ln (x_i!)$

Optimization:

$$\frac{\partial ll(D, \lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

$= 0$

Solution:

$$\lambda_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$= 5.5$

MAP and ML estimates are called the point estimates.
**Example: MAP Inference**

**Example:** \( \mathcal{D} = \{2, 5, 9, 5, 4, 8\} \) is i.i.d. sample from \( \text{Poisson}(\lambda) \), \( \lambda \in \mathbb{R}^+ \)

Assume \( \lambda \) is taken from \( \Gamma(x|k, \theta) \) with parameters \( k = 3 \) and \( \theta = 1 \)

Find \( \lambda \)

**Solution:** Poisson: \( p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \)

Gamma: \( \Gamma(x|k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \), where \( x > 0, k > 0, \text{ and } \theta > 0 \).

**Likelihood:** \( p(\mathcal{D}|\lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!} \)

**Prior:** \( p(\lambda) = \frac{\lambda^{k-1} e^{-\frac{\lambda}{\theta}}}{\theta^k \Gamma(k)} \).
**Example: MAP Inference**

Log-likelihood:

\[
\ln p(\lambda|\mathcal{D}) \propto \ln p(\mathcal{D}|\lambda) + \ln p(\lambda) \\
= \ln \lambda (k - 1 + \sum_{i=1}^{n} x_i) - \lambda (n + \frac{1}{\theta}) - \sum_{i=1}^{n} \ln x_i! - k \ln \theta - \ln \Gamma(k)
\]

We now obtain

\[
\lambda_{MAP} = \frac{k - 1 + \sum_{i=1}^{n} x_i}{n + \frac{1}{\theta}} \\
= 5
\]
**Another Example**

**Example:** \( \mathcal{D} = \{x_i\}_{i=1}^n \) is i.i.d. sample from Gaussian(\( \mu, \sigma^2 \))

Find \( \mu \) and \( \sigma \)

**Solution:** Gaussian: \( p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)

\[
\mu_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i \\
\sigma_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\text{ML}})^2.
\]
The KL divergence between two probability distributions $p(x)$ and $q(x)$ is

$$D_{KL}(p||q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx$$

Assume now the data is generated according to some $p(x|\theta_t)$. We estimated it as $p(x|\theta)$.

Let’s look at the KL divergence

$$D_{KL}(p(x|\theta_t)||p(x|\theta)) = \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{p(x|\theta_t)}{p(x|\theta)} \, dx - \mathbb{E} [\log p(x|\theta)]$$

$$= \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta)} \, dx - \int_{-\infty}^{\infty} p(x|\theta_t) \log \frac{1}{p(x|\theta_t)} \, dx.$$
RELATIONSHIP TO KULLBACK-LEIBLER (KL) DIVERGENCE

\[
\frac{1}{n} \sum_{i=1}^{n} \log p(x_i|\theta) \xrightarrow{\text{a.s.}} \mathbb{E}[\log p(x|\theta)]
\]

when \( n \to \infty \).

**Conclusion:**

When \( n \to \infty \), ML estimation implies \( p(x|\theta_{\text{ML}}) = p(x|\theta_i) \)

This usually implies \( \theta_{\text{ML}} = \theta_i \)
CONDITIONAL DISTRIBUTIONS

Given: a set of observations $D = \{(x_i, y_i)\}_{i=1}^{n}, x_i, y_i \in \mathbb{R}$

Assumption: $X$ and $Y$ are random variables and $p(y|x) = \mathcal{N}(\mu = x, \sigma^2)$

Objective: find $\sigma$
**BAYESIAN APPROACH**

**Idea:** consider posterior risk $R$

$$R = \int_{\mathcal{F}} \ell(f, \hat{f}) \cdot p(f|\mathcal{D})df$$

where $\ell(f, \hat{f})$ is some error function.

Assume $\ell(f, \hat{f}) = (f - \hat{f})^2$ and find best $\hat{f}$

$$\frac{\partial}{\partial \hat{f}} R = 2\hat{f} - 2\int_{\mathcal{F}} f \cdot p(f|\mathcal{D})df$$

$$= 0$$
**Bayesian Approach**

Solution:

\[
\hat{f}_B = \int f \cdot p(f|\mathcal{D})df = \mathbb{E}[F|D = \mathcal{D}]
\]

**Example:** \(\mathcal{D} = \{2, 5, 9, 5, 4, 8\}\) is i.i.d. sample from \(\text{Poisson}(\lambda_0), \lambda \in \mathbb{R}^+\)

Assume \(\lambda_0\) is taken from \(\Gamma(x|k, \theta)\) with parameters \(k = 3\) and \(\theta = 1\)

Estimate \(\lambda_0\).
PARAMETER ESTIMATION FOR MIXTURES OF DISTRIBUTIONS

**Given:** a set of observations \( D = \{x_i\}_{i=1}^n, \ x_i \in \mathcal{X} \)

\[
p(x|\theta) = \sum_{j=1}^{m} w_j p(x|\theta_j).
\]

where \( \theta = (w_1, w_2, \ldots, w_m, \theta_1, \theta_2, \ldots, \theta_m) \)

**Example:** Consider a mixture of \( m = 2 \) exponential distributions.

\[
p(x|\theta_j) = \lambda_j e^{-\lambda_j x}, \text{ where } \lambda_j > 0
\]

\[
p(x|\lambda_1, \lambda_2, w_1, w_2) = w_1 \cdot \lambda_1 e^{-\lambda_1 x} + w_2 \cdot \lambda_2 e^{-\lambda_2 x}
\]

where \( \lambda_1, \lambda_2 > 0, w_1, w_2 \geq 0, \) and \( w_1 = 1 - w_2 \)
PARAMETER ESTIMATION FOR MIXTURES OF DISTRIBUTIONS

Likelihood:

\[ p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \]

\[ = \prod_{i=1}^{n} \left( \sum_{j=1}^{m} w_j p(x_i|\theta_j) \right) \]

\( p(D|\theta) \) has \( O(m^n) \) terms. It can be calculated in \( O(mn) \) time as a log-likelihood.

How can we find \( \theta \)? Is there a closed-form solution?
IDEA #1

Suppose we know what data point is generated by what mixing component.

That is, \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} \) is an i.i.d. sample from some distribution \( p(x, y) \), where \( y \in \mathcal{Y} = \{1, 2, \ldots, m\} \) specifies the mixing component.

\[
p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(x_i, y_i|\theta) \\
= \prod_{i=1}^{n} p(x_i|y_i, \theta)p(y_i|\theta) \\
= \prod_{i=1}^{n} w_{y_i} p(x_i|\theta_{y_i}),
\]

where \( w_j = P(Y = j) \).
### IDEA #1

Log-likelihood:

\[
\log p(\mathcal{D}|\theta) = \sum_{i=1}^{n} (\log w_{y_i} + \log p(x_i|\theta_{y_i})) \\
= \sum_{j=1}^{m} n_j \log w_j + \sum_{i=1}^{n} \log p(x_i|\theta_{y_i}),
\]

where \( n_j \) is the number of data points in \( \mathcal{D} \) generated by the \( j \)-th mixing component.

**Constrained optimization:** Let’s first find \( w \)

\[
L(w, \alpha) = \sum_{j=1}^{m} n_j \log w_j + \alpha \left( \sum_{j=1}^{m} w_j - 1 \right)
\]

where \( \alpha \) is the Lagrange multiplier.
IDEA #1

Set \( \frac{\partial}{\partial w_k} L(w, \alpha) = 0 \) for every \( k \in \mathcal{Y} \) and \( \frac{\partial}{\partial \alpha} L(w, \alpha) = 0 \). Solve it.

It follows that \( w_k = -\frac{n_k}{\alpha} \) and \( \alpha = -n \).

\[
w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k),
\]

where \( I(\cdot) \) is the indicator function.

To find all \( \theta_j \), we need to get concrete; i.e., \( p(x|\theta_j) = \lambda_j e^{-\lambda_j x} \).

\[
\frac{\partial}{\partial \lambda_k} \sum_{i=1}^{n} \log p(x_i|\lambda_{y_i}) = 0,
\]

for each \( k \in \mathcal{Y} \).
IDEA #1

Thus, assuming an exponential distribution we obtain that

$$\lambda_k = \frac{n_k}{\sum_{i=1}^{n} I(y_i = k) \cdot x_i},$$

for each $k \in \mathcal{Y}$.

Recall that

$$w_k = \frac{1}{n} \sum_{i=1}^{n} I(y_i = k)$$

If the mixing component designations $\mathbf{y}$ are known, the parameter estimation is greatly simplified.
Idea #2

Suppose we know the $\theta$ but not the mixing component designations.

Looks like clustering, right? Let’s see. Express

$$p(y|\mathcal{D}, \theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta)$$

$$= \prod_{i=1}^{n} \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)}$$

and subsequently find the best configuration out of $m^n$ possibilities.

Data is i.i.d. so $y_i$ can be estimated separately. The MAP estimate for $y_i$

$$\hat{y}_i = \arg \max_{y_i \in \mathcal{Y}} \left\{ \frac{w_{y_i} p(x_i|\theta_{y_i})}{\sum_{j=1}^{m} w_j p(x_i|\theta_j)} \right\}$$
COMBINE THE TWO IDEAS (ITERATIONS)

1. Assume $\theta$ is known, call it $\theta^{(0)}$
2. Compute $y^{(0)}$ using $\theta^{(0)}$ as known
3. Compute $\theta^{(1)}$ using $y^{(0)}$ as known
4. Compute $y^{(1)}$ using $\theta^{(1)}$ as known
5. ... (until convergence)
**Classification Expectation Maximization (CEM)**

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$

2. Calculate $y_i^{(0)} = \arg \max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(0)} p(x_i | \lambda_k^{(0)})}{\sum_{j=1}^{m} w_j^{(0)} p(x_i | \lambda_j^{(0)})} \right\}$ for $\forall i \in \{1, 2, \ldots, n\}$

3. Set $t = 0$

4. Repeat until convergence

   (a) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} I(y_i^{(t)} = j)$

   (b) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} I(y_i^{(t)} = k)}{\sum_{i=1}^{n} I(y_i^{(t)} = k) \cdot x_i}$

   (c) $y_i^{(t+1)} = \arg \max_{k \in \mathcal{Y}} \left\{ \frac{w_k^{(t)} p(x_i | \lambda_k^{(t)})}{\sum_{j=1}^{m} w_j^{(t)} p(x_i | \lambda_j^{(t)})} \right\}$

   (d) $t = t + 1$

5. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$
**EXPECTATION-MAXIMIZATION (EM) ALGORITHM**

Problems with the CEM formulation:

1. We want to estimate $\theta$
2. We do not necessarily need to compute $\mathbf{y}$

Main idea for the EM algorithm:

1. Take step $t$ and assume $\theta^{(t)}$ is known
2. Maximize $\mathbb{E}[p(\mathcal{D}, \mathbf{Y}|\theta)|\theta^{(t)}]$ to calculate $\theta(t + 1)$
**EXPECTATION-MAXIMIZATION (EM) ALGORITHM**

Expected log-likelihood of the complete data over the posterior distribution for $y$ assuming $\theta^{(t)}$ is true:

$$
\mathbb{E}[\log p(D, Y|\theta)|\theta^{(t)}] = \begin{cases} 
\sum_y \log p(D, y|\theta)p(y|D, \theta^{(t)}) & y: \text{discrete} \\
\int_y \log p(D, y|\theta)p(y|D, \theta^{(t)}) dy & y: \text{continuous}
\end{cases}
$$

$$
\theta^{(t+1)} = \arg \max_\theta \left\{ \mathbb{E}[\log p(D, Y|\theta)|\theta^{(t)}] \right\}
$$
EXPECTED-MAXIMIZATION (EM)

1. Initialize $\lambda_k^{(0)}$ and $w_k^{(0)}$ for $\forall k \in \mathcal{Y}$

2. Set $t = 0$

3. Repeat until convergence

(a) $p_Y(k|x_i, \theta^{(t)}) = \frac{w_k^{(t)} p(x_i|\lambda_k^{(t)})}{\sum_{j=1}^{m} w_j^{(t)} p(x_i|\lambda_j^{(t)})}$ for $\forall (i, k)$

(b) $w_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_Y(k|x_i, \theta^{(t)})$

(c) $\lambda_k^{(t+1)} = \frac{\sum_{i=1}^{n} p_Y(k|x_i, \theta^{(t)})}{\sum_{i=1}^{n} x_i p_Y(k|x_i, \theta^{(t)})}$

(d) $t = t + 1$

4. Report $\lambda_k^{(t)}$ and $w_k^{(t)}$ for $\forall k \in \mathcal{Y}$
**EXPECTATION-MAXIMIZATION (EM) ALGORITHM**

E-step: evaluate $p(y|D, \theta^{(t)})$

M-step: $\theta^{(t+1)} = \arg \max_\theta \{ \mathbb{E}[\log p(D, Y|\theta)|\theta^{(t)}] \}$
HOW DID WE ARRIVE AT THIS SOLUTION

1. Try to maximize likelihood $p(D|\theta)$
   
   (a) can be difficult even as log-likelihood; e.g., we get log of a sum of some function of parameters $\theta$ that is unfriendly to differentiation

2. Recognize we have some unobserved or “hidden” variables
   
   (a) mixture case: we figured out that there is a “class label” vector $y$ so we can see the complete data as set of pairs $\{(x_i, y_i)\}_{i=1}^n$

3. Attempt to maximize the likelihood of complete data $p(D, y|\theta)$
   
   (a) cannot do because vector $y$ is unobserved
HOW DID WE ARRIVE AT THIS SOLUTION

4. Think of an iterative process and assume we have $\theta^{(t)}$ as an approximation of $\theta_{ML}$ in step $t$. New goal: find $\theta^{(t+1)}$ of the next step $(t+1)$ that is a little better than $\theta^{(t)}$ from step $t$.

   (a) good news: we can compute the posterior of unobserved data $p(y|D, \theta^{(t)})$ since $D$ and $\theta^{(t)}$ are given.

   (b) This will become the E-step.

5. To find $\theta^{(t+1)}$, try to maximize the expected likelihood of the complete data $\mathbb{E}[p(D, Y|\theta)|D, \theta^{(t)}]$, where we integrate over $p(y|D, \theta^{(t)})$

   (a) this is still hard as we have to work with products instead of sums

6. Try to maximize the expected log-likelihood of the complete data $\mathbb{E}[\log p(D, Y|\theta)|D, \theta^{(t)}]$

   (a) good news: we get expressions that can be simplified so we can compute $\theta^{(t+1)}$ by maximizing $\mathbb{E}[\log p(D, Y|\theta)|D, \theta^{(t)}]$

   (b) This will become the M-step.
HOW DID WE ARRIVE AT THIS SOLUTION

7. The EM algorithm iterates the E-step with the M-step.

8. Prove that maximizing \( \mathbb{E}[\log p(D, Y|\theta)|D, \theta^{(t)}] \) maximizes \( p(D|\theta) \)

   (a) good news: it can be done, but it is not obvious so it had to be done.
   (b) bad news: we can only prove local maximization as the likelihood function is not convex.
RECAP OF REASONING

1. Try to maximize likelihood $p(\mathcal{D}|\theta)$

2. Recognize we have some unobserved or “hidden” variables

3. Attempt to maximize the likelihood of complete data $p(\mathcal{D}, \mathbf{y}|\theta)$

4. Think of an iterative process and assume we have $\theta^{(t)}$ as an approximation of $\theta_{ML}$ in step $t$. New goal: find $\theta^{(t+1)}$ of the next step $(t + 1)$ that is a little better than $\theta^{(t)}$ from step $t$.

5. To find $\theta^{(t+1)}$, try to maximize the expected likelihood of the complete data $\mathbb{E}[p(\mathcal{D}, \mathbf{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$, where we integrate over $p(\mathbf{y}|\mathcal{D}, \theta^{(t)})$

6. Try to maximize the expected log-likelihood of the complete data $\mathbb{E}[\log p(\mathcal{D}, \mathbf{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$

7. The EM algorithm iterates the E-step with the M-step.

8. Prove that maximizing $\mathbb{E}[\log p(\mathcal{D}, \mathbf{Y}|\theta)|\mathcal{D}, \theta^{(t)}]$ maximizes $p(\mathcal{D}|\theta)$
WHY EM WORKS

$$\log p(D|\theta^{(t+1)}) - \log p(D|\theta^{(t)}) = \log \frac{p(D|\theta^{(t+1)})}{p(D|\theta^{(t)})}$$
\[ \log p(D|\theta^{(t+1)}) - \log p(D|\theta^{(t)}) = \log \frac{p(D|\theta^{(t+1)})}{p(D|\theta^{(t)})} \]

Marginalize

\[ = \log \int_y \frac{p(D, y|\theta^{(t+1)})}{p(D|\theta^{(t)})} dy \]
**WHY EM WORKS**

\[
\log p(D|\theta^{(t+1)}) - \log p(D|\theta^{(t)}) = \log \frac{p(D|\theta^{(t+1)})}{p(D|\theta^{(t)})} \\
= \log \int_y \frac{p(D, y|\theta^{(t+1)})}{p(D|\theta^{(t)})} dy \\
= \log \int_y \frac{p(D, y|\theta^{(t+1)})}{p(D, y|\theta^{(t)})} p(y|\mathcal{D}, \theta^{(t)}) dy
\]

Apply product rule
WHY EM WORKS

\[
\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) = \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})}
\]

\[
= \log \int_{\mathbf{y}} \frac{p(\mathcal{D}, \mathbf{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\mathbf{y}
\]

\[
= \log \int_{\mathbf{y}} \frac{p(\mathcal{D}, \mathbf{y}|\theta^{(t+1)})}{p(\mathcal{D}, \mathbf{y}|\theta^{(t)})} p(\mathbf{y}|\mathcal{D}, \theta^{(t)}) d\mathbf{y}
\]

Apply Jensen’s inequality

\[
\geq \int_{\mathbf{y}} \log \frac{p(\mathcal{D}, \mathbf{y}|\theta^{(t+1)})}{p(\mathcal{D}, \mathbf{y}|\theta^{(t)})} p(\mathbf{y}|\mathcal{D}, \theta^{(t)}) d\mathbf{y}
\]
$\log p(\mathcal{D}|\theta^{(t+1)}) - \log p(\mathcal{D}|\theta^{(t)}) = \log \frac{p(\mathcal{D}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})}$

$= \log \int_{\mathbf{y}} \frac{p(\mathcal{D}, \mathbf{y}|\theta^{(t+1)})}{p(\mathcal{D}|\theta^{(t)})} d\mathbf{y}$

$= \log \int_{\mathbf{y}} \frac{p(\mathcal{D}, \mathbf{y}|\theta^{(t+1)})}{p(\mathcal{D}, \mathbf{y}|\theta^{(t)})} p(\mathbf{y}|\mathcal{D}, \theta^{(t)}) d\mathbf{y}$

$\geq \int_{\mathbf{y}} \log \frac{p(\mathcal{D}, \mathbf{y}|\theta^{(t+1)})}{p(\mathcal{D}, \mathbf{y}|\theta^{(t)})} p(\mathbf{y}|\mathcal{D}, \theta^{(t)}) d\mathbf{y}$

Rewrite

$= \mathbb{E}[\log p(\mathcal{D}, \mathbf{Y}|\theta^{(t+1)})|\mathcal{D}, \theta^{(t)}] - \mathbb{E}[\log p(\mathcal{D}, \mathbf{Y}|\theta^{(t)})|\mathcal{D}, \theta^{(t)}]$